박 사 학 위 논 문
Ph.D. Dissertation

구성비 데이터를 위한 커널 방법과 차원 축소

Kernel Methods for Compositional Data and
Dimensionality Reduction

2024

박 준 영 (朴 俊映 Park, Junyoung)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

구성비 데이터를 위한 커널 방법과 차원 축소

2024

박 준 영

한 국 과 학 기 술 원

수리과학과

구성비 데이터를 위한 커널 방법과 차원 축소

박 준 영

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2023년 12월 13일

심사위원장 박 철 우 (인)

심 사 위 원 안 정 연 (인)

심 사 위 원 전 용 호 (인)

심 사 위 원 전 현 호 (인)

심 사 위 원 정 연 승 (인)

Kernel Methods for Compositional Data and Dimensionality Reduction

Junyoung Park

Major Advisor: Cheolwoo Park
Co-Advisor: Jeongyoun Ahn

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematical Sciences

Daejeon, Korea
December 13, 2023

Approved by

Cheolwoo Park
Professor of Mathematical Sciences

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DMAS 박준영. 구성비 데이터를 위한 커널 방법과 차원 축소. 수리과학과 . 2024년. 103+vi 쪽. 지도교수: 박철우, 안정연. (영문 논문)
Junyoung Park. Kernel Methods for Compositional Data and Dimensionality Reduction. Department of Mathematical Sciences . 2024. 103+vi pages. Advisor: Cheolwoo Park, Jeongyoun Ahn. (Text in English)

초 록

구성비 데이터의 분석은 최근 인간 마이크로바이옴 연구에서의 중요성 때문에 특히 주목받고 있다. 이러한 데이터 세트는 고차원적이고 상당수의 0으로 구성되어 있기 때문에 기존의 방법론으로는 어려움을 겪는 경우가 많다. 이를 극복하기 위해 0을 자연스럽게 처리하는 커널 방법을 사용하여 이 문제에 접근하고 고차원성을 완화하기 위한 차원 축소 방법들을 개발하였다. 이 논문에서는 구성비 데이터에 커널 방법을 활용하는 세 가지 프로젝트를 소개한다.

프로젝트 1에서는 0 값들을 치환한 후 로그-비율 변환을 수행하는 만연한 접근 방식이 데이터의 기하 형상에 심각한 왜곡을 초래한다는 것을 보여준다. 대안으로, 기하학적 고려 사항에 기반한 커널 방법을 사용하여 제로 치환이 필요 없는 방법을 제안한다. 프로젝트 2에서는 구성비 데이터의 변수 선택 방법론을 커널에 기반하여 제시하고, 부분 컴포지션보다 병합을 사용해야 한다고 주장한다. 프로젝트 3에서는 두 번째 프로젝트의 방법론을 확장하여 병합에 대한 보다 완화된 접근 방식을 통해 구성비 데이터의 새로운 차원 축소 방법을 개발한다.

핵 심 낱 말 구성비 데이터, 변수 선택, 병합, 커널 방법, 충분 차원 축소

Abstract

Compositional data analysis has been garnering more focus, particularly due to its significance in human microbiome studies. Traditional techniques often struggle with recent data sets as they are high-dimensional and constituted of a significant proportion of zeros. We approach this problem using kernel methods, which naturally handle zeros in data, and develop dimension reduction methods to alleviate the curse of dimensionality and enhance interpretability in subsequent analyses. In this thesis, we introduce three projects utilizing kernel methods for compositional data.

In Project 1, we demonstrate that the prevalent approach of log-ratio transformation, performed after zero-replacement, produces significant distortions in the marginal distribution of data. Instead, we suggest employing kernel methods based on geometric considerations, eliminating the need for zero replacements. In Project 2, we propose a kernel-based variable selection method of compositional data, arguing the use of amalgamation over subcomposition. In Project 3, we extend the methodology from the second project to develop a novel method for reducing the dimension of compositional data through a more relaxed version of amalgamation.

Keywords Compositional data, variable selection, amalgamation, kernel methods, sufficient dimension reduction

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Compositional data are multivariate data with nonnegative values in which only the relative proportions of the components are meaningful. They are frequently normalized to sum to unity to analyze the data consistently. Thus, the domain of compositional data with $d+1$ variables is the d-dimensional simplex $\Delta^d \subset \mathbb{R}^{d+1}$:

$$\Delta^d = \left\{ (x_0, x_1, \ldots, x_d) \,\middle|\, \sum_{i=0}^{d} x_i = 1,\ x_i \geq 0,\ \forall i \right\}.$$

Compositional data appear in many scientific applications, for instance, geochemical compositions [115], word compositions in texts [69], and composition of immune cells [129]. Among various examples, this research is primarily motivated by microbiome data, which indicates the relative abundance of microbes that live in or on the human body.

The human microbiome studies have gained much attention since it is linked to various diseases and health-related attributes in humans [40, 49, 125]. Modern high-throughput sequencing technologies, such as 16S ribosomal RNA (rRNA) gene sequencing, have enabled generating the raw number of microbiomes from collected samples. Due to the varying amounts of DNA count across different samples, this data must be regarded as compositional [59]. In addition, microbiome data notably exhibits that the number of microbial taxa is often much higher than the available sample size, i.e., high dimensionality and that a significant portion of data are zeros [68]. Therefore, most microbiome data obtained in practice are located mainly on the boundary of a high-dimensional simplex. All these aspects - compositional structure, high dimensionality, excess zeros - pose significant challenges as they must be addressed simultaneously.

The compositional structure, imposing the constant sum constraint, results in spurious negative correlations in the data [19, 84]. That is, each component of a composition is inevitably affected by a change of other components, and this phenomenon would yield uninterpretable results if classical multivariate methods were applied blindly to the data. An overwhelmingly dominant approach to overcome this problem is to take *log-ratio transformations* to compositional data, which is proposed by Aitchison [3]. The log-ratio transformations naturally address the relative nature of compositional data by dealing with ratios directly, and the logarithm is applied to spread out the ratio values to the Euclidean space. Thus, after applying these transformations, one can apply traditional statistical methods in the Euclidean space. Furthermore, his argument establishes a homeomorphism between the interior of the simplex,

$$\Delta^d_{>0} = \left\{ (x_0, x_1, \ldots, x_d) \,\middle|\, \sum_{i=0}^{d} x_i = 1,\ x_i > 0,\ \forall i \right\},$$

called a *positive simplex*, and the whole Euclidean space $\mathbb{R}^d$. This one-to-one correspondence naturally pulls back the linear vector space structure to the positive simplex, called the *Aitchison geometry* [6], which has compelled researchers for a long time.

However, the log-ratio approaches for compositional data are not directly applicable for data with zero values since both logarithm and ratio computations are not able to deal with zeros. As microbiome data, which is of the most recent interest, have a significant number of zeros, this is a crucial weakness to the log-ratio approaches. In order to overcome this drawback, it has been a common practice to replace zeros with small positive values. There are some widely accepted heuristics for zero replacements; see

[42] for example, but one can also create countless ways to do this. It is important to note that there is no consensus for zero replacement, and many studies have shown that the data analysis results are often sensitive to the choice of zero replacement strategy, when paired with log-ratio transformations [67, 89]. We will address this issue in Chapter 2 that this sensitivity is an unavoidable, essential geometric problem.

In this thesis, we propose to employ various kernel methods to avoid such zero problems in the dominant approach. Applying kernel methods directly on the compositional domain does not require zero replacements, so it does not damage data. In Chapter 2, we prove that kernel methods based on the embedding to reproducing kernel Hilbert space (RKHS) consistently apply with the compositional nature, *scale invariance* [3]. Within the kernel framework, we also address the high dimensionality of recent data. We develop an effective variable selection method and a general dimension reduction method in Chapter 4 and Chapter 5, both of which are based on the *conditional covariance operator* of RKHSs. These works are generalizations of the kernel dimension reduction works for Euclidean data [20, 34]; we extend their original method for orthogonal projections to arbitrary nonlinear projections, without sacrificing theoretical guarantees. We now describe brief summaries of those three projects in the following sections.

1.1 Kernel Methods for Compositional Data with Many Zeros

Addressing the zero problem in the log-ratio approaches has been a longstanding problem of compositional data literature. There are many heuristics and imputation methods for replacing zeros deemed reasonable [71, 72], but the inconsistency of data analytic results depending on the choice of replacement methods have continuously been reported [67, 78, 89]. Therefore, one must examine the effects of the replaced zeros when using the log-ratio approach, which further complicates the data analysis process and interpretation of the results.

In Chapter 2, we first show that the dominant approach, zero replacement followed by log-ratio transformations, is geometrically improper, leading to anomalous distortions in the marginal distribution of data. Based on a simple geometric intuition, we demonstrate that, in terms of Aitchison geometry, zero replacement is not a small data alteration and the sensitivity to the choice of zero replacement is essentially unavoidable. Therefore, the inconsistency of the zero replacements is a fundamental, unavoidable problem when paired with log-ratio transformations.

To circumvent this problem, a new approach that departs from the log-ratio framework is needed. As an alternative, we propose to apply kernel methods on the compositional domain directly while addressing the geometric structure of compositional data. The relative structure of compositional data displays a unique geometric structure; they are in fact represented by nonnegative *radial lines*, and every statistical method has to be applied consistently along with the points on the radial lines. This is also known as the scale invariance principle [3], and we prove that RKHS embeddings enjoy this scale invariance property. Having known this fact, it is clear that kernel methods obviously overcome the zero problem of log-ratio approaches; it does not require zero replacements, so it does not damage data.

Furthermore, we suggest that the spherical representation of compositional data may provide a better domain for applying kernel methods. Such representation is obtained via *radial transformation*, a perspective which has long been neglected in the literature after debate of Watson and Aitchison in 1989-1992 [96]. Since dot-product kernels on the hypersphere have long deeply studied in the literature [99, 91, 39], we can take advantages of their known properties, such as known decay rates for the eigenvalues of kernels. Experiments with kernel principal component analysis reveal that the conventional

methodology, log-ratio transformation after zero replacement, disperses the data with zeros erratically, whereas our proposed method based on radial transformation remains stable. Chapter 2 is a joint work with Changwon Yoon, Cheolwoo Park, and Jeongyoun Ahn, and is published in the *International Conference of Machine Learning* in 2022 [80].

1.2 Variable Selection Method for Compositional Data

In many recently available compositional data such as microbiome data, the number of variables often far exceeds the number of available samples. We are often interested in particular taxa that determine the state of human disease or health, so variable selection is one of the most relevant task in high-dimensional compositional data analysis. However, existing methods often rely on log-ratio transformations [63, 90], which may distort data and lead to unreliable selections as will be pointed out in Chapter 2. To avoid this issue, we propose an effective kernel-based variable selection method for compositional data in Chapter 4. This work extends the kernel dimension reduction (KDR) work for Euclidean data [20, 34] that leverages the *conditional covariance operator*. Unlike traditional kernel methods, it maintains interpretability by measuring conditional independence *after* data projections, thereby offering a criterion for achieving *sufficient dimension reduction* (SDR) [60]. This work and the next project shares the same theoretical background, whose detail is demonstrated in Chapter 3 separately.

In addressing variable selection for compositional data, our work highlights a crucial but overlooked issue in the traditional approaches. In most cases, renormalization is conducted after choosing a subset of variables since the selected variables retain only relative information. This process, renormalization after selection, is called a subcomposition, which has been regarded as a fundamental operation of compositional data [6]. However, we uncover that building predictive models based on even accurate subcompositions could severely deteriorate performance, as renormalization unnecessarily eliminates the relative information of variables to the *total*. To rectify this, we propose a straightforward but powerful strategy, *amalgamating* [3] all unselected variables into an additional coordinate. By adding this dummy coordinate, the relative information of selected variables to the total is preserved and the selection result directly locates in a lower dimensional simplex.

Combining our amalgamation-based framework and the generalized theory of conditional covariance operator, we build an effective variable selection algorithm that aims for SDR of compositional data. The idea of the proposed method is essentially minimization of the conditional covariance operator as developed in Chapter 3, but we slightly change the theoretical environment to efficiently perform the variable selection in this chapter. We develop the theory in discrete settings, and then overcome the computational infeasibility of the proposed discrete optimization in high dimensions by continuously relaxing our algorithm and performing projected gradient descent. As a result of relaxation, we eventually obtain *importance weights* of variables, which are rounded to result in a discrete variable selection. Experiments in a variety of situations exhibits the superior performance of our method than log-ratio-based methods. The content of Chapter 4 is a joint work with Jeongyoun Ahn and Cheolwoo Park, and is published in *International Conference of Machine Learning* in 2023 [81].

1.3 Interpretable Dimensionality Reduction for Compositional Data

Although we develop a kernel approach to compositional data analysis in Chapter 2 that does not distort zero values in data, direct applications of such kernel methods to high-dimensional compositional data suffer from the curse of dimensionality. This problem exacerbates further for data with higher variances; for instance, see Song and Chen [106] for decreasing power of kernel two sample tests as the variability of data increases. As our primary motivating data, microbiome data exhibits extreme dispersion [68], we need to find a solution that alleviates the curse of dimensionality.

Despite the need for dimension reduction methods of compositional data, rigorous exploration in this area has been limited due to their intricate relative structures. Furthermore, most existing methods grapple with weaknesses in the interpretability of reduction results or accuracy issues stemming from zero values. For example, simple linear projections of the original count data break the compositional structure [117] in general, and linear projections of log-ratio-transformed variables are not clearly interpreted as compositional data again. This highlights a need for new dimension reduction methods that fulfill two requirements: interpretability and ability to deal with zeros adequately.

In Chapter 5, we introduce a novel dimension reduction framework for compositional data that results in lower dimensional compositions, building on the concept of amalgamation. Our new framework is more flexible than rigid amalgamation while maintaining the crucial interpretation of aggregating similar compositional variables. We approach this dimension reduction by seeking an optimal SDR projection as in the previous variable selection work. Here, the combination of our new framework and the SDR relation represents an intuitive objective: aggregation of the original variables based on their *functional similarity*.

To achieve a desirable dimension reduction within our framework, we extensively generalize the KDR method of Fukumizu et al. [34]. Although we apply this extended theory only to compositional data, our generalized theory shows a clear potential to be applied to other data with a specified class of dimension reduction functions. To emphasize its potential, we elaborate all the theoretical details with a fully general language in a separate Chapter 3.

Applying the theoretical result developed in Chapter 3 with some algorithmic developments tailored to compositional data, we find via experiments that our proposed algorithm works extremely well in general. By setting the target dimension of the algorithm to three variables, our method exhibits an unprecedented graphical exploration tool for compositional data through ternary plots. We can fully understand the projected variables in terms of the original variables, and the visualizations of projected data naturally exhibit relative behavior, or interactions, of projected variables. The content of Chapters 3 and 5 are joint works with Jeongyoun Ahn and Cheolwoo Park.

1.4 Organization

This thesis is organized as follows. In Chapter 2, we demonstrate the details of our proposal to employing kernel methods to compositional data. Chapter 3 devotes to generalize the kernel dimension reduction theory to arbitrary structured projections, being a theoretical basis for the subsequent chapters. We point out a crucial problem of subcomposition in Chapter 4, and we propose an amalgamation-based variable selection framework, approached by a slightly modified version of the method developed in Chapter 3. In Chapter 5, we propose a new composition-to-composition dimension reduction frame-

work, which produces an unprecedented compelling graphical exploration tool in the compositional data literature. Finally, we provide several further research directions in Chapter 6.

Each of the following chapters in this dissertation is a version of the author's publications or works in progress. Because they are independently studied with different objectives, each chapter may have its own introduction and conclusion, and the notations and definitions may differ across the chapters.

Chapter 2. Kernel Methods for Radial Transformed Compositional Data with Many Zeros

2.1 Introduction

Compositional data are multivariate nonnegative data carrying only *relative* information of components. They are often normalized to have a constant total sum, typically one, so that the data with $d+1$ variables reside in a compact subset of the Euclidean space:

$$\Delta^d = \left\{ (x_1, \cdots, x_{d+1}) \in \mathbb{R}^{d+1} \middle| \sum_{i=1}^{d+1} x_i = 1,\ x_i \geq 0,\ \forall i \right\},$$

called a *simplex*. Here, the superscript d denotes the topological dimension of the simplex.

Data comprised of compositions are ubiquitous in many scientific fields: geochemical composition of rocks or soils in earth science; proportions of various micro-organisms at different sea depths in marine science; portfolio allocation in finance, just to name a few. Among them, our primary motivating examples are microbiome data consisting of relative abundance of microbes, whose analyses have recently been spotlighted in medical research [37], thanks to the emerging scientific and public interests in human gut microbiomes that are associated with many diseases and health-related attributes of humans and animals. Notable characteristics of microbiome data are that the number of bacterial taxa is typically much higher than the available sample size, i.e., high dimension, low sample size, and that a significant portion, about 50 – 80%, of data are zeros [42]. Those zeros make microbiome data locate mainly on the boundary of a high-dimensional simplex.

The compositional aspect of the data poses challenges to statistical data analysis. Due to the constant sum constraint, each component of a composition is inevitably affected by other components. To be specific, they have spurious negative correlations [84, 19]. This would yield uninterpretable results if classical multivariate methods are applied blindly to the data. An overwhelmingly dominant approach to overcome this problem is to take log-ratio transformations to compositional data, which is proposed by Aitchison [3]. There are three types of such transformations, *additive*, *centered*, and *isometric log-ratio transformations*, all of which send compositional data to the Euclidean space. After applying one of these transformations, one may use traditional multivariate statistical methods in the Euclidean space.

However, the log-ratio methods are not readily applicable for data with many zeros because logarithm and ratio computations in the transforms do not allow zero values. Indeed, the log-ratio transformations are forced to deal only with data on the *open simplex*:

$$\mathcal{S}^d = \left\{ (x_1, \cdots, x_{d+1}) \in \mathbb{R}^{d+1} \middle| \sum_{i=1}^{d+1} x_i = 1,\ x_i > 0,\ \forall i \right\},$$

and they cannot manage the boundary points essentially. In order to apply the log-ratio methods for data on the simplex boundary, researchers have suggested perturbing the data slightly so that they all fit into $\mathcal{S}^d$, e.g., by substituting zeros with small positive values and then re-normalizing them to sum to one. There are countless ways to do this. See Martín-Fernández et al. [70, 71] for some widely-used substitution methods. For comparisons of various zero replacement algorithms, see Rasmussen et al. [89], Lubbe et al. [67]. Nonetheless, it has been repeatedly reported that analysis results and subsequent

scientific conclusions are often sensitive to the choice of the zero replacement strategy. This further complicates the data analysis process and interpretation of the results, since one must examine the effects of the replaced zeros.

2.1.1 Main Contributions

We first point out in Section 2.2 that the underlying geometry of the above log-ratio approaches, so-called the Aitchison geometry [6], *enlarges* the dependence on the zero replacement methods. Moreover, it will be demonstrated that the log-ratio approach with zero replacement distorts the intrinsic structure of the data substantially, which implies that this approach is theoretically unjustifiable. It motivates us to find an alternative way to manage zeros in compositional data.

The objective of the current study is to show that kernel methods, including classical kernel approaches [100, 101, 111] and kernel mean embedding methods [45, 131, 77] can be applied to compositional data containing many zeros, by considering an alternative transform of the data, which is *radial transformation*. It will be seen that this transformation preserves the relative ratio information in the composition and does not require zero substitutions, thus more appropriate than the log-ratio approaches in handling such data.

The new domain for the transformed compositional data is a hypersphere, where a rich class of kernels is available. We establish a theoretical framework for kernel methods in this new domain by proving multi-level equivalences between domains before and after the transformation. We also give a list of isotropic kernels with desirable properties such as universality. Therefore, this work enables practitioners to employ kernel-based learning tools such as kernel principal component analysis (PCA) and maximum mean discrepancy to compositional data in a theoretically justifiable fashion. Also, as the computational cost of most kernel methods is $O(n^2)$ once the gram matrix is calculated, the proposed method effectively provides a solution to the curse of dimensionality in analyzing high-dimensional compositional data.

2.1.2 Related Works

A number of works have endeavored to manage zero values of compositional data without replacing them in order to honor the essential zeros [70] of data. One popular method is to take square-root transformation and then use the theory of directional statistics [95, 112, 120]. Butler & Glasbey [15] proposed a latent Gaussian model on the simplex that does not require data transformation, but their approach ignores the relative structure of compositional data. Zadora et al. [130] and Bear & Billheimer [12] modeled the probability of zero values separately with logratio-based distributions, and so did Tsagris & Stewert [119] but with the Dirichlet distribution on the open simplex.

We note that our geometric treatment, the radial transformation for compositional data, has been considered in geological literature a few decades ago in Watson and Philip [124], followed by an exchange of papers and letters to the editor between Watson and Aitchison, published in Mathematical Geology from 1989-1992. There had been aggressive rebuttals to each other during the exchange; see Section 3 of Scealy and Welsh [96] and references therein for a summary of their arguments. A main reason for Aitchison's disapproval of the radial transformation was that the angular distance is not subcompostionally dominant [5]. However, our proposed method does not require interpreting the distance of data but only embeds data into a larger space where this criticism is irrelevant.

After moving onto the hyperspheres we are able to take advantage of fruitful library of kernels.

Kernels on hyperspheres, especially isotropic types, have been broadly studied in the literature [39, 91, 99]. Isotropic kernels are also known as dot-product kernels, which are depend only on the dot product of inputs. It is known that the decay of eigenvalues of kernels is related to the performance of learning with kernels, and much is known for dot-product kernels on spheres; we refer to Scetbon and Harchaoui [97] for recent exposition.

Traditional kernel methods are based on performing numerous linear approaches inside the high-dimensional feature space via kernel embeddings, called the kernel trick [101]. In recent years, it has been recognized that we can also embed probability distributions on the domain into the feature space. This is called the kernel mean embedding, which are broadly applicable in arbitrary domains with appropriate kernels. Using the kernel mean embedding, Gretton et al. [45] proposed a non-parametric two-sample test based on the distance between probability measures, and Balasubramanian et al. [10] proposed a goodness-of-fit test with discussions of minimax optimality. See Muandet et al. [77] for a comprehensive review of mean embedding and other numerous applications. Universal or characteristic kernels should be used for the mean embedding methods; Micchelli et al. [75] and Sriperumbudur et al. [109] characterized them in various cases.

Finally, we note that a key geometric motivation of this work is shared with our previous work [56], in which we interpret the compositional domain as a hypersphere modded out by a reflection group action and use spherical harmonics theory to construct finite-dimensional polynomial kernels. However, in this work, we propose an intuitive radial transformation and consider a general class of kernels, which is computationally much more attractive.

2.1.3 Organization of the Chapter

Section 2.2 demonstrates that the log-ratio approaches produce geometric distortions to the data. Then, we briefly review kernel methods and the pull-back construction of function spaces in Section 2.3. In Section 2.4, we propose a radial transformation with the equivalence property that rationalizes the analysis of the compositional data on the nonnegative part of a hypersphere. Section 2.5 briefly reviews well-known dot-product kernels on the hyperspheres with their universality. We take the example of kernel PCA in Section 2.6 to showcase the benefits of the proposed idea and the data distortions in log-ratio approaches through experiements. We give conclusions and discussions of this chapter in Section 2.7 with some future research directions. Supplementary materials in Section 2.8 provide detailed information on experiments as well as extensive additional experimental results.

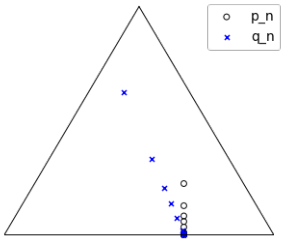
2.2 Geometric Limitations

In this section, we take a deeper look at the Aitchison geometry [6] on the positive simplex $\Delta^d_{>0}$ and reveal an anomalous, counter-intuitive behavior near the boundary of the simplex. Clearly an underlying premise of zero replacement is that it causes *negligible alteration* in the data. However, in the following we discuss how that cannot happen under the log-ratio scheme.

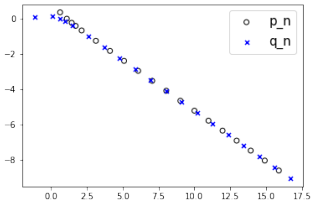
The centered log-ratio (clr) transformation is defined by

$$\mathrm{clr}(x) = \left(\log \frac{x_0}{x'}, \dots, \log \frac{x_d}{x'}\right) \in \mathbb{R}^{d+1}$$

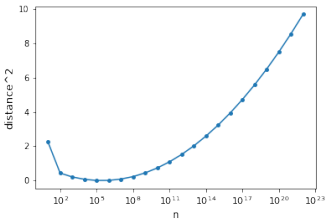
for all $x \in \Delta^d_{>0}$, where $x' = (x_0 \cdots x_d)^{1/(d+1)}$. It is a homeomorphism between $\Delta^d_{>0}$ and a hyperplane in $\mathbb{R}^{d+1}$, so it transfers the linear structure and the inner product defined on the hyperplane to the open



(a) Ternary plot of p_n and q_n in Δ^2



(b) Plot of $\mathrm{ilr}(p_n)$ and $\mathrm{ilr}(q_n)$ in $\mathbb{R}^2$.



(c) Squared Aitchison distance $\|\mathrm{ilr}(p_n) - \mathrm{ilr}(q_n)\|^2_{\mathbb{R}^2}$

Figure 2.1: Demonstration of Aitchison geometry using two sequences p_n and q_n in Δ^2 converging to a common point on the simplex boundary, shown in (a). The two sequences of ilr-transformed points in $\mathbb{R}^2$ shown in (b) are divergent, which is also verified by that the squared Aitchison distance between p_n and q_n is divergent, as shown in (c).

simplex $\Delta^d_{>0}$. The geometry of clr-transformed data is called the Aitchison geometry. In this geometry, it is crucial to note that points close to the boundary of $\Delta^d_{>0}$ are far from the origin (the center of the simplex) since $\log y$ diverges to $\pm\infty$ as $y \to 0$ or $y \to \infty$. The problem occurs here; if a real dataset is concentrated on a boundary point, as in microbiome data, the Aitchison geometry views the data as *diverging to infinity*.

To be more specific, let us consider the following two sequences on $\Delta^2_{>0}$:

$$p_n = \left(\frac{2}{3} - \frac{1}{n}, \frac{2}{n}, \frac{1}{3} - \frac{1}{n}\right), \quad \text{and} \quad q_n = \left(\frac{2}{3} - \frac{6}{n^{1.1}}, \frac{7}{n^{1.1}}, \frac{1}{3} - \frac{1}{n^{1.1}}\right),$$

with $n \geq 9$. Note that both sequences converge to the same point $(2/3, 0, 1/3)$, as displayed in the ternary plot in Figure 2.1(a). Thus *they are almost the same* for all sufficiently large n. We then pass through these sequences via the isometric log-ratio (ilr) transformation [27], which maps Aitchison geometry of $\Delta^2_{>0}$ to $\mathbb{R}^2$ isometrically. The ilr-transformed sequences are displayed in Figure 2.1(b), where the points exactly indicate the Aitchison geometry of p_n and q_n, as the ilr transformation preserves the inner product. Here, the points $\mathrm{ilr}(p_n)$ and $\mathrm{ilr}(q_n)$ continue to move from left to right as n increases, indicating that both sequences *diverge* as $n \to \infty$. We also check how the similarity of p_n and q_n are changed by ilr transformation. To see their relative distance in Aitchison geometry, we calculate the distance $\|\mathrm{ilr}(p_n) - \mathrm{ilr}(q_n)\|^2_{\mathbb{R}^2}$ and plot them in Figure 2.1(c). We can see that the distance between the two sequences is clearly diverging toward infinity, although the original sequences get close to each other.

This example tells us that the Aitchison geometry tends to *amplify a tiny movement near the boundary* of the simplex. Another interpretation is that points close to the boundary are *close to infinity*, and the replacement of zeros in the Aitchison geometry is like towing points at infinity to a finite position. Consequently, the configuration of log-ratio transformed data are critically dependent on which zero replacement method is used. Since there are countless ways of replacing zeros, it may not be possible at all to find an appropriate representation of the data in this way. Moreover, the inconsistent interpretation of the data subject to the zero replacement method makes the results of statistical analysis unreliable. It is also clear that if there are more zeros or the dimension is higher, these problems exacerbate even further. In summary, the log-ratio approach with zero replacement is theoretically unjustifiable due to the faulty representation of the data geometry.

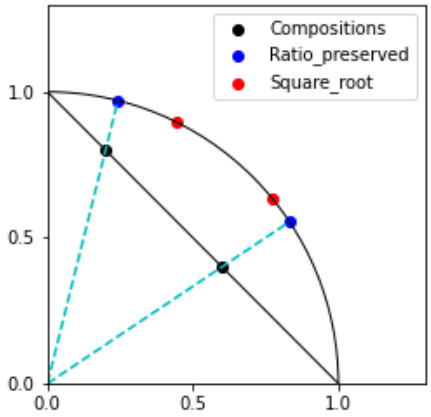


Figure 2.2: Comparison of the compositional radial vectors (dashed lines) and the square-root transformed points (red points) on the unit circle. The relative ratios of the red points are different from other points.

2.2.1 Square-Root Transformation

The square-root transformation of compositional data have also been considered as a transformaion-based approach to compositional data. It sends the point $x = (x_0, \ldots, x_d) \in \Delta^d$ to $(\sqrt{x_0}, \ldots, \sqrt{x_d})$ so that the transformed point lies in the first orthant of the hypersphere $\mathbb{S}^d$. Since this transformation does not suffer from zeros in the data, it has been frequently appeared in the literature, though less popular [95, 96].

Although approaches based on this strategy are theoretically well-justified, we point out one crucial disadvantage of the square-root transformation. Compositional data consist of relative ratios represent in fact the corresponding *radial vectors* (see Section 2.4 for details). Thus, the most natural representation of the composition on a sphere is where the radial vector intersects the sphere. However, the square-root transform produces a different point, which implies that it distorts the original composition. Figure 2.2 illustrates it in the case of $d = 1$, where blue dots represent the transformed points from our radial transformation and the red dots, clearly not preserving the ratios, are from the square-root transformation.

2.3 Theory of Kernels and Pull-Back Construction

Here we briefly review the general theory of kernel methods and summarize a few important definitions. We denote by $\mathcal{X}$ the sample space of observations and assume that it is *compact* to avoid unnecessary theoretical remarks.

2.3.1 RKHS and the Associated Feature Map

By a *kernel*, we mean a real-valued *continuous*, positive definite and symmetric function defined on $\mathcal{X} \times \mathcal{X}$ throughout the chapter. Once a kernel K is given, there exists an associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ and a *feature map* $\Phi_K : \mathcal{X} \to \mathcal{H}_K$ which maps $x \in \mathcal{X}$ to a function

$$\Phi_K(x)(\cdot) := K(x, \cdot)$$

on $\mathcal{X}$ [101]. We omit the subscripts K if there is no confusion in notations. The Hilbert space $\mathcal{H}_K$ is endowed with the inner product $\langle \cdot, \cdot \rangle$ which has the *reproducing property*

$$\langle f, \Phi(x) \rangle = f(x), \quad \forall f \in \mathcal{H}_K, \tag{2.1}$$

which in turn implies $\langle \Phi(x), \Phi(y) \rangle = K(x, y)$. The space $\mathcal{H}_K$ is the closed span of the image of the feature map Φ, i.e., $\mathcal{H}_K = \overline{\text{span}\{\Phi(x) | x \in \mathcal{X}\}}$, and also called the *feature space*. Kernel-based learning means that we map the data via Φ and then apply various learning methods in $\mathcal{H}_K$. Multifarious linear methods, such as PCA, kernel ridge regression (KRR), and support vector machines (SVM), can be "kernelized" without explicitly specifying Φ since both methods depend only on the inner product of the original data.

In order to improve the performance of linear methods in $\mathcal{H}_K$, it is often required that the RKHS $\mathcal{H}_K$ is *large enough* so that the transformed data are linearly analyzable. The corresponding notion to the largeness is *universal kernels*, the kernels with the property that $\mathcal{H}_K$ is dense in $C(\mathcal{X})$ where $C(\mathcal{X})$ is the space of continuous functions on $\mathcal{X}$. Note that $\mathcal{H}_K \subseteq C(\mathcal{H}_K)$ since our kernel K is continuous. Universal kernels play several central roles in kernel methods. For example, Steinwart [110] proved the consistency of SVM using universal kernels, together with examples of universal dot-product kernels on $\mathbb{R}^d$.

2.3.2 Kernel Mean Embedding of Probability Distributions

Identifying each data point $x \in \mathcal{X}$ with the Dirac probability measure δ_x centered on x, one can extend the domain of the feature map $\Phi_K : \mathcal{X} \to \mathcal{H}_K$ to the set of *probability measures* on $\mathcal{X}$. The extended map is called the *kernel mean embedding*, and the mean embedding of a probability measure $\mathbb{P}$ with respect to K is defined by

$$\mu_{\mathbb{P}}(\cdot) := \int_{\mathcal{X}} K(x, \cdot) \, d\mathbb{P}(x).$$

Under the aforementioned assumptions on $\mathcal{X}$ and K, it is known that $\mu_{\mathbb{P}} \in \mathcal{H}_K$, and it has the *generalized reproducing property*

$$\mathbb{E}_{X \sim \mathbb{P}} [f(X)] = \langle \mu_{\mathbb{P}}, f \rangle \tag{2.2}$$

for all $f \in \mathcal{H}_K$ [105].

The kernel K is said to be *characteristic* if the corresponding mean embedding μ is injective. Characteristic kernels play an essential role in the theory of mean embedding because they ensure that

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K} = 0 \text{ if and only if } \mathbb{P} = \mathbb{Q}$$

for all probability measures $\mathbb{P}, \mathbb{Q}$ on $\mathcal{X}$. Here, the distance $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K}$ is called the *maximum mean discrepancy* (MMD), whose empirical estimate can be used for non-parametric two-sample test. Gretton et al. [45] showed that all universal kernels are characteristic, and thus we focus on universal kernels in this work.

2.3.3 Pull-Back of RKHS

Let $\mathcal{Y}$ be another domain of observations, and let $\varphi : \mathcal{Y} \to \mathcal{X}$ be any (continuous) function. We consider transferring an RKHS $\mathcal{H}_K$ defined on the original domain $\mathcal{X}$ through φ. The resulting space is called the *pull-back* along φ. See Section 5.4 of Paulsen and Raghupathi [82] for the proofs of the following results.

Given a kernel K defined on $\mathcal{X}$, let $K \circ \varphi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the *pull-back* of K along φ, denote the function given by

$$K \circ \varphi(s,t) = K(\varphi(s), \varphi(t)) \tag{2.3}$$

for all $s, t \in \mathcal{Y}$. One can readily show that $K \circ \varphi$ is positive definite and symmetric, and therefore it is a *kernel* on $\mathcal{Y}$. Hence the kernel $K \circ \varphi$ defines an RKHS $\mathcal{H}_{K\circ\varphi}$ of functions on $\mathcal{Y}$, called the pull-back of $\mathcal{H}_K$ along φ. The following theorem gives a full characterization of the members of $\mathcal{H}_{K\circ\varphi}$.

Theorem 2.1 ([82]). *The elements of the RKHS $\mathcal{H}_{K\circ\varphi}$ on $\mathcal{Y}$, which is generated by a kernel $K \circ \varphi$, are completely described as*

$$\mathcal{H}_{K\circ\varphi} = \{f \circ \varphi \mid f \in \mathcal{H}_K\}.$$

Furthermore, the norm of any function $g \in \mathcal{H}_{K\circ\varphi}$ is associated with the original RKHS norm of $\mathcal{H}_K$ as

$$\|g\|_{\mathcal{H}_{K\circ\varphi}} = \min_{f \in \mathcal{H}_K} \{\|f\|_{\mathcal{H}_K} \mid g = f \circ \varphi\}.$$

Theorem 2.1 establishes a well-defined linear map $\varphi^* : \mathcal{H}_K \to \mathcal{H}_{K\circ\varphi}$ given by $\varphi^*(f) = f \circ \varphi$, called the *pull-back map* of φ. Typically, we consider the case where $\mathcal{Y}$ is a subset of $\mathcal{X}$ and φ is the canonical inclusion map of $\mathcal{Y} \subseteq \mathcal{X}$. We denote by $K|_{\mathcal{Y}} = K \circ \varphi$ in this case. Then $\mathcal{H}_{K|_{\mathcal{Y}}}$ is just a set of restrictions of functions in $\mathcal{H}_K$ to $\mathcal{Y}$. The pull-back construction will be instrumental in formulating our theoretical frameworks in Sections 2.4 and 2.5.

2.4 Radial Transformation and Equivalence of Function Spaces

As we showed inadequacies of traditional methods for compositional data with zeros in Section 2.2, we suggest a new alternative approach in the present section. We propose to use RKHS embeddings of compositional domains, together with the radial viewpoint of compositional data mentioned briefly in Section 2.2.1. We first point out that there are equivalent expressions of compositional data along the radial direction, and then prove that function-theoretic and RKHS approaches to these expressions are, in fact, equivalent. Therefore, it is natural to look for the most convenient domain for analysis, and we claim that the hypersphere meets these needs. Since compositional data are mostly normalized onto the simplex, we define a radial transformation sending data on the simplex to the hypersphere and proceed with our main results.

2.4.1 Ratio-Preserving Radial Transformation

Recall that compositional data consist only of relative information, which is scale-invariant. This invariance implies that the ratio information is inherent in the corresponding *radial vectors*, thus the radial vectors possess the core of compositions. Taking this viewpoint into account, we can interpret the simplicial expression of compositional data as the intersection of nonnegative radial vectors and a linear manifold.

From this radial interpretation, we realize that other representations of compositional data are possible depending on the choice of intersection manifold. For example, we may choose hyperspheres or hypercubes, which would yield hyperspherical or hypercubical expression of compositional data respectively. We already saw in Figure 2.2 that the blue dots, the intersection of the circle and the radial vectors, equivalently represent the corresponding compositional data on the simplex Δ^1. Then the following two questions naturally arise:

(a) Are the data analysis results independent of the choice of representations?

(b) If so, which representation of compositional data is most convenient for computations and expected to give satisfying results in general?

We show that the first question is affirmative *for function-theoretic* or *kernel-based* analyses in the following subsections. The equivalence of RKHS embeddings on various domains of compositional data is derived by the pull-back construction in Section 2.3.3. Note that computations of pull-back kernels are often unnecessarily complicated in practice, and thus it is preferable to fix an appropriate domain on which various easily computable and well-studied kernels exist. For the second question, we claim that the hyperspherical expression is best for kernel learnings as there are a plethora of easily computable kernels on hyperspheres with *desirable decay of eigenvalues* [97]. Because understanding the decay of kernel eigenvalues is important for low-dimensional interpretation of the results from kernel learnings, we believe that the sphere is the safest domain in this regard.

Along these lines, we define a *radial transformation* $\psi : \Delta^d \to \mathbb{S}^d_{\geq 0}$ by

$$\psi(x) = \frac{x}{\|x\|_2} \quad \text{for all} \quad x \in \Delta^d,$$

where $\mathbb{S}^d_{\geq 0}$ denotes the nonnegative part of $\mathbb{S}^d$. In the following subsections, we prove the equivalences of two types of function spaces that answer the question (a). The statements and proofs are written in terms only of ψ, but they can be immediately generalized to arbitrary homeomorphic transforms along the radial direction.

2.4.2 Function-Theoretic Equivalence

Note that ψ is continuous, and we readily obtain the continuous inverse $\pi : \mathbb{S}^d_{\geq 0} \to \Delta^d$ of ψ, where $\pi(y) = y/\|y\|_1$. Thus, the domains Δ^d and $\mathbb{S}^d_{\geq 0}$ are homeomorphic, i.e., they are *topologically equivalent*. This equivalence leads to a well-known identification of spaces of continuous functions, stated as follows.

Proposition 2.2. *The homeomorphism ψ induces an isometric isomorphism of function spaces*

$$C(\Delta^d) \cong C(\mathbb{S}^d_{\geq 0}).$$

Hence, function-theoretic analysis on the space $C(\Delta^d)$ is equivalent to the corresponding analysis on $C(\mathbb{S}^d_{\geq 0})$. For example, if one wants to find a continuous function on Δ^d that interpolates the given data, it suffices to find the corresponding one on $\mathbb{S}^d_{\geq 0}$ based on the equivalence.

2.4.3 Equivalence of RKHS Embeddings

We also verify that the radial transformation ψ induces the equivalence between RKHS embeddings on Δ^d and $\mathbb{S}^d_{\geq 0}$. Let K be a kernel defined on $\mathbb{S}^d_{\geq 0}$ and let $K \circ \psi$ denote the pull-back along ψ given by (2.3). The pull-back map $\psi^* : \mathcal{H}_K \to \mathcal{H}_{K\circ\psi}$ defined in Section 2.3.3 establishes the following equivalence.

Theorem 2.3. *$\psi^* : \mathcal{H}_K \to \mathcal{H}_{K\circ\psi}$ is an isometric isomorphism of Hilbert spaces. Furthermore, the feature maps associated to K and $K \circ \psi$ are compatible with ψ^* in the sense that the following diagram commutes.*

$$\begin{array}{ccc} \Delta^d & \xrightarrow{\Phi_{K\circ\psi}} & \mathcal{H}_{K\circ\psi} \\ \downarrow{\scriptstyle \psi} & & \uparrow{\scriptstyle \psi^*} \\ \mathbb{S}^d_{\geq 0} & \xrightarrow{\Phi_K} & \mathcal{H}_K \end{array}$$

The diagram expresses that $\psi^* \Phi_K \psi(x) = \Phi_{K\circ\psi}(x)$ for all $x \in \Delta^d$. Note that the vertical maps are invertible so that the feature maps Φ_K and $\Phi_{K\circ\psi}$ *describe each other*. It implies that they are essentially equivalent and that any method using the kernel feature map applied in either of the two domains gives the same result via pull-back kernels.

Proof of Theorem 2.3. By the reproducing property (2.1), for all $f \in \mathcal{H}_K$, for a finite linear combination $\sum_i K(x_i, \cdot)$ we have

$$\begin{aligned}\left\langle \psi^*(f), \sum_i \psi^* K(x_i, \cdot) \right\rangle_{\mathcal{H}_{K\circ\psi}} &= \sum_i \psi^*(f)(\psi^{-1}(x_i)) \\ &= \left\langle f, \sum_i K(x_i, \cdot) \right\rangle_{\mathcal{H}_K}.\end{aligned}$$

As the finite linear combinations $\sum_i K(x_i, \cdot)$ are dense in $\mathcal{H}_K$, it follows that

$$\langle f, g \rangle_{\mathcal{H}_K} = \langle \psi^*(f), \psi^*(g) \rangle_{\mathcal{H}_{K\circ\psi}}$$

for all $f, g \in \mathcal{H}_K$, which proves ψ^* is an isometric isomorphism. The commutativity of the diagram is readily seen from simple evaluations. □

The kernel feature maps in Theorem 2.3 can be generalized to kernel mean embeddings. Let $\mathcal{P}(\mathcal{X})$ denote the space of Borel probability measures on $\mathcal{X}$. The function $\psi : \Delta^d \to \mathbb{S}^d_{\geq 0}$ extends to a function $\psi_* : \mathcal{P}(\Delta^d) \to \mathcal{P}(\mathbb{S}^d_{\geq 0})$ of spaces of probability measures, called the *push-forward* of ψ; see, e.g., Section 3.6 of Bogachev and Ruas [13]. Then the generalization of Theorem 2.3 is stated as follows.

Theorem 2.4. *The following diagram*

$$\begin{array}{ccc} \mathcal{P}(\Delta^d) & \longrightarrow & \mathcal{H}_{K\circ\psi} \\ \big\downarrow{\scriptstyle \psi_*} & & {\scriptstyle \psi^*}\big\uparrow \\ \mathcal{P}(\mathbb{S}^d_{\geq 0}) & \longrightarrow & \mathcal{H}_K \end{array}$$

is commutative where the horizontal maps are kernel mean embeddings.

Proof of Theorem 2.4. It is straightforward from the definition of the push-forward map ψ_* and the generalized reproducing property (2.2). To elaborate, let μ and ν denote the kernel mean embeddings of Δ^d and $\mathbb{S}^d_{\geq 0}$, respectively. For a probability measure $\mathbb{P}$ on Δ^d, we need to show

$$\mu_{\mathbb{P}} = \psi^* \nu_{\psi_* \mathbb{P}},$$

so it is enough to check the right hand side satisfies the generalized reproducing property of $\mu_{\mathbb{P}}$. For $f = \psi^* g \in \mathcal{H}_{K\circ\psi}$, we have

$$\begin{aligned}\langle \psi^* \nu_{\psi_* \mathbb{P}}, f \rangle_{\mathcal{H}_{K\circ\psi}} &= \langle \nu_{\psi_* \mathbb{P}}, g \rangle_{\mathcal{H}_K} && \text{(Theorem 2.3)} \\ &= \mathbb{E}_{X \sim \psi_* \mathbb{P}}[g(X)] && \text{(2.2)} \\ &= \mathbb{E}_{X \sim \mathbb{P}}[f(X)] && \text{(Change of variables)}\end{aligned}$$

The uniqueness of Riesz representer of Hilbert spaces finishes the proof; the diagram commutes. □

We conclude from Theorems 2.3 and 2.4 that all results obtained by kernel methods on Δ^d can be obtained by applying the corresponding methods on $\mathbb{S}^d_{\geq 0}$. This will allow us to analyze compositional data using various *well-studied kernels on the hypersphere* $\mathbb{S}^d$. From here on, we equate Δ^d and $\mathbb{S}^d_{\geq 0}$ and call them *compositional domains*.

Table 2.1: A parametric family of isotropic kernels on $\mathbb{S}^d$ and their universality. The parameters γ and β are positive real numbers, and p is a positive integer. For Matérn kernels, θ stands for $\langle x, y\rangle$, and K_ν is the modified Bessel function of the second kind of order $\nu \in (0, \frac{1}{2}]$.

Kernels	$K(x, y)$	Universal
Linear	$\langle x, y\rangle$	$\times$
Polynomial	$(\gamma\langle x, y\rangle + 1)^p$	$\times$
Gaussian	$\exp\left(-\gamma\|x - y\|_2^2\right)$	$\circ$
von-Mises	$\exp\left(\gamma\langle x, y\rangle\right)$	$\circ$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\theta}{\gamma}\right) K_\nu\left(\frac{\theta}{\gamma}\right)$	$\circ$
Rational quadratic	$(\|x - y\|_2^2 + \gamma^2)^{-\beta}$	$\circ$

2.5 Kernels on Compositional Domains

Having discussed the equivalence between compositional domains, here we study what kernels can be used for the analysis of compositional data. As $\mathbb{S}^d_{\geq 0}$ is a subset of the hypersphere $\mathbb{S}^d$, it is natural to utilize the restriction of kernels on $\mathbb{S}^d$. Their RKHS embeddings are expressed as pull-backs discussed in Section 2.3.3. We start with reviewing examples of kernels on hyperspheres.

2.5.1 Isotropic Kernels on $\mathbb{S}^d$

An alternative name of the dot-product kernel is the isotropic kernel. To be specific, a kernel $K : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$ is said to be *isotropic* if there exists a function $k : [0, \pi] \to \mathbb{R}$ such that

$$K(x, y) = k(\arccos\langle x, y\rangle) \quad \forall x, y \in \mathbb{S}^d,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual dot product in $\mathbb{R}^{d+1}$. Hence, the values of isotropic kernels on $\mathbb{S}^d$ depend only on the *geodesic distance*, or equivalently, on the angle of two input variables. Gneiting [39] provides an extensive survey of these kernels.

Isotropic kernels on spheres have been studied for a long time since Schoenberg [99], and they are broadly used in directional data analysis. For a recent example, see Balasubramanian et al. [10] for goodness-of-fit tests on $\mathbb{S}^d$ with the Gaussian kernel. Note that the Gaussian kernel fits to the definition of the isotropic kernel.

2.5.2 Universal and Characteristic Kernels

As mentioned in Section 2.3.2, the universality or characteristicity of kernels is required to apply kernel mean embedding methods properly. Micchelli et al. [75] provide a complete characterization of isotropic universal kernels on $\mathbb{S}^d$ that have strictly positive coefficients in the Gegenbauer expansions. It is proved that various broadly-used kernels on spheres are universal, and thus it suffices to check the following theorem to utilize them on the compositional domain.

Theorem 2.5. *Let K be a kernel on $\mathbb{S}^d$.*

(i) If K is universal, then the restriction $K|_{\mathbb{S}^d_{\geq 0}}$ is universal.

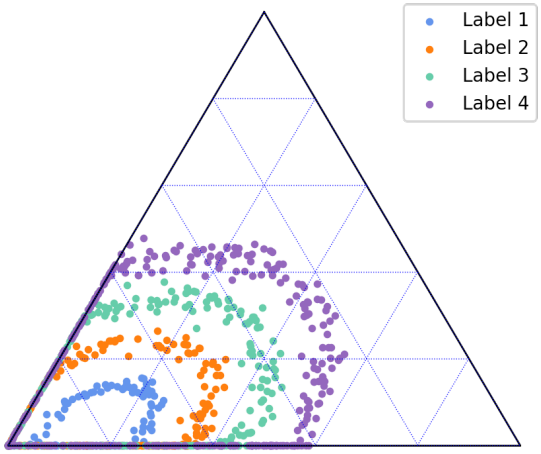


Figure 2.3: Ternary plot of the simulated data with $d = 2$.

(ii) If K is characteristic, then the restriction $K|_{\mathbb{S}^d_{\geq 0}}$ is characteristic.

The proof of Theorem 2.5(i) can be found in, for example, Lemma 4.55 of Steinwart and Christmann [111]. We state the characteristicity in Theorem 2.5(ii) for completeness, although universality is sufficient in practice. Since it is readily proved from the generalized reproducing property (2.2), the proof is omitted here.

Proof. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on $\mathbb{S}^d_{\geq 0}$ and let $i : \mathbb{S}^d_{\geq 0} \to \mathbb{S}^d$ denote the inclusion map. Suppose $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \mathbb{E}_{X\sim\mathbb{Q}}[f(X)]$ for all $f \in \mathcal{H}_{K|_{\mathbb{S}^d_{\geq 0}}}$. By change of variables formula and the pullback theorem 2.1, this implies that

$$\mathbb{E}_{X\sim i_*\mathbb{P}}[f(X)] = \mathbb{E}_{X\sim i_*\mathbb{Q}}[f(X)] \quad \text{for all } f \in \mathcal{H}_K,$$

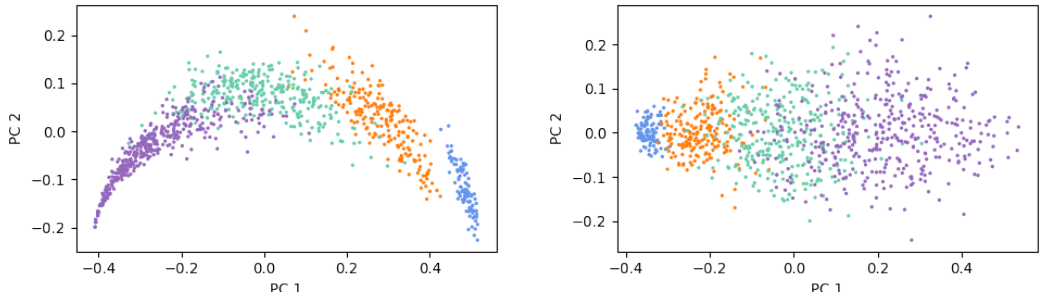
where i_* denotes the pushforward map of measures. We have $i_*\mathbb{P} = i_*\mathbb{Q}$ since K is characteristic, and this implies that $\mathbb{P} = \mathbb{Q}$. □

We summarize some well-known and easily computable isotropic kernels on $\mathbb{S}^d$ in Table 2.1. Their universality properties are also marked for their use in mean embedding methods.

2.6 Empirical Examples

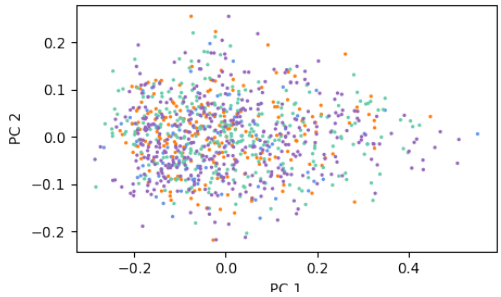
2.6.1 Illustrative Examples

First, we generate simulated compositional data with many zeros to illustrate the effectiveness of the proposed method. Data with sample size 1000 are generated using random samples from d-dimensional multivariate normal distribution with zero mean vector and identity covariance matrix, and then normalized to have a radius of one. After that, four different radius values are multiplied to create four subgroups. The size of each subgroup is proportional to the radius and Gaussian noise with variance inversely proportional to the radius is added. Then we make the data compositional by applying a linear transformation and projecting the points outside of the simplex to the boundary. The detailed description of the data generation process is in the supplementary material, Section 2.8. The simulated data have

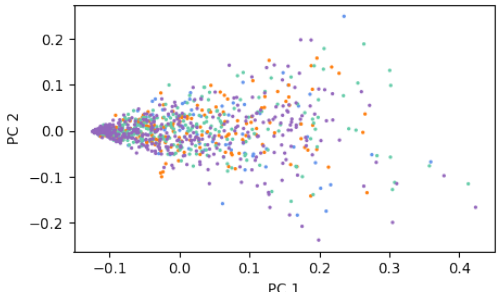


(a) radial transform ($\gamma = 150$)

(b) radial transform ($\gamma = 60$)



(c) clr transform ($\gamma = 0.001$)



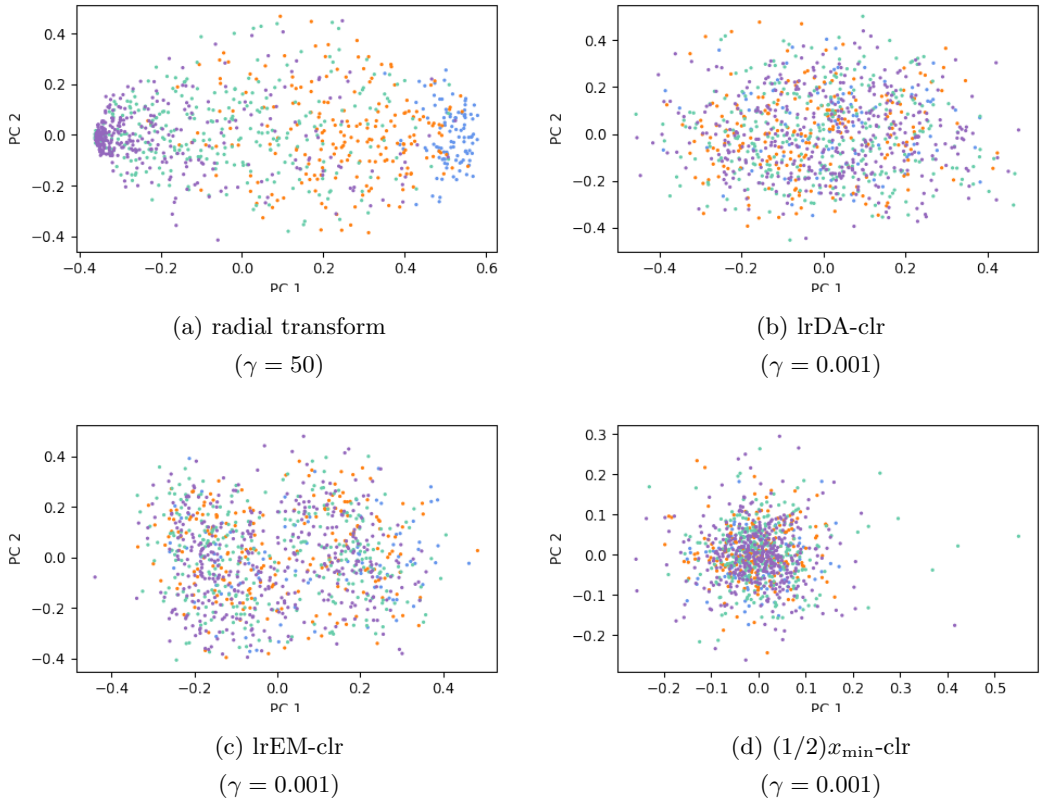
(d) clr transform ($\gamma = 0.005$)

Figure 2.4: Projection plots from kernel PCA with Gaussian kernel using the radial transformed data in (a) and (b), and the clr transformed data in (c) and (d). Each color corresponds to the label of the data shown in Figure 2.3. Here, γ indicates the parameter for Gaussian kernel.

about 40% of zero values. For illustration, Figure 2.3 displays the simulated data for $d = 2$ on a ternary plot. In the actual analysis, we use $d = 100$ and $d = 15$.

In order to see the difference in the geometries of the proposed transformation and the log-ratio method, we implement kernel PCA to the simulated data with $d = 100$. Note that in this case all data points are on the boundary of the simplex. Figure 2.4 shows projection plots from kernel PCA using Gaussian kernel with two different values of the parameter γ, for both the radial transformed data and the clr transformed data. For the clr transformation, we replace the zeros with $(1/2)x_{\min}$ where $x_{\min}$ is the minimum positive value of each composition. It can be clearly seen that the radial transform preserves the separation of the four groups, and particularly in (b) we see that the variance information of the groups is well retained in the embedded space. On the other hand, all meaningful characteristics in the original compositional data disappear in the clr transformed data, as seen in (c) and (d). It is well known that the geometry of the embedded space heavily depends on the kernel parameter even within the same kernel [2]. In this regard, we should point out that the results from the clr transform never become like (a) or (b), regardless of the parameter. It should be also noted that polynomial kernel with $p = 3$, $\gamma = .1$ and von-Mises kernel with $\gamma = 10$ on the radial transformed data yield a similar result to Figure 2.4(b). We refer to the supplementary Section 2.8.2 for use of other kernels and parameters.

We also examine how different zero replacement methods can produce different Aitchison geometry. We implement three methods for the clr transformation, which are lrDA, lrEM, and simple replacement of $(1/2)x_{\min}$, and compared them with the radial transformed data in Figure 2.5. The results of lrDA



(a) radial transform ($\gamma = 50$)

(b) lrDA-clr ($\gamma = 0.001$)

(c) lrEM-clr ($\gamma = 0.001$)

(d) $(1/2)x_{\min}$-clr ($\gamma = 0.001$)

Figure 2.5: Demonstration on how different zero replacement methods can yield vastly different results. For simulated compositional data with $d = 15$, kernel PCA with Gaussian kernel is implemented for radial transformed data in (a) and for clr transformed data in (b)–(d) based on three different zero replacement methods.

and lrEM are produced by the R package `zCompositions` [79]. Due to the computational limitation of the R package, we use the simulated data with $d = 15$ for this figure. Note that in our simulation setting, the lower the dimension is, the more overlap the subgroups have. We can see from Figures 2.5(b)-(d) that kernel PCA with any of the three zero replacement methods fails to distinguish the subgroups, and that the overall shapes of the projected data are quite different from one another. On the contrary, kernel PCA with the radial transformed data in (a) is able to distinguish the subgroups much better.

2.6.2 Quantitative Evaluation of the Proposed Method

We then provide a quantitative assessment of kernel PCA on new synthetic data and real-world data examples. The eigenvalues of the Gram matrix, denoted by $\lambda_1, \ldots, \lambda_n$, are used to measure the effectiveness of kernel PCA. Note that as in linear PCA, eigenvalues of the Gram matrix can be interpreted as the amount of information that each principal component (PC) holds. Thus the number of PCs that are necessary to account for, say 90% of the variability in the data, is calculated as the smallest m such that $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{n} \lambda_i \geq .9$. The smaller this number is, the more efficient dimension reduction we can achieve by kernel PCA. We implement kernel PCA with the Gaussian kernel after the radial and the clr transformation, where we replace zeros with $(1/2)x_{\min}$ before the clr transformation.

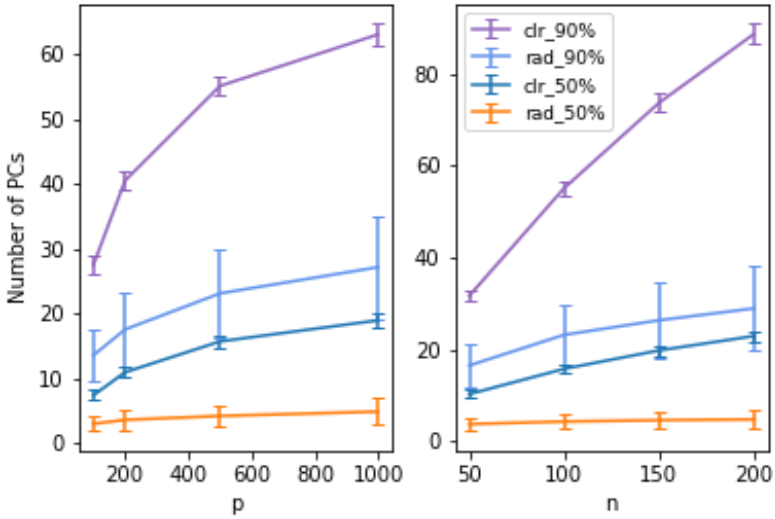


Figure 2.6: Number of PCs needed to capture the variability in synthetic data. The sample size is fixed at $n = 100$ for the left panel and the dimension is $p = 500$ on the right panel.

Synthetic Data

We simulate high-dimensional compositional data following Te Beest et al. [114] with slight modifications to reflect a much higher percentage of zeros in real-world microbiome datasets. The data are generated as a matrix of counts X, whose (i, j)-entry is drawn from a negative binomial distribution with mean μ_{ij} and variance $\mu_{ij} + \mu_{ij}^2$, where each μ_{ij} is modeled by a log-linear model

$$\log \mu_{ij} = a_i + t_j + b_j x_i,$$

$i = 1, \ldots, n$, $j = 1, \ldots, p$ where n is the number of samples and p is the number of taxa. The term a_i reflects the size of total counts and is drawn from $N(-1.5, 1)$, t_j reflects the abundance of taxon j and is drawn from $N(-0.5, 2)$, and x_i is a binary 0-1 variable representing two different treatment groups of equal size with the effect size b_j on taxon j. Twenty percent of p taxa are made differentially abundant at random with equal probability, being either up- or down-regulated by setting b_j to be $\log 3$ or $-\log 3$. After data generation, taxa present in less than five samples are considered meaningless and removed.

Typically, the simulated data have about $69.5 \pm 1.5\%$ of zeros which is more or less similar to real data examples in Table 2.2. Figure 2.6 displays the means and standard errors of the number of PCs needed to explain 50% and 90% of the total variance based on 100 replications. It can be seen that the radial transformation shows far better performance than the clr transformation in all cases. From the perspective of Section 2.2, the result indicates that zero-replacements in Aitchison geometry *disperse* data erratically. This also underpins the poor projection plots of clr transformed data in Section 2.6.1.

Real Data Examples

We also analyze real-world microbiome datasets, whose availability is listed in the supplementary Section 2.8.4. Their attributes such as n, p, and the percentage of zeros are presented in Table 2.2, with the number of PCs needed to explain 50%, and 90% of the total variation of the data using the radial or clr transformation, respectively. From Table 2.2, it is evidenced that the radial transform (rad-50% or rad-90%) shows better performance than the clr transform (clr-50% or clr-90%) with respect to the efficiency of dimension reduction.

Table 2.2: Number of PCs needed for real data examples.

Dataset	n	p	Zero proportion	rad-50%	rad-90%	clr-50%	clr-90%
Arumugam et al. [8]	280	553	67%	1	8	1	18
Carrieri et al. [16][1]	1200	186	58%	4	15	24	111
Carrieri et al. [16][1]	278	186	69%	4	16	19	82
Charlson et al. [18]	60	856	89%	4	17	9	39
Gimblet et al. [35]	632	1860	95%	1	8	8	142
Hayden et al. [46]	1279	643	93%	5	35	22	128
Schiffer et al. [98]	381	780	76%	3	17	3	52

[1] The article provides two datasets, one from the Canada cohort (first) and the other from the UK cohort (second).

2.7 Conclusions and Discussions

In this chapter, we showed that it is possible to use kernel-based learning for compositional data via radial transformation and pointed out that the traditional log-ratio approaches might lose their justification when applied to the compositional data with high proportion of zeros. We also provided an appropriate mathematical framework for theoretical justification and demonstrated the idea with examples. We believe that many scientific questions regarding compositional data will be answered by newly enabled statistical inference and analysis using kernels, such as graphical models, hypothesis testing, and regression models.

A unique feature of microbiome data is that each variable in the composition, namely bacterial taxon, corresponds to a node in the phylogenetic tree. One of the most common ways to define a distance between two microbiome compositions is to measure the β-diversity based on the tree [65], which is called the UniFrac distance. Principal coordinates analysis, equivalently multi-dimensional scaling, is then used to obtain the leading eigenspace to find the best low-dimensional representation of the data. It is straightforward to see that the UniFrac distance matrix essentially plays the same role as the kernel matrix in kernel PCA. Then it is natural to wonder about the properties of this "UniFrac kernel", which can be an interesting direction for future research.

2.8 Supplementary Materials

This section provides supplementary materials for this chapter. We describe the detailed process for our simulated data generation and provide additional experimental results that support our discussions in Section 2.6. We also give data availability for real data experiments conducted in Section 2.6.2.

2.8.1 Simulated Data Generation Process

This section covers the detailed description for the generation of simulated data in Section 2.6.1.

The experimental data are generated on the d-dimensional simplex Δ^d with a hyperspherical shape and four clusters with different radii. Each cluster is generated through an identical procedure but with a different radii. We describe the detailed steps as follows:

Step 1 (Spherical generation). Let r denote the primary radius assigned to each cluster. Initially, r is set as $1, 2, 3, 4$, and we generate, for each cluster, $100 \times r$ random samples drawn from the multivariate normal distribution $N(\mathbf{0}_d, \mathbf{I}_d)$, labeled differently by their cluster. Then, we normalize them onto the hyperspheres $\mathbb{S}^d$ and add Gaussian noises with $N(\mathbf{0}_d, (0.01/r)\mathbf{I}_d)$ respectively. Hence, each cluster's sample size is proportional to the radius, whereas the Gaussian noise is inversely proportional to the radius.

Step 2 (Scaling and shifting). The scale parameter $\dfrac{r}{10\sqrt{d/4 + 0.5}}$ is multiplied to each cluster. Then, we linearly shift the data by adding the d-dimensional vector $[0.15/(d-1), \cdots, 0.15/(d-1), 0.04/(d-1)]$.

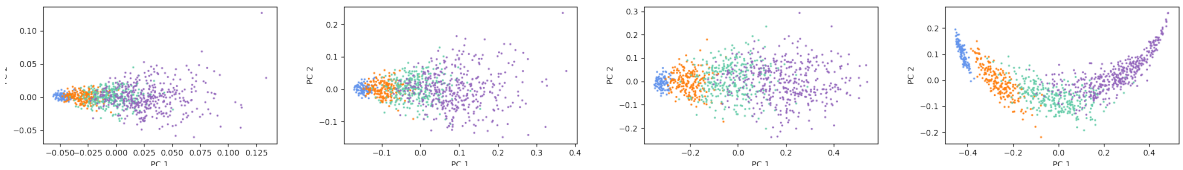
Step 3 (Projection to the simplex Δ^d). Among the whole data, we replace the component values below zero by zero; i.e., they are projected to the boundary of Δ^d. Until now, the generated data live in $\mathbb{R}^d$. Finally, we project the resulting data to $\Delta^d \subset \mathbb{R}^{d+1}$ by creating the last coordinate have the value of $1-$ (sum of the other components). Note that the last sum never exceeds 1 due to our appropriately chosen scale parameters and the variance of the Gaussian noise.

2.8.2 Kernel PCA with Various Kernels and Parameters Using Radial Transformed and clr-Transformed Data

In this section, we present additional results from kernel PCA with various kernels and parameters regarding Figure 2.4 in the body of the chapter. We use the same radial transformed and clr transformed data with $(1/2)x_{\min}$ zero replacements. For kernels, Gaussian, polynomial, and von-Mises kernels are used. For the polynomial kernel, the degree $p = 3$ is used. The parameter γ ranges from 1 to 100 for the radial transformed data, and from 0.0001 to 0.01 for the clr transformed data. The difference in ranges is due to the different magnitudes in the transformed data.

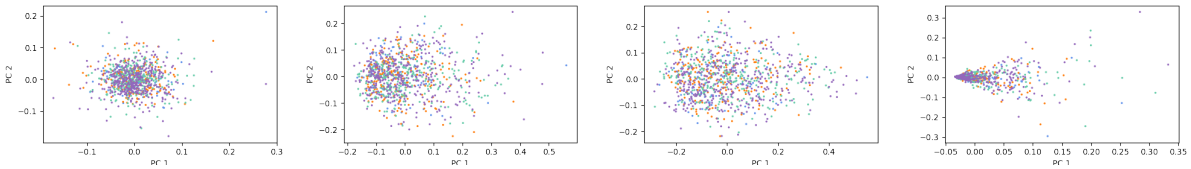
Gaussian kernel

Kernel PCA projection plots using Gaussian kernel for the radial transformed data are given in Figure 2.7. For kernel PCA results with the clr transformed data, see Figure 2.8. We still observe with the various parameter choices that the obvious manifold pattern in our dataset is distorted in clr-transformed data after zero replacement. The displayed parameters are chosen based on the plots being clearly visible.



(a) radial transform ($\gamma = 1$) (b) radial transform ($\gamma = 10$) (c) radial transform ($\gamma = 50$) (d) radial transform ($\gamma = 100$)

Figure 2.7: Projection plots from kernel PCA with Gaussian kernel using the radial transformed data by various values of parameter.



(a) clr transform ($\gamma = 0.0001$)

(b) clr transform ($\gamma = 0.0005$)

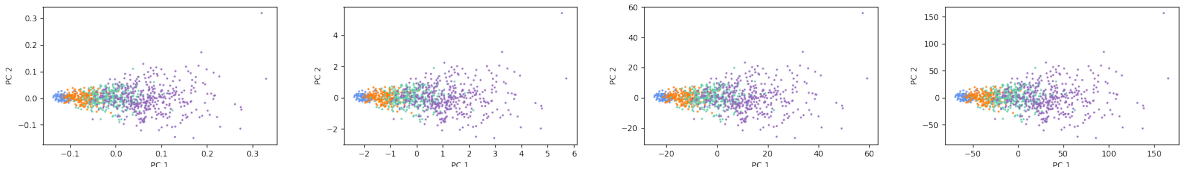
(c) clr transform ($\gamma = 0.001$)

(d) clr transform ($\gamma = 0.01$)

Figure 2.8: Projection plots from kernel PCA with Gaussian kernel using the clr transformed data by various values of parameter.

Polynomial kernel

With the same criteria but the kernel is changed, we display the results in Figure 2.9 and Figure 2.10. The results are very similar to the Gaussian kernel case.



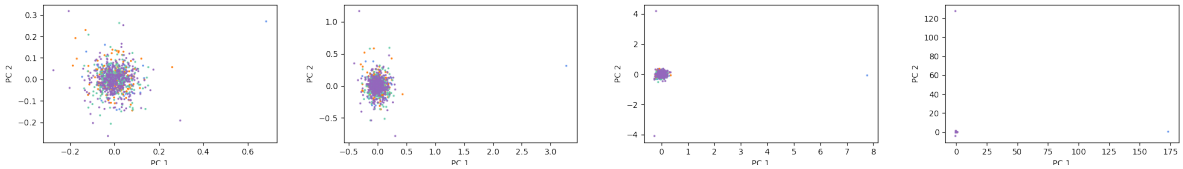
(a) radial transform ($\gamma = 1$)

(b) radial transform ($\gamma = 10$)

(c) radial transform ($\gamma = 50$)

(d) radial transform ($\gamma = 100$)

Figure 2.9: Projection plots from kernel PCA with polynomial kernel using the radial transformed data by various values of parameter.



(a) clr transform ($\gamma = 0.0001$)

(b) clr transform ($\gamma = 0.0005$)

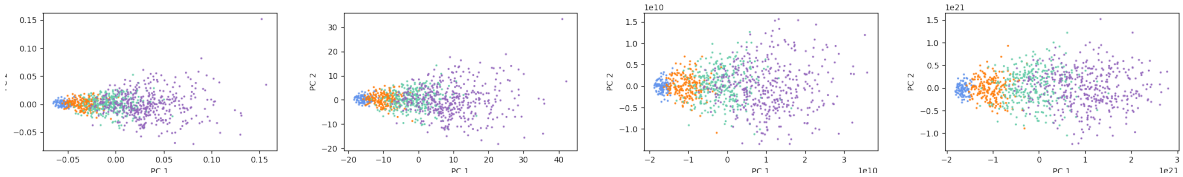
(c) clr transform ($\gamma = 0.001$)

(d) clr transform ($\gamma = 0.01$)

Figure 2.10: Projection plots from kernel PCA with polynomial kernel using the clr transformed data by various values of parameter.

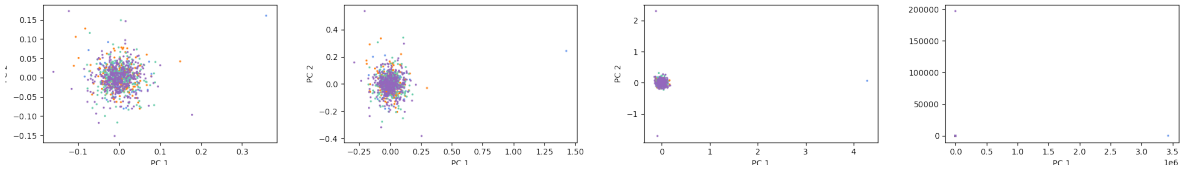
von-Mises kernel

The results using von-Mises kernel are displayed in Figure 2.11 and Figure 2.12. The results for the radial transformed case is similar to the other kernels. Since the von-Mises kernel takes exponential of inner products, the numerical results are very unstable after clr transformation.



(a) radial transform ($\gamma = 1$) (b) radial transform ($\gamma = 10$) (c) radial transform ($\gamma = 50$) (d) radial transform ($\gamma = 100$)

Figure 2.11: Projection plots from kernel PCA with von-Mises kernel using the radial transformed data by various values of parameter.



(a) clr transform ($\gamma = 0.0001$) (b) clr transform ($\gamma = 0.0005$) (c) clr transform ($\gamma = 0.001$) (d) clr transform ($\gamma = 0.01$)

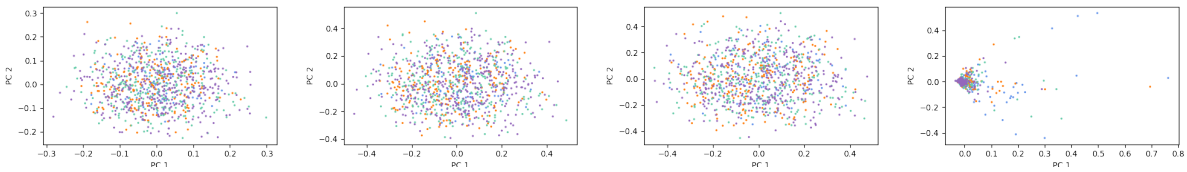
Figure 2.12: Projection plots from kernel PCA with von-Mises kernel using the clr transformed data by various values of parameter.

2.8.3 Kernel PCA with Various Kernels and Parameters Using lrDA and lrEM Zero Replacement Methods

In this section, we present additional results from kernel PCA with various kernels and parameters regarding to Figure 2.5. We use the same lrDA-clr and lrEM-clr transformed data. Again, the degree is $p = 3$ for the polynomial kernel.

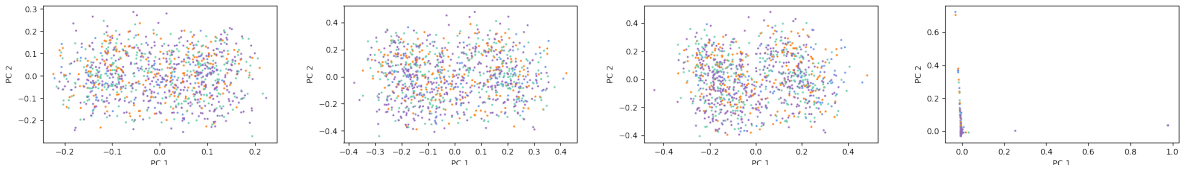
Gaussian kernel

Using the Gaussian kernel with various parameters, the kernel PCA projection plots for the lrDA-clr transformed data is given in Figure 2.13 and for the lrEM-clr transformed data is given in Figure 2.14. We can observe that changing the zero replacement method does not prevent the data distortion problem pointed out in Section 2.2.



(a) lrDA-clr ($\gamma = 0.0001$) (b) lrDA-clr ($\gamma = 0.0005$) (c) lrDA-clr ($\gamma = 0.001$) (d) lrDA-clr ($\gamma = 0.01$)

Figure 2.13: Projection plots from kernel PCA with Gaussian kernel using the lrDA-clr transformed data by various values of parameter.

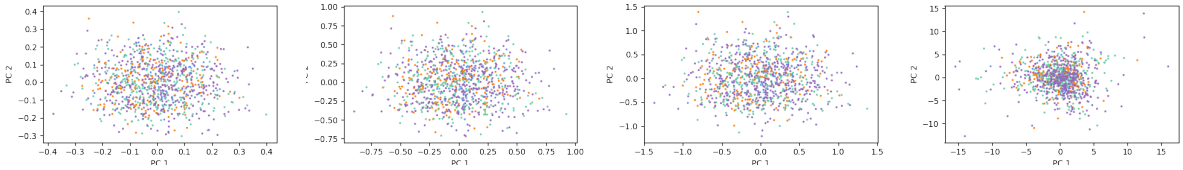


(a) lrEM-clr ($\gamma = 0.0001$) (b) lrEM-clr ($\gamma = 0.0005$) (c) lrEM-clr ($\gamma = 0.001$) (d) lrEM-clr ($\gamma = 0.01$)

Figure 2.14: Projection plots from kernel PCA with Gaussian kernel using the lrEM-clr transformed data by various values of parameter.

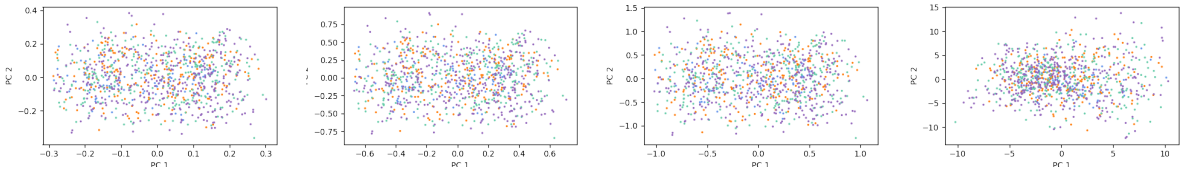
Polynomial kernel

We obtain similar results to the Gaussian kernel case, given in Figure 2.15 and Figure 2.16.



(a) lrDA-clr ($\gamma = 0.0001$) (b) lrDA-clr ($\gamma = 0.0005$) (c) lrDA-clr ($\gamma = 0.001$) (d) lrDA-clr ($\gamma = 0.01$)

Figure 2.15: Projection plots from kernel PCA with polynomial kernel using the lrDA-clr transformed data by various values of parameter.

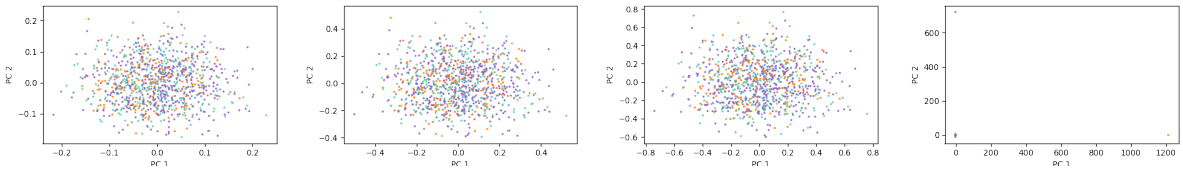


(a) lrEM-clr ($\gamma = 0.0001$) (b) lrEM-clr ($\gamma = 0.0005$) (c) lrEM-clr ($\gamma = 0.001$) (d) lrEM-clr ($\gamma = 0.01$)

Figure 2.16: Projection plots from kernel PCA with polynomial kernel using the lrEM-clr transformed data by various values of parameter.

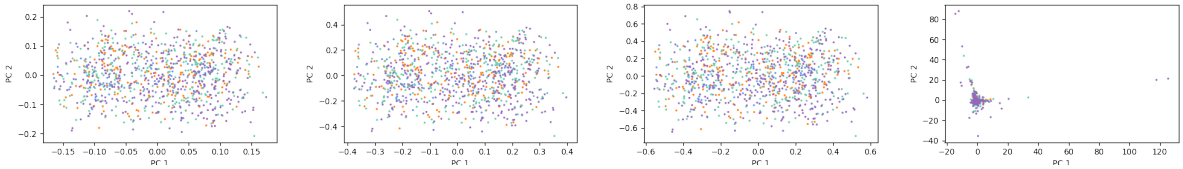
von-Mises kernel

As before, the kernel PCA fails to capture the manifold structure of the simulated data, and the numerical results are fairly unstable due to the exponential computation. See Figure 2.17 and Figure 2.18 for the results.



(a) lrDA-clr ($\gamma = 0.0001$) (b) lrDA-clr ($\gamma = 0.0005$) (c) lrDA-clr ($\gamma = 0.001$) (d) lrDA-clr ($\gamma = 0.01$)

Figure 2.17: Projection plots from kernel PCA with von-Mises kernel using the lrDA-clr transformed data by various values of parameter.



(a) lrEM-clr ($\gamma = 0.0001$) (b) lrEM-clr ($\gamma = 0.0005$) (c) lrEM-clr ($\gamma = 0.001$) (d) lrEM-clr ($\gamma = 0.01$)

Figure 2.18: Projection plots from kernel PCA with von-Mises kernel using the lrEM-clr transformed data by various values of parameter.

2.8.4 Data availability

Specific data availability for real data examples in section 2.6.2 are summarized in Table 2.3.

Table 2.3: Data availability for real data examples.

Dataset	Data Source
Hayden et al. [46]	'BONUS-CF (WGS)' dataset from MicrobiomeDB.org
Gimblet et al. [35]	'Experimental cutaneous leishmaniasis' dataset from MicrobiomeDB.org
Arumugam et al. [8]	'enterotype' dataset in R package phyloseq
Carrieri et al. [16]	Supplementary material of the referenced article
Charlson et al. [18]	'throat.otu.tab' dataset in R package GUniFrac
Schiffer et al. [98]	'vaginal.otu.tab' dataset in R package GUniFrac

Chapter 3. Conditional Covariance Operator of RKHS and Generalized Kernel Dimension Reduction

3.1 Introduction

In Chapter 2, we demonstrated that a wide range of kernel methods, from classical ones like kernel principal component analysis (KPCA) [100], kernel fisher discriminant analysis (KFDA) [76], kernel ridge regression (KRR), and support vector machines (SVM) [101] to recent advances in kernel mean embedding methods such as kernel two-sample test via maximal mean discrepancy (MMD)[45] and Hilbert-Schmidt independence criterion (HSIC) [44], apply successfully to compositional data with taking care of geometric structures. These kernel methods on the compositional domain naturally treat the prevalent zero values in compositional data in practice, providing a more rigorous class of compositional data analysis methods than the traditional log-ratio approach.

However, the application of these kernel methods to practical datasets is not without challenges; notably, an issue of *high dimensionality* arises. This is particularly evident in microbiome data, our motivating dataset, where the number of variables often significantly exceeds the number of samples. It is known that numerous kernel methods with a popular choice of kernels suffer from the curse of dimensionality; for example, Donhauser et al. [23] demonstrated that the kernel ridge regression with rotationally invariant kernels can only fit low-degree polynomials in high dimensions, and Ramdas et al. [88] addressed the decreasing power of the kernel two-sample test with Gaussian and Laplace kernels in high dimensions. Furthermore, because the common data standardization process destroys the structure of compositional data, we cannot apply standardization before using kernel methods with popular distance-based kernels, resulting in numerical biases in high-variance coordinates. As a result, applying those kernel methods directly to high-dimensional compositional data may lead to suboptimal performances.

One simple and intuitive way of solving this problem is to develop kernel-based dimension reduction methods that retain as much data information as possible. In addition, it is desirable to have *interpretable* dimension reduction results in terms of original variables since it is vital in many biological applications. Classical RKHS methods, such as KPCA or KFDA, fail to produce interpretable results since the variable information is lost during the first RKHS embedding. In contrast, the kernel dimension reduction (KDR) method of Fukumizu et al. [34] offers an interpretable dimension reduction in Euclidean space because their method applies RKHS embedding *after* orthogonal projection of data. In particular, the KDR method aims for linear *sufficient dimension reduction* (SDR) [60], representing that the data projection does not change the conditional distribution of the response variable:

$$Y \perp\!\!\!\perp X \mid B^T X,$$

where X is a vector of covariates, Y is a response, and B is an orthogonal projection matrix. This framework of linear SDR primarily focuses on finding the best orthogonal projection spanning an SDR *subspace*. However, since orthogonal projections of compositional data do not retain their compositional nature, there is a need for a more suitable approach. It is crucial to first define a class of suitable and interpretable dimension reduction maps designed for compositional data to retain interpretability. In Chapters 4 and 5, we shall define particularly structured projection maps for compositional data to accomplish this. Considering their unique relative structures, we will describe appropriate classes of

variable selections and dimension reduction maps.

Instead, this chapter focuses on expanding the KDR theory to encompass arbitrary measurable projections of particular interest. The main advantage of the KDR approach over other existing nonlinear sufficient dimension reduction methods is that it maintains the interpretable structure of the dimension reduction functions. Lee et al. [53] also proposed two generalized nonlinear SDR methods, generalized sliced inverse regression (GSIR) and generalized sliced average variance estimator (GSAVE), but these methods also sacrifice the interpretability of the projection result, even though their methods can be applied to compositional data without replacing zeros. In contrast, we will see throughout this chapter that the KDR approach is capable of generalizing to the broader family of specifically structured dimension reductions, not confined to orthogonal projections. This signifies that the simple intuition behind the KDR theory generalizes well: finding a dimension reduction map p that minimizes the conditional covariance of Y given $p(X)$ will result in a desirable dimension reduction.

In particular, we generalize that minimizing the conditional covariance operator of Y given $p(X)$ among $p \in \mathcal{F}$ also approaches SDR in Section 3.4, where $\mathcal{F}$ is a certain family of dimension reduction functions for X. Furthermore, if the family $\mathcal{F}$ satisfies some continuity assumptions and is equipped with a metric that makes it a compact metric space, we prove in Section 3.6 that the empirical M-estimator for our generalized kernel dimension reduction is consistent. Such families $\mathcal{F}$ will include the Stiefel manifold of orthogonal matrices and our compositional dimension reductions will be defined in Chapter 5. We elaborate on all the details for our generalization process because the original theory makes use of some unique properties of orthogonal matrices. As a result, we will confirm that our generalized theory holds surprisingly under milder assumptions than originally assumed in Fukumizu et al. [34]. We should also note that our theory is formulated in general terms of nonlinear dimension reductions: its applicability is not limited to compositional data.

The rest of this chapter is outlined as follows. We first extensively review the theory of kernel mean embeddings and their interpretation via function-valued integration in Section 3.2. We define cross-covariance operators of RKHS in Section 3.3 and interpret them as a covariance of embedded random variables in RKHS. Section 3.4 introduces the theory of conditional covariance operator and prove our generalized version of the KDR method, as well as some discussions on the generalized unsupervised KDR framework. We provide a thorough computation process for the empirical estimate of the conditional covariance operator in Section 3.5. Then, we prove that the M-estimator of generalized kernel dimension reduction is statistically consistent in Section 3.6. Section 3.7 concludes this chapter with discussions on potential applications of our generalized result.

3.2 Kernel Mean Embedding and Function-Valued Integrations

Kernel mean embedding is pivotal in the analysis of probability measures with the refined structure of reproducing kernel Hilbert spaces. Once we embed a random variable using an RKHS embedding, taking the expectation of the embedded variable turns out to be an embedding of the original probability measure. To describe this precisely, we need notions of RKHS-valued integrations and random variables defined on an RKHS. As the kernel mean embedding and its rigorous foundations will play a crucial role in our proposed method and theory, we briefly review their definitions and essential properties that will be needed later.

Throughout the chapter, we let $\mathcal{X}$ denote a *compact* subset of Euclidean space and let $\mathcal{Y}$ denote the domain of responses. The compactness of $\mathcal{X}$ is not restrictive, often assumed in the literature, and fit our

purpose: the compositional domain Δ^m is compact. As $\mathcal{X}$ is compact, it is natural to assume that $\mathcal{Y}$ is also compact. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be kernels (recall that we assumed kernels are continuous, symmetric, and positive definite in Chapter 2) on $\mathcal{X}$ and $\mathcal{Y}$, respectively. These kernels uniquely define the associated RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively, with the reproducing property (2.1). We denote the associated feature maps by $\phi(x) = k_{\mathcal{X}}(x, \cdot) \in \mathcal{H}_{\mathcal{X}}$ and $\psi(y) = k_{\mathcal{Y}}(y, \cdot) \in \mathcal{H}_{\mathcal{Y}}$.

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a joint random vector, with the joint distribution $\mathbb{P}_{XY}$. We also denote the marginal distributions of X and Y by $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. Since we always deal with *continuous* kernels on compact domains, we have the following boundedness:

$$\mathbb{E}_X[k_{\mathcal{X}}(X, X)] < \infty \quad \text{and} \quad \mathbb{E}_Y[k_{\mathcal{Y}}(Y, Y)] < \infty. \tag{3.1}$$

Note that most kernel methods in practice uses continuous and bounded kernels, so our assumption is very weak. Such a minimal assumption ensures that the corresponding RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are continuously embedded in the $L^2(\mathbb{P}_X)$ and $L^2(\mathbb{P}_Y)$, respectively, since

$$\mathbb{E}_X[f(X)^2] = \mathbb{E}_X[\langle f, k_{\mathcal{X}}(X, \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}^2] \leq \|f\|_{\mathcal{H}_{\mathcal{X}}}^2 \mathbb{E}_X[k_{\mathcal{X}}(X, X)]$$

for all $f \in \mathcal{H}_{\mathcal{X}}$. In addition, the boundedness (3.1) ensures the existence of *kernel mean embeddings*,

$$\mu : \mathcal{P}(\mathcal{X}) \to \mathcal{H}_{\mathcal{X}}, \qquad \mathbb{P} \mapsto \mu_{\mathbb{P}} \in \mathcal{H}_{\mathcal{X}} : x \mapsto \int_{\mathcal{X}} k_{\mathcal{X}}(x, z) d\mathbb{P}(z), \tag{3.2}$$

where $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$ and $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. Indeed, given a probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, the linear functional $\mathcal{L}_{\mathbb{P}}(f) := \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ on $\mathcal{H}_{\mathcal{X}}$ is bounded because

$$\begin{aligned} |\mathcal{L}_{\mathbb{P}}(f)| \leq \mathbb{E}_{X\sim\mathbb{P}}[|f(X)|] = \mathbb{E}_{X\sim\mathbb{P}}[\langle f, k_{\mathcal{X}}(X, \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}] &\leq \|f\|_{\mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k_{\mathcal{X}}(X, X)}] \\ &\leq \|f\|_{\mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X\sim\mathbb{P}}[k_{\mathcal{X}}(X, X)], \end{aligned}$$

here the Hölder's inequality is applied twice. Then, the Riesz representer $\rho_{\mathbb{P}} \in \mathcal{H}_{\mathcal{X}}$ of $\mathcal{L}_{\mathbb{P}}$ exists and satisfies, for $k_{\mathcal{X}}(x, \cdot) \in \mathcal{H}_{\mathcal{X}}$,

$$\rho_{\mathbb{P}}(x) = \langle \rho_{\mathbb{P}}, k_{\mathcal{X}}(x, \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_{X\sim\mathbb{P}}[k_{\mathcal{X}}(x, X)] = \int_{\mathcal{X}} k_{\mathcal{X}}(x, z) d\mathbb{P}(z) = \mu_{\mathbb{P}}(x).$$

Therefore, $\mu_{\mathbb{P}} = \rho_{\mathbb{P}}$ is a well defined element in $\mathcal{H}_{\mathcal{X}}$ for every probability measure $\mathbb{P}$ on $\mathcal{X}$, representing an embedding of $\mathbb{P}$ into $\mathcal{H}_{\mathcal{X}}$ by a kernel function $k_{\mathcal{X}}$.

3.2.1 Bochner Integral and Their Basic Properties

The integration in the definition (3.2) of the mean embedding $\mu_{\mathbb{P}}$ can be written instead as a *function-valued integration*,

$$\mu_{\mathbb{P}}(\cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\cdot, z) d\mathbb{P}(z) \in \mathcal{H}_{\mathcal{X}}. \tag{3.3}$$

Here, $k_{\mathcal{X}}(\cdot, z)$ is an element of $\mathcal{H}_{\mathcal{X}}$, and the association $z \mapsto k_{\mathcal{X}}(\cdot, z) \in \mathcal{H}_{\mathcal{X}}$ can be thought of as a function-valued continuous function $f : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}$ (recall that $k_{\mathcal{X}}$ is continuous), where $\mathcal{H}_{\mathcal{X}}$ is equipped with a Borel σ-field. The foundation of such a function-valued integration is established with a more general language of Banach spaces since only completeness and the property of norms are required to develop the relevant generalized theory. It is called a *Bochner integral*, and we briefly review its construction and properties in this subsection. See Section 2.6 of Hsing and Eubank [48] for detailed information and related results.

Let $(\mathcal{X}, \mathcal{B}, \nu)$ be a measure space and let f be a function on $\mathcal{X}$ that takes values in a Banach space $\mathbb{B}$, equipped with a Borel σ-field. We can proceed a similar construction of Lebesgue integration of scalar-valued functions as follows.

Definition 3.1 (Simple functions)**.** A function $f : \mathcal{X} \to \mathbb{B}$ is called *simple* if its range consists of only finitely many points. If we write the range as $\mathrm{ran}(f) = \{h_1, \ldots, h_n\} \subset \mathbb{B}$ and $E_j = f^{-1}(\{h_j\})$, then f is represented as a finite $\mathbb{B}$-linear combination of indicator functions

$$f(x) = \sum_{j=1}^{n} \mathbb{1}_{E_j}(x) h_j \in \mathbb{B},$$

where $\mathbb{1}_{E_i}$ denotes the indicator function.

Every simple function is represented uniquely as above, called the *standard representation* [30] of f. Note that the E_j above are measurable if the function f is measurable.

Definition 3.2 (Bochner integral of simple measurable functions)**.** If a simple measurable function f has standard representation $f(x) = \sum_{j=1}^{n} \mathbb{1}_{E_j}(x) h_j$ and it satisfies $\nu(E_j) < \infty$ for all j, then f is said to be *Bochner integrable* with its *Bochner integral*

$$\int_{\mathcal{X}} f \, d\nu = \sum_{j=1}^{n} \nu(E_j) h_j.$$

Extension of the Bochner integral of simple functions to the general measurable functions can be stated with the ordinary scalar-valued Lebesgue integration as follows.

Definition 3.3 (Bochner integrable functions)**.** A measurable function $f : \mathcal{X} \to \mathbb{B}$ is called *Bochner integrable* if there exists a sequence $\{f_k\}$ of simple Bochner integrable functions such that

$$\lim_{k \to \infty} \int_{\mathcal{X}} \|f_k - f\|_{\mathbb{B}} \, d\nu = 0.$$

In this case, we define the Bochner integral of f by

$$\int_{\mathcal{X}} f \, d\nu := \lim_{k \to \infty} \int_{\mathcal{X}} f_k \, d\nu. \tag{3.4}$$

The right hand side of the definition (3.4) is well-defined, which is a straightforward consequence of the completeness of the Banach space $\mathbb{B}$. If $\mathbb{B}$ is a separable Hilbert space, then its Bochner integrability is readily checked using the following theorem, which we omit the proof.

Theorem 3.4 (see [48, Theorem 2.6.5])**.** *If $\mathbb{B}$ is a separable Hilbert space, $f : \mathcal{X} \to \mathbb{B}$ is measurable, and $\int_{\mathcal{X}} \|f\|_{\mathbb{B}} \, d\nu < \infty$, then f is Bochner integrable.*

Recall that our working space is a compact domain $\mathcal{X}$ with a continuous kernel $k_{\mathcal{X}}$ so that the corresponding RKHS $\mathcal{H}_{\mathcal{X}}$ is always separable. This is a consequence of the famous Mercer's theorem [74] that explicitly describes a countable orthonormal basis of $\mathcal{H}_{\mathcal{X}}$ by eigenfunctions of a corresponding kernel integral operator. Then, since the function $z \mapsto k_{\mathcal{X}}(\cdot, z) \in \mathcal{H}_{\mathcal{X}}$ is continuous and $\int_{\mathcal{X}} \|k_{\mathcal{X}}(\cdot, z)\|_{\mathcal{H}_{\mathcal{X}}} \, d\mathbb{P}(z) = \int_{\mathcal{X}} \sqrt{k_{\mathcal{X}}(z, z)} \, d\mathbb{P}(z) < \infty$ is integrable for all probability measures $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, the definition of $\mu_{\mathbb{P}}$ using a Bochner integral in (3.3) is well-defined, justifying a notation $\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}}[k_{\mathcal{X}}(\cdot, X)] \in \mathcal{H}_{\mathcal{X}}$. This definition also gives an interpretation that $\mu_{\mathbb{P}}$ is indeed a *mean* of the feature embedding of the random variable $X \in \mathcal{X}$ with the distribution $\mathbb{P}$. This perspective will be further clarified in the next subsection.

We end this subsection with presenting a useful and intuitive result on linear transformation of Bochner integrable functions. Since the $\mathbb{B}$-valued Bochner integral is essentially an infinite linear combination of elements in $\mathbb{B}$, continuous linear transformations will be compatible with the integrations:

Theorem 3.5 ([48, Theorem 3.1.7])**.** *Let $\mathbb{B}_1, \mathbb{B}_2$ be Banach spaces and let $f : \mathcal{X} \to \mathbb{B}_1$ be a Bochner integrable function. If $L : \mathbb{B}_1 \to \mathbb{B}_2$ is a bounded linear transformation of Banach spaces, then Lf is also Bochner integrable and*

$$L\left(\int_{\mathcal{X}} f \, d\nu\right) = \int_{\mathcal{X}} Lf \, d\nu.$$

Proof. Letting a sequence $\{f_k\}$ of simple functions whose integration converges to the Bochner integral $\int_{\mathcal{X}} f \, d\nu$, it is straightforward to check the equality based on Definition 3.3. □

3.2.2 Random Elements in a Hilbert Space and the Central Limit Theorem

Once we have a random variable $X \in \mathcal{X}$ from a probability space $(\Omega, \mathcal{F}, P)$, we may consider passing X through the RKHS feature map $\phi : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}$ defined by a kernel $k_{\mathcal{X}}$ on $\mathcal{X}$. As the composition $k_{\mathcal{X}}(X, \cdot) = \phi \circ X : \Omega \to \mathcal{H}_{\mathcal{X}}$ is Borel measurable, such a random function $k_{\mathcal{X}}(X, \cdot)$ can be viewed as a Hilbert space-valued random variable. It is then natural to expect that we may generalize the notion and related theories of random variables to those defined on Hilbert spaces. We give a simple but essential generalization and its large sample theory as follows.

Definition 3.6. Let $\mathcal{H}$ be a Hilbert space equipped with the Borel σ-field generated by its inner product. Given a probability space $(\Omega, \mathcal{F}, P)$, a *random element* χ on $\mathcal{H}$ is a measurable map $\chi : \Omega \to \mathcal{H}$. We simply denote this as $\chi \in \mathcal{H}$.

Clearly, the RKHS embedding $k_{\mathcal{X}}(X, \cdot)$ of a random variable X is a random element of $\mathcal{H}_{\mathcal{X}}$, which will be our central interest. If χ is a random element of a separable Hilbert space, then Theorem 3.4 suggests to define their integration.

Definition 3.7. If $\mathcal{H}$ is separable and $\mathbb{E}[\|\chi\|_{\mathcal{H}}] < \infty$, the *mean element* of χ is defined as the Bochner integral

$$\mathbb{E}[\chi] := \int_{\Omega} \chi \, dP \in \mathcal{H}.$$

Note that Theorem 3.5 implies that we may *interchange* the expectation and the inner product: for all $f \in \mathcal{H}$,

$$\langle \mathbb{E}[\chi], f \rangle_{\mathcal{H}} = \mathbb{E}\left[\langle \chi, f \rangle_{\mathcal{H}}\right]. \tag{3.5}$$

This also indicates that $\mathbb{E}[\chi]$ is the Riesz representer of the bounded linear functional $f \mapsto \mathbb{E}\left[\langle \chi, f \rangle_{\mathcal{H}}\right]$ on $\mathcal{H}$, whose boundedness is established by the condition $\mathbb{E}[\|\chi\|_{\mathcal{H}}] < \infty$.

In case $\chi = k_{\mathcal{X}}(X, \cdot) \in \mathcal{H}_{\mathcal{X}}$, where X is a random variable with its distribution $\mathbb{P}_X$, we have

$$\mathbb{E}[\chi] = \mathbb{E}[k_{\mathcal{X}}(X, \cdot)] = \int_{\Omega} k_{\mathcal{X}}(X, \cdot) \, dP = \int_{\mathcal{X}} k(x, \cdot) \, d\mathbb{P}_X(x)$$

by the change of variables formula. Therefore,

$$\mathbb{E}[\chi] = \mathbb{E}_{X \sim \mathbb{P}_X}[k_{\mathcal{X}}(X, \cdot)] = \mu_{\mathbb{P}_X},$$

that is, the mean element of $k_{\mathcal{X}}(X, \cdot)$ coincides with the mean embedding of $\mathbb{P}_X$.

It is known that elementary large sample theories, such as the central limit theorem (CLT), also hold for random elements of a Hilbert space. We state the CLT below, whose proof is mostly similar to the case of univariate random variables.

Theorem 3.8 (Central Limit Theorem [48, Theorem 7.7.6]). *Let $\chi_1, \chi_2, \ldots,$ be independent and identically distributed random elements of a separable Hilbert space $\mathcal{H}$. If $\mathbb{E}\|\chi_1\|_{\mathcal{H}} = 0$ and $\mathbb{E}\|\chi_1\|_{\mathcal{H}}^2 < \infty$, then*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \chi_i \xrightarrow{d} \mathbb{G},$$

where $\mathbb{G}$ is a zero mean Gaussian random element in $\mathcal{H}$ with its covariance operator $\mathbb{E}[\chi_1 \otimes \chi_1]$. The superscript d over the arrow indicates convergence in distribution.

Here, the Gaussian random element $\mathbb{G}$ above satisfies that $(\langle \mathbb{G}, f\rangle_{\mathcal{H}}, \langle \mathbb{G}, g\rangle_{\mathcal{H}})$ is a bivariate normal random vector with the cross-covariance equal to $\mathbb{E}[\langle \chi_1, f\rangle_{\mathcal{H}} \langle \chi_1, g\rangle_{\mathcal{H}}]$. One meaningful consequence of this CLT is that we have the weak law of large numbers with the order $O_p(1/\sqrt{n})$. Here the symbol $X_n = O_p(Y_n)$ represents that X_n/Y_n is bounded in probability.

3.3 Cross-Covariance Operators and Hilbert-Schmidt Operators

Having equipped with the functional machinery shown in the previous section, we introduce another essential ingredient of our theory, the cross-covariance operator of RKHSs. While the kernel mean embedding represents embedding probability distributions into RKHS, the cross-covariance operator represents independence structure between two probability distributions. Given that RKHSs are rich enough, such an operator is surprisingly capable of discriminating the independence of two distributions. As we describe below, the construction and the properties of the cross-covariance operator will show notable similarity to the covariance matrix of the Euclidean joint random vectors.

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random vector with the joint distribution $\mathbb{P}_{XY}$, and denote $\mathbb{P}_X$ and $\mathbb{P}_Y$ by their marginal distributions. Using the kernel feature embeddings $\phi(x) = k_{\mathcal{X}}(x, \cdot) \in \mathcal{H}_{\mathcal{X}}$ and $\psi(y) = k_{\mathcal{Y}}(y, \cdot) \in \mathcal{H}_{\mathcal{Y}}$, we can embed the marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ into the RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively, which are essentially means of the embedded random variables of RKHSs. We specifically denote such kernel mean embeddings by $m_X = \mathbb{E}_X[k_{\mathcal{X}}(X, \cdot)] = \mathbb{E}_X[\phi(X)]$ and $m_Y = \mathbb{E}_Y[k_{\mathcal{Y}}(Y, \cdot)] = \mathbb{E}_Y[\psi(Y)]$. Note that such mean elements indeed *generate* associated means, meaning that, for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$,

$$\langle f, m_X\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}[f(X)], \text{ and } \langle g, m_Y\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}[g(Y)].$$

Going further, we can also generate the covariance of evaluations of X and Y using a linear operator of RKHSs. The *cross-covariance operator* of (X, Y), $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$, is defined by the following adjoint relations; for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$,

$$\begin{aligned} \langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])\,(g(Y) - \mathbb{E}_Y[g(Y)])\right] \\ &= \mathrm{Cov}\left[f(X), g(Y)\right]. \end{aligned} \tag{3.6}$$

That is, the operator Σ_{YX} generates every possible cross-covariance of evaluations of X and Y via the member of RKHSs. The Riesz representation theorem uniquely defines a linear operator satisfying the relation (3.6) for all f and g. If $Y = X$, then we say $\Sigma_{XX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}$ the *covariance operator*. Note that $\Sigma_{YX}^* = \Sigma_{XY}$, where the $*$ indicates the adjoint of the operator, and thus the covariance operator Σ_{XX} is self-adjoint. Using the reproducing property, one can rewrite the above relation as

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{X,Y}\left[\langle f, \phi(X) - m_X\rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \psi(Y) - m_Y\rangle_{\mathcal{H}_{\mathcal{Y}}}\right]. \tag{3.7}$$

Table 3.1: Similarity between the Euclidean and RKHS notions in computations arose in equations (3.7) and (3.8).

Euclidean notions	RKHS notions
$X,\ Y$: random variables	$\phi(X),\ \psi(Y)$: random elements of RKHS
$\mathbb{E}[X],\ \mathbb{E}[Y]$: means	$m_X,\ m_Y$: mean elements
C_{YX}: cross-covariance matrix	Σ_{YX}: cross-covariance operator
$\alpha \in \mathbb{R}^m$: vector as a dual function $\langle \alpha, \cdot \rangle_{\mathbb{R}^m}$	$f \in \mathcal{H}_\mathcal{X}$: RKHS function
$\alpha^T X = \langle \alpha, X \rangle_{\mathbb{R}^m}$: dual evaluation of $\alpha \in \mathbb{R}^m$	$\langle f, \phi(X) \rangle_{\mathcal{H}_\mathcal{X}} = f(X)$: evaluation of $f \in \mathcal{H}_\mathcal{X}$.

It is worthwhile to notice that the equation (3.7) parallels the corresponding relation of the Euclidean cross-covariance matrix. If $\mathcal{X}$ and $\mathcal{Y}$ were the subsets of $\mathbb{R}^m$ and $\mathbb{R}^k$, respectively, and if we denote the cross-covariance matrix of Y and X by C_{YX}, then for all $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^k$, we have

$$\beta^T C_{YX} \alpha = \mathbb{E}_{X,Y}\left[\beta^T (Y - \mathbb{E}[Y])(X - E[X])^T \alpha\right] = \mathrm{Cov}[\alpha^T X, \beta^T Y]. \tag{3.8}$$

Here, the computations in the equations (3.7) and (3.8) exhibit a vast similarity and we summarize those correspondences in Table 3.1.

We may also explicitly represent the evaluations of the operator Σ_{YX}. If we put $g = \psi(y) = k_\mathcal{Y}(y, \cdot)$ in (3.6), we get the evaluation of $\Sigma_{YX} f$ at $y \in \mathcal{Y}$:

$$\begin{aligned}(\Sigma_{YX} f)(y) = \langle \psi(y), \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} &= \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])(\psi(y)(Y) - \mathbb{E}_Y[\psi(y)(Y)])\right] \\ &= \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])(k_\mathcal{Y}(y, Y) - \mathbb{E}_Y[k_\mathcal{Y}(y, Y)])\right] \\ &= \mathrm{Cov}[f(X), k_\mathcal{Y}(y, Y)].\end{aligned} \tag{3.9}$$

The covariance operator Σ_{YX} is particularly useful when the RKHSs $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are rich enough so that it can generate more possible covariances $\mathrm{Cov}[f(X), g(Y)]$. Recall that, in the Euclidean case, X and Y are uncorrelated if $C_{YX} = 0$. More can be said for the covariance operators on rich RKHSs, which is reminiscent of the result that uncorrelated Gaussian random vectors are independent.

Proposition 3.9. *If $\mathcal{H}_\mathcal{X} + \mathbb{R}$ is dense in $L^2(\mathbb{P}_X)$ and if $\mathcal{H}_\mathcal{Y} + \mathbb{R}$ is dense in $L^2(\mathbb{P}_Y)$, then*

$$\Sigma_{YX} = 0 \ \Leftrightarrow\ X \perp\!\!\!\perp Y.$$

Proof. Observe first that, in L^2 spaces,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)],\ \forall f \in L^2(\mathbb{P}_X), \forall g \in L^2(\mathbb{P}_Y)\ \Leftrightarrow\ X \perp\!\!\!\perp Y$$

since these L^2 spaces contain all the indicator functions $\mathbb{1}_A$, where the A are measurable. Thus, it suffices to show that if $\Sigma_{YX} = 0$, then for any fixed pair of L^2 functions $f \in L^2(\mathbb{P}_X)$ and $g \in L^2(\mathbb{P}_Y)$, we have $\mathrm{Cov}[f(X), g(Y)] = 0$. Pick sequences of functions $\{f_n + c_n\}_{n=1}^\infty \subset \mathcal{H}_\mathcal{X} + \mathbb{R}$ and $\{g_n + d_n\}_{n=1}^\infty \subset \mathcal{H}_\mathcal{Y} + \mathbb{R}$ that converge to f and g in L^2-norms, respectively. Here, $f_n \in \mathcal{H}_\mathcal{X}$ and $g_n \in \mathcal{H}_\mathcal{Y}$, and we can exclude the constants c_n and d_n in our argument by considering the variances,

$$\mathrm{Var}[f(X) - f_n(X)] \to 0, \quad \text{and} \quad \mathrm{Var}[g(Y) - g_n(Y)] \to 0.$$

Using the RKHS functions f_n and g_n, we can write

$$\begin{aligned}\mathrm{Cov}[f(X), g(Y)] &= \mathrm{Cov}[f(X) - f_n(X) + f_n(X), g(Y) - g_n(Y) + g_n(Y)] \\ &= \mathrm{Cov}[f(X) - f_n(X), g(Y) - g_n(Y)] + \mathrm{Cov}[f - f_n(X), g_n(Y)] \\ &\quad + \mathrm{Cov}[f_n(X), g(Y) - g_n(Y)] + \mathrm{Cov}[f_n(X), g_n(Y)],\end{aligned}$$

and note that $\mathrm{Cov}[f_n(X), g_n(Y)] = \langle g_n, \Sigma_{YX} f_n \rangle_{\mathcal{H}_\mathcal{Y}} = 0$ since $\Sigma_{YX} = 0$. Letting $n \to \infty$, the right hand side of the above goes to 0 by the Cauchy-Schwarz ineqaulity, which finishes the proof. □

Note that, using an empirical estimate for the covariance operator Σ_{YX} that will be introduced in Section 3.5, one can construct an independence test for two different random variables, called the *Hilbert Schmidt independence criterion* (HSIC) [44]. Here, the key ingredient of the result is that the RKHSs can estimate all the $L^2(\mathbb{P}_X)$ and $L^2(\mathbb{P}_Y)$ functions up to constants. It is thus natural to ask when the RKHS are rich enough to achieve such approximations, and we give two relevant definitions on richness of RKHS, independent of the given random variables X and Y.

Definition 3.10. Given a kernel $k_\mathcal{X}$ on the compact domain $\mathcal{X}$, we say the kernel $k_\mathcal{X}$ or the corresponding RKHS $\mathcal{H}_\mathcal{X}$ is

- *universal* if $\mathcal{H}_\mathcal{X}$ is a dense subspace of the space $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$, and
- *characteristic* if the mean embedding map $\mu : \mathcal{P}(\mathcal{X}) \to \mathcal{H}_\mathcal{X}$, $\mathbb{P} \mapsto \mu_\mathbb{P}$ is injective.

Since we consider only continuous kernels, the corresponding RKHS consists only of continuous functions, which verifies the well-definedness of the universal kernels. The characteristic RKHS indicates it is rich enough to distinguish all the probability measures on $\mathcal{X}$, and it is known that all universal kernels are characteristic [45]. Many popular kernels used in practice, such as Gaussain and Laplace kernels, are universal; Table 2.1 provides more examples of kernels on sphere with their universality. For further information, see Micchelli et al. [75] for extensive characterizations of universal kernels. Finally, we present a fact that characteristic kernels always satisfy the constraint of the Proposition 3.9.

Proposition 3.11 (see [34, Proposition 5])**.** *Let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ be an RKHS on $\mathcal{X}$. Then*

$$k \text{ is characteristic} \iff \mathcal{H}_\mathcal{X} + \mathbb{R} \text{ is dense in } L^2(\mathbb{P}) \text{ for all } \mathbb{P} \in \mathcal{P}(\mathcal{X}).$$

Characteristic kernels will play an essential role in our theory of nonlinear kernel dimension reduction in the next sections. Before developing the theory, we present another viewpoint of covariance operators that will be useful in technical proofs in the following subsection.

3.3.1 Hilbert-Schmidt Operator and Covariance of Random Elements

The cross-covariance operator introduced in this section is in fact a generalized covariance of a random element on a Hilbert space. Similarly to the kernel mean embedding, the essential properties of the covariance matrix extend to the covariance element with a greater generality. We briefly present such a general construction and two associated measures of the covariance operator.

Recall that the cross-covariance matrix C_{YX} of random vectors $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^k$, which is essentially a linear map $\mathbb{R}^m \to \mathbb{R}^k$, can be viewed as an element of $\mathbb{R}^k \otimes \mathbb{R}^m = \mathbb{R}^{k \times m}$. We can proceed a parallel construction for two random elements on Hilbert spaces, while we need a well-defined notion of tensor product of Hilbert spaces and their relation to linear operators.

Let $\{e_i\}_{i=1}^{\infty}$ be a complete orthonormal system (CONS) of $\mathcal{H}_1$ and let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded operator of separable Hilbert spaces. T is called *Hilbert-Schmidt* if it satisfies $\sum_{i=1}^{\infty} \|Te_i\|_{\mathcal{H}_2} < \infty$, and it is well known that this sum is independent of the choice of CONS of $\mathcal{H}_1$. We denote the class of Hilbert-Schmidt operators by $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)_{HS}$. The space of Hilbert-Schmidt operators is endowed with an inner product $\langle T_1, T_2 \rangle_{HS} := \sum_{i=1}^{\infty} \langle T_1 e_i, T_2 e_i \rangle_{H_2}$ for all $T_1, T_2 \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)_{HS}$, making it also a Hilbert space with the *Hilbert-Schmidt norm*, $\|T\|_{HS}^2 := \sum_{i=1}^{\infty} \|Te_i\|_{\mathcal{H}_2}^2$. If $\mathcal{H}_2 = \mathcal{H}_1$ so that $T : \mathcal{H}_1 \to \mathcal{H}_1$ is a bounded operator, we define another notion of *trace*, given by

$$\mathrm{Tr}(T) := \sum_{i=1}^{\infty} \langle Te_i, e_i \rangle_{\mathcal{H}_1},$$

which is again independent of the choice of CONS $\{e_i\}$. If $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)_{HS}$, it is clear that $\mathrm{Tr}(T^*T) = \|T\|_{HS}^2$.

On the other hand, one can define a tensor product $\mathcal{H}_2 \otimes \mathcal{H}_1$ of Hilbert spaces that is a completion of the space of finite linear combinations of simple tensors $g \otimes f$, $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, endowed with the inner product $\langle g_1 \otimes f_1, g_2 \otimes f_2 \rangle_{\mathcal{H}_2 \otimes \mathcal{H}_1} = \langle g_1, g_2 \rangle_{\mathcal{H}_2} \langle f_1, f_2 \rangle_{\mathcal{H}_1}$ that linearly extends. A simple tensor $g \otimes f$ is also called as rank-1 tensor, which defines a linear operator $g \otimes f : \mathcal{H}_1 \to \mathcal{H}_2$ that maps $h \mapsto g\langle f, h \rangle_{\mathcal{H}_1} \in \mathcal{H}_2$. Clearly, $\|g \otimes f\|_{HS} = \|g\|_{\mathcal{H}_1} \|f\|_{\mathcal{H}_2} = \|g \otimes f\|_{\mathcal{H}_2 \otimes \mathcal{H}_1}$ by Parseval's identity. It is known that such an association linearly extends to an isomorphism of Hilbert spaces [77]:

$$\mathcal{H}_2 \otimes \mathcal{H}_1 \to \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)_{HS}, \quad \sum_{i=1}^{\infty} g_i \otimes f_i \mapsto \left(h \mapsto \sum_{i=1}^{\infty} g_i \langle f_i, h \rangle_{\mathcal{H}_1} \right).$$

Therefore, we identify all the elements of the tensor product space $\mathcal{H}_2 \otimes \mathcal{H}_1$ as a Hilbert-Schmidt operator in $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)_{HS}$.

Using these notions, tensor product and the Hilbert-Schmidt operator on separable Hilbert spaces, we define the *cross-covariance* of random elements $\chi_2 \in \mathcal{H}_2$ and $\chi_1 \in \mathcal{H}_1$ by

$$\Sigma_{\chi_2 \chi_1} := \mathbb{E}\left[(\chi_2 - \mathbb{E}[\chi_2]) \otimes (\chi_1 - \mathbb{E}[\chi_1]) \right] \in \mathcal{H}_2 \otimes \mathcal{H}_1.$$

This is well-defined as long as the expectation is Bochner integrable, which is readily checked using Theorem 3.4. Once we have a well-defined $\Sigma_{\chi_2 \chi_1}$, it defines a Hilbert-Schmidt operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ following the above discussion. The explicit computation can be derived from Theorem 3.5. Since the expectation and inner products are interchangeable, we can explicitly describe the evaluations of $\Sigma_{\chi_2 \chi_1}$ as

- $\Sigma_{\chi_2 \chi_1} f = \mathbb{E}[(\chi_2 - \mathbb{E}[\chi_2]) \langle \chi_1 - \mathbb{E}[\chi_1], f \rangle_{\mathcal{H}_1}]$ for all $f \in \mathcal{H}_1$, and
- $\langle g, \Sigma_{\chi_2 \chi_1} f \rangle_{\mathcal{H}_2} = \mathbb{E}[\langle \chi_2 - \mathbb{E}[\chi_2], g \rangle_{\mathcal{H}_2} \langle \chi_1 - \mathbb{E}[\chi_1], f \rangle_{\mathcal{H}_1}]$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$.

Coming back to the RKHS setting, $\mathcal{H}_1 = \mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_2 = \mathcal{H}_{\mathcal{Y}}$, $\chi_1 = k_{\mathcal{X}}(X, \cdot) = \phi(X)$, and $\chi_2 = k_{\mathcal{Y}}(Y, \cdot) = \psi(Y)$ as before, we see that the latter equality coincides with the equation (3.7). Therefore, the cross-covariance operator Σ_{YX} defined in (3.6) is the same as the cross-covariance of random elements $\psi(Y)$ and $\phi(X)$, which is Hilbert-Schmidt. It immediately follows that

$$\|\Sigma_{YX}\|_{HS}^2 = \|\mathbb{E}[(\psi(Y) - m_Y) \otimes (\phi(X) - m_X)]\|_{\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}}^2 \tag{3.10}$$

where $m_X = \mathbb{E}[\phi(X)]$ and $m_Y = \mathbb{E}[\psi(Y)]$ are the mean elements. Also, using Parseval's identity and Theorem 3.5, we readily have

$$\mathrm{Tr}(\Sigma_{XX}) = \mathbb{E}[\|\phi(X) - m_X\|_{\mathcal{H}_{\mathcal{X}}}^2].$$

3.4 Conditional Covariance Operator and Generalization of Kernel Dimension Reduction

Using the mean and covariance of random elements, we are ready to define the conditional covariance operator on an RKHS. Although conditional covariance does not generally behave well for arbitrary joint random vectors, its construction and pleasing properties of the Gaussian random vectors successfully generalize to the RKHS environment with great generality. Similar to the covariance operator, the conditional covariance operator determines conditional independence under the richness of RKHSs.

Based on such a powerful property, Fukumizu et al. [32, 34] developed a famous kernel dimension reduction (KDR) method for supervised, sufficient dimension reduction in Euclidean space. The central intuition for their method is to find a dimension reduction that minimizes the conditional covariance of the response variable. The KDR method performs surprisingly well for data with a small sample size, though it is only developed for seeking orthogonal projections. However, the key intuition behind the KDR method suggests that it should be clearly generalized to arbitrary forms of dimension reductions of interest. We will verify that the theory can be greatly generalized without relying on the specific properties of orthogonal matrices.

Section 3.4.1 begins with reviewing the theory of conditional covariance operator on RKHS and how its properties correspond to the conditional covariance matrix of Gaussian random vectors. Then, we generalize the linear KDR theory of Fukumizu et al. [34] to encompass nonlinear projection maps. The main theoretical result, minimizing the trace of conditional covariance operator approaches sufficient dimension reduction (SDR), will remain valid and can be readily proved using measure-theoretic languages. As a result, we adequately formulate a generalized KDR algorithm at the population level. To address the complexity of the parameter choice problem of kernels, we delve into this issue within the scope of response variables in Section 3.4.2. We will observe that employing a linear kernel for univariate and continuous response variables can greatly streamline the parameter selection process without significantly diminishing theoretical efficacy. Finally, we will explore the generalization of the unsupervised KDR work of Wang et al. [121] in Section 3.4.3 with deeper discussions on the interpretation of their proposed framework.

3.4.1 Conditional Covariance Operator and Sufficient Dimension Reduction

Consider the scenario where (X, Y) forms a joint Gaussian random vector. In such cases, the conditional covariance of Y given X is traditionally defined as $C_{YY|X} = C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}$, provided that the covariance matrix C_{XX} is invertible. This invites the notion of similarly defining the conditional covariance operator. However, one must carefully address the issue of invertibility with the covariance operator Σ_{XX}, which is not always invertible. To circumvent this, we give an alternative definition of the conditional covariance operator without operator inversion. This approach utilizes the following result of Baker [9], which presents a concept akin to correlation defined on RKHS.

Theorem 3.12 (Baker [9, Theorem 1])**.** *There exists a bounded operator $V_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ such that*

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}, \quad \|V_{YX}\| \leq 1, \quad \textit{and} \quad V_{YX} = P_{\overline{\mathrm{ran}}(\Sigma_{YY})} V_{YX} P_{\overline{\mathrm{ran}}(\Sigma_{XX})}, \tag{3.11}$$

where $\|\cdot\|$ is the operator norm, $\overline{\mathrm{ran}}(\cdot)$ is the closure of the range, and $P_\mathcal{N}$ denotes the projection operator onto the subspace $\mathcal{N}$ of a Hilbert space. Furthermore, the operator V_{YX} is unique up to the relations (3.11).

Note that the correlation operator V_{YX} is defined without inverting the marginal covariances Σ_{XX} and Σ_{YY}. In case Σ_{XX} is invertible, the relation $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \Sigma_{YY}^{1/2}V_{YX}V_{XY}\Sigma_{XX}^{1/2}$ holds. Given the constant presence of the correlation operator, we can define the conditional covariance operator by leveraging this concept.

Definition 3.13. The *conditional covariance operator* $\Sigma_{YY|X} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ of Y given X is defined by

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YY}^{1/2}V_{YX}V_{XY}\Sigma_{YY}^{1/2}.$$

Remark. An alternative to employing the correlation operator for defining the conditional covariance operator is using the Moore-Penrose inverse $\Sigma_{XX}^\dagger$ of Σ_{XX}, which is defined in Hsing and Eubank [48, Section 3.5]. Because the Moore-Penrose inverse is defined only on $\mathrm{ran}(\Sigma_{XX})$, we need an *assumption* $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$ that assures the well-definedness of the operation $\Sigma_{XX}^\dagger\Sigma_{XY}$. With this assumption, it is apparent that the equality $\Sigma_{YY}^{1/2}V_{YX}V_{XY}\Sigma_{YY}^{1/2} = \Sigma_{YX}\Sigma_{XX}^\dagger\Sigma_{XY}$ holds, and thus we could have defined

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^\dagger\Sigma_{XY},$$

which appears more intuitive. This approach is adopted in, for example, Li and Song [57]. However, the precondition $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$ may not hold in general, even though it is regarded as a mild assumption because we always have $\mathrm{ran}(\Sigma_{XY}) \subseteq \overline{\mathrm{ran}}(\Sigma_{XX})$. We follow our more general construction using correlation operators since one can construct simple examples that easily violate the condition $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$; e.g., see Example 6.8 of Klebanov et al. [51].

We investigate some computational properties of the conditional covariance operator. It is natural to expect that some computational properties of the Gaussian conditional covariance matrix $C_{YY|X}$ may generalize to the operator setting. If $(X, Y) \in \mathbb{R}^m \times \mathbb{R}^k$ are joint Gaussian random variables, the covariance matrix of the conditional random variable $Y|X = x$ is exactly the conditional covariance matrix $C_{YY|X}$; that is,

$$\mathrm{Var}(b^TY|X) = b^TC_{YY|X}b, \quad \forall\, b \in \mathbb{R}^k,$$

where the left hand side is independent of the realizations $X = x$. To make the equality deterministic, we may also write

$$\mathbb{E}[\mathrm{Var}(b^TY|X)] = b^TC_{YY|X}b, \quad \forall\, b \in \mathbb{R}^k. \tag{3.12}$$

Also, using the fact that $Y - AX$, $A \in \mathbb{R}^{k\times m}$, is also Gaussian, one can easily derive

$$\min_{A\in\mathbb{R}^{k\times m}} \|\widetilde{Y} - A\widetilde{X}\|_2^2 = \mathrm{Tr}(C_{YY|X})$$

where $\widetilde{Y} = Y - \mathbb{E}[Y]$ and $\widetilde{X} = X - \mathbb{E}[X]$. Replacing Y with b^TY results in

$$\min_{a\in\mathbb{R}^m} \left|b^T\widetilde{Y} - a^T\widetilde{X}\right|^2 = b^TC_{YY|X}b = \mathrm{Var}(b^TY|X), \quad \forall\, b \in \mathbb{R}^k. \tag{3.13}$$

The equations (3.13) and (3.12) successfully extends to the similar properties of the conditional covariance operator $\Sigma_{YY|X}$, equations (3.14) and (3.15), respectively.

Proposition 3.14 ([34, Propositions 2 and 3])**.** *For any $g \in \mathcal{H}_\mathcal{Y}$, we have*

$$\langle g, \Sigma_{YY|X}g\rangle = \inf_{f\in\mathcal{H}_\mathcal{X}} \mathrm{Var}(g(Y) - f(X)). \tag{3.14}$$

If $\mathcal{H}_\mathcal{X} + \mathbb{R}$ is dense in $L^2(\mathbb{P}_X)$ (e.g. characteristic), we further have

$$\langle g, \Sigma_{YY|X}g\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]]. \tag{3.15}$$

Proof. Define $\mathcal{E}_g(f) = \mathbb{E}_{X,Y}\left|(g(Y) - \mathbb{E}_Y[g(Y)]) - (f(X) - \mathbb{E}_X[f(X)])\right|^2 = \mathrm{Var}(g(Y) - f(X))$. Observe that

$$\begin{aligned}
\mathcal{E}_g(f) &= \langle g, \Sigma_{YY} g\rangle_{\mathcal{H}_\mathcal{Y}} - 2\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} + \langle f, \Sigma_{XX} f\rangle_{\mathcal{H}_\mathcal{X}} \\
&= \|\Sigma_{XX}^{1/2} f\|_{\mathcal{H}_\mathcal{X}}^2 - 2\langle V_{XY}\Sigma_{YY}^{1/2} g, \Sigma_{XX}^{1/2} f\rangle_{\mathcal{H}_\mathcal{X}} + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}}^2 \\
&= \|\Sigma_{XX}^{1/2} f - V_{XY}\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}}^2 - \|V_{XY}\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 \\
&= \|\Sigma_{XX}^{1/2} f - V_{XY}\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 + \langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_\mathcal{Y}}.
\end{aligned}$$

It is thus obvious that $\inf_{f\in\mathcal{H}_\mathcal{X}} \mathcal{E}_g(f) \geq \langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_\mathcal{Y}}$. From the fact that $\mathrm{ran}(V_{XY}) \subseteq \overline{\mathrm{ran}}(\Sigma_{XX}) = \overline{\mathrm{ran}}(\Sigma_{XX}^{1/2})$ (Theorem 3.12), $\exists f^* \in \mathcal{H}_\mathcal{X}$ such that $\Sigma_{XX}^{1/2} f^*$ arbitrarily closely approximates $V_{XY}\Sigma_{YY}^{1/2} g$ in $\mathcal{H}_\mathcal{X}$, which means that

$$\inf_{f\in\mathcal{H}_\mathcal{X}} \|\Sigma_{XX}^{1/2} f - V_{XY}\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 = 0, \text{ proving the equality (3.14).}$$

For the second equality, we rewrite the equality (3.14) as

$$\begin{aligned}
\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_\mathcal{Y}} &= \inf_{f\in\mathcal{H}_\mathcal{X}} \mathrm{Var}[g(Y) - f(X)] \\
&= \inf_{f\in\mathcal{H}_\mathcal{X}} \{\mathrm{Var}_X[\mathbb{E}[g(Y) - f(X)|X]] + \mathbb{E}_X[\mathrm{Var}[g(Y) - f(X)|X]\} \\
&= \inf_{f\in\mathcal{H}_\mathcal{X}} \mathrm{Var}_X[\mathbb{E}[g(Y)|X] - f(X)] + \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]].
\end{aligned}$$

Since the regression function $\varphi(\cdot) = \mathbb{E}[g(Y)|X = \cdot\,]$ is $L^2(\mathbb{P}_X)$ ($\because \mathbb{E}[\mathbb{E}[g(Y)|X]^2] \leq \mathbb{E}[\mathbb{E}[g(Y)^2|X]] = \mathbb{E}[g(Y)^2] < \infty$), we can approximate φ given the richness of $\mathcal{H}_\mathcal{X}$; that is, for any $\varepsilon > 0$, there exists $f \in \mathcal{H}_\mathcal{X}$ such that $\|\varphi - (f + c)\|_{L^2(\mathbb{P}_X)}^2 \leq \varepsilon$ for some constant $c \in \mathbb{R}$. As $\mathrm{Var}[\varphi(X) - f(X)] \leq \|\varphi - (f + c)\|_{L^2(\mathbb{P}_X)}^2 \leq \varepsilon$ we conclude that $\inf_{f\in\mathcal{H}_\mathcal{X}} \mathrm{Var}_X[\mathbb{E}[g(Y)|X] - f(X)] = 0$, which completes the proof. □

Suppose that $\mathcal{X} \subset \mathbb{R}^d$ is an d-dimensional compact domain of predictors and let $\mathcal{Z} \subseteq \mathbb{R}^m$, $m \leq d$, be a target domain of dimension reduction on which we are given another RKHS $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$. Let $p : \mathcal{X} \to \mathcal{Z}$ be any measurable map that indicates a general nonlinear dimension reduction.

We are then ready to present our generalized theory of kernel dimension reduction. The original KDR theory is stated using the unique structure of orthogonal matrix $B \in \mathbb{R}^{d\times m}$, which can embed the projected variable $B^T X$ into the original space as $BB^T X \in \mathcal{X}$. We avoid the need of this embedding by stating our theorem in terms of the variable Z. We will also see that the σ-field viewpoint of the conditional expectation does not require any properties of the dimension reduction map p in contrast to the original proof of Fukumizu et al. [34], which used the orthogonal complement of the orthogonal projection matrix B.

Theorem 3.15. *Suppose that $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ is dense in $L^2(\mathbb{P}_X)$ (e.g., $k_\mathcal{X}$ is universal), and let $Z = p(X)$, where $p : \mathcal{X} \to \mathcal{Z}$ is a measurable map. Then,*

$$\Sigma_{YY|Z} \succeq \Sigma_{YY|X},$$

where the inequality $\succeq$ stands for the partial order of self-adjoint operators. If we further assume that $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ and $(\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ are characteristic, then

$$\text{the equality } \Sigma_{YY|Z} = \Sigma_{YY|X} \text{ holds if and only if } Y \perp\!\!\!\perp X \mid Z.$$

Remark. In this result, the role of the kernel $k_\mathcal{X}$ on the original domain $\mathcal{X}$ is only to provide a lower bound of the conditional covariance operator after projection, $\Sigma_{YY|Z}$.

Proof. Note first that $\mathcal{H}_\mathcal{X}$ is dense in $L^2(\mathbb{P}_X)$ whenever $k_\mathcal{X}$ is universal since $C(\mathcal{X})$ is dense in $L^2(\mathbb{P}_X)$. For $g \in \mathcal{H}_\mathcal{Y}$, Proposition 3.14 and the L^2-density of $\mathcal{H}_\mathcal{X}$ implies that

$$\begin{aligned}\langle g, \Sigma_{YY|Z} g\rangle_{\mathcal{H}_\mathcal{Y}} &= \inf_{h \in \mathcal{H}_\mathcal{Z}} \mathrm{Var}(g(Y) - h(Z)) \\ &= \inf_{f \in \mathcal{H}_\mathcal{X}^p} \mathrm{Var}(g(Y) - f(X)) \quad (\because Z = p(X)) \\ &\leq \inf_{f \in \mathcal{H}_\mathcal{X}} \mathrm{Var}(g(Y) - f(X)) = \langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_\mathcal{Y}},\end{aligned}$$

which proves the inequality. Here, the space $\mathcal{H}_\mathcal{X}^p$ is the pullback of $\mathcal{H}_\mathcal{Z}$ along the map p, which is a subspace of $L^2(\mathbb{P}_X)$; see (3.23) for its definition. To show the equality case under the characteristicity of RKHSs, we use another equality of Proposition 3.14:

$$\langle g, (\Sigma_{YY|Z} - \Sigma_{YY|X}) g\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}[\mathrm{Var}(g(Y)|Z)] - \mathbb{E}[\mathrm{Var}(g(Y)|X)].$$

As the law of total variance implies that

$$\mathrm{Var}(g(Y)|Z) = \mathbb{E}[\mathrm{Var}(g(Y)|X, Z)|Z] + \mathrm{Var}(\mathbb{E}[g(Y)|X, Z]|Z),$$

we have

$$\begin{aligned}\mathbb{E}[\mathrm{Var}(g(Y)|Z)] &= \mathbb{E}[\mathbb{E}[\mathrm{Var}(g(Y)|X, Z)|Z]] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[g(Y)|X, Z]|Z)] \\ &= \mathbb{E}[\mathrm{Var}(g(Y)|X, Z)] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y|X, Z]|Z)].\end{aligned}$$

Since the inclusion of σ-fields $\sigma(Z) \subset \sigma(X)$ holds clearly, we have

$$\begin{aligned}\Sigma_{YY|Z} = \Sigma_{YY|X} &\iff \mathbb{E}[\mathrm{Var}(\mathbb{E}[g(Y)|X]|Z)] = 0, \ \forall g \in \mathcal{H}_\mathcal{Y} \\ &\iff \mathrm{Var}(\mathbb{E}[g(Y)|X]|Z) = 0 \text{ almost surely}, \forall g \in \mathcal{H}_\mathcal{Y} \\ &\iff \mathbb{E}[g(Y)|X] = \mathbb{E}[g(Y)|Z] \text{ almost surely}, \forall g \in \mathcal{H}_\mathcal{Y}.\end{aligned}$$

The assumption that $\mathcal{H}_\mathcal{Y}$ is characteristic ensures that for all measurable set $A \subset \mathcal{Y}$, the indicator function $\mathbb{1}_A$ is approximated by $\mathcal{H}_\mathcal{Y}$-functions up to a constant; i.e.,

$$\mathbb{E}[g(Y)|X] = \mathbb{E}[g(Y)|Z] \text{ a.s.}, \forall g \in \mathcal{H}_\mathcal{Y} \iff \mathbb{P}_{Y|X} = \mathbb{P}_{Y|Z},$$

where the last equality is equivalent to the SDR $Y \perp\!\!\!\perp X \mid Z$. □

Suppose that we are given a family $\mathcal{F}$ of measurable dimension reduction maps p from $\mathcal{X}$ to $\mathcal{Z}$. The family $\mathcal{F}$ should be problem-specific; for ordinary Euclidean data, Fukumizu et al. [32, 34] took $\mathcal{F}$ as a Stiefel manifold of orthogonal matrices. We will take other families tailored to compositional data in Chapter 4 and Chapter 5 in which we develop variable selection and dimension reduction algorithms. Theorem 3.15 indicates that minimizing the conditional covariance operator $\Sigma_{YY|p(X)}$ over $p \in \mathcal{F}$ attacks SDR. Here, the operator $\Sigma_{YY|p(X)}$ coincides with $\Sigma_{YY|Z}$ in Theorem 3.15, but this notation clearly indicates its dependence on the projection map $p : \mathcal{X} \to \mathcal{Z}$. Applying the trace (Section 3.3.1) to the operator $\Sigma_{YY|p(X)} \preceq \Sigma_{YY}$, we may describe our objective function for SDR in a computationally easier form:

$$\arg\min_{p \in \mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)}). \tag{3.16}$$

Since $\Sigma_{YY|p(X)} \succeq \Sigma_{YY|X}$ and the positive operator of trace zero is zero, the trace equality $\mathrm{Tr}(\Sigma_{YY|p(X)}) = \mathrm{Tr}(\Sigma_{YY|X})$ also implies SDR. Therefore, the optimization problem (3.16) always achieves SDR if the class $\mathcal{F}$ contains the identity map $id_\mathcal{X}$. The empirical estimator to this optimization problem will be given in Section 3.5, and their large sample consistency will be studied in Section 3.5.

3.4.2 Kernel Choice and Sufficient Dimension Reduction for Conditional Mean

To compute the minimization (3.16), we need to decide what kernels we will use on the domain of projected data $\mathcal{Z}$ and the domain of responses $\mathcal{Y}$. As remarked below Theorem 3.15, computation of the kernel $k_{\mathcal{X}}$ on the original domain $\mathcal{X}$ is not needed in solving our optimization problem. Since popular kernels used in practice are defined up to a parametric family, we will encounter a complex parameter choice problem in solving the problem (3.16). Fortunately, we will see in this subsection that the kernel choice problem for the labels $\mathcal{Y}$ can be significantly simplified.

As shown in Theorem 3.15, we should choose a characteristic kernel $k_{\mathcal{Y}}$ on the domain $\mathcal{Y}$ to target SDR. We first consider a multi-class response problem on the discrete domain $\mathcal{Y} = \{y^{(1)}, \dots, y^{(k)}\}$. Gaussian kernel or Laplace kernel can be natural candidates after embedding $\mathcal{Y} \subset \mathbb{R}$, but simpler kernels are also possible depending on the type of the labels. The *delta kernel*,

$$k_{\mathcal{Y}}(y, y') := \delta_{y,y'} = \begin{cases} 1 & \text{if } y = y' \\ 0 & \text{otherwise} \end{cases}$$

is universal, since $\mathcal{H}_{\mathcal{Y}} = C(\mathcal{Y}) = \mathbb{R}^k$ in this case, so our theory applies. The delta kernel is not only computationally simpler than Gaussian or Laplace kernels, but it also fits our intuition: we often want to regard different labels as *totally dissimilar*. It is commonly adopted when researchers consider kernels on the discrete domain of labels [107, 20, 11]. There are also experimental supports that the delta kernel is empirically better than Gaussian kernels for the labels; see, for example, Yamada et al. [126].

For continuous responses, we can make a similar simplification in the univariate case by adjusting our theoretical basis, Theorem 3.15. In case $\mathcal{Y} \subset \mathbb{R}$, we take the *linear kernel* $k_{\mathcal{Y}}(y, y') = yy'$ as suggested by Chen et al. [20]. Note that the linear kernel results in a non-characteristic RKHS $\mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\vee}$, the dual space of $\mathbb{R}$, so minimizing the conditional covariance $\Sigma_{YY|p(X)}$ may not lead to SDR. Nonetheless, the RKHS $\mathcal{H}_{\mathcal{Y}}$ includes the identity function $id_{\mathcal{Y}}$ in this case, which gives a satisfactory relaxed result of dimension reduction. The following result is a slightly corrected form of Corollary 3 of Chen et al. [20].

Proposition 3.16. *Let $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, $p : \mathcal{X} \to \mathcal{Z}$, and $Z = p(X)$ as in Theorem 3.15. Suppose that $(\mathcal{H}_{\mathcal{Z}}, k_{\mathcal{Z}})$ is characteristic, $\mathcal{Y} \subset \mathbb{R}$, and we take the linear kernel $k_{\mathcal{Y}}(y, y') = yy'$. Then the trace equality $\mathrm{Tr}(\Sigma_{YY|X}) = \mathrm{Tr}(\Sigma_{YY|Z})$ implies $\mathbb{E}[Y|X] = \mathbb{E}[Y|Z]$ almost surely; that is, knowing Z is sufficient for predicting Y.*

Proof. In the proof of Theorem 3.15, we have seen that

$$\langle g, (\Sigma_{YY|Z} - \Sigma_{YY|X})g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}[\mathrm{Var}(\mathbb{E}[g(Y)|X]|Z)]$$

given that $\mathcal{H}_{\mathcal{Z}}$ is characteristic. Since the trace of operators on $\mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\vee}$ is computed only with the identity function $id_{\mathcal{Y}}$, putting $g = id_{\mathcal{Y}}$ above gives

$$0 = \mathrm{Tr}(\Sigma_{YY|Z} - \Sigma_{YY|X}) = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y|X]|Z)],$$

which implies $\mathrm{Var}(\mathbb{E}[Y|X]|Z) = 0$ a.s., so $\mathbb{E}[Y|X] = \mathbb{E}[Y|Z]$ almost surely. □

Remark. The equality $\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)]$ is also known as sufficient dimension reduction for *conditional mean*, studied in Cook and Li [22]. If we *assume* that the conditional mean $\mathbb{E}[Y|X]$ completely determines the conditional distribution $\mathbb{P}_{Y|X}$; i.e., $Y \perp X \,|\, \mathbb{E}[Y|X]$, then the equality $\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)]$ suffices to guarantee $Y \perp\!\!\!\perp X \,|\, p(X)$. This assumption is common in the statistics literature, for instance, the additive error models $Y = \mathbb{E}[Y|X] + \varepsilon$ with $X \perp\!\!\!\perp \varepsilon$ satisfies the assumption.

Another advantage of the linear kernel on $\mathcal{Y} \subset \mathbb{R}$ is that we can understand our trace objective $\mathrm{Tr}(\Sigma_{YY|p(X)})$ as a form of kernel ridge regression (KRR) with an intercept. This is done by adding constants to the RKHS $\mathcal{H}_\mathcal{Z}$ as follows

Corollary 3.17 (Corollary 4 of Chen et al. [20])**.** *We have*

$$\mathrm{Tr}(\Sigma_{YY|Z}) = \inf_{f \in \mathcal{H}_\mathcal{Z} + \mathbb{R}} \mathbb{E}[(Y - f(Z))^2]. \tag{3.17}$$

Proof. Recall that $\mathrm{Tr}(\Sigma_{YY|Z}) = \inf_{f \in \mathcal{H}_\mathcal{Z}} \mathrm{Var}(Y - f(Z))^2$. It is straightforward to check that this equals to the above least squares error. □

Although the right hand side of (3.17) is not exactly the form of the kernel ridge regression, we will see in Section 3.5 that the Tikhonov-type regularization in our empirical objective of $\mathrm{Tr}(\Sigma_{YY|Z})$ plays a similar role as the regularization term in the KRR.

3.4.3 Unsupervised Generalized Kernel Dimension Reduction

If no labels are available in the data, we encounter an unsupervised problem. In such situations, we can identify the label with the input data itself, setting $\mathcal{Y} = \mathcal{X}$ and $Y = X$. To continue the powerful framework of sufficient dimension reduction, Wang et al. [121] introduced the following conditional independence relation as an objective for unsupervised dimension reduction:

$$X \perp\!\!\!\perp \widetilde{X} \,|\, p(X), \quad p \in \mathcal{F}. \tag{3.18}$$

Here, p represents a measurable map for dimension reduction, and $\widetilde{X}$ is an i.i.d. random variable with the same distribution $\mathbb{P}_X$ as X. Letting $Z = p(X)$, the conditional independence (3.18) indicates that the σ-field $\sigma(X)$ is contained in the *completion* of $\sigma(Z)$, denoted $\overline{\sigma(Z)}$, with the completion being relative to the ambient σ-field of the probability space [50, Corollary 8.11]. Thus, the relation (3.18) implies that all information pertaining to X is almost surely addressed by the information in $Z = p(X)$. Given the clear inclusion of $\sigma(Z) \subseteq \sigma(X)$, we may write the unsupervised SDR relation as simply $\overline{\sigma(Z)} = \overline{\sigma(X)}$. Consequently, we could equivalently interpret (3.18) as

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)] \ \text{ almost surely}$$

for all integrable random variables Y.

Applying Theorem 3.15, we similarly formulate the unsupervised dimension reduction problem over a class $\mathcal{F}$ as

$$\arg\min_{p \in \mathcal{F}} \mathrm{Tr}(\Sigma_{XX|p(X)}),$$

whose empirical estimate will be given in Section 3.5 as

$$\arg\min_{p \in \mathcal{F}} \mathrm{Tr}(G_X (G_{p(X)} + n\varepsilon_n I_n)^{-1}).$$

Here, G_X and $G_{p(X)}$ are the centered kernel Gram matrices of data $x_1, \ldots, x_n$ and projections $p(x_1), \ldots, p(x_n)$, respectively. Wang et al. [121] also argue that minimizing this empirical objective is akin to maximizing the HSIC[44]-based objective, expressed as $\mathrm{Tr}(G_X G_{p(X)})$ over $p \in \mathcal{F}$. This approach is noted for being more computationally efficient than the SDR-based approach, and the authors claim that their experimental results of these two approaches are also similar. However, their theoretical validation for such equivalence is confined to uniformly distributed data on the sphere, a hardly encountered case in

practice. Furthermore, a recent result of Liu and Ruan [64, Proposition 2] shows that the maximizer of the HSIC-based objective cannot be theoretically guaranteed to achieve informative dimension reduction, as they construct counterexamples in case $\mathcal{F}$ is a class of variable importance weights. Specifically, it might overlook a crucial signal variable, an issue that does not arise in the SDR-based objective, which has a statistical guarantee as will be discussed in Section 3.6. Consequently, the suggested equivalence of [121] fails in general circumstances. Thus, we primarily focus on our generalized SDR-based objective in this thesis.

3.5 Computing Empirical Estimates of Dimension Reduction

We now give an empirical estimate for our trace object, $\mathrm{Tr}(\Sigma_{YY|p(X)})$, where $p : \mathcal{X} \to \mathcal{Z}$ is an arbitrary measurable dimension reduction map. The natural empirical estimates are generated from the embedded empirical probability measures on RKHSs because the covariance operators are described in terms of expectations on RKHSs. To enable a stable and intuitive computation, we will apply a Tikhonov regularization to the component $\Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ and show how the regularization $(\Sigma_{XX} + \varepsilon I)^{-1}$ acts similarly to the penalization in kernel ridge regression. Then, assuming $\mathcal{F}$ is a parametric family of functions that is a compact metric space, we will finally be able to solve the optimization problems in Chapters 4 and 5. Note that $\mathcal{F}$ was taken as a Stiefel manifold of orthogonal matrices in Fukumizu et al. [34], and we generalize this to obtain an interpretable SDR of compositional data.

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ be sampled i.i.d. from the joint distribution of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Let $\phi_i = k_{\mathcal{X}}(x_i, \cdot) \in \mathcal{H}_{\mathcal{X}}$ and $\psi_i = k_{\mathcal{Y}}(y_i, \cdot) \in \mathcal{H}_{\mathcal{Y}}$ be embedded functions of data in RKHSs. Recall that the cross-covariance operator Σ_{XY} were defined as

$$\Sigma_{YX} = \mathbb{E}[(\psi(Y) - m_Y) \otimes (\phi(X) - m_X)] \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}},$$

where $\phi(X) = k_{\mathcal{X}}(X, \cdot) \in \mathcal{H}_{\mathcal{X}}$ and $\psi(Y) = k_{\mathcal{Y}}(Y, \cdot) \in \mathcal{H}_{\mathcal{Y}}$ are random elements. Recall also that $m_X = \mathbb{E}[\phi(X)]$ and $m_Y = \mathbb{E}[\psi(Y)]$, so their natural empirical estimates should be $\widehat{m}_X = \frac{1}{n}\sum_{i=1}^n \phi_i$ and $\widehat{m}_Y = \frac{1}{n}\sum_{i=1}^n \psi_i$. Thus we may define a natural empirical estimate for Σ_{YX} via the sample average,

$$\widehat{\Sigma}_{YX}^{(n)} := \frac{1}{n}\sum_{i=1}^n (\psi_i - \widehat{m}_Y) \otimes (\phi_i - \widehat{m}_X),$$

which is called the *empirical cross-covariance operator*. It is immediately seen that the empirical estimate $\widehat{\Sigma}_{YX}$ generates empirical covarinaces:

$$\langle g, \widehat{\Sigma}_{YX}^{(n)} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \frac{1}{n}\sum_{i=1}^n g(y_i) f(x_i) - \left(\frac{1}{n}\sum_{i=1}^n g(y_i)\right)\left(\frac{1}{n}\sum_{i=1}^n f(x_i)\right),$$

for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$.

Using the empirical estimate $\widehat{\Sigma}_{YX}^{(n)}$ for Σ_{YX}, we may define the empirical version of conditional covariance operator $\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2}$. Recall that $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{XX}^{1/2}$ if the covariance operator Σ_{XX} is invertible, as discussed in Section 3.4. By adopting the Tikhonov-type regularization to invert the covariance Σ_{XX}, we define the *empirical conditional covariance operator* as

$$\widehat{\Sigma}_{YY|X}^{(n)} := \widehat{\Sigma}_{YY}^{(n)} - \widehat{\Sigma}_{YX}^{(n)} (\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{\Sigma}_{XY}^{(n)}, \tag{3.19}$$

where $\varepsilon_n > 0$, $\varepsilon_n \to 0$ as the number of samples n goes to infinity, and $I : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}$ is the identity operator. Note that the parameter ε_n works similarly as a regularization parameter of kernel ridge

regression, which we will cover at the end of this section. For brevity, we will drop the superscript (n) for all emipricial operators if there is no confusion.

Although the empirical covariance operators are still defined on the (possibly) infinite-dimensional Hilbert spaces, their numerical measures, such as the Hilbert-Schmidt norm or the trace, can be expressed in finite terms. To simplify the notations, we denote the centered data embeddings by $\widetilde{\psi}_i := \psi_i - \widehat{m}_Y \in \mathcal{H}_\mathcal{Y}$ and $\widetilde{\phi}_i := \phi_i - \widehat{m}_X \in \mathcal{H}_\mathcal{X}$, so the operator $\widehat{\Sigma}_{YX}$ is written in short as $\widehat{\Sigma}_{YX} = \frac{1}{n}\sum_{k=1}^n \widetilde{\psi}_k \otimes \widetilde{\phi}_k$. The reproducing property shows $\widetilde{\phi}_i(x_k) = k_\mathcal{X}(x_i, x_k) - \frac{1}{n}\sum_{l=1}^n k_\mathcal{X}(x_l, x_k)$ and $\widetilde{\psi}_j(y_k) = k_\mathcal{Y}(y_j, y_k) - \frac{1}{n}\sum_{l=1}^n k_\mathcal{Y}(y_l, y_k)$, and similarly we may compute the inner products as, for all $i, j \in [n] := \{1, \dots, n\}$,

$$\begin{aligned}\langle \widetilde{\phi}_i, \widetilde{\phi}_j \rangle_{\mathcal{H}_\mathcal{X}} = k_\mathcal{X}(x_i, x_j) - \frac{1}{n}\sum_{k=1}^n k_\mathcal{X}(x_i, x_k) - \frac{1}{n}\sum_{l=1}^n k_\mathcal{X}(x_l, x_j) \\ + \frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n k_\mathcal{X}(x_l, x_k),\end{aligned}$$

and the similar result holds for $\langle \widetilde{\psi}_i, \widetilde{\psi}_j \rangle_{\mathcal{H}_\mathcal{Y}}$. This implies that, by denoting $K_X = \big(k_\mathcal{X}(x_i, x_j)\big)$ the kernel Gram matrix and $\mathbb{H} = I - \frac{1}{n}\mathbb{1}\mathbb{1}^T$ the centering matrix, where $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^n$, the above computation is described as $\langle \widetilde{\phi}_i, \widetilde{\phi}_j \rangle_{\mathcal{H}_\mathcal{X}} = (\mathbb{H}K_X\mathbb{H})_{i,j}$. The matrix $\mathbb{H}K_X\mathbb{H}$ is often called as a centered Gram matrix, and written in short as G_X. If we define $G_Y = \mathbb{H}K_Y\mathbb{H}$ similarly, we have

$$\langle \widetilde{\phi}_i, \widetilde{\phi}_j \rangle_{\mathcal{H}_\mathcal{X}} = (G_X)_{i,j} \quad \text{and} \quad \langle \widetilde{\psi}_i, \widetilde{\psi}_j \rangle_{\mathcal{H}_\mathcal{Y}} = (G_Y)_{i,j}.$$

We can then evaluate $\widehat{\Sigma}_{YX}$ at the sample embeddings,

$$\begin{aligned}\widehat{\Sigma}_{YX}\widetilde{\phi}_i &= \frac{1}{n}\sum_{k=1}^n \Big(\widetilde{\psi}_k \otimes \widetilde{\phi}_k\Big)\widetilde{\phi}_i \\ &= \frac{1}{n}\sum_{k=1}^n \langle \widetilde{\phi}_k, \widetilde{\phi}_i \rangle_{\mathcal{H}_\mathcal{X}} \widetilde{\psi}_k \\ &= \frac{1}{n}\sum_{k=1}^n (G_X)_{k,i}\widetilde{\psi}_k.\end{aligned}$$

Therefore, letting the spanning systems $\mathcal{B}_X = \{\widetilde{\phi}_i\}_{i=1}^n$ and $\mathcal{B}_Y = \{\widetilde{\psi}_i\}_{i=1}^n$ of subspaces of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, respectively, the empirical cross-covariance operator $\widehat{\Sigma}_{YX}$ from $\operatorname{span}(\mathcal{B}_X)$ to $\operatorname{span}(\mathcal{B}_Y)$ is represented by a centered Gram matrix G_X,

$$_{\mathcal{B}_Y}[\widehat{\Sigma}_{YX}]_{\mathcal{B}_X} = \frac{1}{n}G_X.$$

Here, the notation $_{\mathcal{B}_Y}[\widehat{\Sigma}_{YX}]_{\mathcal{B}_X}$ means that its jth column represents the vector $[\widehat{\Sigma}_{YX}\widetilde{\psi}_j]_{\mathcal{B}_Y}$, where a vector $[\varphi]_{\mathcal{B}_Y}$, for $\varphi \in \operatorname{span}(\mathcal{B}_Y)$, satisfies $\sum_{i=1}^n ([\varphi]_{\mathcal{B}_Y})_i\widetilde{\psi}_i = \varphi$. Note that these vector representations are not unique since the spanning systems $\mathcal{B}_X$ and $\mathcal{B}_Y$ are *linearly dependent*, and we are particularly interested in the representations related to the centered Gram matrices. One can obtain similar representations of $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$, and furthermore it is easy to compute that

$$_{\mathcal{B}_X}[(\widehat{\Sigma}_{XX} + \varepsilon_n I)^{-1}]_{\mathcal{B}_X} = n(G_X + n\varepsilon_n I_n)^{-1}.$$

These computational results are summarized in the following proposition:

Proposition 3.18. *We have the following matrix representations*

$$\begin{aligned}&_{\mathcal{B}_X}[\widehat{\Sigma}_{XX}]_{\mathcal{B}_X} = \frac{1}{n}G_X, \quad _{\mathcal{B}_Y}[\widehat{\Sigma}_{YY}]_{\mathcal{B}_Y} = \frac{1}{n}G_Y \\ &_{\mathcal{B}_Y}[\widehat{\Sigma}_{YX}]_{\mathcal{B}_X} = \frac{1}{n}G_X, \quad _{\mathcal{B}_X}[(\widehat{\Sigma}_{XX} + \varepsilon_n I)^{-1}]_{\mathcal{B}_X} = n(G_X + n\varepsilon_n I_n)^{-1}.\end{aligned}$$

Therefore, the empirical conditional covariance operator $\Sigma_{YY|X}$ has the following matrix representation

$$\begin{aligned}{}_{\mathcal{B}_Y}[\widehat{\Sigma}_{YY|X}]_{\mathcal{B}_Y} &= \frac{1}{n}\left\{G_Y - G_X(G_X + n\varepsilon_n I_n)^{-1}G_Y\right\} \\ &= \varepsilon_n(G_X + n\varepsilon_n I_n)^{-1}G_Y.\end{aligned}$$

Computing the trace of the empirical operator $\widehat{\Sigma}_{YY|X}$ based on this matrix representation requires extra verification because of the linear dependency of the spanning systems $\mathcal{B}_X = \{\widetilde{\phi}_i\}_{i=1}^n$ and $\mathcal{B}_Y = \{\widetilde{\psi}_i\}_{i=1}^n$. Recall that the trace of a positive self-adjoint operator A on a Hilbert space $\mathcal{H}$ is defined by

$$\mathrm{Tr}(A) = \sum_{i=1}^{\infty}\langle h_i, Ah_i\rangle_{\mathcal{H}},$$

where $\{h_i\}_{i=1}^{\infty}$ is a CONS of $\mathcal{H}$. Since the trace is independent of the choice of CONS, it immediately follows that $\mathrm{Tr}(A)$ is computed inside any closed subspace of $\mathcal{H}$ that contains $\overline{\mathrm{ran}}(A)$ by taking a CONS of such a subspace. Since we have shown above that $\overline{\mathrm{ran}}(\widehat{\Sigma}_{YY|X}) \subseteq \mathrm{span}(\mathcal{B}_Y)$, we have $\mathrm{Tr}(\widehat{\Sigma}_{YY|X}) = \mathrm{Tr}(\widehat{\Sigma}_{YY|X}\big|_{\mathrm{span}(\mathcal{B}_Y)})$. Let $\mathcal{E} = \{e_1, \ldots, e_k\}$ be an orthonormal basis for the subspace $\mathrm{span}(\mathcal{B}_Y) \subset \mathcal{H}_{\mathcal{Y}}$ so that

$$\mathrm{Tr}(\widehat{\Sigma}_{YY|X}) = \sum_{i=1}^{k}\langle e_i, \widehat{\Sigma}_{YY|X}e_i\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathrm{Tr}(_{\mathcal{E}}[\widehat{\Sigma}_{YY|X}]_{\mathcal{E}}).$$

Fixing a matrix representation ${}_{\mathcal{B}_Y}[I_{\mathcal{Y}}]_{\mathcal{E}}$ of the identity operator $I_{\mathcal{Y}} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$, a simple linear algebra shows that

$$\begin{aligned}\mathrm{Tr}({}_{\mathcal{B}_Y}[\widehat{\Sigma}_{YY|X}]_{\mathcal{B}_Y}) &= \mathrm{Tr}({}_{\mathcal{B}_Y}[I_{\mathcal{Y}}]_{\mathcal{E}} \cdot {}_{\mathcal{E}}[\widehat{\Sigma}_{YY|X}]_{\mathcal{E}} \cdot {}_{\mathcal{E}}[I_{\mathcal{Y}}]_{\mathcal{B}_Y}) \\ &= \mathrm{Tr}({}_{\mathcal{E}}[\widehat{\Sigma}_{YY|X}]_{\mathcal{E}} \cdot {}_{\mathcal{E}}[I_{\mathcal{Y}}]_{\mathcal{B}_Y} \cdot {}_{\mathcal{B}_Y}[I_{\mathcal{Y}}]_{\mathcal{E}}) \\ &= \mathrm{Tr}({}_{\mathcal{E}}[\widehat{\Sigma}_{YY|X}]_{\mathcal{E}} \cdot {}_{\mathcal{E}}[I_{\mathcal{Y}}]_{\mathcal{E}}) \\ &= \mathrm{Tr}({}_{\mathcal{E}}[\widehat{\Sigma}_{YY|X}]_{\mathcal{E}})\end{aligned}$$

since ${}_{\mathcal{E}}[I_{\mathcal{Y}}]_{\mathcal{E}}$ is uniquely determined as the identity matrix I_k. Therefore, we conclude that our regularized empirical estimate is given by

$$\mathrm{Tr}(\widehat{\Sigma}_{YY|X}) = \varepsilon_n \,\mathrm{Tr}((G_X + n\varepsilon_n I_n)^{-1}G_Y). \tag{3.20}$$

By replacing X with the projected variable $p(X)$, $p \in \mathcal{F}$, we propose to attack SDR by solving the following empirical problem:

$$\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\widehat{\Sigma}_{YY|p(X)}) = \arg\min_{p\in\mathcal{F}} \mathrm{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1}G_Y). \tag{3.21}$$

Here, the multiplier ε_n is removed in the right hand side because it is constant when n is fixed. We will discuss how to solve this optimization problem in different tasks in Chapter 4 and Chapter 5. We conclude this section with another interesting interpretation of $\mathrm{Tr}(\widehat{\Sigma}_{YY|X})$, though we will not use it in empirical computations.

Remark. Computing the trace of the empirical operator $\widehat{\Sigma}_{YY|X} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ using a CONS of $\mathcal{H}_{\mathcal{Y}}$ directly exhibits an analogy to kernel ridge regression (KRR). Since the RKHS $\mathcal{H}_{\mathcal{Y}}$ we will use is often finite dimensional and very simple as discussed in Section 3.4.2, this viewpoint that we describe below may provide a correct intuition for practical applications.

For any function $g \in \mathcal{H}_{\mathcal{Y}}$, we first compute that

$$\langle g, \widehat{\Sigma}_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \inf_{f \in \mathcal{H}_{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} \left[\left(g(y_i) - \frac{1}{n} \sum_{j=1}^{n} g(y_j) \right) - \left(f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \right]^2 + \varepsilon_n \|f\|^2_{\mathcal{H}_{\mathcal{X}}}.$$

This can be proved similarly to Proposition 3.14. Letting $\{g_i\}$ be a CONS of $\mathcal{H}_{\mathcal{Y}}$, the trace of $\widehat{\Sigma}_{YY|X}$ is described by the sum of the right hand sides above, where g is replaced by the g_i's. In case $\mathcal{Y} \subset \mathbb{R}$ with the linear kernel, as discussed in Section 3.4.2, we obtain a simplified form

$$\mathrm{Tr}[\widehat{\Sigma}_{YY|X}] = \inf_{f \in \mathcal{H}_{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \right]^2 + \varepsilon_n \|f\|^2_{\mathcal{H}_{\mathcal{X}}}$$

by assuming that the y_i are centered. This is similar to the loss function of KRR with the regularization parameter ε_n, and we may write the optimization problem (3.21) as

$$\arg\min_{p \in \mathcal{F}} \mathrm{Tr}(\widehat{\Sigma}_{YY|p(X)}) = \arg\min_{p \in \mathcal{F}} \inf_{f \in \mathcal{H}^p_{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \right]^2 + \varepsilon_n \|f\|^2_{\mathcal{H}^p_{\mathcal{X}}},$$

where $\mathcal{H}^p_{\mathcal{X}}$ is the pullback of the RKHS $\mathcal{H}_{\mathcal{Z}}$ along the map $p : \mathcal{X} \to \mathcal{Z}$ as defined in (3.23). Therefore, roughly speaking, our empirical optimization problem (3.21) can be seen as finding an optimal projection p that results in the lowest KRR-like loss, in case $\mathcal{Y} \subset \mathbb{R}$ and we use a linear kernel.

3.6 Theoretical Properties of the Dimension Reduction Estimator

This section is dedicated to stating and proving the consistency of the empirical estimator (3.21) of dimension reduction. An empirical solution $\hat{p}^{(n)}$ of (3.21), computed from n samples, will be shown to converge to a population minimizer p of (3.16) under a decay condition on ε_n, where the convergence is stated in terms of a metric defined on the family $\mathcal{F}$. We will prove that most results stated in terms of orthogonal projection matrices in Fukumizu et al. [34] successfully extend to our generalized version of kernel dimension reduction. Furthermore, we will clarify some of the underpinning assumptions of the original theory and show that one of their assumptions is redundant.

Section 3.6.1 will present essential results with the underlying assumptions, showing that our generalized method requires few constraints and applies to almost all situations in practice. We will show the intuitive compatibility result of covariance operators and the pullback operation in Section 3.6.2. Finally, we will give the main proof of the results in Section 3.6.3.

3.6.1 Large Sample Convergence

Recall that our domains $\mathcal{X}$ and $\mathcal{Y}$ are compact topological spaces, with $\mathcal{X}$ being a subset of $\mathbb{R}^d$. We also assume that $\mathcal{Y}$ can be topologically embedded in an Euclidean space. The compact domain $\mathcal{Z} \subset \mathbb{R}^m$, $m \leq d$, is a target domain of dimension reduction on which a characteristic RKHS $(\mathcal{H}_{\mathcal{Z}}, k_{\mathcal{Z}})$ is defined. We build our theory on the class $\mathcal{F}$, which consists of measurable maps $p : \mathcal{X} \to \mathcal{Z}$ that will be interpretable in practical applications. We assume $\mathcal{F}$ to be equipped with a metric ρ, making it a compact metric space. To ensure that any approximation $\hat{p} \in \mathcal{F}$ of a function $p \in \mathcal{F}$ indeed approximates dimension reductions of data, i.e., $\hat{p}(x)$ is close to $p(x)$ for every $x \in \mathcal{X}$, we naturally impose the following assumption:

Assumption 3.1. For each $x \in \mathcal{X}$, the evaluation functional at x, $p \mapsto p(x)$ is continuous on $\mathcal{F}$.

Note that the classes $\mathcal{F}$ that meet Assumption 3.1 encompass the Stiefel manifolds with their geodesic distance. This assumption guarantees that our empirical solution set

$$\arg\min_{p \in \mathcal{F}} \operatorname{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y)$$

is nonempty, since $\mathcal{F}$ is compact and $k_{\mathcal{Z}}$ is continuous on $\mathcal{Z}$.

To fulfill the technical requirements of our proof of the consistency result to hold, we make the following additional weak assumptions regarding the kernel $k_{\mathcal{Z}}$, the class $\mathcal{F}$, and the data distribution:

Assumption 3.2. There exists a measurable function $\varphi : \mathcal{X} \to \mathbb{R}$ satisfying $\mathbb{E}[\varphi(X)^2] < \infty$ and the Lipschitz condition

$$\|k_{\mathcal{Z}}(p_1(x), \cdot) - k_{\mathcal{Z}}(p_2(x), \cdot)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \varphi(x)\, \rho(p_1, p_2)$$

holds for all $x \in \mathcal{X}$ and $p_1, p_2 \in \mathcal{F}$.

Assumption 3.3. For each $y \in \mathcal{Y}$, the conditional probability density function $f_{X|Y}(x|y)$ of $X|Y = y$ exists, and it is continuous in x, bounded in y, and measurable in y.

Assumption 3.2 is a version of the assumption (A-3) of Fukumizu et al. [32] for our broader setup and will play a crucial role in establishing the uniform convergence of empirical estimates. A typical scenario where this assumption holds is when $k_{\mathcal{Z}}$ is an l^2-radial kernel, i.e., $k_{\mathcal{Z}}(z_1, z_2) = h(\|z_1 - z_2\|^2)$ for some $h : \mathbb{R} \to \mathbb{R}$, with the property that h is Lipschitz continuous. This scenario includes the popular Gaussian kernel and the rational quadratic kernel. Letting $C > 0$ a Lipschitz constant such that $|h(s) - h(t)| \leq C|s - t|,\ \forall s, t \in \mathbb{R}$, we have

$$\begin{aligned} \|k_{\mathcal{Z}}(p_1(x), \cdot) - k_{\mathcal{Z}}(p_2(x), \cdot)\|^2_{\mathcal{H}_{\mathcal{Z}}} &= 2h(0) - 2h(\|p_1(x) - p_2(x)\|^2) \\ &\leq 2C\|p_1(x) - p_2(x)\|^2. \end{aligned}$$

If there is a function φ on $\mathcal{X}$ such that

$$\|p_1(x) - p_2(x)\| \leq \varphi(x)\rho(p_1, p_2),$$

meaning that the evaluation at x is also Lipschitz continuous on $\mathcal{F}$ with the constant $\varphi(x)$, we obtain a similar form to Assumption 3.2 up to a constant multiplication. In many instances, including cases involving the Stiefel manifold and our choice in Chapter 5, the family $\mathcal{F}$ of dimension reductions will have bounded $\varphi(x)$ on the compact domain $\mathcal{X}$, thereby achieving Assumption 3.2.

Assumption 3.3 refines the assumption (A-1) presented in Fukumizu et al. [34], where the original assumption was stated as our Lemma 3.19 on continuity. They addressed that such a continuity is attainable if the conditional probability distribution $Y|X = x$ is continuous in x, where they used the advantageous properties of orthogonal matrices. However, as pointed out in Ackerman et al. [1], there are instances where the continuity of a conditional distribution $Y|X = x$ in the conditioning variable x is violated, especially when the conditioned variable Y is discrete. Therefore, we propose replacing the somewhat intricate continuity assumption (A-1) of the reference with our more straightforward Assumption 3.3, along with our accompanying Lemma 3.19. We postpone the proof of this lemma to Section 3.6.3. Note that whether Y is continuous or discrete, Assumption 3.3 is more apparent and acceptable than the original argument.

Lemma 3.19. *Given Assumptions 3.1 and 3.3, the function*

$$p \mapsto \mathbb{E}[\mathbb{E}[g(Y)|p(X)]^2]$$

is continuous on $\mathcal{F}$ for all $g \in \mathcal{H}_\mathcal{Y}$.

It is worth noting that we removed the assumption (A-2) of Fukumizu et al. [34] since it can be weakened and subsumed to our ground assumption that $\mathcal{H}_\mathcal{Z}$ is characteristic. We will demonstrate this later in Section 3.6.2 and Section 3.6.3.

Finally, under Assumptions 3.1-3.3, we have the following desirable consistency result even in our expanded framework.

Theorem 3.20. *Suppose that $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ is characteristic, and that the regularization parameter ε in* (3.21) *satisfies*

$$\varepsilon_n \to 0 \quad \text{and} \quad n^{1/2}\varepsilon_n \to \infty \quad \text{as} \quad n \to \infty. \tag{3.22}$$

Let $\hat{p}^{(n)}$ be a member of the nonempty empirical solution set (3.21)*. Then, the set of optimal population solutions,* $\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)})$*, is nonempty. Furthermore, we have the convergence*

$$\mathrm{Tr}\left(\widehat{\Sigma}^{(n)}_{YY|\hat{p}^{(n)}(X)}\right) \to \mathrm{Tr}(\Sigma_{YY|p'(X)})$$

in probability, where p' is a population solution in $\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)})$*.*

Note that the theorem could be stated directly with the solution set $\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)})$. However, we avoided describing it that way because it requires additional, unnecessary arguments, particularly since the set arg min may not necessarily be measurable.

3.6.2 Pulling Back to the Original Domain

Before giving a detailed proof for Theorem 3.20, we make a remark on how we pull back all the notions defined in terms of the target domain to the original domain $\mathcal{X}$. In fact, Fukumizu et al. [34] stated all the results using the pullbacked covariance operators and RKHSs, but they did not recognize that the covariance operators are coherent with pullbacks, made them impose the redundant assumption (A-2) in Section 4. We elaborate its detail as follows.

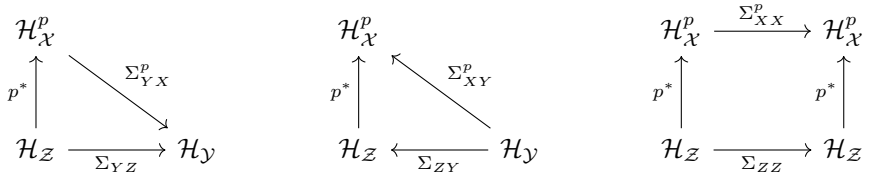
We can pull back the RKHS $\mathcal{H}_\mathcal{Z}$ along the map $p : \mathcal{X} \to \mathcal{Z}$ to the space of functions on $\mathcal{X}$ via composition:

$$\mathcal{H}^p_\mathcal{X} := \{f : \mathcal{X} \to \mathbb{R} \mid f = g \circ p \text{ for some } g \in \mathcal{H}_\mathcal{Z}\}, \tag{3.23}$$

called the *pullback* of $\mathcal{H}_\mathcal{Z}$ along the map p. It is known that the pullback space $\mathcal{H}^p_\mathcal{X}$ is also an RKHS on $\mathcal{X}$ with the *pullback kernel* $k^p_\mathcal{X}(x, x') := k_\mathcal{Z}(p(x), p(x'))$, and is isomorphic to the orthogonal complement of the vanishing space of p, $\{g \in \mathcal{H}_\mathcal{Z} \mid g \circ p = 0\}^\perp \subset \mathcal{H}_\mathcal{Z}$ [94, Theorem 2.9]. The isomorphism is naturally given by the restriction of the *pullback operator* $p^* : \mathcal{H}_\mathcal{Z} \to \mathcal{H}^p_\mathcal{X}$, sending $g \in \mathcal{H}_\mathcal{Z}$ to $g \circ p \in \mathcal{H}^p_\mathcal{X}$.

Let Σ^p_{YX}, Σ^p_{XY}, and Σ^p_{XX} denote the cross-covariance operators with respect to the pullback kernel $k^p_\mathcal{X}$, and define $\Sigma^p_{YY|X}$ similarly. We also define the pullbacked empirical operators $\widehat{\Sigma}^p_{**}$ similarly. It is natural to expect that these operators defined upon the space $\mathcal{H}^p_\mathcal{X}$ are coherent with the original covariance operators depending on $\mathcal{H}_\mathcal{Z}$, such as Σ_{YZ} and $\Sigma_{YY|Z}$, where $Z = p(X)$. Also, since those conditional covariance operators are sending $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}_\mathcal{Y}$, we instinctively expect that $\Sigma_{YY|Z} = \Sigma^p_{YY|X}$. The following result demonstrates that all these expectations are correct, enabling us to work on the convenient domain depending on situations.

Proposition 3.21. *The pullback operator $p^* : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}^p_{\mathcal{X}}$ pulls back the covariance operators coherently, meaning that the following diagrams are commutative.*



These commutativities also hold for the empirical operators $\widehat{\Sigma}_{Z*}$ and their pullbacks $\widehat{\Sigma}^p_{X*}$.

Proof. Write $\phi(Z) = k_{\mathcal{Z}}(Z, \cdot) \in \mathcal{H}_{\mathcal{Z}}$ and $\psi(Y) = k_{\mathcal{Y}}(Y, \cdot) \in \mathcal{H}_{\mathcal{Y}}$ as before, and let $\widetilde{\phi}(Z) = \phi(Z) - m_Z$ and $\widetilde{\psi}(Y) = \psi(Y) - m_Y$, where the m_* denote the mean embeddings. Recall that the covariance operator Σ_{YZ} is written as

$$\Sigma_{YZ} = \mathbb{E}[\widetilde{\psi}(Y) \otimes \widetilde{\phi}(Z)] \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Z}}.$$

Since Z is the image of X under the map $p : \mathcal{X} \to \mathcal{Z}$, we can explicitly pullback the $\phi(Z)$ as

$$p^*\phi(Z) = k_{\mathcal{X}}(p(X), p(\cdot)) = k^p_{\mathcal{X}}(X, \cdot) \in \mathcal{H}^p_{\mathcal{X}},$$

and m_Z is pullbacked similarly as

$$p^* m_Z = p^*\mathbb{E}[\phi(Z)] = \mathbb{E}[p^*\phi(Z)] = \mathbb{E}[k^p_{\mathcal{X}}(X, \cdot)] \in \mathcal{H}^p_{\mathcal{X}}$$

by Theorem 3.5 since the pullback operator p^* is bounded. Therefore, the cross-covariance operator $\Sigma^p_{YX} : \mathcal{H}^p_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ can be written as

$$\Sigma^p_{YX} = \mathbb{E}[\widetilde{\psi}(Y) \otimes p^*\widetilde{\phi}(Z)] \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}^p_{\mathcal{X}},$$

which establishes the commutative equality $\Sigma^p_{YX} p^* = \Sigma_{YZ}$. The remaining results, including the commutativity of empirical operators, are proved similarly, and we omit the proof. □

One important consequence of the preceding proposition is that, given a positive constant $\varepsilon_n > 0$, we have

$$\widehat{\Sigma}_{YY|p(X)} = \widehat{\Sigma}^p_{YY|X} \text{ on } \mathcal{H}_{\mathcal{Y}}.$$

We can check this equality by a repeated applications of the commutativity result:

$$\begin{aligned}
\widehat{\Sigma}^p_{YX}(\widehat{\Sigma}^p_{XX} + \varepsilon_n I)^{-1}\widehat{\Sigma}^p_{XY} &= \widehat{\Sigma}^p_{YX}(\widehat{\Sigma}^p_{XX} + \varepsilon_n I)^{-1} p^* \widehat{\Sigma}_{ZY} \\
&= \widehat{\Sigma}^p_{YX} p^* (\widehat{\Sigma}_{ZZ} + \varepsilon_n I)^{-1}\widehat{\Sigma}_{ZY} \\
&= \widehat{\Sigma}_{YZ}(\widehat{\Sigma}_{ZZ} + \varepsilon_n I)^{-1}\widehat{\Sigma}_{ZY}.
\end{aligned}$$

To prove the similar result on the population operators $\Sigma_{YY|p(X)}$ and $\Sigma^p_{YY|X}$, we have to additionally pullback the correlation operators V_{YZ} and V_{ZY}. The following result is also intuitively clear, but requires the uniqueness property of the correlation operators.

Proposition 3.22. *Let V_{YZ} and V^p_{YX} be the correlation operators satisfying*

$$\Sigma_{YZ} = \Sigma^{1/2}_{YY} V_{YZ} \Sigma^{1/2}_{ZZ} \quad \text{and} \quad \Sigma^p_{YX} = \Sigma^{1/2}_{YY} V^p_{YX} (\Sigma^p_{XX})^{1/2}.$$

Then, the correlation operators are also coherent with the pullback operator p^*, that is, we have

$$V_{YZ} = V^p_{YX} p^* \quad \text{and} \quad V^p_{XY} = p^* V_{ZY}.$$

Proof. We prove the similar commutativity to Proposition 3.21 for the square-root operators $\Sigma_{ZZ}^{1/2}$ and $(\Sigma_{XX}^p)^{1/2}$. Note first that $\overline{\mathrm{ran}}(\Sigma_{ZZ}) \subseteq (\ker p^*)^\perp \cong \mathcal{H}_{\mathcal{X}}^p$ since, for all $h \in \mathcal{H}_{\mathcal{Z}}$ and $l \in (\ker p^*)^\perp$, we have

$$\langle l, \Sigma_{ZZ} h \rangle_{\mathcal{H}_{\mathcal{Z}}} = \mathrm{Cov}[h(Z), l(p(X))] = 0.$$

Considering the spectral decomposition with a CONS $\{e_i\} \subset \overline{\mathrm{ran}}(\Sigma_{ZZ})$,

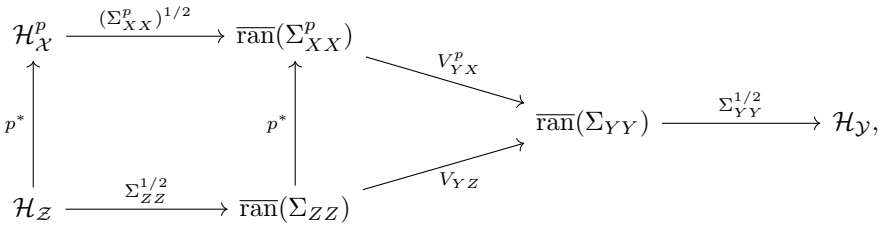
$$\Sigma_{ZZ} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i \in \mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{Z}},$$

we have the following equality from Proposition 3.21:

$$\Sigma_{XX}^p = \sum_{i=1}^{\infty} \lambda_i p^*(e_i) \otimes p^*(e_i) \in \mathcal{H}_{\mathcal{X}}^p \otimes \mathcal{H}_{\mathcal{X}}^p. \tag{3.24}$$

Since the pullback operator behaves as an isometry on $(\ker p^*)^\perp$, we conclude that $\langle p^*(e_i), p^*(e_j) \rangle_{\mathcal{H}_{\mathcal{X}}^p} = \delta_{ij}$, meaning that the equality Equation (3.24) is a spectral decomposition of Σ_{XX}^p. It is then straightforward to see that $p^* \Sigma_{ZZ}^{1/2} = (\Sigma_{XX}^p)^{1/2} p^*$ using the spectral decompositions of the square-root operators.

Based on such a commutativity, we may draw the following diagram



where the square on the left hand side is commutative. The commutativity of the center triangle follows from the uniqueness of the correlation operator V_{YZ}, described in Theorem 3.12. Therefore, we have the equality $V_{YZ} = p^* V_{YX}^p$, and the other equality can also be proven in the same way. □

Consequently, we have the equality $V_{YX}^p V_{XY}^p = V_{YZ} V_{ZY}$, implying that

$$\Sigma_{YY|p(X)} = \Sigma_{YY|X}^p \quad \text{as an operator} \quad \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}. \tag{3.25}$$

Recall that we stated our theory in terms of the projected random variable $Z = p(X)$, as this allows us to leverage the fixed RKHS $\mathcal{H}_{\mathcal{Z}}$ and its characteristic properties. This perspective is particularly beneficial in proving Lemma 3.30, where we demonstrate that the assumption (A-2) of Fukumizu et al. [34] is not essential. However, we also notice that in some instances, such as when proving Lemma 3.29, it is more practical to work within the original domain $\mathcal{X}$ when we take computational advantages of L^2 spaces. Regarding L^2 spaces, the space $L^2(\mathbb{P}_Z)$ varies as p changes, whereas $L^2(\mathbb{P}_X)$ remains fixed. Hence, it is worth recognizing the flexibility of our result of this subsection, which enables us to alternate between these two perspectives based on the requirements of the task at hand. We will continue the related and detailed discussions in the upcoming subsection, where we prove our main results.

3.6.3 Proof of the Consistency Result

In this section, our approach mostly aligns with the theoretical flow established by Fukumizu et al. [34]. The uniform convergence of Proposition 3.23 establishes the desired consistency, and the ingredients for verifying this uniform convergence are similar. We thoroughly analyze the continuity results, which are essential elements of the uniform convergence, within our broader class $\mathcal{F}$ as we no longer have

powerful properties of orthogonal matrices. Through this process, we will explore how Assumptions 3.1, 3.2, and 3.3 essentially substitute for the roles played by orthogonal matrices in the original theory. This exploration will enable us to effectively broaden the KDR framework to encompass a more diverse and interpretable range of dimension reduction functions.

Proposition 3.23. *Given the same assumptions as in Theorem 3.20, the trace of the population conditional covariance operator* $\mathrm{Tr}(\Sigma_{YY|p(X)})$ *is continuous on* $\mathcal{F}$*, and*

$$\sup_{p\in\mathcal{F}}\left|\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) - \mathrm{Tr}(\Sigma_{YY|p(X)})\right| \to 0, \qquad \text{as } n\to\infty, \text{ in probability.}$$

It is straightforward to prove the consistency of our estimator using this uniform convergence.

Proof of Theorem 3.20 given Proposition 3.23. The following proof is a standard theory of M-estimators. Note that the continuity of $\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)})$ under Assumption 3.1 is addressed in Section 3.6.1. Hence, the solutions sets

$$\arg\min_{p\in\mathcal{F}}\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) \quad\text{and}\quad \arg\min_{p\in\mathcal{F}}\mathrm{Tr}(\Sigma_{YY|p(X)})$$

are nonempty as the class $\mathcal{F}$ is compact. Pick minimizers $\hat{p}^{(n)}$ and p' from each solution set, respectively.

For a positive number $\varepsilon > 0$, there exists a large number $N > 0$ such that

$$\sup_{p\in\mathcal{F}}\left|\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) - \mathrm{Tr}(\Sigma_{YY|p(X)})\right| \le \varepsilon_n$$

for all $n \ge N$ at least probability $1-\varepsilon_n$. Then, using such a uniform convergence twice, we obtain the following two inequalities:

$$\begin{aligned}
&\mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|\hat{p}^{(n)}(X)}\Big) \le \mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|p'(X)}\Big) \le \mathrm{Tr}\big(\Sigma_{YY|p'(X)}\big) + \varepsilon_n, \text{ and}\\
&\mathrm{Tr}\big(\Sigma_{YY|p'(X)}\big) \le \mathrm{Tr}\big(\Sigma_{YY|\hat{p}^{(n)}(X)}\big) \le \mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|\hat{p}^{(n)}(X)}\Big) + \varepsilon_n,
\end{aligned}$$

with probability $\ge 1-\varepsilon_n$. This implies, with the same probability,

$$\left|\mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|\hat{p}^{(n)}(X)}\Big) - \mathrm{Tr}\big(\Sigma_{YY|p'(X)}\big)\right| \le \varepsilon_n$$

that concludes the proof of the desired consistency. □

In what follows we give the proof of Proposition 3.23 through a series of lemmas. The sequence

$$\sup_{p\in\mathcal{F}}\left|\mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|p(X)}\Big) - \mathrm{Tr}(\Sigma_{YY|p(X)})\right|$$

is decomposed into two parts,

$$\sup_{p\in\mathcal{F}}\left|\mathrm{Tr}\Big(\widehat{\Sigma}^{(n)}_{YY|p(X)}\Big) - \mathscr{T}_{\varepsilon_n}(p)\right| \quad\text{and}\quad \sup_{p\in\mathcal{F}}\left|\mathscr{T}_{\varepsilon_n}(p) - \mathrm{Tr}(\Sigma_{YY|p(X)})\right| \tag{3.26}$$

where $\mathscr{T}_{\varepsilon_n}(p) = \mathrm{Tr}\big(\Sigma_{YY} - \Sigma_{Yp(X)}(\Sigma_{p(X)p(X)} + \varepsilon_n I)^{-1}\Sigma_{p(X)Y}\big)$ is the regularized population conditional covariance, viewed as a function on $\mathcal{F}$. Note that the first part indicates the uniform convergence between the empirical conditional covariance and the regularized population version, and the second part shows that the regularized conditional covariance uniformly converges to the actual conditional covariance. We prove the uniform convergence of the first part using Lemma 3.24 and Lemma 3.25 and then prove the second part using Lemma 3.28, Lemma 3.29, and Lemma 3.30. For the following, see Section 3.3.1 for the definition of the Hilbert-Schmidt norm of operators of Hilbert spaces.

Lemma 3.24 (Fukumizu et al. [34, Lemma 8])**.** *For fixed n and $p \in \mathcal{F}$, letting $Z = p(X)$, we have*

$$\begin{aligned}&\left|\mathrm{Tr}\left(\widehat{\Sigma}^{(n)}_{YY|Z}\right) - \mathrm{Tr}\big(\Sigma_{YY} - \Sigma_{YZ}(\Sigma_{ZZ} + \varepsilon_n I)^{-1}\Sigma_{ZY}\big)\right| \\ &\quad \le \frac{1}{\varepsilon_n}\Big\{(\|\widehat{\Sigma}_{YZ}\|_{HS} + \|\Sigma_{YZ}\|_{HS})\|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{HS} + \mathrm{Tr}(\Sigma_{YY})\|\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\|\Big\} + \left|\mathrm{Tr}(\widehat{\Sigma}_{YY} - \Sigma_{YY})\right|.\end{aligned}$$

See the reference for the proof of this lemma. Since the spectral decomposition of self-adjoint operators shows that the operator norm is bounded by the Hilbert-Schmidt norm, $\|\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\| \le \|\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\|_{HS}$, it is enough to show the following lemma to guarantee the convergence of the first part of the uniform convergence. The following proof corrects some minor mistakes in the original proof and is simpler with our non-pullbacked covariance operators.

Lemma 3.25. *Under the Lipschitz condition of Assumption 3.2, we have*

$$\sup_{p\in\mathcal{F}} \|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{HS}, \ \sup_{p\in\mathcal{F}} \|\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\|_{HS}, \quad \textit{and} \quad \left|\mathrm{Tr}(\widehat{\Sigma}_{YY} - \Sigma_{YY})\right|$$

are of $O_p\big(\frac{1}{\sqrt{n}}\big)$ as $n \to \infty$, where Z denotes the p-dependent variable $p(X)$.

Proof. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ denote a random i.i.d. sample from the joint distribution of (X, Y). Write the centered random elements of $\mathcal{H}_\mathcal{Z}$ and $\mathcal{H}_\mathcal{Y}$ as (we slightly change our notation here)

$$\begin{aligned}\phi(p) &= k_\mathcal{Z}(p(X), \cdot) - \mathbb{E}[k_\mathcal{Z}(p(X), \cdot)], \qquad \psi = k_\mathcal{Y}(Y, \cdot) - \mathbb{E}[k_\mathcal{Y}(Y, \cdot)], \\ \phi_i(p) &= k_\mathcal{Z}(p(X_i), \cdot) - \mathbb{E}[k_\mathcal{Z}(p(X), \cdot)], \qquad \psi_i = k_\mathcal{Y}(Y_i, \cdot) - \mathbb{E}[k_\mathcal{Y}(Y, \cdot)].\end{aligned}$$

Then, $\phi, \phi_1, \dots, \phi_n$ and $\psi, \psi_1, \dots, \psi_n$ are also i.i.d. with zero mean. We can write the objectives using these new notations as

$$\begin{aligned}\mathrm{Tr}(\widehat{\Sigma}_{YY} - \Sigma_{YY}) &= \frac{1}{n}\sum_{i=1}^n \left\|\psi_i - \frac{1}{n}\sum_{j=1}^n \psi_j\right\|^2_{\mathcal{H}_\mathcal{Y}} - \ \mathbb{E}\|\psi\|^2_{\mathcal{H}_\mathcal{Y}} = \frac{1}{n}\sum_{i=1}^n \|\psi_i\|^2_{\mathcal{H}_\mathcal{Y}} - \mathbb{E}\|\psi\|^2_{\mathcal{H}_\mathcal{Y}} - \left\|\frac{1}{n}\sum_{i=1}^n \psi_i\right\|^2_{\mathcal{H}_\mathcal{Y}}, \\ \|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{HS} &= \left\|\frac{1}{n}\sum_{i=1}^n \left(\psi_i - \frac{1}{n}\sum_{j=1}^n \psi_j\right) \otimes \sum_{i=1}^n \left(\phi_i(p) - \frac{1}{n}\sum_{j=1}^n \phi_j(p)\right) - \mathbb{E}[\psi \otimes \phi(p)]\right\|_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} \\ &\le \left\|\frac{1}{n}\sum_{i=1}^n (\psi_i \otimes \phi_i(p) - \mathbb{E}[\psi \otimes \phi(p)])\right\|_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} + \left\|\frac{1}{n}\sum_{i=1}^n \psi_i\right\|_{\mathcal{H}_\mathcal{Y}} \left\|\frac{1}{n}\sum_{i=1}^n \phi_i(p)\right\|_{\mathcal{H}_\mathcal{Z}},\end{aligned}$$

and $\|\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\|_{HS}$ is expressed similarly as $\|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{HS}$ by replacing $\psi \to \phi$. For the trace $\mathrm{Tr}(\widehat{\Sigma}_{YY} - \Sigma_{YY})$, we reaily compute that

$$\left|\mathrm{Tr}(\widehat{\Sigma}_{YY} - \Sigma_{YY})\right| \le \left|\frac{1}{n}\sum_{i=1}^n \|\psi_i\|^2_{\mathcal{H}_\mathcal{Y}} - \mathbb{E}\|\psi\|^2_{\mathcal{H}_\mathcal{Y}}\right| + \left\|\frac{1}{n}\sum_{i=1}^n \psi_i\right\|^2_{\mathcal{H}_\mathcal{Y}},$$

resulting in the order $O_p\big(\frac{1}{\sqrt{n}}\big)$ from the central limit theorem (Theorem 3.8).

To prove convergence for the Hilbert-Schmidt norms, we first note that

$$\|\phi_i(p)\|^2_{\mathcal{H}_\mathcal{Z}} = \langle k_\mathcal{Z}(p(X_i), \cdot) - \mathbb{E}[k_\mathcal{Z}(p(X), \cdot)], k_\mathcal{Z}(p(X_i), \cdot) - \mathbb{E}[k_\mathcal{Z}(p(X), \cdot)]\rangle_{\mathcal{H}_\mathcal{Z}} \le 4C^2,$$

where C is a large constant such that $k_\mathcal{Z} \le C^2$. Similar computation is possible for ψ; we say $\|\psi_i\|_{\mathcal{H}_\mathcal{Y}} \le 2C'$. Then, for different projections $p_1, p_2 \in \mathcal{F}$, we have

$$\|\psi_i \otimes \phi_i(p_1) - \psi_i \otimes \phi_i(p_2)\|_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} = \|\psi_i\|_{\mathcal{H}_\mathcal{Y}} \|\phi_i(p_1) - \phi_i(p_2)\|_{\mathcal{H}_\mathcal{Z}} \le 2C'\varphi(X_i)\, d(p_1, p_2)$$

since

$$\begin{aligned}\|\phi(p_1) - \phi(p_2)\|_{\mathcal{H}_\mathcal{Z}} &\leq \|k_\mathcal{Z}(p_1(X), \cdot) - k_\mathcal{Z}(p_2(X), \cdot)\|_{\mathcal{H}_\mathcal{Z}} + \|\mathbb{E}[k_\mathcal{Z}(p_1(X), \cdot)] - \mathbb{E}[k_\mathcal{Z}(p_2(X), \cdot)]\|_{\mathcal{H}_\mathcal{Z}} \\ &= \|k_\mathcal{Z}(p_1(X), \cdot) - k_\mathcal{Z}(p_2(X), \cdot)\|_{\mathcal{H}_\mathcal{Z}} + \mathbb{E}[\|k_\mathcal{Z}(p_1(X), \cdot) - k_\mathcal{Z}(p_2(X), \cdot)\|_{\mathcal{H}_\mathcal{Z}}] \\ &\leq 2\varphi(X)\, d(p_1, p_2)\end{aligned}$$

by Assumption 3.2. We also similarly obtain

$$\begin{aligned}\|\phi(p_1) \otimes \phi(p_1) - \phi(p_2) \otimes \phi(p_2)\|_{\mathcal{H}_\mathcal{Z}} &\leq \{\|\phi(p_1)\|_{\mathcal{H}_\mathcal{Z}} + \|\phi(p_2)\|_{\mathcal{H}_\mathcal{Z}}\}\|\phi(p_1) - \phi(p_2)\|_{\mathcal{H}_\mathcal{Z}} \\ &\leq 4C\varphi(X)\, d(p_1, p_2).\end{aligned}$$

These observations are all we need to apply Proposition 3.26 that will be described subsequently. As the requirements (3.27) are fulfilled, the following proposition completes the proof. □

Proposition 3.26 (see [34, Proposition 15])**.** *Let $\mathcal{H}$ be a Hilbert space. Suppose that $X, X_1, \ldots, X_n$ are i.i.d random variables on $\mathcal{X}$, and suppose that $F : \mathcal{X} \times \mathcal{F} \to \mathcal{H}$ is a Borel measurable map. If*

$$\begin{aligned}&\sup_{p\in\mathcal{F}} \|F(x,p)\|_\mathcal{H} < \infty \quad \text{for all} \quad x \in \mathcal{X}, \ \text{and} \\ &\|F(x,p_1) - F(x,p_2)\|_\mathcal{H} \leq \varphi(x)\, d(p_1,p_2) \quad \text{for all} \quad p_1, p_2 \in \mathcal{F},\end{aligned} \tag{3.27}$$

for some $\varphi \in L^2(\mathbb{P}_X)$, then we have the following uniform rate

$$\sup_{p\in\mathcal{F}} \left\| \frac{1}{n}\sum_{i=1}^n (F(X_i,p) - \mathbb{E}[F(X,p)])) \right\|_\mathcal{H} = O_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{as} \quad n \to \infty.$$

The proof of this proposition involves some empirical process theory in Hilbert spaces and is given in the appendix of Fukumizu et al. [34]. We need to fix the Hilbert space we work on to apply this proposition, so we work on $\mathcal{H}_\mathcal{Z}$ rather than the pullbacked space $\mathcal{H}^p_\mathcal{X}$ here. From Lemma 3.24 and Lemma 3.25, we immediately obtain the following rate of uniform convergence:

Corollary 3.27. *Under Assumption 3.2 and the condition* (3.22) *on the regularization parameter ε_n, we have*

$$\sup_{p\in\mathcal{F}} \left| \mathrm{Tr}\left(\widehat{\Sigma}^{(n)}_{YY|p(X)}\right) - \mathscr{T}_{\varepsilon_n}(p) \right| = O_p\left(\frac{1}{\varepsilon_n\sqrt{n}}\right)$$

as $n \to \infty$.

We now prove the uniform convergence on the population level, the second part of (3.26). We begin with the pointwise convergence of the function $\mathscr{T}_\varepsilon(p)$ for each fixed $p \in \mathcal{F}$. Since the variable of interest is fixed as $Z = p(X)$, the proof of the following lemma is identical to Lemma 11 of Fukumizu et al. [34].

Lemma 3.28 ([34, Lemma 11])**.** *$\mathscr{T}_\varepsilon(p) \to \mathrm{Tr}[\Sigma_{YY|p(X)}]$ as $\varepsilon \to 0$, for each $p \in \mathcal{F}$.*

We then establish the continuity results to derive uniform convergence. Here, the continuity Assumption 3.1 plays a crucial role. It is convenient to work with the pullbacked notions in the following lemma.

Lemma 3.29. *Under Assumption 3.1, $\mathscr{T}_\varepsilon(p)$ is continuous on $\mathcal{F}$ if $\varepsilon > 0$.*

Proof. It suffices to show that the mapping

$$p \mapsto \mathrm{Tr}(\Sigma_{YZ}(\Sigma_{ZZ} + \varepsilon I)^{-1}\Sigma_{ZY}) = \mathrm{Tr}(\Sigma^p_{YX}(\Sigma^p_{XX} + \varepsilon I)^{-1}\Sigma^p_{XY})$$

is continuous, where the equality was studied in Section 3.6.2. Let $\{g_i\}$ be a CONS of $\mathcal{H}_\mathcal{Y}$ so that

$$\mathrm{Tr}(\Sigma^p_{YX}(\Sigma^p_{XX}+\varepsilon I)^{-1}\Sigma^p_{XY}) = \sum_{i=1}^{\infty}\langle g_i, \Sigma^p_{YX}(\Sigma^p_{XX}+\varepsilon I)^{-1}\Sigma^p_{XY}g_i\rangle_{\mathcal{H}_\mathcal{Y}}.$$

The dominated convergence theorem ensures that it suffices to show, for any fixed $g \in \mathcal{H}_\mathcal{Y}$, the mapping

$$p \mapsto \langle g, \Sigma^p_{YX}(\Sigma^p_{XX}+\varepsilon I)^{-1}\Sigma^p_{XY}g\rangle_{\mathcal{H}_\mathcal{Y}}$$

is continuous. As the kernels are bounded, we have the natural inclusions of function spaces $\mathcal{I}^p_X : \mathcal{H}^p_\mathcal{X} \to L^2(\mathbb{P}_X)$ and $\mathcal{I}_Y : \mathcal{H}_\mathcal{Y} \to L^2(\mathbb{P}_Y)$. Also, the explicit formula of Σ_{YX} in (3.9) suggests that the covariance operators are naturally seen as integral operators. For instance, if we define $J^p_{YX} : L^2(\mathbb{P}_X) \to L^2(\mathbb{P}_Y)$ as

$$J^p_{YX}(f)(y) = \mathrm{Cov}[f(p(X)), k_\mathcal{Y}(y, Y)], \quad \forall f \in L^2(\mathbb{P}_X),$$

we have the commutative relation $\mathcal{I}_Y\Sigma^p_{YX} = J^p_{YX}\mathcal{I}^p_X$. We similarly get $\mathcal{I}^p_X\Sigma^p_{XX} = J^p_{XX}\mathcal{I}^p_X$, and thus $\mathcal{I}^p_X(\Sigma^p_{XX}+\varepsilon I)^{-1} = (J^p_{XX}+\varepsilon I)^{-1}\mathcal{I}^p_X$. Using these relations, we have

$$\langle g, \Sigma^p_{YX}(\Sigma^p_{XX}+\varepsilon I)^{-1}\Sigma^p_{XY}g\rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{Cov}\big[g(Y), \big((J^p_{XX}+\varepsilon I)^{-1}J^p_{XY}\big)(X)\big]$$

and we can solve the continuity problem in L^2 spaces with the L^2 operator norms.

Letting $X' \sim X$ an i.i.d. variable and for $p_1, p_2 \in \mathcal{F}$, we compute as

$$\begin{aligned}
\|(J^{p_1}_{XY} - J^{p_2}_{XY})g\|^2_{L^2(\mathbb{P}_Y)} &= \mathbb{E}_{X'}\big[\mathrm{Cov}_{X,Y}[k^{p_1}_\mathcal{X}(X,X') - k^{p_2}_\mathcal{X}(X,X'), g(Y)]^2\big] \\
&\le \mathbb{E}_{X'}[\mathrm{Var}_X[k^{p_1}_\mathcal{X}(X,X') - k^{p_2}_\mathcal{X}(X,X')] \cdot \mathrm{Var}[g(Y)]] \\
&\le \mathbb{E}_{X,X'}\Big[\big\{k_\mathcal{Z}(p_1(X),p_1(X')) - k_\mathcal{Z}(p_2(X),p_2(X'))\big\}^2\Big]\|g\|^2_{L^2(\mathbb{P}_Y)}.
\end{aligned}$$

By bounded convergence theorem and the *continuity of evaluations* (Assumption 3.1), the last expectation goes to zero as $d(p_1,p_2) \to 0$, proving the continuity of $p \mapsto J^p_{XY}$. Then, the continuity of $(J^p_{XX}+\varepsilon I)^{-1}$ is proved using the continuity of J^p_{XX} as

$$\begin{aligned}
\big\|(J^{p_1}_{XX}+\varepsilon I)^{-1} - (J^{p_2}_{XX}+\varepsilon I)^{-1}\big\| &= \big\|(J^{p_1}_{XX}+\varepsilon I)^{-1}(J^{p_2}_{XX} - J^{p_1}_{XX})(J^{p_2}_{XX}+\varepsilon I)^{-1}\big\| \\
&\le \frac{1}{\varepsilon^2}\|J^{p_2}_{XX} - J^{p_1}_{XX}\|,
\end{aligned}$$

which finishes the proof. □

The final ingredient of the desired uniform convergence is that the population objective $\mathrm{Tr}(\Sigma_{YY|p(X)})$ is continuous on $\mathcal{F}$. Working on $\mathcal{H}_\mathcal{Z}$ of the target domain, we can perform much simpler proof than that in the reference. As mentioned in Section 3.6.1, we need the condition that $\mathcal{H}_\mathcal{Z}$ is characteristic and Lemma 3.19.

Lemma 3.30. *Under Assumptions 3.1 and 3.3,* $\mathrm{Tr}\big(\Sigma_{YY|p(X)}\big)$ *is a continuous function on* $\mathcal{F}$*, provided that the RKHS* $\mathcal{H}_\mathcal{Z}$ *is characteristic.*

Proof. By the same argument as the preceding lemma, it is enough to show that the mapping $p \mapsto \langle g, \Sigma_{YY|p(X)}g\rangle_{\mathcal{H}_\mathcal{Y}}$ is continuous on $\mathcal{F}$, for any $g \in \mathcal{H}_\mathcal{Y}$. Since $\mathcal{H}_\mathcal{Z}$ is characteristic, Proposition 3.14 implies that

$$\begin{aligned}
\langle g, \Sigma_{YY|p(X)}g\rangle_{\mathcal{H}_\mathcal{Y}} &= \mathbb{E}[\mathrm{Var}[g(Y)|p(X)]] \\
&= \mathbb{E}[g(Y)^2] - \mathbb{E}[\mathbb{E}[g(Y)|p(X)]].
\end{aligned}$$

Lemma 3.19 therefore concludes the proof, which is a consequence of Assumptions 3.1 and 3.3. □

We give our proof of Lemma 3.19 below, which states that the mapping $p \mapsto \mathbb{E}[\mathbb{E}[g(Y)|p(X)]^2]$ is continuous on $\mathcal{F}$ for all $g \in \mathcal{H}_\mathcal{Y}$.

Proof of Lemma 3.19. Recall that every function $g \in \mathcal{H}_\mathcal{Y}$ is continuous and bounded on $\mathcal{Y}$ as we work on the compact domain $\mathcal{Y}$ and continuous kernels. This also holds for the kernel choices discussed in Section 3.4.2.

We first prove that, under Assumption 3.3, the conditional expectation $\mathbb{E}[g(Y)|X = x]$ is continuous on a $\mathbb{P}_X$-measure one subset of $\mathcal{X}$. Let $f_{X|Y}(x|y)$ be a conditional density function as given in the assumption. Note that the conditional density of Y given X may not exist because we allow Y can be a discrete random variable. Letting $\mathfrak{B}_\mathcal{Y}$ denote the Borel σ-field of $\mathcal{Y}$, the measure-theoretic version of the *Bayes theorem* (the conditional density of Y given X may not exist because Y may be a discrete variable) gives a regular conditional distribution $\kappa : \mathcal{X} \times \mathfrak{B}_\mathcal{Y} \to [0, 1]$ defined by

$$\kappa(x, B) = \begin{cases} \dfrac{\int_B f_{X|Y}(x|y)\, \mathbb{P}_Y(dy)}{\int_\mathcal{Y} f_{X|Y}(x|y)\, \mathbb{P}_Y(dy)} & \text{if the denominator is positive and finite,} \\ \text{arbitrary } \mu_x \in \mathcal{P}(\mathcal{Y}) & \text{otherwise.} \end{cases}$$

Here, the set N of elements $x \in \mathcal{X}$ for which the denominator $f_X(x) = \int_\mathcal{Y} f_{X|Y}(x|y)\, \mathbb{P}_Y(dy)$ is zero or infinite is a $\mathbb{P}_X$-null set. To verify this, for a measurable set $A \subseteq \mathcal{X}$, observe that

$$\begin{aligned} \mathbb{P}(X \in A) &= \int_\mathcal{Y} \left(\int_A f_{X|Y}(x|y) dx \right) \mathbb{P}_Y(dy) \\ &= \int_A \left(\int_\mathcal{Y} f_{X|Y}(x|y) \mathbb{P}_Y(dy) \right) dx = \int_A f_X(x) dx \leq 1 \end{aligned}$$

by Fubini's theorem. Therefore, $f_X(x)$ becomes a marginal density function for X, which verifies that $N = \{x \,|\, f_X(x) = 0\} \cup \{x \,|\, f_X(x) = \infty\}$ is a $\mathbb{P}_X$-null set. Then, for all $u \in \mathcal{X} \setminus N$, we compute that

$$\mathbb{E}[g(Y)|X = x] = \int_\mathcal{Y} g(y) \kappa(x, dy) = \frac{\int_\mathcal{Y} g(y) f_{X|Y}(x|y)\, \mathbb{P}_Y(dy)}{\int_\mathcal{Y} f_{X|Y}(x|y)\, \mathbb{P}_Y(dy)},$$

so the bounded convergence ensures that $\mathbb{E}[g(Y)|X = x]$ is continuous on $\mathcal{X} \setminus N$.

We then restrict our functions $p \in \mathcal{F}$ to the continuity set $\mathcal{X} \setminus N$ of $\mathbb{E}[g(Y)|X = x]$. Letting $p_1 = p|_{\mathcal{X} \setminus N}$ be the restriction, note that the σ-fields $\sigma(p(X))$ and $\sigma(p_1(X))$ are the same up to their *completion* with respect to the ambient probability space [50, Chapter 8] because $p^{-1}(B)$ and $p_1^{-1}(B)$ only differ by a null set, for all Borel sets $B \subseteq \mathcal{Z}$. A standard measure-theoretic argument [50, Exercise 1.9] and the almost sure uniqueness of the conditional expectation shows that $\mathbb{E}[g(Y)|p(X)] = \mathbb{E}[g(Y)|p_1(X)]$ almost surely on $\mathcal{X}$. Therefore,

$$\mathbb{E}[\mathbb{E}[g(Y)|p(X)]^2] = \mathbb{E}[\mathbb{E}[g(Y)|p_1(X)]^2],$$

which verifies that *we may assume* $\mathbb{E}[g(Y)|X = x]$ *is continuous on* $\mathcal{X}$ by replacing $\mathcal{X}$ with $\mathcal{X} \setminus N$. Here, we keep the metric on $\mathcal{F}$ after the replacement. Note that $\mathcal{X} \setminus N$ may not be compact, but the compactness of $\mathcal{X}$ is not required in the following arguments.

To complete the continuity proof, it suffices to show that the mapping

$$p \mapsto \mathbb{E}[g(Y)|p(X) = p(x)]$$

is continuous for all $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$, using the bounded convergence theorem. To describe the conditional expectation $\mathbb{E}[g(Y)|p(X) = p(x)]$ in terms of $\mathbb{E}[g(Y)|X = \cdot\,]$ without additional structural information

on p, we need an aid of the *disintegration theorem* [17]. The theorem guarantees that, for every $p_*\mathbb{P}_X$-a.e. $z \in \mathcal{Z}$, where p_* denotes the pushforward map of measures along the function $p : \mathcal{X} \to \mathcal{Z}$, there exists a Borel probability measure $\mathbb{P}_{X,z}$ on $\mathcal{X}$ that lives in the fiber $p^{-1}(z)$; i.e., $\mathbb{P}_{X,z}(E) = \mathbb{P}_{X,z}(E \cap p^{-1}(z))$ for all Borel sets $E \subseteq \mathcal{X}$. This disintegration measure gives rise to the following integral representation

$$\mathbb{E}[g(Y)|p(X) = p(x)] = \int_{\mathcal{X}} \mathbb{E}[g(Y)|X = u]\, d\mathbb{P}_{X,p(x)}(u)$$

for all $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$. Letting $\mathcal{Z}_0 \subseteq \mathcal{Z}$ a subset on which the disintegration measures $\mathbb{P}_{X,z}$, $z \in \mathcal{Z}_0$, are defined, Tjur [116, Section 4] proved that the mapping from $\mathcal{Z}_0$ to $\mathcal{P}(\mathcal{X})$, sending $z \mapsto \mathbb{P}_{X,z}$, is continuous in the weak topology of probability measures. Recall that $\mathbb{E}[g(Y)|X = u]$ is continuous on $\mathcal{X}$ and is furthermore bounded as $g \in \mathcal{H}_\mathcal{Y}$ is a bounded function. Therefore, the weak convergence and the continuity of the evaluation $p \mapsto p(x)$ ensure that the mapping $p \mapsto \mathbb{E}[g(Y)|p(X) = p(x)]$ is continuous in $p \in \mathcal{F}$, which finishes the proof. □

We finally establish the uniform convergence of the second part of (3.26), which completes the proof of Proposition 3.23.

Corollary 3.31. *Suppose that $\varepsilon_n \to 0$ as $n \to \infty$. Under Assumptions 3.1 and 3.3, we have*

$$\sup_{p \in \mathcal{F}} \left| \mathscr{T}_{\varepsilon_n}(p) - \mathrm{Tr}(\Sigma_{YY|p(X)}) \right| \to 0, \quad \textit{as} \quad n \to \infty.$$

Proof. We have shown in Lemma 3.29 and Lemma 3.30 that $\mathscr{T}_{\varepsilon_n}(p)$ and $\mathrm{Tr}(\Sigma_{YY|p(X)})$ are continuous on the compact set $\mathcal{F}$. Since the regularization exhibits monotonicity, $\mathscr{T}_\varepsilon(p) \geq \mathscr{T}_{\varepsilon'}(p)$ whenever $\varepsilon \geq \varepsilon'$, we have the desired uniform convergence by Theorem 7.13 of Rudin [93]. □

3.7 Conclusions and Discussions

This chapter established a generalization of the linear kernel dimension reduction theory on the Stiefel manifold to other specifically structured projection maps based on function-theoretic backgrounds and intuitions on conditional covariance operators. We proved that the empirical estimator of this generalized kernel dimension reduction is statistically consistent under mild continuity assumptions on the family of projections. Furthermore, we have refined some of the theoretical assumptions of the original theory by demonstrating the compatible nature of the conditional covariance operator and the pullback of RKHS.

This advancement will pave the way for interpretable dimension reduction in compositional data analysis. The forthcoming chapters will further illustrate this potential: Chapter 4 will introduce two classes of $\mathcal{F}$, the class of compositional variable selections and continuously relaxed importance weights, and Chapter 5 will explore the classes of amalgamations and their continuous extensions, with a particular focus on the latter. We also anticipate that we may naturally restrict the class of projections according to extrinsic information of data; for example, a genetic similarity constraint to $\mathcal{F}$ based on phylogenetic tree structure.

Beyond this, we envision the adaptation of our generalized framework to a wide range of structured datasets. Another intriguing possibility lies in exploring $\mathcal{F}$ as a parameterized class within a fixed-architecture neural network. These prospects represent compelling future research directions.

Lastly, we discuss the computational complexity of the KDR objective. As seen in the equation (3.21), the single computation of the trace of the empirical conditional covariance operator requires at

least $O(n^3)$ operations due to the matrix inversion. While this is not a critical issue for microbiome datasets, which typically have fewer than 1,000 samples, broadening the application of our framework will necessitate efforts to reduce computational complexity. One possible approach is employing low-rank approximations to kernel Gram matrices with the Woodbury identity for matrix inversion, as suggested by Chen et al. [20]. However, this may compromise the statistical guarantees demonstrated in Section 3.6, thus posing a new research problem to assess consistency under these low-rank approximations.

Chapter 4. Kernel Sufficient Dimension Reduction and Variable Selection for Compositional Data via Amalgamation

4.1 Introduction

Compositional data are multivariate data that consist of nonnegative values in which only the relative proportions of the components are meaningful. They are frequently normalized to sum to unity. Thus, compositional data with $d+1$ variables lie on the d-dimensional simplex $\Delta^d \subset \mathbb{R}^{d+1}$:

$$\Delta^d = \left\{ (x_0, \ldots, x_d) \middle| \sum x_i = 1,\ x_i \geq 0,\ \forall i \right\}.$$

This type of data appears commonly in many applications; for example, chemical compositions of honey in food science, mineral compositions of rocks in geology, and composition of product categories of customers' internet shopping carts. This research is primarily inspired by microbiome data, which measure the relative abundance of microbes that live in or on a human or animal's body.

Microbiomes have recently received much attention in medical research due to their association with various diseases and health-related attributes in humans [49, 40]. Modern sequencing technologies, such as 16S rRNA gene sequencing, are used to quantify the raw numbers of microbiomes. However, as the total number of counts varies greatly amongst the samples, the raw count data obtained in this manner must be viewed as compositional [59]. In addition, microbiome data often exhibit *high dimensionality* and contain *excess zeros*, i.e., there are much higher numbers of microbial taxa than available samples, and a large percentage, about 50% to 90%, of counts are zero [68]. Identifying relevant variables is a common and important task in the study of microbiome data because most taxa are unlikely to be associated with the response of interest [54]. Accurately chosen microbial variables can be used in subsequent analyses such as prediction with reduced computational cost and increased interpretability.

Despite the necessity of variable selection for high-dimensional sparse compositional data, there are few approaches that rigorously perform it. As pointed out in Susin et al. [113], the main difficulty lies in how to account for the compositional nature of the data, i.e., the spurious negative correlation due to the sum-to-one constraint [84]. Dominantly popular approaches to compositional variable selection are based on the log-ratio transformation designed to address this spurious correlation problem, which is sometimes referred to as CoDA (Compositional Data analysis) methods [3]; we will introduce some of them in Section 4.1.2.

However, these methods have a clear drawback that they cannot handle zeros in the data directly due to the log transformations, even though most of the compositional data dealt with today contain a large proportion of zeros. Researchers have then replaced zeros with small positive values, but the results of data analysis have been inconsistent depending on how the zeros are replaced [67]. More importantly, Park et al. [80] reveal that the *combination* of zero replacement and log-ratio transforms inevitably yields unexpected distortions in the data. They demonstrate how even very basic manifold structures of compositional data can be broken by such a combination of data translations, compromising the accuracy of subsequent data analyses. The challenges regarding such inconsistency and distortion have been widely documented in a variety of contexts including variable selection [78].

4.1.1 Our Contributions

This work presents a new variable selection framework for compositional data. It provides a solution to the two primary challenges in dealing with modern compositional data: high dimensionality and abundance of zeros. Our method does not rely on log-ratio transformation, thereby successfully overcoming the issues of inconsistency and distortion mentioned above. Inspired by Park et al. [80] who advocate the use of kernel methods for compositional data with a compelling geometric argument, our proposed approach is rooted in the existing kernel dimension reduction research by Fukumizu et al. [34, 32], Chen et al. [20], which will be briefly reviewed in Section 4.3.

In Section 4.2, we show that a nontrivial critical problem occurs when defining the reduced set of variables in compositional data. A process called *amalgamation* is suggested as a solution to this problem, based on which we propose a variable selection algorithm in Section 4.4. The proposed method aims to achieve sufficient dimension reduction (SDR) so that the (compositional) variables and the response become independent conditioning on the projected covariates onto the SDR subspace [60]. Minimizing the trace of the kernel conditional covariance operator after variable selection with amalgamation is shown to yield a consistent SDR. In the compositional context, this means that all information in the covariates relevant to the response is contained in some amalgamation of the original composition.

We also clarify the type of kernels to be used in classification and regression problems respectively, in order to ensure the SDR property. It is revealed that the linear kernel commonly used for regression is not universal and thus yields SDR under a rather restrictive population model. This finding corrects some results in Chen et al. [20].

Finally, we demonstrate the performance of the proposed method with synthetic and real microbiomes data in Section 4.5 and conclude the chapter with discussions in Section 4.6.

4.1.2 Related Works

Variable selection methods using kernels. Various kernel measures on probability distributions have been used in the literature to achieve adequate variable selections. For example, the Hilbert-Schmidt Independence Criterion (HSIC) [44] is applied to obtain maximal dependence between variables and the response, the conditional covariance operator is used to obtain conditional independence for SDR [20], and the Maximum Mean Discrepancy (MMD) [45] is applied to find marginally different variables between two samples. Among them, HSIC seems to be used more frequently, from the greedy algorithm of Song et al. [107] to continuously relaxed algorithms with regularizations [73, 126].

Several studies have also been conducted to test the significance of the selected variables using kernels. These methods are based on Lee et al. [52]'s pioneering work on *post-selection inference* (PSI), and kernel-based approaches have been successfully developed within this framework [127, 128, 62, 31]. However, because these kernel-based PSI algorithms utilize HSIC or MMD, which focus on marginal distributions of individual variables, they may not be suitable for compositional data due to the spurious correlation issue.

Variable selection methods for compositional data. A majority of variable selection methods in the literature are based on log-ratio transformations. Among them, the constrained lasso approach to the log-transformed data, in particular, has been extensively studied, where the constraint reflects the ratio computations and may further reflect grouping or tree structures [63, 103, 123, 66]. Rivera-Pinto et al. [90] alternatively propose a forward selection process using the log-ratio balance [27].

Recently, there has been a growing consensus on how to address the zero problem in compositional

data analysis, leading to the development of methods that do not use log-ratio transforms. Tomassi et al. [117] propose a likelihood-based SDR for compositional data as well as a variable selection method. However, their non-log-ratio approach uses a linear projection of the raw count matrix, whose structure is hardly interpretable. Wang [122] proposes a test for the differential abundance of each taxon based on a multinomial model for the count matrix. These methods are yet based on the assumption that the count data are drawn from specific distributions such as multinomial or Poisson distributions, whereas our proposed method does not impose such assumptions on the underlying distribution.

4.2 Compositional Variable Selection via Amalgamation

In many cases, dimension reduction does not end with identifying a subspace or a subset of relevant variables. The main goal of variable selection is mostly to improve the performance and interpretability of a predictive model. Thus it is necessary to attain a dimension-reduced dataset that is suitable for subsequent analyses. In the context of compositional data, it is crucial that the dimension-reduced data are also compositional. Intuitively, there are two ways of achieving this, namely *sub-composition* and *amalgamation* [3].

The sub-composition approach is simpler, in that it just *re-normalizes* the selected variables to make a composition. This method is widely used in practice because taking sub-compositions can be considered as orthogonally projecting data in the log-ratio geometry; see, for example, Section 4.6 of Pawlowsky-Glahn et al. [83]. However, the toy example below demonstrates that this approach may not yield learnable data.

Consider the following toy microbiome data $X = (X_0, X_1, X_2, X_3) \in \Delta^3$ with four covariates. Let $Y \in \{0, 1\}$ be a binary variable indicating the presence of a disease and assume that the *deficiency* of two taxa $\mathcal{S} = \{X_0, X_1\}$ causes the disease. Let $(x, 1)$ and $(x', 0)$ be two samples from (X, Y) with

$$x = (0.01, 0.01, 0.38, 0.6) \quad \text{and} \quad x' = (0.4, 0.4, 0.1, 0.1).$$

Suppose some variable selection is carried out and the variables in $\mathcal{S}$ are correctly selected. Then, both sub-compositions $x_{\mathcal{S}}$ and $x'_{\mathcal{S}}$ are $(0.5, 0.5)$ with different labels, which are unsuitable for further investigation. This is because relative abundance to the *total* is lost when taking sub-compositions. This problem exacerbates when there are many zeros in the data, which is almost always the case in microbiome studies. If the absence of taxa in $\mathcal{S}$ causes the disease, then the disease group's sub-composition will probably be entirely zero, which cannot be made compositional.

In contrast, amalgamation is an intuitive process to reduce the dimensionality of compositions. It is commonly used to organize microbiomes according to the phylogenetic tree structure [59]. The procedure involves defining integers c_i such that

$$0 = c_0 < c_1 < \cdots < c_{m+1} = d + 1,$$

and then taking $z_j = x_{c_j} + \cdots + x_{c_{j+1}-1}$, $j = 0, \ldots, m$, so that the resulting vector $(z_0, \ldots, z_m)$ lies on the lower dimensional simplex Δ^m. However, even though compositional data are frequently obtained through an amalgamation process, it has been hardly used for data analysis, because it is incompatible with the dominant log-ratio approaches [83]. In particular, amalgamations do not behave like linear projections in log-ratio geometry. Recent studies have attempted to reconcile amalgamation with log-ratio methods [41, 43], arguing that amalgamation yields better interpretation and is essential for certain types of compositional data, such as in geochemistry and mineralogy.

In this work, we argue that the controversy surrounding amalgamation becomes irrelevant in kernel methods in the sense of Park et al. [80] and amalgamation is the most valid way to perform dimension reduction or variable selection of compositional data. We state the variable selection framework as follows: if $\mathcal{S} = \{s_1, \ldots, s_m\} \subset \{0, \ldots, d\}$ is a subset of variables, then we propose to identify the projection map $p_{\mathcal{S}} : \Delta^d \to \Delta^m$,

$$p_{\mathcal{S}}(x_0, \ldots, x_d) = \left(x_{s_1}, \ldots, x_{s_m}, \sum_{j \notin \mathcal{S}} x_j \right). \tag{4.1}$$

By including a dummy variable that aggregates all unselected variables, this special case of amalgamation is intuitive and overcomes the issue of sub-composition discussed earlier, preserving information on the relative abundance to the total.

4.3 Sufficient Dimension Reduction and Variable Selection with Kernels

This section provides an overview on the principle of SDR and kernel variable selection derived from kernel dimension reduction. The latter discussion is largely credited to Chen et al. [20], who adopt the kernel dimension reduction (KDR) method of Fukumizu et al. [34] for the variable selection purpose.

4.3.1 Sufficient Dimension Reduction

Let (X, Y) be a joint random variable with a joint distribution $P_{X,Y}$ defined on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is a domain of covariates and $\mathcal{Y}$ is a domain of response. The general dimension reduction problem is described as finding a pair $(\mathcal{Z}, p)$ of a lower dimensional domain $\mathcal{Z} \subset \mathbb{R}^m$, $m \leq d$, and a projection map $p : \mathcal{X} \to \mathcal{Z}$ such that the variable $p(X)$ has enough information about Y. In case $p(X)$ retains all the relevant information of Y, it is called *sufficient dimension reduction* (SDR) and is theoretically defined as

$$P_{Y|p(X)} = P_{Y|X}, \text{ or equivalently, } Y \perp\!\!\!\perp X \,|\, p(X) \tag{4.2}$$

where $P_{Y|*}$ denotes conditional probability distribution of Y given $*$. This is a general scheme that makes no assumptions about the distribution of (X, Y), and has been extensively studied in the literature; for a recent reference, see Li [55]. Early studies on SDR tend to find the map p that achieves (4.2) among the orthogonal projections onto linear subspaces. However, because the orthogonal projections do not generally send simplex to simplex, the nonlinear SDR theory [53] is more relevant to our purpose.

The conditional mean function $E[Y|X]$, rather than the entire dependence structure $P_{Y|X}$, is frequently of interest in statistical problems. Then the dimension reduction aims to achieve a weaker condition, i.e., the maximum predictive ability using $p(X)$:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)] \Leftrightarrow Y \perp\!\!\!\perp \mathbb{E}[Y|X] \,|\, p(X). \tag{4.3}$$

This assumption is called *sufficient dimension reduction for conditional mean*, which is by definition a special case of SDR. Intuitively, (4.3) means that it is enough for predicting Y, and it becomes equivalent to SDR under certain assumptions. For example, statistical models often assume that the conditional mean has all information on $P_{Y|X}$; that is, $Y \perp\!\!\!\perp X \,|\, \mathbb{E}[Y|X]$. This is known as *location regression* [22], and it includes the additive error models $Y = f(X) + \epsilon$ with $X \perp\!\!\!\perp \epsilon$. Under this assumption, it is straightforward to see that (4.3) implies (4.2).

4.3.2 RKHS and Conditional Covariance Operator

While many SDR approaches are available, Fukumizu et al. [32, 34] propose to use kernel measures of conditional independence, which has often exhibited empirical success. In what follows, we present theory and remarkable properties of the conditional covariance operator of RKHSs, proposed originally by Baker [9].

Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ denote positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$ satisfying the boundedness condition in means:

$$\mathbb{E}_X[k_{\mathcal{X}}(X,X)] < \infty \quad \text{and} \quad \mathbb{E}_Y[k_{\mathcal{Y}}(Y,Y)] < \infty. \tag{4.4}$$

Note that (4.4) ensures that the corresponding RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are continuously embedded in $L^2(P_X)$ and $L^2(P_Y)$, respectively, and ensures the existence of mean embedding maps $P \mapsto \mu_P := \mathbb{E}_P[k(W,\cdot)] \in \mathcal{H}$, where P denotes an arbitrary probability measure [77]. If the mean embedding map of an RKHS $(\mathcal{H}, k)$ is injective, it is called *characteristic*.

The *cross-covariance operator* of (X,Y), $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$, is defined by the adjoint relations

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])\,(g(Y) - \mathbb{E}_Y[g(Y)]) \right]$$

for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. If Y is equal to X, then Σ_{XX} is called the *covariance operator*. It induces a unique bounded operator $V_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ such that

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$$

with $\|V_{YX}\| \leq 1$ [9]. This is called the *normalized cross-covariance operator* (NOCCO), which resembles the correlation in classical statistics [33]. It helps to define the following conditional covariance operator without worrying about the invertibility of Σ_{XX}:

Definition 4.1. The *conditional covariance operator* $\Sigma_{YY|X} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ of Y given X is defined by

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2}.$$

When Σ_{XX} is invertible, it immediately follows that

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY},$$

analogous to the well-known multivariate Gaussian case.

The following two results from Fukumizu et al. [34] provide insights into the meaning of the conditional covariance operator. The former shows its role in assessing the predictive ability of Y given X, and the latter reveals that $\Sigma_{YY|X}$ indeed captures the conditional variance of Y given X.

Proposition 4.2. *For any $g \in \mathcal{H}_{\mathcal{Y}}$, we have*

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \inf_{f \in \mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X,Y}\left| (g(Y) - \mathbb{E}_Y[g(Y)]) - (f(X) - \mathbb{E}_X[f(X)]) \right|^2.$$

If $\mathcal{H}_{\mathcal{X}} + \mathbb{R}$ is dense in $L^2(P_X)$, then

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]]. \tag{4.5}$$

Note that the condition of (4.5) always holds when $k_{\mathcal{X}}$ is bounded and characteristic [34]. This means that the injectivity of the mean embedding map ensures the richness of RKHS up to a constant sum. There is another notion of richness of RKHS $(\mathcal{H}, k)$, called *universality*. When the domain $\mathcal{X}$ is compact and k is continuous, we say that $(\mathcal{H}, k)$ is universal if $\mathcal{H}$ is dense in the space of continuous functions $C(\mathcal{X})$. There are numerous universal kernels used in practice, such as Gaussian or Laplace kernels, and it is known that every universal kernel is characteristic [45].

4.3.3 Kernel Feature Selection (KFS) via Minimization of Conditional Covariance

Motivated by (4.5), Fukumizu et al. [34] and Chen et al. [20] show that minimizing the trace of the conditional covariance operator *after projection* achieves SDR. The problem of finding suitable projections is formulated as follows. For any vector $x \in \mathbb{R}^d$ and any subset $\mathcal{S} \subseteq \{1, 2, \ldots, d\}$, let $x_{\mathcal{S}}$ be the vector with components $(x_{\mathcal{S}})_i = x_i$ if $i \in \mathcal{S}$ and $(x_{\mathcal{S}})_i = 0$ otherwise. Then the objective for variable selection is to find $\mathcal{S}$ such that

$$\underset{\mathcal{S} \subseteq \{1,\ldots,d\}}{\operatorname{argmin}} \operatorname{Tr}(\Sigma_{YY|X_{\mathcal{S}}}), \tag{4.6}$$

where $\operatorname{Tr}(\cdot)$ denotes the trace of a self-adjoint operator.

It is important to note that this approach is essentially different from traditional RKHS methods for dimension reduction. Well-known RKHS methods such as kernel PCA or kernel Fisher discriminant analysis [76], first map the data into an RKHS and then carry out low-dimensional projections within the high-dimensional RKHS. This initial embedding process inevitably leads to an interpretation loss with respect to the original variables. On the other hand, the KFS methods [34, 20] *first project* the data (or select the variables) in a way that preserve interpretability, and then use kernel measures to evaluate the validity of the projection.

4.4 Proposed method

This section describes our kernel variable selection method for compositional data using the amalgamation in (4.1). Given n i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n)$ of the random variables $(X, Y) \in \Delta^d \times \mathcal{Y}$, our task is to find a subset $\mathcal{S} = \{s_1, \ldots, s_m\} \subset \{0, \ldots, d\}$ of variables whose projection $p_{\mathcal{S}}(X) = (X_{s_1}, \ldots, X_{s_m}, \sum_{j \notin \mathcal{S}} X_j) \in \Delta^m$ best represents the outcome Y.

4.4.1 Construction of RKHS

The proposed method first *lift*s the data by adding an extra zero coordinate to X, i.e., we set $\widetilde{X} = (X, 0) \in \Delta^{d+1}$. This lifting process does not affect the theory but will simplify the notations. Let $\mathcal{X} = \Delta^{d+1}$ be the extended domain where lifted compositions reside and define, by abusing notations, $p_{\mathcal{S}} : \mathcal{X} \to \Delta^m$ by $p_{\mathcal{S}}(x') = (x'_{s_1}, \ldots, x'_{s_m}, \sum_{j \notin \mathcal{S}} x'_j)$. That is, we also lift the projection map $p_{\mathcal{S}}$ to satisfy $p_{\mathcal{S}}(\widetilde{x}) = p_{\mathcal{S}}(x)$. Then, define a right inverse $i_{\mathcal{S}} : \Delta^m \to \mathcal{X}$ of $p_{\mathcal{S}}$, given by $i_{\mathcal{S}}(z_1, \ldots, z_{m+1}) = x$ with $x_j = 0$ for $j \notin \mathcal{S}$, $x_{s_i} = z_i$, and $x_{d+1} = z_{m+1}$. One can readily check that $p_{\mathcal{S}} \circ i_{\mathcal{S}}(z) = z$ for all $z \in \Delta^m$. Finally, we identify $\widetilde{X} = X$ and redefine the notation $X_{\mathcal{S}}$ of the selection result by

$$X_{\mathcal{S}} = i_{\mathcal{S}} \circ p_{\mathcal{S}}(X), \quad X \in \mathcal{X}. \tag{4.7}$$

Let $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ be an RKHS on $\mathcal{X} = \Delta^{d+1}$, and let $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ be an RKHS on $\mathcal{Y}$. The embedding $i_{\mathcal{S}}$ defined above gives rise to a *pullback kernel* $k_{\mathcal{S}}$ on Δ^m defined by

$$k_{\mathcal{S}}(z, w) = k_{\mathcal{X}}(i_{\mathcal{S}}(z), i_{\mathcal{S}}(w)).$$

Defining a kernel on the codomain Δ^m in this way has the advantage that it can cover all possible values of the target dimension m, and that the RKHS of $k_{\mathcal{S}}$, denoted by $\mathcal{H}_{\mathcal{S}}$, can naturally interact with functions on $\mathcal{X}$. The interactions can be stated as the following lemma:

Lemma 4.3. *There is another RKHS $(\mathcal{H}, k)$ on $\mathcal{X}$ that is isomorphic to $(\mathcal{H}_\mathcal{S}, k_\mathcal{S})$ on Δ^m. Furthermore, if $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ is universal, then so is $(\mathcal{H}_\mathcal{S}, k_\mathcal{S})$.*

The kernel k is given so that $k(x, x') = k_\mathcal{X}(x_\mathcal{S}, x'_\mathcal{S})$; we provide the proof in the supplementary materials. According to the lemma, we can conduct all of our theoretical analysis on the projected domain Δ^m, including those requiring universality, within the function space on $\mathcal{X}$, $L^2(P_X)$. Meanwhile, using $\mathcal{H}_\mathcal{S}$ has an explicit interpretation of the function space on the projected domain, as will be seen in Corollary 4.6.

4.4.2 SDR and Conditional Covariance Operator

From the discussion above, we can derive a theorem that parallels Theorem 2 in Chen et al. [20] and Theorem 4 in Fukumizu et al. [34]:

Theorem 4.4. *Let $\Sigma_{YY|X_\mathcal{S}}$ denote the conditional covariance operator with the kernel k given in Lemma 4.3. Then, if $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ is universal and $(\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ is characteristic, we have*

$$\Sigma_{YY|X} \preceq \Sigma_{YY|X_\mathcal{S}},$$

where the equality is attained if and only if $Y \perp\!\!\!\perp X \,|\, X_\mathcal{S}$. Here, the inequality $\preceq$ stands for the partial order of self-adjoint operators.

The inequality part follows immediately from Proposition 4.2. However, proving the equality condition needs exhaustive work due to the new projection $X_\mathcal{S}$ in (4.7). We give a full proof in the supplementary materials, Section 4.7. Note that the universality of $\mathcal{H}_\mathcal{X}$ is imposed for simplicity and interpretability, and it may be relaxed to being characteristic.

Theorem 4.4 implies that the trace of self-adjoint operators has the following relation

$$\mathrm{Tr}(\Sigma_{YY|X}) \leq \mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}})$$

for all subsets of variables $\mathcal{S}$. Thus the variable selection problem for compositional data can be stated as

$$\underset{\mathcal{S} \subseteq \{0,\ldots,d\}}{\mathrm{argmin}} \ \mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}}), \tag{4.8}$$

which is a compositional version of (4.6) with $X_\mathcal{S}$ defined in (4.7). Note that the trace equality $\mathrm{Tr}(\Sigma_{YY|X}) = \mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}})$ implies SDR since the operator $\Sigma_{YY|X_\mathcal{S}} - \Sigma_{YY|X}$ is nonnegative and self-adjoint.

Based on this main result, we now consider the choice of the kernel $k_\mathcal{Y}$. For binary or multi-class classification tasks with $\mathcal{Y} = \{y_1, \ldots, y_k\} \subset \mathbb{R}$, we can use the *delta kernel* $k_\mathcal{Y}(y, y') = \delta_{y,y'}$, which is equal to 1 when $y = y'$ and 0 otherwise. Note that the delta kernel is universal on the discrete domain $\mathcal{Y}$ so the aforementioned theory applies. The relative advantage of the delta kernel over the Gaussian kernel has been mentioned by Yamada et al. [126] who investigate the performance of HSIC-Lasso under the two kernel choices.

For regression problems, Chen et al. [20] argue that one can use the *linear kernel* for a univariate response. However, we discover that Corollaries 3 and 4 in their work contain minor errors and the conclusions are overstated. Even though these errors do not preclude the practical application of the method, we give a corrected version below for clarity. The proofs are provided in the supplementary Section 4.7.

Let $\mathcal{Y} = \mathbb{R}$ and define $k_{\mathcal{Y}}$ as the linear kernel $k_{\mathcal{Y}}(y, y') = yy'$. It should be noted that the RKHS $\mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\vee}$ is not characteristic so Theorem 4.4 cannot be applied to ensure the full SDR, which is claimed in Corollary 3 of Chen et al. [20]. Nonetheless, the presence of the identity function $id_{\mathcal{Y}}$ in $\mathcal{H}_{\mathcal{Y}}$ leads to a weaker result, which is the SDR for conditional mean:

Proposition 4.5. *If $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ is universal, $\mathcal{Y} = \mathbb{R}$, and if $k_{\mathcal{Y}}$ is the linear kernel on $\mathcal{Y}$, then the trace equality $\mathrm{Tr}(\Sigma_{YY|X}) = \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}})$ implies $\mathbb{E}[Y|X] = \mathbb{E}[Y|X_{\mathcal{S}}]$, the SDR for conditional mean.*

That is, in the case of univariate regression with the linear kernel, solving (4.8) achieves the *SDR for conditional mean*. However, Corollary 3 of Chen et al. [20] inaccurately states that it achieves the full SDR. If we further assume the location regression model on the population (Section 4.3.1), we then obtain the full SDR:

$$\mathrm{Tr}(\Sigma_{YY|X}) = \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}) \;\Leftrightarrow\; Y \perp\!\!\!\perp X \,|\, X_{\mathcal{S}}.$$

Using the linear kernel on $\mathcal{Y} = \mathbb{R}$ has another advantage of characterizing the trace of the conditional covariance operator as the *minimized variance of prediction error after projection*. Thus solving (4.8) is equivalent to finding a subset $\mathcal{S}$ that minimizes this variance:

Corollary 4.6. *Under the assumptions of Proposition 4.5,*

$$\mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}) = \inf_{f \in C(\Delta^m)} \mathrm{Var}_{X,Y}[Y - f(p_{\mathcal{S}}(X))].$$

If we assume on the population that there exists a continuous function f on Δ^m such that the response is expressed as

$$Y = f(p_{\mathcal{S}}(X)) + \epsilon,\; X \perp\!\!\!\perp \epsilon,\; \text{and } \mathbb{E}[\epsilon] = 0, \tag{4.9}$$

then Corollary 4.6 is equivalently stated in terms of the mean squared error:

$$\mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}) = \inf_{f \in C(\Delta^m)} \mathbb{E}_{X,Y}(Y - f(p_{\mathcal{S}}(X)))^2.$$

This is the form asserted in Corollary 4 of Chen et al. [20], implicitly assuming (4.9).

4.4.3 Variable Selection Algorithm

The solution set of (4.8) is always nonempty since the whole data X achieves the minimum. For practical purposes, it is natural to limit the number of variables we want to select, and this is written as

$$\operatorname*{argmin}_{|\mathcal{S}| \le m} \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}). \tag{4.10}$$

Solving (4.10) will result in a variable selection that is nearly SDR (classification) or SDR for conditional mean (univariate regression). The remaining procedure for solving this objective is similar to that of Chen et al. [20] and we briefly illustrate it here.

For $(x_1, y_1), \ldots, (x_n, y_n) \in \Delta^d \times \mathcal{Y}$, we first lift them into $\mathcal{X} \times \mathcal{Y}$ as described before. Then the empirical estimate of $\mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}})$ is defined by

$$\begin{aligned} \mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY|X_{\mathcal{S}}}) &= \mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY} - \hat{\Sigma}^{(n)}_{YX_{\mathcal{S}}}(\hat{\Sigma}^{(n)}_{X_{\mathcal{S}}X_{\mathcal{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}^{(n)}_{X_{\mathcal{S}}Y}) \\ &= \epsilon_n \,\mathrm{Tr}(G_Y(G_{X_{\mathcal{S}}} + n\epsilon_n I_n)^{-1}), \end{aligned} \tag{4.11}$$

where the $\hat{\Sigma}^{(n)}_{**}$ are empirical estimates of covariance operators, G_Y and $G_{X_{\mathcal{S}}}$ are centered Gram matrices, and ϵ_n is a regularization parameter. Here, letting $\mathbb{H} = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, $\mathbb{1} = (1, \cdots, 1) \in \mathbb{R}^n$, the centered version of a gram matrix K is defined by $G = \mathbb{H}K\mathbb{H}$.

Note that the delta kernel we use in the classification case is equivalent to the *linear kernel* $k_{\mathcal{Y}}(y, y') = \langle y, y' \rangle$ on the one-hot encoded domain $\mathcal{Y} = \{y \in \{0,1\}^k \mid \sum_i y_i = 1\} \subset \mathbb{R}^k$. Hence, we fix $k_{\mathcal{Y}}$ by the linear kernel for the classification or univariate regression case. Then the Gram matrix K_Y is $\mathbf{Y}\mathbf{Y}^T$, where $\mathbf{Y}$ is the matrix of sample responses on rows. We assume, without loss of generality, that the mean of each column of $\mathbf{Y}$ is zero, resulting in $G_Y = \mathbf{Y}\mathbf{Y}^T$. Then, the minimization of (4.11) is stated as

$$\min_{|\mathcal{S}| \leq m} \operatorname{Tr}(\mathbf{Y}^T (G_{X_{\mathcal{S}}} + n\epsilon_n I_n)^{-1} \mathbf{Y}), \tag{4.12}$$

which is the empirical version of our objective. In the binary response case, $k = 2$, this is equivalent to using $\mathcal{Y} = \{0, 1\}$ with the linear kernel so that we may reduce the column dimension of $\mathbf{Y}$ to 1.

In the following theorem, we show that a consistency result holds for the *global* optimum of (4.12), justifying that minimizing the empirical estimate will asymptotically achieve the population minimum (4.8). See the supplementary materials, Section 4.7, for the simpler proof using discrete nature than the proof given in Chapter 3.

Theorem 4.7. *Let $\hat{\mathcal{S}}^{(n)}$ be a global optimum that minimizes* (4.12)*. If the regularization parameter ϵ_n satisfies*

$$\epsilon_n \to 0 \quad \textit{and} \quad n^{1/2}\epsilon_n \to \infty \quad \textit{as} \quad n \to \infty,$$

then $\operatorname{Tr}\left(\hat{\Sigma}^{(n)}_{YY|X_{\hat{\mathcal{S}}^{(n)}}}\right) \to \operatorname{Tr}\left(\Sigma_{YY|X_{\mathcal{S}'}}\right)$ *in probability, where* $\mathcal{S}' \in \textit{argmin}_{|\mathcal{S}| \leq m} \operatorname{Tr}(\Sigma_{YY|X_{\mathcal{S}}})$.

A brute-force search of (4.12) is computationally infeasible for high dimensions since the number $\binom{d}{m}$ grows exponentially. We relax this problem to a continuous one that can be solved by the gradient descent method, as similarly done in Chen et al. [20]. Note that we can express $X_{\mathcal{S}}$ as $(w \odot X, 1 - w^T X)$ where $w = (w_0, \ldots, w_d) \in \{0,1\}^{d+1}$ denotes a binary weight vector with $w_i = 1$ if and only if $i \in \mathcal{S}$, and $\odot$ denotes the Hadamard product. Now relaxing the weights to allow continuous values with $0 \leq w_i \leq 1$ and $\sum w_i \leq m$, define

$$X_w := (w \odot X, 1 - w^T X) \in \mathcal{X}.$$

Then our relaxed objective is written as

$$\begin{aligned} \min_{w} \quad & \operatorname{Tr}(\mathbf{Y}^T (G_{X_w} + n\epsilon_n I_n)^{-1} \mathbf{Y}) \\ \text{subject to} \quad & \|w\|_1 \leq m, \ 0 \leq w_i \leq 1, \ \forall i. \end{aligned} \tag{4.13}$$

Given that the kernel $k_{\mathcal{X}}$ is smooth and universal, we can apply projected gradient descent to solve this optimization problem. Although the objective function is nonconvex when typical universal kernels are used, the projected gradient descent algorithm is able to find true signal variables well, as shown in Section 4.5 (see also Ruan et al. [92]). After obtaining an approximated solution $\hat{w}$ via gradient descent, we reconstruct a variable selection $\hat{\mathcal{S}}$ whose corresponding binary vector is closest to $\hat{w}$.

Note that each gradient descent step to equation (4.13) requires $O(n^2 d + n^3)$ computations. This is not a big problem in practice because compositional data typically have a low sample size. The complexity can be reduced further by adopting a low-rank approximation of kernel matrices, such as random Fourier features [87].

4.5 Experiments

This section conducts experiments on synthetic and real microbiome data to assess the performance of the proposed variable selection method under both classification and regression scenarios. We compare

it with two methods, coda-lasso [63, 66] and selbal [90], chosen from the recent survey by Susin et al. [113]. These two methods are based on log-ratio transformation, so we replace zero values in each sample x by $0.5x_{\min}$, where $x_{\min}$ is the minimum positive value of x. We also provide results of other zero replacement methods in the supplementary Section 4.7.1; to do this, we delete columns with fewer than two positive values in all data. We use the R codes provided by Susin et al. [113] for their implementation, and the Python code for our method is available at `https://github.com/pjywang/KVS-CoDa`.

For the proposed method, we use a Gaussian kernel $k_{\mathcal{X}}(x, x') = \exp\left(-\|x - x'\|^2/\sigma^2\right)$ with σ being the standard median pairwise distance between samples. Across all experiments, the regularization parameter ϵ is set to $\epsilon = 0.001$ for classification tasks and $\epsilon = 0.1$ for regression tasks; we find that these values work stably in general. Cross-validation (CV) can also be used in conjunction with classification or regression algorithms.

4.5.1 Synthetic Data

We begin with simulations of microbiome count data proposed by Te Beest et al. [114], which reflect the varying total counts and zero-inflation. The (i, j)th-entry X_{ij} of an $n \times p$ count matrix $\mathbf{X}$ is sampled from a negative binomial distribution with mean μ_{ij} and variance $\mu_{ij} + \mu_{ij}^2$. The mean μ_{ij} follows a log-linear model

$$\log \mu_{ij} = a_i + t_j + e y_i, \tag{4.14}$$

where a_i reflects the total abundance of the ith sample, t_j reflects the abundance of taxon j, e represents the effect size on taxon j, and $y_i \in \{0, 1\}$ indicates whether the ith sample has an effect. The parameters a_i and t_j are drawn from normal distributions, $N(0, 1)$ and $N(0, 2)$, respectively. Only 10% of p taxa are set to be relevant to y_i with $e = \pm \log 5$, where the signs are given with equal probabilities, while the rest of e are set to zero. To ensure that taxa mostly consisting of zeros receive no effect, these 10% relevant taxa are randomly selected from the top 70% of variables with the highest t_j values. The binary response vector $Y = (y_1, \ldots, y_n)$ is set to have the same number of zeros and ones. Finally, the taxa present in fewer than two samples are removed, and the count matrix $\mathbf{X}$ is normalized so that each row sums to 1. This model generates approximately 50% of zeros in the data.

We first generate data with fixed $(n, p) = (200, 100)$ so that only ten taxa retain effects. We then apply the variable selection algorithms with the desired number of selected variables $m \in \{5, 10, \ldots, 40\}$. Because the lasso algorithm does not specify the number of variables to be chosen, we perform coda-lasso on with the tuning parameter ranged in $[0.01, 0.2]$, and the *best performance* among the models that select $m, m + 1, \ldots, m + 4$ variables is recorded. This process obviously *favors* coda-lasso, as it inflates its power. Nevertheless, its performance is inferior to our method.

For the second experiment, we fix $p = 100$ and vary $n \in \{200, 400, \ldots, 1000\}$ to examine the convergence to the true number of variables with effects. In this case, we set the proposed and selbal algorithms to pick the true number of variables, $m = 10$. We again perform coda-lasso as described above, and record the best performance among the models that choose $m^* \in \{10, 11, \ldots, 14\}$ variables. We run these two experiments 50 times; the results are shown in Figure 4.1.

As illustrated in the figure, the proposed method clearly outperforms the log-ratio methods on average. The left panel shows an increasing probability of selecting true signal variables as we select more variables in the algorithm. Note that selbal fails to exhibit such a phenomenon because its forward selection algorithm often terminates before achieving the upper bound m. In contrast, the proposed method achieves the bound in most cases. In the right panel, we observe that the power of the proposed

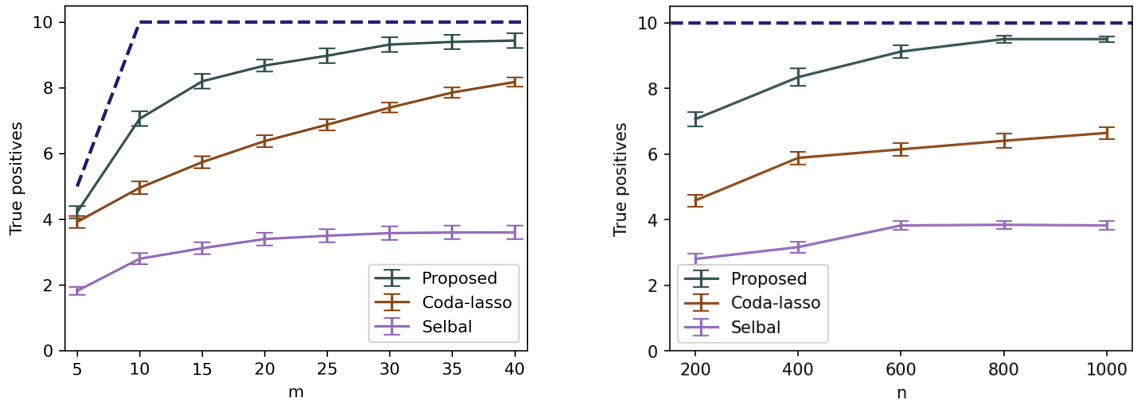


Figure 4.1: Variable selection results from 50 runs of synthetic data. The y-axis denotes the number of correctly selected features. The maximum number of true variables can be chosen by algorithms is indicated by the top dotted line. The x-axis of the left panel denotes the desired number $m \in \{5, 10, \ldots, 40\}$ of variables selected by algorithms, while the x-axis of the right panel denotes the sample size $n \in \{200, 400, \ldots, 1000\}$. The average numbers of selected variables $\pm$ standard error are shown for each method. Note that the result of coda-lasso is displayed *in its favor*.

method increases as the sample size grows and converges to the true value of 10. The log-ratio methods do not exhibit clear convergence to the true value, and the power of selbal does not even increases as n grows.

Varying Zero Proportions. By adjusting the means of the parameters a_i and t_j in the log-linear model (4.14), we may generate similar synthetic data with different zero proportions. Suppose a_i and t_j are drawn from $N(a, 1)$ and $N(t, 2)$, respectively. Setting (a, t) as $(2.2, 1.5)$, $(1, 0.5)$, $(0, 0)$, and $(-1.1, -0.5)$ yields the generate data to contain about 10%, 30%, 50%, and 70% of zeros, respectively. Table 4.1 reports the results with $(n, p) = (500, 100)$. The proposed method clearly outperforms the other methods and shows consistent power over a wide range of zero proportions. The performance is slightly weakened when the zero proportion is 70%, which is a natural consequence of the data generation process. Since the data are generated as nonnegative counts, the signal of data shrinks as the ratio of zeros increases because the effect size $e = \log 5$ is fixed.

In contrast, the other two log-ratio methods, coda-lasso and selbal, exhibit highly *inconsistent results* as the zero proportion changes. When the zero proportion is less than 50%, these methods perform unexpectedly poorly. This is probably due to the data distortions caused by zero replacement and log transformation [80] are more severe when the zero rates are moderately low. This issue is alleviated slightly if the zero proportion increases, as the model (4.14) generates a larger number of columns with mostly zeros. Such columns are similarly impacted by zero replacement and log transformation, making it easier for supervised learning methods to rule these irrelevant columns out. It would be worthwhile to observe if this inconsistent behavior maintains for other synthetic data settings, and we leave this as future work.

Table 4.1: Average numbers of true variables selected from 50 runs of synthetic data with different zero proportions. The data has ten true variables, and the parameter m is set to 10. The tuning parameter of coda-lasso is set to select between 10 and 14 variables. All standard errors range between 0.1 and 0.2.

Zero %	10%	30%	50%	70%
Proposed	9.26	9.1	9.12	8.46
Selbal	0.78	1.74	3.36	4.22
Coda-lasso	2.32	3.7	5.76	6.3

Table 4.2: Prediction accuracy of each variable selection method on the BMI dataset. Results are shown in terms of mean ± one standard error of the estimated MSEs over ten repetitions.

	Estimated MSE		
Methods	$m = 3$	$m = 5$	$m = 10$
Proposed	$28.90 \pm .048$	$28.87 \pm .037$	$28.99 \pm .072$
Selbal	33.03 ± 1.66	32.91 ± 1.79	34.64 ± 1.85
Coda-lasso	$29.29 \pm .297$ (selects 0 to 8 variables)		

4.5.2 BMI Microbiomes Data

We also evaluate our proposed method with the body mass index (BMI) dataset [125], which has been repeatedly analyzed with the constrained lasso approaches [63, 103, 123]. The dataset consists of 98 gut microbiome samples with BMI information and organized into 87 genera. For the purpose of comparison in the supplementary Section 4.7.1, 10 genera that appeared in only one sample are removed. As a result, the data have 77 dimensions with 68.6% of zero values.

To obtain the estimated prediction error for this regression problem, we run ten repetitions of randomly split five-fold CV. For selbal and the proposed method, we use $m \in \{3, 5, 10\}$. While selbal and coda-lasso are integrated within prediction modeling, the proposed method requires a separate regression analysis to assess its prediction ability. We radially transform the chosen amalgamation onto the sphere and then apply kernel ridge regression (KRR) with the Gaussian kernel [80]. All tuning parameters, including the Gaussian width of KRR and regularization parameters of KRR and lasso, are chosen based on the five-fold CV within the training set.

Table 4.2 lists the estimated mean squared errors (MSE) over ten runs of CV. As can be observed, the proposed method compares favorably with log-ratio methods, achieving the smallest MSE and variance. The choice of $m = 3, 5$ is comparable to the fact that only four genera are selected in Lin et al. [63] and Shi et al. [103]. However, the selected genera from our method fairly differ from coda-lasso. Given the prediction accuracy and results presented in Section 4.5.1, our result should be considered more reasonable.

4.6 Conclusion and Future Works

This work proposes a new variable selection framework for compositional data based on amalgamation. The proposed method aims to achieve SDR by minimizing the conditional covariance of the response given selected covariates. Also, the statistical consistency of the proposed method is provided. It is broadly applicable to general compositional data and does not impose strong assumptions on the underlying probability distributions. Finally, the proposed approach is shown to exhibit consistent results and outperform existing log-ratio approaches in both synthetic and real-world experiments.

An interesting implication of the present research is that amalgamation may have many more applications than have been previously considered for compositional data analysis. Amalgamation would not be a justifiable practice for general Euclidean data, however, the intrinsic nature of compositional data makes it a valid option for reducing the complexity of the data. For instance, in the dimension reduction context, we may extend the search space to include all possible amalgamations of the variables, which we leave as future work.

The optimization problem of the kernel-based dimension reduction and variable selection is nonconvex and susceptible to local optima. However, recent work by Ruan et al. [92] finds that with l_1 kernels, the stationary points of gradient descent are nonetheless able to select the true signal variables. It is worthwhile to examine if this result extends to our amalgamation-based situation.

4.7 Supplementary Materials

This section presents additional experimental results with different choices of zero replacement methods and give proofs of the omitted theoretical results.

4.7.1 Comparison to Other Zero Replacement Methods

While our method does not substitute zero values of compositional data, the other log-ratio methods compared in Section 4.5 produce different results depending on how the zeros are replaced [67]. Therefore, in this section, we provide additional experimental results using two other zero replacement methods: 1sum (which adds one pseudocount; e.g., see Brill et al. [14]) and the geometric Bayesian multiplicative (gbm) replacement [72]. The gbm method requires data to have at least two positive values at each column and is implemented by the R package `zCompositions`. The results show that the proposed method still has superior performance and that the $0.5x_{\min}$ replacement is not a bad choice for coda-lasso and selbal.

Synthetic data

Table 4.3: Mean true positives over 50 runs of synthetic data with varying m and n. The other experimental settings are the same as in Section 4.5. Standard errors range between 0.1 and 0.3

	$n = 200,\ p = 100$				$p = 100,\ m = 10$				
Methods	$m = 10$	$m = 20$	$m = 30$	$m = 40$	$n = 200$	$n = 400$	$n = 600$	$n = 800$	$n = 1000$
proposed	7.06	8.68	9.32	9.44	7.06	8.34	9.12	9.5	9.5
coda-lasso + $0.5x_{\min}$	4.96	6.38	7.40	8.18	4.58	5.88	6.14	6.40	6.64
coda-lasso + 1sum	5.00	6.38	7.42	8.10	4.66	6.22	6.32	6.64	7.08
coda-lasso + gbm	3.84	4.90	6.08	6.98	3.66	4.28	4.54	4.74	4.34
selbal + $0.5x_{\min}$	2.60	3.44	3.64	3.68	2.80	3.16	3.82	3.84	3.82
selbal + 1sum	2.92	3.50	3.74	3.74	2.92	3.28	3.90	3.96	4.00
selbal + gbm	1.56	2.16	2.32	2.36	1.56	2.22	2.44	2.78	2.58

BMI Microbiomes Data

Table 4.4: Estimated MSE over 10 repetitions of cross-validation on the BMI dataset.

	Estimated MSE		
Methods	$m = 3$	$m = 5$	$m = 10$
proposed	$28.90 \pm .048$	$28.87 \pm .037$	$28.99 \pm .072$
selbal + $0.5x_{\min}$	33.03 ± 1.66	32.91 ± 1.79	34.64 ± 1.85
selbal + 1sum	33.46 ± 1.73	33.92 ± 1.85	36.52 ± 1.95
selbal + gbm	32.91 ± 2.00	37.05 ± 3.57	41.16 ± 4.12
coda-lasso + $0.5x_{\min}$	$29.29 \pm .297$ (selects 0 to 8 variables)		
coda-lasso + 1sum	$30.52 \pm .379$ (selects 0 to 16 variables)		
coda-lasso + gbm	$29.05 \pm .440$ (selects 0 to 7 variables)		

4.7.2 Proof of Results

Although most proofs for this chapter are already similarly given in Chapter 3, we elaborate some details of the precise proofs here since the underlying settings of the theory of this chapter are slightly changed and tailored for variable selection.

Proof of Lemma 4.3

The kernel $k_{\mathcal{S}}$ defines another pullback kernel

$$k(x, x') = k_{\mathcal{S}}(p_{\mathcal{S}}(x), p_{\mathcal{S}}(x')) = k_{\mathcal{X}}(x_{\mathcal{S}}, x'_{\mathcal{S}}) \tag{4.15}$$

with the corresponding RKHS $\mathcal{H}$ on $\mathcal{X}$. By pullback theorem of Paulsen and Raghupathi [82], there is a well-defined surjective *pullback map* $p^* : \mathcal{H}_{\mathcal{S}} \to \mathcal{H}$ given by

$$p^*(f) = f \circ p_{\mathcal{S}} \in \mathcal{H}, \quad \forall f \in \mathcal{H}_{\mathcal{S}}.$$

Note that the fact $f \circ p_{\mathcal{S}} \in \mathcal{H}$ is nontrivial and this is where the pullback theorem is used. Recall that $x_{\mathcal{S}} = i_{\mathcal{S}} \circ p_{\mathcal{S}}(x)$ and $p_{\mathcal{S}} \circ i_{\mathcal{S}} = id_{\Delta^m}$. As $p_{\mathcal{S}}$ is *surjective*, the equation (4.15) implies that the pullback map p^* preserves the RKHS inner product; thus, p^* is an isometry. Therefore, the pullback map p^* is an isomorphism $\mathcal{H}_{\mathcal{S}} \cong \mathcal{H}$.

By construction, the embedding $i_{\mathcal{S}} : \Delta^d \to \mathcal{X}$ is a homeomorphism onto its image. This topological embedding $i_{\mathcal{S}}$ allows the codomain Δ^m to be regarded as a subset of $\mathcal{X}$. Then, if $k_{\mathcal{X}}$ is universal, so is $k_{\mathcal{S}}$, as stated in Lemma 4.55 of Steinwart and Christmann [111].

Proof of Theorem 4.4

The proof is almost identical to the proof of Theorem 3.15, except for the settings on the RKHS $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$.

For any $g \in \mathcal{H}_{\mathcal{Y}}$, by Proposition 4.2 we have

$$\begin{aligned}
\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} &= \inf_{f \in \mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X,Y} \left| (g(Y) - \mathbb{E}_Y[g(Y)]) - (f(X) - \mathbb{E}_X[f(X)]) \right|^2 \\
\langle g, \Sigma_{YY|X_{\mathcal{S}}} g \rangle_{\mathcal{H}_{\mathcal{Y}}} &= \inf_{f \in \mathcal{H}} \mathbb{E}_{X,Y} \left| (g(Y) - \mathbb{E}_Y[g(Y)]) - (f(X) - \mathbb{E}_X[f(X)]) \right|^2 .
\end{aligned}$$

Note that our $\mathcal{X}$ is compact Hausdorff, and hence the space $C(\mathcal{X})$ is dense in $L^2(\mu)$ for all probability measures μ on $\mathcal{X}$. It is well-known that $C(\mathcal{X})$ is continuously embedded in L^2, so $\mathcal{H}_{\mathcal{X}}$ is dense in $L^2(\mu)$ for all probability measure μ by universality assumption. As $\mathcal{H}$ is contained in $L^2(P_{\mathcal{X}})$, it immediately follows that

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} \leq \langle g, \Sigma_{YY|X_{\mathcal{S}}} g \rangle_{\mathcal{H}_{\mathcal{Y}}} \quad \text{for all } g \in \mathcal{H}_{\mathcal{Y}},$$

which is exactly the definition of partial order $\preceq$; that is, $\Sigma_{YY|X} \preceq \Sigma_{YY|X_{\mathcal{S}}}$.

For the equality, we consider the counterpart of feature selection $p_{\mathcal{S}^c}(X) \in \Delta^{d-m+2}$ where $\mathcal{S}^c = \{0, \ldots, d\} \setminus \mathcal{S}$. Let $(U, V) = (X_{\mathcal{S}}, X_{\mathcal{S}^c})$. The *primary ingredient* of the proof is that $(X_{\mathcal{S}}, X_{\mathcal{S}^c})$ is in one-to-one correspondence with the original X, rather than the strict equality as in the references (finding this kind of counterpart with one-to-one correspondence may be hard if we take arbitrary projections). Then, by the law of total variance, we have

$$\mathrm{Var}_{Y|U}[g(Y)|U] = \mathbb{E}_{(U,V)|U}[\mathrm{Var}_{Y|U,V}[g(Y)|U, V]|U] + \mathrm{Var}_{(U,V)|U}[\mathbb{E}_{Y|U,V}[g(Y)|U, V]|U]. \tag{4.16}$$

We then take $\mathbb{E}_U$ on both sides. Identifying $U = X_{\mathcal{S}}$ with $p_{\mathcal{S}}(X)$ on Δ^m, the left hand side becomes

$$\mathbb{E}_U[\mathrm{Var}_{Y|U}[g(Y)|U]] = \langle g, \Sigma_{YY|U} g \rangle_{\mathcal{H}_{\mathcal{Y}}}$$

by Proposition 4.2. Based on the one-to-one correspondence and the tower law, the first term of the right-hand side of (4.16) is computed as

$$\begin{aligned}\mathbb{E}_U[\mathbb{E}_{(U,V)|U}[\mathrm{Var}_{Y|U,V}[g(Y)|U,V]|U]] &= \mathbb{E}_{U,V}[\mathrm{Var}_{Y|U,V}[g(Y)|U,V]] \\ &= \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]] \\ &= \langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

Then the equation (4.16) turns into

$$\langle g, (\Sigma_{YY|U} - \Sigma_{YY|X})g\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_U[\mathrm{Var}_{X|U}[\mathbb{E}_{Y|X}[g(Y)|X]|U]]. \tag{4.17}$$

As we have shown that $\Sigma_{YY|X} \preceq \Sigma_{YY|X_\mathcal{S}}$, the LHS is zero if and only if $\Sigma_{YY|X} = \Sigma_{YY|U}$ (note that $g \in \mathcal{H}_\mathcal{Y}$ is arbitrary). On the other hand, the RHS of (4.17) is zero if and only if $\mathrm{Var}_{X|U}[\mathbb{E}_{Y|X}[g(Y)|X]|U] = 0$ for almost every U, which means that

$$\begin{aligned}\mathbb{E}_{Y|X}[g(Y)|X] &= \mathbb{E}_{X|U}[\mathbb{E}_{Y|X}[g(Y)|X]|U] \\ &= \mathbb{E}_{Y|U}[g(Y)|U]\end{aligned}$$

for almost every U, and for every $g \in \mathcal{H}_\mathcal{Y}$. It then follows that the mean embeddings of conditional distributions $P_{Y|X}$ and $P_{Y|U}$ are the same in $\mathcal{H}_\mathcal{Y}$. As $\mathcal{H}_\mathcal{Y}$ is characteristic, we have that $P_{Y|X} = P_{Y|U}$, which is equivalent to $Y \perp\!\!\!\perp X \,|\, U$ ($\because \sigma(U) \subseteq \sigma(X)$).

Proof of Proposition 4.5

Plugging $g = id_\mathcal{Y}$ into the equation (4.17), we have

$$\langle id_\mathcal{Y}, (\Sigma_{YY|X_\mathcal{S}} - \Sigma_{YY|X})id_\mathcal{Y}\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_{X_\mathcal{S}}[\mathrm{Var}_{X|X_\mathcal{S}}[\mathbb{E}_{Y|X}[Y|X]|X_\mathcal{S}]] = 0, \tag{4.18}$$

which *only* implies $\mathbb{E}[Y|X] = \mathbb{E}[Y|X_\mathcal{S}]$. This can imply $Y \perp\!\!\!\perp X|X_\mathcal{S}$ as stated in Chen et al. [20] in case of *location regressions*.

Proof of Corollary 4.6

Since $id_\mathcal{Y}$ forms a complete orthonormal system of $\mathcal{H}_\mathcal{Y}$, we have

$$\mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}}) = \langle id_\mathcal{Y}, \Sigma_{YY|X_\mathcal{S}} id_\mathcal{Y}\rangle_{\mathcal{H}_\mathcal{Y}} = \inf_{f\in\mathcal{H}} \mathbb{E}_{X,Y}((Y - \mathbb{E}_Y[Y]) - (f(X_\mathcal{S}) - \mathbb{E}_{X_\mathcal{S}}[f(X_\mathcal{S})]))^2$$

by Proposition 4.2, where the RHS equals to the variance of $Y - f(p_\mathcal{S}(X))$. Since $\mathcal{H}_\mathcal{S}$ is dense in $C(\Delta^m)$ with uniform convergence norm by universality, we have

$$\mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}}) = \inf_{f\in C(\Delta^m)} \mathrm{Var}_{X,Y}[Y - f(p_\mathcal{S}(X))].$$

Proof of Theorem 4.7

We first state the following *uniform* convergence result:

Proposition 4.8. *If ϵ_n satisfies the asymptotic behavior given in Theorem 4.7,*

$$\sup_{|\mathcal{S}|\leq m} \left|\mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY|X_\mathcal{S}}) - \mathrm{Tr}(\Sigma_{YY|X_\mathcal{S}})\right| \to 0$$

as $n \to \infty$ in probability.

As usual, this uniform convergence implies that the limit of minimums converges to the minimum of the limits:

Proof of Theorem 4.7 given Proposition 4.8. Let $\epsilon > 0$ be a positive real number. There exists a large number $N > 0$ such that

$$\left|\mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY|X_{\mathcal{S}}}) - \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}})\right| < \frac{\epsilon}{2} \quad \text{for all } |\mathcal{S}| \leq m \text{ and for all } n \geq N$$

with probability $\geq 1 - \epsilon$. Let $\mathcal{S}' \in S$ be any global optimum. Then by definition of $\hat{\mathcal{S}}^{(n)}$ we have

$$\mathrm{Tr}\left(\hat{\Sigma}^{(n)}_{YY|X_{\hat{\mathcal{S}}^{(n)}}}\right) \leq \mathrm{Tr}\left(\hat{\Sigma}^{(n)}_{YY|X_{\mathcal{S}'}}\right) \leq \mathrm{Tr}\left(\Sigma_{YY|X_{\mathcal{S}'}}\right) + \frac{\epsilon}{2}$$

and thus

$$\left|\mathrm{Tr}\left(\Sigma_{YY|X_{\hat{\mathcal{S}}^{(n)}}}\right) - \mathrm{Tr}\left(\Sigma_{YY|X_{\mathcal{S}'}}\right)\right| \leq \mathrm{Tr}\left(\Sigma_{YY|X_{\hat{\mathcal{S}}^{(n)}}}\right) - \mathrm{Tr}\left(\hat{\Sigma}^{(n)}_{YY|X_{\hat{\mathcal{S}}^{(n)}}}\right) + \frac{\epsilon}{2} < \epsilon$$

with probability $\geq 1 - \epsilon$ (here, we use the uniform convergence twice). This concludes the desired convergence in probability. □

Note that the proof of Proposition 4.8 requires only pointwise convergence due to discreteness; i.e., it suffices to show pointwise convergence for *each* $\mathcal{S}$. This fact makes proof considerably simpler than originally given in Chapter 3. Proof of such a pointwise convergence can similarly be derived as we have done in Section 3.6.

Chapter 5. Dimension Reduction for Compositional Data with Interpretable Compositional Outcomes

5.1 Introduction

Compositional data, characterized by nonnegative proportions of variables to a whole, are ubiquitous in various scientific domains, including geochemistry, economics, and microbiology. As the numerical values of variables represent only relative, not absolute, information, they are normalized to a d-dimensional unit simplex:

$$\Delta^{d-1} = \left\{ (x_1, \ldots, x_d)^T \in \mathbb{R}^d \,\middle|\, \sum_{i=1}^{d} x_i = 1,\ x_i \geq 0,\ \forall i \right\}.$$

Recently, the surge of human microbiome studies has significantly promoted the research on compositional data analysis. The human microbiome consists of all microorganisms, such as bacteria and viruses, living in or on the human body. They are shown to be closely associated with various human health and diseases [49], including obesity [118, 125], diabetes [61], inflammatory bowel disease [36], and cancer [40, 102]. They are commonly obtained by employing modern high-throughput sequencing technologies, such as 16S ribosomal RNA sequencing [21] and shotgun metagenomic sequencing [85], resulting in raw count numbers of microbiomes. Because the total number of counts varies vastly across the samples, microbiome data are normalized and viewed as compositional [37]. The microbiome compositional data obtained in this manner present high dimensionality and extreme sparsity; that is, the number of variables mostly outnumbers the number of samples, and the data contain a large proportion of zeros.

In high-dimensional data analysis, dimension reduction is always one of the most relevant tasks as it alleviates the curse of dimensionality. Furthermore, in biological applications, it is crucial to have *interpretable* analysis results that can be explained by original variables. Despite its importance, however, existing approaches to dimension reduction of compositional data lack clear interpretability. For microbiome data, a popular method, principal coordinate analysis (PCoA) with microbiome-specific measures of dissimilarity (e.g. Bray-Curtis, UniFrac), produces some appealing visualizations, but they are not described by original variables.

Such an interpretability issue also is not resolved in general compositional data analysis approaches. Classically, compositional data analyses have primarily been conducted within the framework of log-ratio transformation [3] to address the spurious negative correlation [84] arising from the unit-sum constraint. A broad class of linear dimension reduction methods may apply after log-ratio transformations, including principal component analysis [4] and linear discriminant analysis [29]. Then, the resulting lower dimensional coordinates are described by linear combinations of log-ratio transformed variables, which have some interpretations but are not clearly interpretable than a simple linear combination of variables. Furthermore, it is challenging to understand the *interactions* between these unclear lower dimensional coordinates, although they are orthogonal in the transformed Euclidean space.

Besides the interpretability, the log-ratio transformations, in fact, have severe defects. As log-ratio-based approaches cannot compute zero values in data, they require zero replacements followed by renormalization. Although this procedure with high-dimensional statistical machinery seems to provide adequate approaches to high-dimensional compositional data analysis [63, 108], there is a crucial flaw:

data itself being significantly distorted in this procedure (Chapter 2). Another drawback of log-ratio transformations arises with the linear models after transformation, mostly based on the log-contrast model [7]. Such models are commonly assumed in the literature [103, 104], but they are incompatible with *amalgamation*, one of the most natural operations of compositional data [83]. Amalgamation is an aggregation procedure using similarity information of variables; for instance, $(x_1 + x_2, x_3 + x_4)$ is an amalgamation of (x_1, x_2, x_3, x_4). Such an operation is pervasive in microbiome data processing: variables are aggregated based on their genetic hierarchical structure to reduce the number of variables and the sequencing error [59]. Therefore, linear model-based approaches after log-ratio transformations do not consistently model this common practice; i.e., data analysis results may differ across the different taxonomic aggregation levels.

In this chapter, we address all these issues with our new interpretable dimension reduction framework of compositional data, as well as our algorithm that does not rely on log-ratio transformations. Our proposed framework will generalize the amalgamation of compositional data, focusing on its intuitive nature; it simply aggregates similar variables. Although amalgamation has not been of great interest to researchers due to incompatibility with log-ratio methods, very recently, a few amalgamation-based compositional data analysis approaches have emerged in the literature. They commonly articulate the intuitive nature of the amalgamation in practice, while some approaches still rely on log-ratio transformations so that they suffer from zero problems [41, 43]. Furthermore, we will argue in Section 5.2.1 that amalgamation has a crucial property, preservation of relative information to the total, making it an essential choice for an interpretable dimension reduction framework. However, despite the intuitive nature and the crucial property, approaches to finding desirable amalgamation encounter some practical challenges: the aggregating operation is too rigid and discrete. Such discreteness poses two problems: it cannot incorporate weak similarity structures, and finding an optimal amalgamation leads to a computationally infeasible discrete optimization problem. Quinn and Erb [86] formulated an optimization problem of amalgamation, but their proposed genetic algorithm is suboptimal and computationally intensive. Li et al. [58] proposed a novel regression framework with amalgamation-encouraging penalties; nevertheless, their model is confined to linear regression models, and the resulting amalgamation may only capture linear functional similarities.

To solve these rigidity problems and enhance their flexibility, we propose to generalize the amalgamation operation from a combinatorial perspective. By intuitively relaxing such a discrete operation continuously, we will see in Section 5.2.2 that our proposed framework forms another natural class of dimension reductions for compositional data. The generalized framework will also be clearly described by the original variables. Furthermore, during the generalization, our framework preserves two essential aspects of compositional data: it keeps the relative information of the quantities to the total, and the resulting dimension reduction is compositional data again. It should be noted that the latter property automatically provides a solution to the problem of interaction interpretability because compositional data are naturally defined by relative information, which is simply information on interactions between variables. In addition, our composition-to-composition framework provides a new visualization approach to high-dimensional compositional data. We will mainly describe this application in Section 5.5 through real data experiments. The analysis of our visualization will further demonstrate the distinctive and attractive interpretations of our new approach.

To achieve desirable dimension reduction within our framework for compositional data, we propose to generalize the kernel dimension reduction (KDR) approach of Fukumizu et al. [34]. As mentioned in Chapter 2, using kernels adequately deals with prevalent zero values in compositional data and thus does

not rely on log-ratio transformations. The generalized KDR approach is model-free and aims for *sufficient dimension reduction*, and we will introduce a detailed context of why we chose this approach in the following subsection 5.1.1. Having defined the class of generalized compositional dimension reductions, we prove that the empirical estimator for our dimension reduction is consistent, as initially proved using properties of orthogonal matrices in Fukumizu et al. [34]. Our generalization is applicable also for unsupervised problems, as suggested by Wang et al. [121]. We should also emphasize that, though we focus only on the dimension reduction of compositional data, our generalized KDR approach and its theoretical guarantee are not limited to compositional data and show great generality. It could be an interesting research direction to explore other application possibilities.

5.1.1 Sufficient Dimension Reduction and Related Works

Sufficient dimension reduction (SDR), proposed originally by Li [60] with the sliced inverse regression (SIR) approach, aims for the conditional independence relation $Y \perp\!\!\!\perp X \,|\, B^T X$, where X is a vector of d covariates, Y is a response, and B is an $m \times d$ matrix with orthonormal rows and $m \leq d$. Note that this framework no longer applies to compositional data, as orthogonal projections break the proportional structure of data. Lee et al. [53] extensively generalized this framework to arbitrary projection functions p, so the target relation becomes $Y \perp\!\!\!\perp X \,|\, p(X)$. If we put our compositional dimension reduction function $p : \Delta^{d-1} \to \Delta^{m-1}$, which will be defined in Section 5.2.2, to this relation, we obtain an intuitive interpretation within the SDR framework: the target function p reduces dimension via aggregating variables based on their *functional similarity*.

To achieve this compelling dimension reduction of compositional data, we need a method that fulfills two requirements: it can adequately deal with zero values and adapt to our specifically designed interpretable projection functions. Most classical linear SDR methods (see Li [55] for various examples) are confined to finding linear subspaces, so they do not apply directly to compositional data. Lee et al. [53] proposed two nonlinear SDR methods, generalized sliced inverse regression (GSIR) and generalized sliced average variance estimator (GSAVE), but these methods find general nonlinear functions in Hilbert spaces and break interpretations of variables, even though they can be applied to compositional data without replacing zeros. In contrast, the KDR method shows adaptability to other specifically structured functions by replacing their Stiefel manifold optimization with other classes of specific functions. It is also capable of dealing with zeros appropriately, as discussed before. Therefore, we approach our dimension reduction framework in this chapter by expanding the ability of the KDR method.

Recently, Tomassi et al. [117] proposed a likelihood-based inverse regression approach to SDR for compositional data. They gave both log-ratio and non-log-ratio methods, where the former requires zero replacements, and the latter applies to the count data before normalization. The latter method does not require zero replacements, but the orthogonal projections of the original count data do not account for the compositional nature. We compare our proposed method and their approach in Section 5.5.

5.1.2 Main Contribution

The contribution of this chapter is three-fold: (1) we propose a new interpretable composition-to-composition dimension reduction framework as a generalization of amalgamation. The compositional outcome naturally addresses interactions between lower dimension coordinates; (2) to achieve desirable dimension reduction within our framework, we extensively generalize the KDR method to arbitrarily structured dimension-reducing functions and prove the consistency of the empirical dimension reduction

estimator. We also refine the assumptions of the consistency theory originally given in Fukumizu et al. [34], showing that our theory successfully generalizes to broader settings under milder assumptions than originally given; (3) through the experimental results, we showcase that our method provides a new compelling data visualization tool for compositional data through the ternary plot. With the information on the dimension reduction estimate, our analysis gives a clear, intuitive explanation in terms of the original variables.

5.1.3 Outline of the Chapter

In Section 5.2, we propose our dimension reduction framework for compositional data by generalizing the amalgamation with an emphasis on its information-preserving property. To achieve our new compositional dimension reduction, Section 5.3 first extends the ability of the KDR method to a broader class of dimension reduction functions and then applies it to our compositional framework. Section 5.4 is an independent section of compositional data that describes the theoretical guarantee of our generalized KDR estimator. We conduct real data experiments in Section 5.5, which showcases that our method provides an excellent new graphical exploration tool for compositional data. Finally, we discuss conclusions and fruitful directions for future works in Section 5.6.

5.2 Composition-To-Composition Dimension Reduction Framework

While many typical operations of Euclidean data are not applicable to compositional data, there are two natural dimension-reducing operations called subcomposition and amalgamation [3]. In this section, we explore the crucial property of amalgamation and advocate using amalgamation for dimension reduction. Then, as the strict amalgamation is too rigid and unwieldy, we propose a generalized framework for dimension reduction for compositional data that includes amalgamation.

5.2.1 Amalgamation and Relative Information to the Total

In the pioneering work of Aitchison [3] on compositional data, he introduced two fundamental dimension-reducing operations of compositional data: subcomposition and amalgamation. The subcomposition is a simple variable selection process that performs renormalization after selection of variables. Amalgamation is an aggregation-based dimension reduction process of compositional data, which is intuitive and very common in practice. Given a compositional data $x \in \Delta^{d-1}$ and a fixed-order partition $P = \{I_1, \ldots, I_m\}$ of variable indices $\{1, \ldots, d\}$, we have a partition-based aggregation

$$x_P := \left(\sum_{j \in I_1} x_j, \ldots, \sum_{j \in I_m} x_j \right) \in \Delta^{m-1},$$

called the amalgamation of x with respect to the partition P. Such a partition P is typically formed based on extrinsic information of similarity, such as phylogenetic tree for microbiome data. However, the genetic similarity may not exactly indicate the functional similarity of taxa, which is often a crucial objective of microbiome data analysis. Thus, it is desirable to have a data-driven approach to figure out an appropriate partition P that aggregates compositional covariates via functional similarity.

Apart from its interpretability, we also argue an essential and crucial property of the amalgamation, preservation of relative information to the total, which was originally pointed out by Park et al. [81].

Although it is common to perform renormalization after various operations of compositional data [3], we demonstrate that such a process produces a significant issue by considering the following toy microbiome data example. Let $X = (X_1, X_2, X_3, X_4) \in \Delta^3$ be a compositional vector with four taxa, and let $Y \in \{0, 1\}$ be a binary response variable indicating the presence of a disease. Assume that the deficiency of two taxa X_1 and X_2 causes the disease, and let $(x, 1)$ and $(x', 0)$ be two samples from (X, Y) with

$$x = (0.01, 0.02, 0.4, 0.57), \quad \text{and} \quad x' = (0.3, 0.6, 0.05, 0.05).$$

Since the rest of the taxa X_3 and X_4 do not directly affect the disease, one might be tempted to select the first two taxa X_1 and X_2 and form a subcomposition. However, after renormalization, the subcompositions of x and x' on the first two variables become the same value $(1/3, 2/3) \in \Delta^1$ with the different disease labels, which precludes the further analysis based on this variable selection. This problem arises essentially due to the following phenomenon: renormalization erases the relative information of variables to the total, which is also a significant feature of data.

In contrast, amalgamation does not require a renormalization process, so it does not lose the relative information to the total during the dimension reduction. The problem of variable selection above is also solved using amalgamation, where the resulting selection should be represented as $(0.01, 0.02, 0.97)$ and $(0.3, 0.6, 0.1)$ by aggregating the rest variables into one coordinate.

In terms of dimension reduction, keeping the relevant data information is extremely important, so we develop our dimension reduction approach based on amalgamation. The following subsection will devise a new dimension reduction framework for compositional data, still not requiring the renormalization process.

5.2.2 Generalized Amalgamation of Compositional Data

As formulated by Quinn and Erb [86], the set of all amalgamations from Δ^{d-1} to Δ^{m-1}, $m \leq d$, can be represented as the following collection of binary matrices

$$\mathcal{A}_{m,d} := \left\{ A = (a_{ij}) \in \mathbb{R}^{m \times d} \,\middle|\, a_{ij} \in \{0, 1\}, \sum_{i=1}^{m} a_{ij} = 1, \forall j = 1, \ldots, d \right\}$$

with the linear operation, $Ax \in \Delta^{m-1}$ for all $A \in \mathcal{A}_{m,d}$ and $x \in \Delta^{d-1}$. Letting $e_i \in \Delta^{m-1}$ as the standard basis vectors, all zeros but one at the i-th coordinate, we may describe all the elements $A \in \mathcal{A}_{m,d}$ as $A = (e_{a(1)}, \ldots, e_{a(d)})$, where $a : \{1, \ldots, d\} \to \{1, \ldots m\}$ is a function. For example, if $A = (e_1, e_1, e_2, e_3) \in \mathcal{A}_{3,4}$, then $Ax = (x_1 + x_2, x_3, x_4)$ for $x \in \Delta^3$. This operation makes sense if two variables x_1 and x_2 are *strongly similar*: e.g. if they are linearly correlated, meaning that the ratio $(x_1 : x_2)$ is constant, or if $Y = f(X_1, X_2, X_3, X_4) = g(X_1 + X_2, X_3, X_4)$ for some functions f, g and a response Y.

However, we argue that this discrete operation is too strict and rigid, producing limitations in both theoretical and practical viewpoints. For example, consider a similar case in which X_2 weakly affects Y, and its weak effect is linearly correlated with X_1. This may incorporate the case $X \in \Delta^3$ and $Y = f(X_1, X_2, X_3) = g(X_1 + cX_2, X_3)$ for some functions f, g and a constant $c \in (0, 1)$, where X_4 indirectly affect Y via the unit-sum constraint. This type of weak similarity structure is also conceivable in practice, but amalgamation does not capture this relation. There is an additional computational challenge of finding optimal amalgamations $A \in \mathcal{A}_{m,d}$ because the cardinality of the class $\mathcal{A}_{m,d}$ is m^d, which is infeasible for high-dimensional datasets. Quinn and Erb [86] proposed distance-based discrete optimization problems for amalgamation, approached via genetic algorithms, but its performance is suboptimal, and computation tends to be extremely intensive.

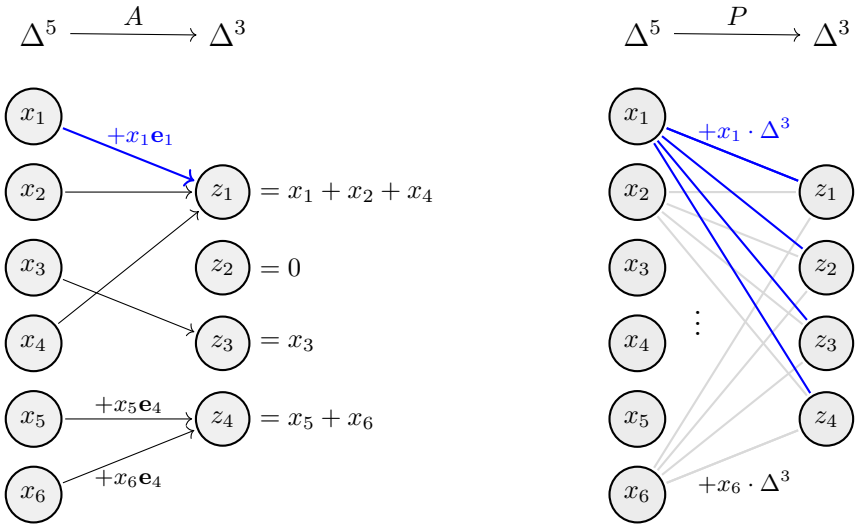


Figure 5.1: Illustrations of the amalgamation and our generalized framework of compositional dimension reduction. The left image shows the amalgamation $A = (e_1, e_1, e_3, e_1, e_4, e_4) \in \mathcal{A}_{4,6}$, and we can interpret this as each x_j is strictly allocated to one of the z_i. The right image visualizes how we can generalize such a discrete allocation continuously. The blue lines indicate x_1 is distributed to the target variables, whereas the blue arrow in the left image indicates a strict allocation.

To circumvent these issues of amalgamation, we propose a new framework of compositional dimension reduction, continuously relaxing the discrete process of amalgamation. In a combinatorial perspective, amalgamation can be viewed as *strict allocations* of the original variables to the target compositional variables. The left image of Figure 5.1 illustrates this idea in case $A = (e_1, e_1, e_3, e_1, e_4, e_4) \in \mathcal{A}_{4,6}$. Here, each x_j is allocated to one of the coordinates e_i with the value x_j, resulting in the amalgamation $Ax = (x_1+x_2+x_4, 0, x_3, x_5+x_6) \in \Delta^3$ after summing up all the allocations. From this viewpoint, we may readily think of adopting a probabilistic perspective. That is, we may *distribute* the original quantities x_j to all the target coordinate, resulting in $x_j \cdot p_j \in x_j \cdot \Delta^3$, where $p_j = (p_{1j}, p_{2j}, p_{3j}, p_{4j})^T \in \Delta^3$. This distribution is illustrated in the right hand side of Figure 5.1. Summing up all these distributions, we have a matrix representation of this operation $x \mapsto Px$, where $P = (p_{ij})$ satisfies $\forall p_{ij} \geq 0$ and $\sum_i p_{ij} = 1$ for all j.

We summarize this generalized framework as the following class for general dimensions ($m \leq d$),

$$\mathcal{M}_{m,d} = \left\{ P = (p_{ij}) \in \mathbb{R}^{m \times d} \,\middle|\, 0 \leq p_{ij} \leq 1,\ \sum_{i=1}^{m} p_{ij} = 1, \forall j = 1, \ldots, d \right\}. \tag{5.1}$$

It is immediately checked that $Px \in \Delta^{m-1}$ for all compositions $x \in \Delta^{d-1}$, thus $\mathcal{M}_{m,d}$ forms a class of composition-to-composition dimension reductions, $\Delta^{d-1} \to \Delta^{m-1}$. It also does not require renormalization and, therefore, does not suffer the information loss problem described in Section 5.2.1. From our discussions, we note that each column P_j of $P \in \mathcal{M}_{m,d}$ represents how the j-th variable x_j is distributed into the lower dimensional coordinates. Also, if the matrix P is binary, it represents an amalgamation. Therefore, we have clear interpretability for our compositional projection matrices $P \in \mathcal{M}_{m,d}$, similar to amalgamation when the columns are nearly binary. We will see in Section 5.5 that this perspective provides an unprecedented visualization of compositional data and its interpretations.

We conclude this section with mentioning a related work on compositional data analysis. Our class

(5.1) also appears in the work of Fiksel et al. [28], where they proposed a linear method for composition-on-composition regression. They focused on the linear regression framework $\mathbb{E}[Y|X] = PX$, where both X, Y are compositional, and they derived that P must be of the form (5.1) as ours. In contrast, we started from amalgamation to achieve dimension reduction and generalized the framework with interpretability. Thus, there are crucial differences in motivations and the tasks, while it is interesting that two different goals approached to the same operation of compositional data.

5.3 Compositional Dimension Reduction with Generalized KDR

As mentioned in Section 5.1.1, we take an SDR approach to achieve desirable compositional dimension reduction within the framework (5.1). Since the projections of $\mathcal{M}_{m,d}$ are interpreted as continuously relaxed aggregations of variables, the SDR relation $Y \perp\!\!\!\perp X \,|\, PX$ indicates a relaxed amalgamation based on the functional similarity between covariates. We propose to generalize the KDR method [34] to achieve this relation, and we begin with some preliminary notions and results.

5.3.1 Conditional Covariance Operator

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random vector with joint distribution $\mathbb{P}_{XY}$, where $\mathcal{X}$ and $\mathcal{Y}$ are compact domains. Since our domain $\mathcal{X}$ will be a compact simplex, it is also reasonable to confine the domain $\mathcal{Y}$ of responses to be compact. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be continuous kernels generating reproducing kernel Hilbert spaces (RKHSs) $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively. As the kernels are continuous, the RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are embedded in the space of continuous functions $C(\mathcal{X})$ and $C(\mathcal{Y})$, respectively.

The *cross-covariance operator* of (X, Y) is a mapping $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ defined uniquely by the following adjoint relations,

$$\begin{aligned} \langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} &= \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])\,(g(Y) - \mathbb{E}_Y[g(Y)]) \right] < \infty \\ &= \mathrm{Cov}\left[f(X), g(Y) \right], \end{aligned} \tag{5.2}$$

for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. We call Σ_{XX} a covariance operator in case $Y = X$. Note that the operator Σ_{YX} captures all the covariances between RKHS evaluations of X and Y. Baker [9] showed that there is also a correlation-analogue, a unique bounded operator $V_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ satisfying

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}, \quad \|V_{YX}\| \leq 1, \quad \text{and} \quad V_{YX} = P_{\overline{\mathrm{ran}}(\Sigma_{YY})} V_{YX} P_{\overline{\mathrm{ran}}(\Sigma_{XX})}, \tag{5.3}$$

where $\overline{\mathrm{ran}}(\Sigma)$ denotes the closure of the range of the operator Σ, and P_W denotes the orthogonal projection onto the subspace W of a Hilbert space.

Based on these concepts, the *conditional covariance operator* on $\mathcal{H}_{\mathcal{Y}}$ is defined by

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2}, \tag{5.4}$$

which coincides with $\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ if the covariance operator Σ_{XX} is invertible. This notion generalizes the conditional covariance matrix of Gaussian random vectors, and Fukumizu et al. [34] demonstrated that computational properties of joint Gaussian variables also generalize to RKHS operators.

Proposition 5.1 ([34, Propositions 2 and 3])**.** *For any $g \in \mathcal{H}_{\mathcal{Y}}$, we have*

$$\langle g, \Sigma_{YY|X} g \rangle = \inf_{f \in \mathcal{H}_{\mathcal{X}}} \mathrm{Var}(g(Y) - f(X)). \tag{5.5}$$

If $\mathcal{H}_\mathcal{X} + \mathbb{R}$ is dense in $L^2(\mathbb{P}_X)$, we further have

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]]. \tag{5.6}$$

Hence, the conditional covariance operator indeed captures the (expected) conditional variance of evaluations $g(Y)$ given X, for all $g \in \mathcal{H}_\mathcal{Y}$, whenever $\mathcal{H}_\mathcal{X} + \mathbb{R}$ is dense in $L^2(\mathbb{P}_X)$. Such a density constraint is equivalent to saying that the RKHS $\mathcal{H}_\mathcal{X}$ is rich enough, and this is satisfied by various kernels $k_\mathcal{X}$, such as *characteristic* and *universal* kernels [34, Proposition 5]. An RKHS $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ is said to be characteristic if the corresponding kernel mean embedding of probability measures on $\mathcal{X}$, $\mathbb{P} \mapsto \mathbb{E}_{X \sim \mathbb{P}}[k_\mathcal{X}(\cdot, X)]$, is one-to-one. It is called universal if $\mathcal{H}_\mathcal{X}$ is dense in $C(\mathcal{X})$ with the uniform topology. It is known that all universal kernels are characteristic [45], and a broad class of popular kernels, such as Gaussian and Laplace, are universal on compact subsets of Euclidean space.

5.3.2 Generalized Kernel Dimension Reduction

We present the generalized theory of kernel dimension reduction in this subsection. Although the original method is developed for orthogonal matrices and exploits some unique properties of orthogonality, its central intuition, minimizing conditional covariance approaches the conditional independence relation of the SDR, generalizes to dimension reduction functions of arbitrary form. We expand the potential power of the KDR method to a broader domain by stating our theory in a fully general language, not confined to compositional data.

Assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact domain of predictors and let $\mathcal{Z} \subset \mathbb{R}^m$, $m \leq d$, be a target compact domain of dimension reduction on which another RKHS $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ is given. We state the following theorem with all *arbitrary* measurable maps $p : \mathcal{X} \to \mathcal{Z}$, and its proof is available in Chapter 3.

Theorem 5.2. *Suppose that $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ is dense in $L^2(\mathbb{P}_X)$ (e.g., $k_\mathcal{X}$ is universal), and let $Z = p(X)$, where $p : \mathcal{X} \to \mathcal{Z}$ is a measurable map. Then,*

$$\Sigma_{YY|Z} \succeq \Sigma_{YY|X}, \tag{5.7}$$

where the inequality $\succeq$ stands for the partial order of self-adjoint operators. If we further assume that $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ and $(\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ are characteristic, then

$$\textit{the equality } \Sigma_{YY|Z} = \Sigma_{YY|X} \textit{ holds if and only if } Y \perp\!\!\!\perp X \,|\, Z.$$

The proof of the original version [32, 34] of Theorem 5.2 uses the property of the matrix $B \in \mathbb{R}^{d \times m}$ with orthonoral columns: the columns can be extended to an orthonormal basis of $\mathbb{R}^d$ so that X can be decomposed into two parts, $B^T X$ and its complementary part. In contrast, we show that such a decomposition property is not actually needed by adopting σ-field languages of conditional expectations [26].

Applying the operator trace to the relation (5.7), we have $\mathrm{Tr}(\Sigma_{YY|p(X)}) \geq \mathrm{Tr}(\Sigma_{YY|X})$ for all p with $\mathrm{Tr}(\Sigma_{YY|p(X)}) = \mathrm{Tr}(\Sigma_{YY|X})$ if and only if $Y \perp\!\!\!\perp X \,|\, p(X)$. Therefore, the generalized KDR algorithm will perform minimization of $\mathrm{Tr}(\Sigma_{YY|p(X)})$ among the dimension reduction functions p of interest. For a particular class $\mathcal{F}$ of functions $p : \mathcal{X} \to \mathcal{Z}$, which may be the Stiefel manifold of Fukumizu et al. [34] or our compositional class $\mathcal{M}_{m,d}$, the population algorithm is written as

$$\arg\min_{p \in \mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)}). \tag{5.8}$$

To ensure this solution set nonempty, we will assume $\mathcal{F}$ is a compact metric space and prove in Section 5.4 that $p \mapsto \Sigma_{YY|p(X)}$ is continuous under weak continuity constraints on $\mathcal{F}$.

When we focus on the trace of conditional covariance operators directly, we may derive a similar result in the special case of $id_{\mathcal{Y}} \in \mathcal{H}_{\mathcal{Y}}$, where id denotes the identity function. This case includes the univariate responses $\mathcal{Y} \subset \mathbb{R}$ with the linear kernel $k_{\mathcal{Y}}(y, y') = yy'$, which is *not* characteristic. The following result compromises little theoretical power but computationally attractive, which generalizes Corollary 3 of Chen et al. [20] with conceptual corrections.

Proposition 5.3. *Let $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, $p : \mathcal{X} \to \mathcal{Z}$, and $Z = p(X)$ as in Theorem 5.2. Suppose that $(\mathcal{H}_{\mathcal{Z}}, k_{\mathcal{Z}})$ is characteristic and $id_{\mathcal{Y}} \in \mathcal{H}_{\mathcal{Y}}$. Then we have $\operatorname{Tr}(\Sigma_{YY|Z}) \geq \operatorname{Tr}(\Sigma_{YY|X})$ with the equality $\operatorname{Tr}(\Sigma_{YY|Z}) = \operatorname{Tr}(\Sigma_{YY|X})$ holds if and only if $\mathbb{E}[Y|X] = \mathbb{E}[Y|Z]$ almost surely.*

Here, the equality $\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)]$ means that knowing $p(X)$ suffices for predicting Y. This is a general version of sufficient dimesion reduction for conditional mean [22, 55] that is slightly weaker than the actual SDR, whereas Chen et al. [20] stated that this is equivalent to SDR. However, it suffices for many practical applications as our primary interest is often estimating the regression function $\mathbb{E}[Y|X]$, rather than the entire distribution $\mathbb{P}_{Y|X}$.

We also mention an application of Theorem 5.2 for unsupervised dimension reduction problems as proposed in Wang et al. [121]. Letting $\mathcal{Y} = \mathcal{X}$ and $Y = X$, we approach the following independence relation by solving the optimization (5.8):

$$X \perp\!\!\!\perp \widetilde{X} \,|\, p(X), \tag{5.9}$$

where $\widetilde{X}$ is an i.i.d. copy of X. We may further investigate this relation as follows. Letting $Z = p(X)$, the independence relation (5.9) is equivalent to the inclusive relation of σ-fields $\sigma(X) \subseteq \overline{\sigma(Z)}$ [50, Corollary 8.11], where $\overline{\sigma(Z)}$ denotes the completion of $\sigma(Z)$ with respect to the ambient σ-field of the probability space. Since then $\overline{\sigma(X)} = \overline{\sigma(Z)}$ holds clearly, we interpret (5.9) as $\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)]$ almost surely for all integrable random variables Y.

5.3.3 Estimating the Generalized KDR and Computational Aspects

Estimating the solution functions of the optimization problem (5.8) requires first to estimate the objective function, $\operatorname{Tr}(\Sigma_{YY|p(X)})$. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ be sampled i.i.d. from the joint distribution $\mathbb{P}_{XY}$. By adopting a Tikhonov-type regularization for the operator inversion of $\Sigma_{p(X)p(X)} : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{Z}}$, we have a similar result of empirical estimate as computed in Fukumizu et al. [34]:

$$\operatorname{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) = \varepsilon_n \operatorname{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y),$$

where ε_n is a regularization parameter converging to zero as $n \to \infty$, $\widehat{\Sigma}^{(n)}_{YY|p(X)} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ is the empirical conditional covariance operator, $G_{p(X)}$ is the centered Gram matrix of kernel $k_{\mathcal{Z}}$ with the projected data $\{p(x_i)\}_{i=1}^n$, and G_Y is the centered gram matrix of $k_{\mathcal{Y}}$. Since we differently derive our estimate without pulling back the RKHS $\mathcal{H}_{\mathcal{Z}}$ to the Hilbert space defined on $\mathcal{X}$, we give detailed computations in Section 3.5.

Letting $\mathcal{F}$ be any class of functions $p : \mathcal{X} \to \mathcal{Z}$ as defined in (5.8), we formulate our empirical estimate for the generalized KDR as

$$\operatorname*{arg\,min}_{p \in \mathcal{F}} \operatorname{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) = \operatorname*{arg\,min}_{p \in \mathcal{F}} \operatorname{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y). \tag{5.10}$$

We will establish the large sample theory and consistency of the estimator (5.10) in Section 5.4. We also have another form of $\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)})$ by directly computing the trace of the operator $\widehat{\Sigma}^{(n)}_{YY|p(X)}$ using a complete orthonormal system (CONS) for $\mathcal{H}_{\mathcal{Y}}$. An interesting case of this direction is $\mathcal{Y} \subset \mathbb{R}$ with the linear kernel $k_{\mathcal{Y}}$, where the RKHS $\mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\vee}$ is spanned by the identity function $id_{\mathcal{Y}}$. In this case, assuming the $\{y_i\}$ are centered, we may write

$$\mathrm{Tr}(\widehat{\Sigma}^{(n)}_{YY|p(X)}) = \inf_{f \in \mathcal{H}_{\mathcal{Z}}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(f(p(x_i)) - \frac{1}{n} \sum_{j=1}^{n} f(p(x_j)) \right) \right]^2 + \varepsilon_n \|f\|^2_{\mathcal{H}_{\mathcal{Z}}}, \tag{5.11}$$

which resembles the minimized loss function of the kernel ridge regression (KRR) after projection by $p \in \mathcal{F}$. Therefore, we can intuitively regard the optimization (5.10) as seeking a function $p \in \mathcal{F}$ that minimizes the KRR residual error after projection.

To compute the objective function $\mathrm{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y)$, we should specify the kernels $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}}$ that may involve two parameter choice problems. Fortunately, the kernel choice for $k_{\mathcal{Y}}$ can be significantly simplified in numerous practical problems. When given a multi-class response problem, $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(k)}\}$, the *delta kernel*

$$k_{\mathcal{Y}}(y, y') := \delta_{y,y'} = \begin{cases} 1 & \text{if } y = y' \\ 0 & \text{otherwise} \end{cases}$$

is a natural choice as it treats different labels as completely dissimilar. It is commonly adopted in various supervised kernel methods [107, 20, 11], and there is also experimental support that the delta kernel empirically outperforms Gaussian kernels in other approaches [126]. The delta kernel is universal, so Theorem 5.2 applies. In case Y is a univariate continuous response, $\mathcal{Y} \subset \mathbb{R}$, we take $k_{\mathcal{Y}}$ as the linear kernel and aim for the SDR for conditional mean, as discussed after Proposition 5.3. Thus, in these circumstances, we only need to decide the kernel parameter for $k_{\mathcal{Z}}$, which we will discuss in Section 5.5 in detail.

The computational complexity of the dimension reduction objective $\mathrm{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y)$ is $O(nl + n^2 m + n^3)$, where l denotes the complexity of computing each projection $p(X)$, assuming that computing $k_{\mathcal{Z}}(p(x_i), p(x_j))$ requires $O(m)$ operations. Recall that m is the dimension of the target domain $\mathcal{Z}$. The dominant term $O(n^3)$ arises from the matrix inversion, and one may consider reducing this by low-rank approximations for kernel matrices, such as random Fourier features [87] or the Nyström method [24]. Since microbiome datasets in practice, which are of our primary interest, typically have a small sample size (mostly $n < 1000$), we do not take this approach in this thesis.

5.3.4 Dimension Reduction Algorithm for Compositional Data

With all of these extended theories and empirical computations in hand, we return to our main goal, the interpretable dimension reduction of compositional data. We set $\mathcal{X} = \Delta^{d-1}$ and $\mathcal{Z} = \Delta^{m-1}$, and recall that our class of dimension reduction functions was

$$\mathcal{M}_{m,d} = \left\{ P = (p_{ij}) \in \mathbb{R}^{m \times d} \,\middle|\, 0 \le p_{ij} \le 1,\ \sum_{i=1}^{m} p_{ij} = 1, \forall j = 1, \ldots, d \right\},$$

where $P \in \mathcal{M}_{m,d}$ sends $x \in \Delta^{d-1}$ to $Px \in \Delta^{m-1}$. Then, the generalized KDR algorithm (5.10) translates into the compositional case as

$$\arg\min_{P \in \mathcal{M}_{m,d}} \mathrm{Tr}((G_{PX} + n\varepsilon_n I_n)^{-1} G_Y). \tag{5.12}$$

We will use the Gaussian kernel for $k_{\mathcal{Z}}$ to satisfy the requirements of Theorem 5.2 and Section 5.4. Then, (5.12) is a matrix optimization problem, which can be optimized by projected gradient descent. Note that each column P_j of $P \in \mathcal{M}_{m,d}$ is a member of the simplex Δ^{m-1}, so the class $\mathcal{M}_{m,d}$ equals to a d-product of simplices $\Delta^{m-1} \times \cdots \times \Delta^{m-1}$. Hence, we can compute the orthogonal projection of an arbitrary matrix $Q \in \mathbb{R}^{m \times d}$ to the set $\mathcal{M}_{m,d}$ via column-wise application of simplex projections, proposed by Duchi et al. [25].

While the projected gradient descent applies to the problem (5.12), it is a nonconvex optimization since the solution matrices exhibit symmetry. That is, if $\hat{P}$ is a solution of (5.12), then all the row permutations of $\hat{P}$ are also contained in $\mathcal{M}_{m,d}$ and minimizes the objective, and this problem occurs also at the population level. Such a symmetry problem is common in deep neural networks and approached by random initializations [38, 47]. We similarly propose to randomly initialize each column of $P \in \mathcal{M}_{m,d}$ by sampling from the uniform distribution on the simplex Δ^{m-1} to break the symmetry. The nonconvexity also naturally occurs from the Gram matrix G_{PX}. Nevertheless, we empirically see that projected gradient descent with random initialization finds a dimension reduction matrix with attractive performance, as will be observed in Section 5.5.

Though we do not proceed further in this thesis, one can similarly approach the unsupervised dimension reduction problem by solving the optimization

$$\underset{P \in \mathcal{M}_{m,d}}{\arg\min} \operatorname{Tr}((G_{PX} + n\varepsilon_n I_n)^{-1} G_X) \tag{5.13}$$

with a characteristic kernel $k_{\mathcal{X}}$ on $\mathcal{X} = \Delta^{d-1}$. Wang et al. [121] suggested that this type of objective function is asymptotically similar to the dependence maximization objective based on the Hilbert-Schmidt independence criterion (HSIC), which is computationally more efficient. However, their argument is confined to uniform distributions of X in the sphere as they exploit rotational symmetry. Furthermore, in general, Liu and Ruan [64] recently proved that such an HSIC-based optimization fails to achieve the SDR relation for the class $\mathcal{F}$ of variable selections, whereas our generalized KDR approach has the theoretical guarantee at the global optimums, as will be seen in the next section.

5.4 Theory of the Generalized KDR Estimator

In this section, we prove that the empirical estimator of our compositional dimension reduction is statistically consistent. Again, to fully describe potentials of our generalized KDR theory, we state all results with a general class $\mathcal{F}$ of functions $p : \mathcal{X} \to \mathcal{Z}$ of compact spaces, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Z} \subset \mathbb{R}^m$. Recall that we assumed $\mathcal{Y}$ also a compact domain of responses. Remarkably, our generalized theory does not require any stringent assumptions except for some continuity assumptions on $\mathcal{F}$ that make optimization at least computable. For completeness of this chapter, we repeat some arguments already stated in Section 3.6.1.

Throughout the section, we assume $\mathcal{F}$ is equipped with a metric ρ, making $\mathcal{F}$ a compact metric space. Since our ultimate interest is the *evaluations* of an estimated solution $\hat{p} \in \mathcal{F}$ but not $\hat{p}$ itself, we naturally impose the following assumption:

Assumption 5.1. For every $x \in \mathcal{X}$, the evaluation functional at x, $p \mapsto p(x)$ is continuous on $\mathcal{F}$.

The following two assumptions are also needed to guarantee the consistency of (5.10):

Assumption 5.2. There exists a measurable function $\varphi : \mathcal{X} \to \mathbb{R}$ satisfying $\mathbb{E}[\varphi(X)^2] < \infty$ and the Lipschitz condition

$$\|k_{\mathcal{Z}}(p_1(x), \cdot) - k_{\mathcal{Z}}(p_2(x), \cdot)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \varphi(x)\, \rho(p_1, p_2)$$

holds for all $x \in \mathcal{X}$ and $p_1, p_2 \in \mathcal{F}$.

Assumption 5.3. For each $y \in \mathcal{Y}$, the conditional probability density function $f_{X|Y}(x|y)$ of $X|Y = y$ exists, and it is continuous in x, bounded in y, and measurable in y.

Assumption 5.1 is clearly satisfied by a broad class of parametric families, including the Stiefel manifold and our compositional family $\mathcal{M}_{m,d}$. It also guarantees that the empirical solution set

$$\arg\min_{p \in \mathcal{F}} \operatorname{Tr}((G_{p(X)} + n\varepsilon_n I_n)^{-1} G_Y)$$

is nonempty.

Assumption 5.2 plays a crucial role in proving uniform convergence of the empirical operator $\widehat{\Sigma}^{(n)}_{YY|p(X)}$ on $\mathcal{F}$. It is an extended version of the assumption (A-3) of Fukumizu et al. [34] for our broader setting. This is satisfied when, for instance, $\mathcal{F}$ satisfies an additional continuity assumption and $k_{\mathcal{Z}}$ is an l^2-radial kernel with the Lipschitz continuity; that is, for some $C > 0$ and $h : \mathbb{R} \to \mathbb{R}$, we have $k_{\mathcal{Z}}(z_1, z_2) = h(\|z_1 - z_2\|^2)$ and $|h(s) - h(t)| \leq C|s - t|$ for all $s, t \in \mathbb{R}$. To see this, observe that

$$\begin{aligned} \|k_{\mathcal{Z}}(p_1(x), \cdot) - k_{\mathcal{Z}}(p_2(x), \cdot)\|^2_{\mathcal{H}_{\mathcal{Z}}} &= 2h(0) - 2h(\|p_1(x) - p_2(x)\|^2) \\ &\leq 2C\|p_1(x) - p_2(x)\|^2 \end{aligned}$$

for all $p_1, p_2 \in \mathcal{F}$. Therefore, if there is a function ψ on $\mathcal{X}$ such that $\|p_1(x) - p_2(x)\| \leq \psi(x)\rho(p_1, p_2)$, i.e., the evaluation functional at x is also Lipschitz continuous on $\mathcal{F}$ with the constant $\psi(x)$, we obtain a similar form to Assumption 5.2. In particular, the Gaussian kernel satisfies this condition, and the Stiefel manifold or our class $\mathcal{M}_{m,d}$ satisfies the Lipschitz condition. Although this case subsumes Assumption 5.1, we separated it because other theories only require continuity of the evaluations.

We may feel free to accept Assumption 5.3. It is satisfied by a broad range of probability distributions even if the random variable Y is discrete. In fact, Assumption 5.3 refines the assumption (A-1) of Fukumizu et al. [34], where they originally assumed that the following mapping

$$p \mapsto \mathbb{E}[\mathbb{E}[g(Y)|p(X)]^2] \tag{5.14}$$

is continuous on $\mathcal{F}$ for all $g \in \mathcal{H}_{\mathcal{Y}}$, in case $\mathcal{F}$ is a Stiefel manifold. They justified their assumption via exploiting unique properties of orthogonal projections, whereas we prove the continuity of (5.14) with a full generality, elaborated in Section 3.6.3. Since the continuity of (5.14) is not a clear fact that can be accepted without doubt (see Ackerman et al. [1] for counterexamples when Y is discrete), we replace the original assumption and prove the continuity for clarity.

On the other hand, we emphasize that we removed the assumption (A-2) of Fukumizu et al. [34] since it can be weakened and subsumed to our almost fixed assumption that $\mathcal{H}_{\mathcal{Z}}$ is characteristic. See Section 3.6.2 and Section 3.6.3 for detailed demonstrations.

Finally, under Assumptions 3.1-3.3, we have the desirable consistency result in our expanded framework:

Theorem 5.4. *Suppose that $(\mathcal{H}_{\mathcal{Z}}, k_{\mathcal{Z}})$ is characteristic, and that the regularization parameter ε in* (3.21) *satisfies*

$$\varepsilon_n \to 0 \quad \textit{and} \quad n^{1/2}\varepsilon_n \to \infty \quad \textit{as} \quad n \to \infty. \tag{5.15}$$

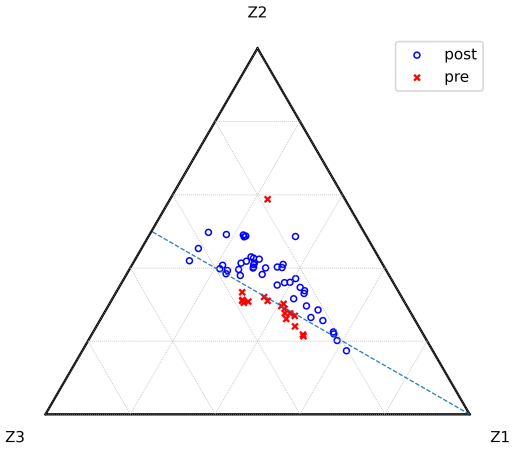


Figure 5.2: Visualization of the dimension reduction result of the skin microbiome data into three variables, Z_1 to Z_3. The dashed line indicates the set of points with $Z_2 = Z_3$. Except one possible outlier in pre-menopausal samples, the menopausal status is discriminated by the relative ratio between Z_2 and Z_3.

Let $\hat{p}^{(n)}$ *be a member of the nonempty empirical solution set* (3.21)*. Then, the set of optimal population solutions,* $\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)})$*, is nonempty. Furthermore, we have the convergence*

$$\mathrm{Tr}\left(\widehat{\Sigma}^{(n)}_{YY|\hat{p}^{(n)}(X)}\right) \to \mathrm{Tr}(\Sigma_{YY|p'(X)})$$

in probability, where p' *is a population solution in* $\arg\min_{p\in\mathcal{F}} \mathrm{Tr}(\Sigma_{YY|p(X)})$*.*

The proof of Theorem 5.4 is organized in Section 3.6, where we mainly follow the lines of the original theory but thoroughly investigate the extendability from orthogonal matrices to general projections.

It should be noted that Theorem 5.4 is independent of the SDR-related results presented in Section 5.3.2. Thus, if there is a function $p \in \mathcal{F}$ such that $\mathbb{P}_{Y|X} = \mathbb{P}_{Y|p(X)}$ or $\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)]$, then the empirical estimator (5.10) is guaranteed to converge to SDR or SDR for conditional mean, provided that additional constraints of Theorem 5.2 or Proposition 5.3 are satisfied. Theorem 5.4 also guarantees the unsupervised case (5.9) whenever there exists $p \in \mathcal{F}$ such that $\overline{\sigma(X)} = \overline{\sigma(p(X))}$.

5.5 Experiments

To demonstrate the usefulness of our new approach to compositional data, we apply the method described in Section 5.3.4 to two real microbiome datasets. Here, we emphasize the visualization ability of our method with the target domain $\mathcal{Z} = \Delta^2$ or Δ^3. Since this provides a new visualization approach to high-dimensional compositional data, we elaborate thoroughly on the interpretability of the proposed method with the first dataset.

For these experiments, we use a Gaussian kernel $k_{\mathcal{Z}}(z, z') = \exp(-\|z - z'\|^2/\sigma^2)$ with σ being the median heuristic measured from the original sample before projection. Also, we fix the regularization parameter to $\varepsilon = 0.001$. Note that these parameters can be fitted alternatively using cross-validation combined with predictive models.

The first dataset we study is skin microbiome data extensively studied by Carrieri et al. [16]. Here, leg skin microbiome samples are obtained from 62 Canadian women to study their association with menopausal status: 44 women in post-menopausal and 18 women in pre-menopausal. The abundances of microbial taxa are measured by 16S rRNA gene sequencing and aggregated to the genus level. As a result, the data consist of $n = 62$ samples and 186 dimensions with about 59% of the data being zeros.

Figure 5.2 displays a three-variable ternary plot obtained by the proposed method with $\mathcal{Z} = \Delta^2$. Although we significantly reduced the data dimension, as observed in the figure, our dimension reduction estimate captures a clear, distinguishable pattern between two classes of menopausal status, except for one sample nearest to the variable Z_2. The dashed line in the figure indicates the line $Z_2 = Z_3$ on the simplex, which seems to work roughly as a decision boundary of two classes. That is, the relative ratio between Z_2 and Z_3 discriminates two classes, whereas Z_1 plays no role in such a discrimination. Note that the principal coordinate analysis plot using the Bray-Curtis dissimilarity, a popular measure of dissimilarity in the microbiome literature, fails to exhibit a distinguishable pattern on this dataset (see Supplementary Figure 5 of Carrieri et al. [16]).

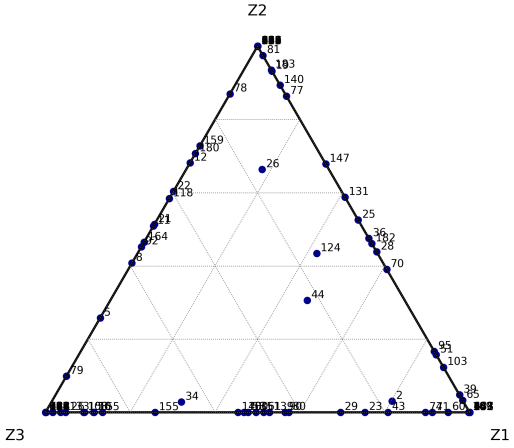


Figure 5.3: Visualization of the estimated columns of the matrix $\widehat{P}$ with their indices. Only five columns are not on the boundary, and the majority of points are located near the vertices.

To reinforce our analysis, we delve deeper into the estimated dimension reduction matrix $\widehat{P}$. As delineated in Section 5.2.2, the columns of $\widehat{P}$ represent how the original variables X_j are distributed to the target variables Z_k. In our result, 125 out of the 186 columns converged to binary vectors, meaning that these 125 (about 67% of the total) taxa are strictly allocated to one of the variables Z_k. That is, we may interpret these taxa as strictly amalgamated: 14 taxa are allocated to Z_1, 63 taxa are allocated to Z_2, and 48 taxa are allocated to Z_3. With a little relaxation, we observe that 158 columns of $\widehat{P}$ have a value greater than 0.7, indicating that about 85% of the taxa are concentrated in one of the Z_k. We visualize the columns of $\widehat{P}$ with their indices in Figure 5.3. From this figure, we further observe that only five taxa are located in the interior of the simplex Δ^2. The boundary points indicate that they are dissimilar to the aggregated taxa into the variable Z_k on the opposite side. Also, since the relative ratio between Z_2 and Z_3 plays a discriminating role between two classes, we may interpret that the taxa near the center of the line segment Z_2Z_3 do not distinguish those classes but are dissimilar to those near Z_1. Finally, the taxa lying on the segment Z_3Z_1 present weak similarity to the discriminant direction of Z_3,

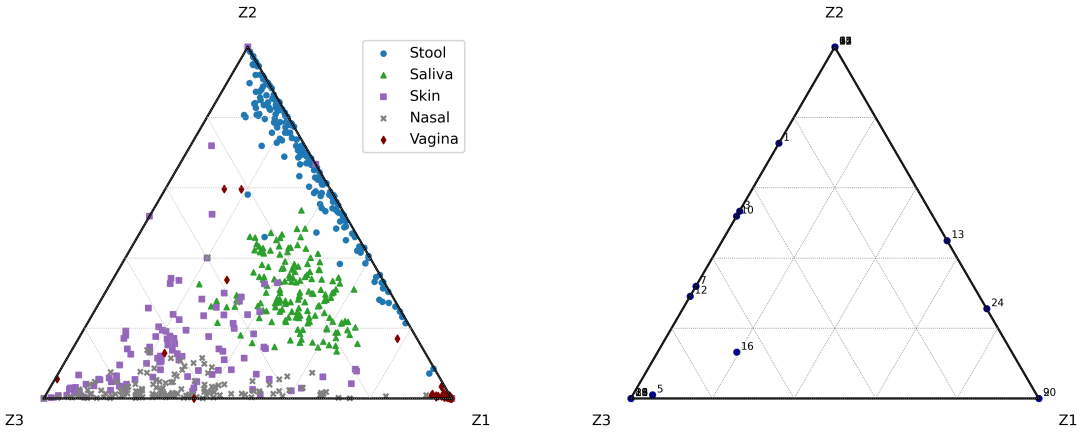


Figure 5.4: Ternary plots displaying the dimension reduction result for the HMP data by the proposed method. The left plot displays the projected data through the estimated dimension reduction matrix, and the right plot displays the columns of the estimated matrix.

whereas they are dissimilar to Z_2. A similar interpretation is possible for the taxa on the segment Z_1Z_2.

We then compare our method to another SDR approach to compositional data proposed by Tomassi et al. [117], a likelihood-based inverse regression based on the fixed-length Poisson graphical model (SDR-FPGM). This method does not require zero replacements as ours, whereas it directly applies to the microbiome count data and seeks orthogonal projections for the unnormalized count variables. We take the same microbiome data they used; the original data was taken from the Human Microbiome Project (HMP) [49]. The data were processed by deleting species with mostly zeros as many as possible, resulting in $d = 27$ taxa with $n = 681$ samples. The processed HMP data have five categories of body site: `nasal`, `saliva`, `skin`, `stool`, and `vagina`.

We first display the ternary plot of the dimension reduction result of our method in Figure 5.4 with visualization for the columns of the estimated matrix. Here, 18 taxa are strictly aggregated into the variables Z_k: two taxa at Z_1, eight taxa at Z_2, and eight taxa at Z_3. The variable Z_1, where two original taxa are amalgamated, discriminates the class `vagina` because the corresponding samples are predominantly located near Z_1. Interestingly, we observe that the relative ratio between Z_2 and Z_3 also distinguishes the remaining four classes. We draw linear decision boundaries for these four classes via dashed lines on the left hand side of Figure 5.5, which makes our observation more apparent. Note that the relative ratio $Z_2 : Z_3$ is constant on each dashed line. Therefore, we may regard the segment Z_2Z_3 as a discriminant direction for these four classes.

Figure 5.5 compares the proposed method to the SDR-FPGM approach of Tomassi et al. [117] via two-dimensional plots. The SDR-FPGM method provides an orthogonal projection of the original count variables, displayed in the right panel. As discussed above, our ternary plot provides clear interpretations through how the original variables are aggregated into the dimension-reduced variables Z_k. Furthermore, our method naturally offers insights into interactions between the dimension-reduced variables Z_k since the result is compositional data again. In contrast, each axis of the SDR-FPGM plot consists of a cumbersome linear combination with both positive and negative coefficients. Nonetheless, one may interpret every single axis solely based on such a linear combination, but it is quite unclear to understand the interactions between two axes simultaneously, even though they are theoretically orthogonal in

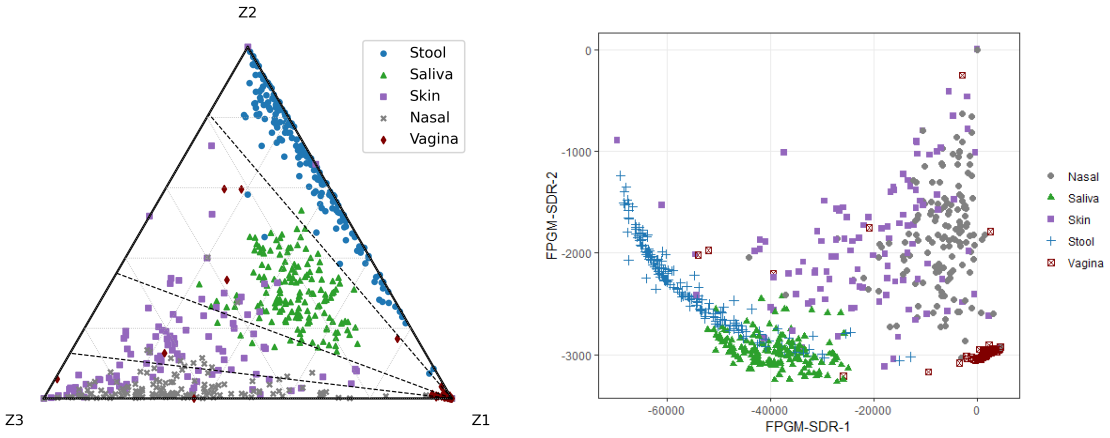


Figure 5.5: Comparison of our method and the SDR-FPGM method of Tomassi et al. [117]. Linear desicion boundaries between the classes except for the class `vagina` is plotted on the left image. On the right panel, we observe similar class separations, but the SDR axes do not provide any clear and intuitive interpretations.

Euclidean space.

To summarize, our proposed method of compositional dimension reduction provides a new visualization for high-dimensional compositional data with surprisingly clear interpretations similar to amalgamation. Furthermore, as the dimension reduction result is compositional again, the resulting variables are interpreted *relatively*: we naturally gain insights into the interactions between resulting variables of dimension reduction, which is unclear in ordinary linear SDR methods.

5.6 Discussions

This chapter proposes a new dimension reduction framework of compositional data, whose result is also compositional and interpreted similarly to the amalgamation. We also generalize the KDR method to achieve the proposed compositional dimension reduction within the SDR framework. The SDR relation combined with our new framework provides a compelling interpretation: aggregation of variables based on their functional similarity. After such an aggregation into lower dimensional compositional coordinates via the proposed method, the result enables a thorough graphical exploration using ternary plots, which naturally exhibit relative information between projected variables. We believe this will provide an attractive option to practitioners dealing with compositional data.

Similarly to the original KDR work of Fukumizu et al. [34], our generalization of the KDR approach also does not infer the target dimension of dimension reduction. Also, developing a distributional theory for the trace objective is still an open problem. On the other hand, we emphasize that our KDR-based approach should not be the only option for achieving our new compositional dimension reduction framework. The attractive experimental results in Section 5.5 encourage the development of various statistical approaches building on our amalgamation-based framework. We leave this direction as a future work.

While our approach in this chapter provides a reasonable dimension reduction, it does not produce prediction functions like other SDR approaches. If one wants to have predictions, we recommend applying

KRR or support vector machines after the dimension reduction of compositional data. The analogy to KRR pointed out in (5.11) suggests we will only need a little parameter tuning for the attached predictive models. However, this intuition needs a justification, and we leave this also as a future work.

Chapter 6. Future Directions

In this chapter, we leave several questions for future research directions. As we proposed new methodological frameworks for compositional data in Chapters 2 and 5, it would be desirable to make further developments based on these frameworks for compositional data analysis.

In Chapter 2, we first approached compositional data using kernel methods to adequately deal with increasingly available high-dimensional, sparse datasets in practice. Here, we suggested using the radial transformation, which has been neglected for a few decades and provides a natural viewpoint for compositional data since it does not break the original proportional interpretation. We proposed to apply kernel methods after radial transformation, which does not replace zeros so that they do not distort the original data. However, this approach suffers from the curse of dimensionality, a problem further exacerbated in overdispersed datasets such as microbiome data. Thus, based on the radial perspective, we ask the following question:

Question 1. Can we perform a high-dimensional data analysis on the nonnegative orthant of the sphere? Can we preserve the interpretability of the variables during the analysis?

Chapter 3 presents a generalization of the KDR method, and it is applied to variable selection in Chapter 4 and dimension reduction in Chapter 5. Despite their large sample consistence and pleasing empirical power, our KDR-based approaches do not provide statistical significance to those selected or aggregated variables. It naturally leads to the following question:

Question 2. Is there a viable approach to infer the selected or aggregated variables statistically? We may imagine a post-selection inference or a post-reduction inference for this purpose.

On the other hand, we are interested in a variety of potential applications of our generalization, not confined to compositional data.

Question 3. Is there another data with a reasonable class $\mathcal{F}$ of dimension reductions to which the generalized KDR method can be applied?

Also, to expand its potential, we need to address the following two questions on the computational complexity and optimization guarantee:

Question 4. Can we develop a scalable algorithm for the generalized KDR method? Can it be compatible with the theoretical guarantee derived in Section 3.6?

Question 5. Does the nonconvex KDR objective function have good stationary points of the gradient descent? (see Ruan et al. [92] for the corresponding result in the variable selection case)

In Chapter 5, we proposed a new dimension reduction framework for compositional data, extending the concept of amalgamation. Combined with the SDR framework, our proposed dimension reduction performs similarity-based variable aggregations, where similarity means the functional similarity of predictor variables. On the other hand, other types of similarity exist, such as correlations between variables or genetic similarity. Therefore, we can imagine the integrations of our framework with the other types of similarity.

Question 6. Can we develop adequate methods for compositional dimension reduction based on correlations or genetic similarity?

Finally, as SDR-based approaches do not produce prediction directly, we ask for other approaches for predictive models that can be integrated with the proposed dimension reduction framework.

Question 7. Can we develop an approach for compositional data that learns the right dimension reduction within our framework and produces predictions simultaneously?

Bibliography

[1] N. L. Ackerman, C. E. Freer, and D. M. Roy. On the computability of conditional probability. *Journal of the ACM (JACM)*, 66(3):1–40, 2019.

[2] J. Ahn. A stable hyperparameter selection for the Gaussian RBF kernel for discrimination. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(3):142–148, 2010.

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

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

[5] J. Aitchison. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4): 365–379, 1992.

[6] J. Aitchison. Principles of compositional data analysis. *Lecture Notes-Monograph Series*, pages 73–81, 1994.

[7] J. Aitchison and J. Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984.

[8] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473:174–180, 2011.

[9] C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[10] K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *J. Mach. Learn. Res.*, 22:1–1, 2021.

[11] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

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

[13] V. I. Bogachev and M. A. S. Ruas. *Measure Theory*, volume 1. Springer, 2007.

[14] B. Brill, A. Amir, and R. Heller. Testing for differential abundance in compositional counts data, with application to microbiome studies. *The Annals of Applied Statistics*, 16(4):2648–2671, 2022.

[15] A. Butler and C. Glasbey. A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(5):505–520, 2008.

[16] A. P. Carrieri, N. Haiminen, S. Maudsley-Barton, L.-J. Gardiner, B. Murphy, A. E. Mayes, S. Paterson, S. Grimshaw, M. Winn, C. Shand, et al. Explainable ai reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific reports*, 11(1):4565, 2021.

[17] J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

[18] E. S. Charlson, J. Chen, R. Custers-Allen, K. Bittinger, H. Li, R. Sinha, J. Hwang, F. D. Bushman, and R. G. Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*, 5(12):e15216, 2010.

[19] F. Chayes. On correlation between variables of constant sum. *Journal of Geophysical Research*, 65 (12):4185–4193, 1960.

[20] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan. Kernel feature selection via conditional covariance minimization. *Advances in Neural Information Processing Systems*, 30, 2017.

[21] J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. Ribosomal database project: data and tools for high throughput rrna analysis. *Nucleic acids research*, 42(D1):D633–D642, 2014.

[22] R. D. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

[23] K. Donhauser, M. Wu, and F. Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

[24] P. Drineas, M. W. Mahoney, and N. Cristianini. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005.

[25] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

[26] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[27] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[28] J. Fiksel, S. Zeger, and A. Datta. A transformation-free linear regression for compositional outcomes and predictors. *Biometrics*, 78(3):974–987, 2022.

[29] P. Filzmoser, K. Hron, and M. Templ. Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 27:585–604, 2012.

[30] G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[31] T. Freidling, B. Poignard, H. Climente-González, and M. Yamada. Post-selection inference with hsic-lasso. In *International Conference on Machine Learning*, pages 3439–3448. PMLR, 2021.

[32] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[33] K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2), 2007.

[34] K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

[35] C. Gimblet, J. S. Meisel, M. A. Loesche, S. D. Cole, J. Horwinski, F. O. Novais, A. M. Misic, C. W. Bradley, D. P. Beiting, S. C. Rankin, L. P. Carvalho, E. M. Carvalho, P. Scott, and E. A. Grice. Cutaneous leishmaniasis induces a transmissible dysbiotic skin microbiota that promotes skin inflammation. *Cell Host & Microbe*, 22(1):13–24.e4, 2017.

[36] K. L. Glassner, B. P. Abraham, and E. M. Quigley. The microbiome and inflammatory bowel disease. *Journal of Allergy and Clinical Immunology*, 145(1):16–27, 2020.

[37] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

[38] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[39] T. Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4): 1327–1349, 2013.

[40] B. Goodman and H. Gardner. The microbiome and cancer. *The Journal of pathology*, 244(5): 667–676, 2018.

[41] M. Greenacre. Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Applied Computing and Geosciences*, 5:100017, 2020.

[42] M. Greenacre. Compositional data analysis. *Annual Review of Statistics and its Application*, 8: 271–299, 2021.

[43] M. Greenacre, E. Grunsky, and J. Bacon-Shone. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Computers & Geosciences*, 148:104621, 2021.

[44] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[45] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[46] H. Hayden, A. Eng, C. Pope, M. Brittnacher, A. Vo, E. Weiss, K. Hager, B. Martin, D. Leung, S. Heltshe, E. Borenstein, S. Miller, and L. Hoffman. Fecal dysbiosis in infants with cystic fibrosis is associated with early linear growth failure. *Nature Medicine*, 26:215–221, 2020.

[47] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[48] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

[49] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. Fitzgerald, R. S. Fulton, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[50] O. Kallenberg. *Foundations of Modern Probability*. Probability theory and stochastic modelling. Springer, 3rd edition, 2021. ISBN 9783030618728.

[51] I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

[52] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[53] K.-Y. Lee, B. Li, and F. Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

[54] S. Lee, S. Jung, J. Lourenco, D. Pringle, and J. Ahn. Resampling-based inferences for compositional regression with application to beef cattle microbiomes. *Statistical Methods in Medical Research*, 32 (1):151–164, 2022.

[55] B. Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018.

[56] B. Li and J. Ahn. Reproducing kernels and new approaches in compositional data analysis. *arXiv preprint arXiv:2205.01158*, 2022.

[57] B. Li and J. Song. Nonlinear sufficient dimension reduction for functional data. *Annals of Statistics*, 45(3):1059–1095, 2017.

[58] G. Li, Y. Li, and K. Chen. It's all relative: Regression analysis with compositional predictors. *Biometrics*, 79(2):1318–1329, 2023.

[59] H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[60] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[61] W.-Z. Li, K. Stirling, J.-J. Yang, and L. Zhang. Gut microbiota and diabetes: From correlation to causality and mechanism. *World journal of diabetes*, 11(7):293, 2020.

[62] J. N. Lim, M. Yamada, W. Jitkrittum, Y. Terada, S. Matsui, and H. Shimodaira. More powerful selective kernel tests for feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 820–830. PMLR, 2020.

[63] W. Lin, P. Shi, R. Feng, and H. Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

[64] K. Liu and F. Ruan. A self-penalizing objective function for scalable interaction detection. *arXiv preprint arXiv:2011.12215*, 2020.

[65] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight. Unifrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5(2):169–172, 2011.

[66] J. Lu, P. Shi, and H. Li. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, 75(1):235–244, 2019.

[67] S. Lubbe, P. Filzmoser, and M. Templ. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, 210:104248, 2021.

[68] K. C. Lutz, S. Jiang, M. L. Neugent, N. J. De Nisco, X. Zhan, and Q. Li. A survey of statistical methods for microbiome data analysis. *Frontiers in Applied Mathematics and Statistics*, 8:884810, 2022.

[69] E. Marine-Roig and B. Ferrer-Rosell. Measuring the gap between projected and perceived destination images of catalonia using compositional analysis. *Tourism management*, 68:236–249, 2018.

[70] J. A. Martín-Fernández, J. Palarea-Albaladejo, and R. A. Olea. Dealing with zeros. *Compositional Data Analysis: Theory and applications*, pages 43–58, 2011.

[71] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688–2704, 2012.

[72] J.-A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2): 134–158, 2015.

[73] M. Masaeli, J. G. Dy, and G. M. Fung. From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 751–758, 2010.

[74] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 83(559):69–70, 1909.

[75] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

[76] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. IEEE, 1999.

[77] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2): 1–141, 2017.

[78] J. T. Nearing, G. M. Douglas, M. G. Hayes, J. MacDonald, D. K. Desai, N. Allward, C. Jones, R. J. Wright, A. S. Dhanani, A. M. Comeau, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):1–16, 2022.

[79] J. Palarea-Albaladejo and J. Martin-Fernandez. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

[80] J. Park, C. Yoon, C. Park, and J. Ahn. Kernel methods for radial transformed compositional data with many zeros. In *International Conference on Machine Learning*, pages 17458–17472. PMLR, 2022.

[81] J. Park, J. Ahn, and C. Park. Kernel sufficient dimension reduction and variable selection for compositional data via amalgamation. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27034–27047. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/park23a.html`.

[82] V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge university press, 2016.

[83] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[84] K. 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.

[85] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[86] T. P. Quinn and I. Erb. Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. *NAR genomics and bioinformatics*, 2(4):lqaa076, 2020.

[87] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

[88] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[89] C. L. Rasmussen, J. Palarea-Albaladejo, M. S. Johansson, P. Crowley, M. L. Stevens, N. Gupta, K. Karstad, and A. Holtermann. Zero problems with compositional data of physical behaviors: a comparison of three zero replacement methods. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1):1–10, 2020.

[90] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. Balances: a new perspective for microbiome analysis. *MSystems*, 3(4):e00053–18, 2018.

[91] A. Ron and X. Sun. Strictly positive definite functions on spheres in Euclidean spaces. *Mathematics of Computation*, 65(216):1513–1530, 1996.

[92] F. Ruan, K. Liu, and M. I. Jordan. Taming nonconvexity in kernel feature selection—favorable properties of the laplace kernel. *arXiv preprint arXiv:2106.09387*, 2021.

[93] W. Rudin. Principles of mathematical analysis. *3rd ed.*, 1976.

[94] S. Saitoh, Y. Sawano, et al. *Theory of reproducing kernels and applications*. Springer, 2016.

[95] J. Scealy and A. 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.

[96] J. Scealy and A. Welsh. Colours and cocktails: Compositional data analysis 2013 lancaster lecture. *Australian & New Zealand Journal of Statistics*, 56(2):145–169, 2014.

[97] M. Scetbon and Z. Harchaoui. A spectral analysis of dot-product kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 3394–3402. PMLR, 2021.

[98] L. Schiffer, R. Azhar, L. Shepherd, M. Ramos, L. Geistlinger, C. Huttenhower, J. B. Dowd, N. Segata, and L. Waldron. HMP16SData: Efficient access to the human microbiome project through bioconductor. *American Journal of Epidemiology*, 188(6):1023–1026, 2019.

[99] I. Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1):96–108, 1942.

[100] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[101] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[102] G. D. Sepich-Poore, L. Zitvogel, R. Straussman, J. Hasty, J. A. Wargo, and R. Knight. The microbiome and human cancer. *Science*, 371(6536):eabc4552, 2021.

[103] P. Shi, A. Zhang, and H. Li. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016.

[104] P. Shi, Y. Zhou, and A. R. Zhang. High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. *Biometrika*, 109(2):405–420, 2022.

[105] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[106] H. Song and H. Chen. Generalized kernel two-sample tests. *Biometrika*, page asad068, 11 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad068. URL `https://doi.org/10.1093/biomet/asad068`.

[107] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

[108] A. Srinivasan, L. Xue, and X. Zhan. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics*, 77(3):984–995, 2021.

[109] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

[110] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

[111] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[112] M. A. Stephens. Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203, 1982.

[113] A. Susin, Y. Wang, K.-A. Lê Cao, and M. L. Calle. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2):lqaa029, 2020.

[114] D. E. Te Beest, E. H. Nijhuis, T. W. Möhlmann, and C. J. Ter Braak. Log-ratio analysis of microbiome data with many zeroes is library size dependent. *Molecular Ecology Resources*, 21(6): 1866–1874, 2021.

[115] M. Templ, P. Filzmoser, and C. Reimann. Cluster analysis applied to regional geochemical data: problems and possibilities. *Applied geochemistry*, 23(8):2198–2213, 2008.

[116] T. Tjur. *A constructive definition of conditional distributions*. Institute of Mathematical Statistics, University of Copenhagen, 1975.

[117] D. Tomassi, L. Forzani, S. Duarte, and R. M. Pfeiffer. Sufficient dimension reduction for compositional data. *Biostatistics*, 22(4):687–705, 2021.

[118] L. Trasande, J. Blustein, M. Liu, E. Corwin, L. Cox, and M. Blaser. Infant antibiotic exposures and early-life body mass. *International journal of obesity*, 37(1):16–23, 2013.

[119] M. Tsagris and C. Stewart. A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412, 2018.

[120] H. Wang, Q. Liu, H. M. Mok, L. Fu, and W. M. Tse. A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179(2):459–468, 2007.

[121] M. Wang, F. Sha, and M. Jordan. Unsupervised kernel dimension reduction. *Advances in neural information processing systems*, 23, 2010.

[122] S. Wang. Robust differential abundance test in compositional data. *Biometrika*, 110(1):169–185, 2022.

[123] T. Wang and H. Zhao. Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11(2):771–791, 2017.

[124] D. Watson and G. Philip. Measures of variability for geological data. *Mathematical Geology*, 21: 233–254, 1989.

[125] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[126] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, 26(1):185–207, 2014.

[127] M. Yamada, Y. Umezu, K. Fukumizu, and I. Takeuchi. Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 152–160. PMLR, 2018.

[128] M. Yamada, D. Wu, Y.-H. H. Tsai, I. Takeuchi, R. Salakhutdinov, and K. Fukumizu. Post selection inference with incomplete maximum mean discrepancy estimator. *In International Conference on Learning Representations*, 2019.

[129] J. Yoo, Z. Sun, M. Greenacre, Q. Ma, D. Chung, and Y. M. Kim. A guideline for the statistical analysis of compositional data in immunology. *Communications for Statistical Applications and Methods*, 29(4):453–469, 2022.

[130] G. Zadora, T. Neocleous, and C. Aitken. A two-level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*, 55(2):371–384, 2010.

[131] W. Zaremba, A. Gretton, and M. Blaschko. B-tests: Low variance kernel two-sample tests. *arXiv preprint arXiv:1307.1954*, 2013.

Acknowledgments in Korean

먼저 지도교수님이신 박철우 교수님과 안정연 교수님께 깊은 감사의 말씀을 드립니다. 통계학을 거의 공부하지 않던 제가 전공을 바꿔 연구를 시작할 수 있는 기회를 주셨고, 기초적인 지식이 부족하여 답답한 부분들이 많으셨을 텐데도 지속해서 지원해주시고 방향을 잡아 주셔서 연구를 하나씩 해낼 수 있었습니다. 부족함이 많은데도 연구에 집중하지 못하며 성실하지 못할 때도 잦았으나, 매주 시간 내주시어 토론과 지도를 해주신 덕분에 연구를 어떻게 해야 하는지 점점 배워나갈 수 있었습니다. 아직도 박사라는 학위가 실감이 안 날 정도로 기본적인 부분들에서조차 부족함을 자주 느끼지만, 한편으로는 통계학 분야에서 연구를 어떻게든 조금씩 해 나갈 수 있는 사람이 이제는 된 것 같습니다. 배운 것들을 밑거름 삼아 앞으로도 더 많이 발전할 수 있도록 노력하는 삶을 살겠습니다.

바쁘신 와중에도 흔쾌히 학위논문 심사를 맡아주신 전현호 교수님, 정연승 교수님, 그리고 멀리 서울에서 와주신 전용호 교수님께도 감사의 말씀을 드립니다. 심사 중에 주셨던 의견과 질문들이 진행 중인 연구에 큰 도움이 되었으며, 모르던 것들을 새롭게 발견하고 더 발전하는 계기가 되었습니다.

이전 지도교수님이셨던 곽시종 교수님께도 감사의 마음을 전해 드립니다. 비록 진로 변경의 욕구를 강하게 느껴 교수님의 품을 떠나게 되었지만, 항상 교수님을 보면서 제가 중장년층이 되었을 때 교수님처럼 멋있게 연구자로서 살고 싶다는 꿈을 키웠습니다. 연구 분야를 변경한 뒤에는 산업에서 무언가를 기여하고 싶은 마음도 컸었지만, 시간이 흐르고 다시 연구자의 꿈을 가지며 제가 교수님께 많은 영향을 받았음을 느낄 수 있었습니다. 꿈을 이루기에 부족한 사람이 되지 않도록 항상 정진하겠습니다.

긴 대학원 과정 동안 저 자신을 새롭게 발견하기도 하고, 인생의 목표에 대해 많은 생각을 정립할 수 있었습니다. 단순히 흥미에 강하게 이끌려 순수 수학 연구자의 꿈을 꾸다가도, 제가 하는 일들이 세상에 조금 더 직접적으로 도움이 되어 그것을 인정받을 수 있는 삶을 동경하던 저 자신도 발견할 수 있었습니다. 연구 분야를 변경하며 새롭게 시작하던 시기에는 그저 정신 없이 꿈이랄게 크게 없는 시기를 길게 보냈습니다. 돌이켜보면 최근 3년 정도는 정말 혼란스러운 시기를 보냈던 것 같지만, 그 시기에 고민을 털어놓고 함께 대화할 수 있었던 친구들이 있어 큰 행운이었습니다. 특히 진로 변경의 시기에 말동무가 자주 되어준 덕상이에게 고마운 마음을 전하며, 연구자의 꿈을 이어나갈까에 대한 고민의 시기에 공감하며 즐겁게 이야기를 나누어준 정호에게 감사를 전합니다. 명확한 목표 없이 흘러가는 대로 살던 시기에 나타나 다시 꿈을 가지고 나아가는 데에 큰 계기와 힘이 되어준 해리에게도 고맙다는 말을 전합니다.

제 첫 연구를 함께한 창원이, 그리고 짧은 시간이지만 연구실 생활 즐겁게 함께한 후배 친구들 정민이와 본우에게도 고마움을 전합니다. 가끔 나누는 대화들이 일상의 소소한 즐거움이었습니다. 저와 가깝게 지내며 덕분에 외롭지 않을 수 있었던, 다른 모든 분에게도 감사의 말씀을 드립니다.

대학원 생활 동안 잘하고 싶은 일이 생기면 현재의 실력보다 원하는 성과의 수준만 많이 높아 조급했던 경우가 잦았습니다. 최근에도 다시 꿈을 설정하고 나아가려다 보니 스스로에게 점점 가혹해졌고 부침도 많았지만, 다행히 크게 지치지 않고 버텨낸 저에게 역시 고맙다고 말해주고 싶습니다. 곁에서 항상 함께하며 저에게 응원과 활력을 주어 즐겁게 버틸 수 있도록 도와준 해리에게도 다시 한번 고맙다는 말을 전합니다. 앞으로 더 큰 발전을 위해서는 조금 더 여유를 가지고 장기적으로 꾸준한 발전을 꾀해보려고 합니다. 필요한 능력을 충분히 쌓아 사회에 이로운 연구를 남기는 연구자가 되겠습니다.

마지막으로, 저를 항상 격려하고 지원해주시는 부모님께 감사의 말씀을 드립니다. 지금의 진로를 개발할 수 있도록 어릴 때 제 교육에 많은 신경을 써주셨던 어머니께 감사드리며, 금융 산업의 최전선에 계시며 사회의 본질에 대해 많은 지혜를 알려주신 아버지께 또한 감사드립니다. 제가 여전히 수학을 좋아하며 밀접한 학문을 즐겁게 연구하고 있는 것은 어머니의 덕분이며, 대학원 과정 중 진로를 변경하면서도 결국 적응해내던 과정에는 아버지께 배운 지혜가 큰 도움이 되었습니다. 앞으로 학교를 벗어나 더 넓은 사회에서도 적응하고 성장하며 즐거운 삶을 살아가겠습니다.

Curriculum Vitae in Korean

이 름: 박 준 영

생 년 월 일: 1994년 11월 08일

출 생 지: 부산

주 소: 대전광역시 서구 만년남로3번길 86-7, 301호

학 력

2010. 3. – 2013. 2. 한성과학고등학교

2013. 3. – 2018. 2. 고려대학교 수학과 (학사)

2018. 3. – 2024. 2. 한국과학기술원 수리과학과 (박사)

경 력

2018. 3. – 2024. 2. 한국과학기술원 수리과학과 조교

학 회 활 동

1. **Park, J.**, Ahn, J., and Park, C., *Interpretable Composition-To-Composition Dimension Reduction via Conditional Covariance Operator*, 2023 Winter Conference, the Korean Statistical Society, Seoul (Korea), December, 2023.
2. **Park, J.**, Ahn, J., and Park, C., *Kernel Sufficient Dimension Reduction and Variable Selection for Compositional Data via Amalgamation*, 40th International Conference on Machine Learning (ICML), Honolulu, HI (USA), July, 2023.
3. **Park, J.**, Ahn, J., and Park, C., *Kernel Sufficient Dimension Reduction and Variable Selection for Compositional Data via Amalgamation*, 2023 Summer Conference, the Korean Statistical Society, Busan (Korea), June, 2023.
4. **Park, J.**, Yoon, C., Park, C., and Ahn, J., *Kernel methods for radial transformed compositional data with many zeros*, 39th International Conference on Machine Learning (ICML), Baltimore, MD (USA), July, 2022.
5. **Park, J.**, Yoon, C., Park, C., and Ahn, J., *Kernel methods for radial transformed compositional data with many zeros*, 2022 Summer Conference, the Korean Statistical Society, Seoul (Korea), June, 2022.

연 구 업 적

1. **Park, J.**, Ahn, J., and Park, C. (2023), Kernel Sufficient Dimension Reduction and Variable Selection for Compositional Data via Amalgamation, In *International Conference on Machine Learning* (ICML), pp. 27034-27047, PMLR.
2. Kang, I., Choi, H., Yoon, Y.-J., **Park, J.**, Kwon, S.-S., and Park, C. (2023), Frechet Distance-Based Cluster Analysis for Multi-Dimensional Functional Data, *Statistics and Computing*, 33(4), 75.
3. **Park, J.**, Yoon, C., Park, C., and Ahn, J. (2022), Kernel Methods for Radial Transformed Compositional Data with Many Zeros, In *International Conference on Machine Learning* (ICML), pp. 17458-17472, PMLR.