Kernel Methods for Radial Transformed Compositional Data with Many Zeros

Junyoung Park [1] Changwon Yoon [2] Cheolwoo Park [1] Jeongyoun Ahn [2]

Abstract

Compositional data analysis with a high proportion of zeros has gained increasing popularity, especially in chemometrics and human gut microbiomes research. Statistical analyses of this type of data are typically carried out via a log-ratio transformation after replacing zeros with small positive values. We should note, however, that this procedure is geometrically improper, as it causes anomalous distortions through the transformation. We propose a radial transformation that does not require zero substitutions and more importantly results in essential equivalence between domains before and after the transformation. We show that a rich class of kernels on hyperspheres can successfully define a kernel embedding for compositional data based on this equivalence. To the best of our knowledge, this is the first work that theoretically establishes the availability of the extensive library of kernel-based machine learning methods for compositional data. The applicability of the proposed approach is demonstrated with kernel principal component analysis.

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 (Gloor et al., 2017), 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 (Greenacre, 2021). 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 (Pearson, 1897; Chayes, 1960). 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 (1982). 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

[1] Department of Mathematical Sciences, KAIST, Daejeon, Korea [2] Department of Industrial & Systems Engineering, KAIST, Daejeon, Korea. Correspondence to: Jeongyoun Ahn <jyahn@kaist.ac.kr>, Cheolwoo Park <parkcw2021@kaist.ac.kr>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

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. (2011; 2012) for some widely-used substitution methods. For comparisons of various zero replacement algorithms, see Rasmussen et al. (2020); Lubbe et al. (2021). 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.

1.1. Main Contributions

We first point out in Section 2 that the underlying geometry of the above log-ratio approaches, so-called the Aitchison geometry (Aitchison, 1994), *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 (Schölkopf et al., 1998; 2002; Steinwart & Christmann, 2008) and kernel mean embedding methods (Gretton et al., 2012; Zaremba et al., 2013; Muandet et al., 2017) 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.

1.2. Related Works

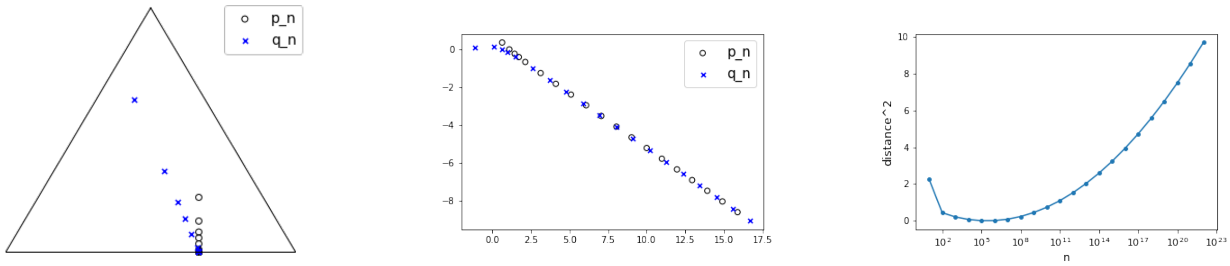
Works on non-substituting zero values. A number of works have endeavored to manage zero values of compositional data without replacing them in order to honor the true zeros (Martín-Fernández et al., 2011). One popular method is to take square-root transformation and then use the theory of directional statistics (Stephens, 1982; Wang et al., 2007; Scealy & Welsh, 2011). Butler & Glasbey (2008) proposed a latent Gaussian model on the simplex, requiring no data transformation. Zadora et al. (2010) and Bear & Billheimer (2016) modeled the probability of zero values separately with logratio-based distributions, and so did Tsagris & Stewart (2018) but with the Dirichlet distribution on the open simplex.

Works on radial transformation. We note that the radial transformation for compositional data has been considered in geological literature a few decades ago in Watson & Philip (1989), 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 & Welsh (2014) and references therein for a summary of their arguments. A main reason for Aitchison's disapproval of the transformation was that the angular distance is not subcompostionally dominant (Aitchison, 1992). 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.

Works on isotropic kernels on hyperspheres. Kernels on hyperspheres, especially isotropic types, have been broadly studied in the literature (Schoenberg, 1942; Ron & Sun, 1996; Gneiting, 2013). 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 & Harchaoui (2021) for recent exposition.

Works on kernel mean embeddings. The kernel mean embedding of distributions has attracted much attention in recent years as they are broadly applicable in arbitrary domains with appropriate kernels. Using the mean embedding, Gretton et al. (2012) proposed a non-parametric two-sample test based on the distance between probability measures, and Balasubramanian et al. (2021) proposed a goodness-of-fit test with discussions of minimax optimality. See Muandet et al. (2017) 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. (2006) and Sriperumbudur et al. (2011) characterized them in various cases.

Finally, we note that a key geometric motivation of this work is shared with our previous work (Li & Ahn, 2022),



(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 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).

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.

1.3. Organization of the Paper

Section 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 3. In Section 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. In Section 5, we briefly review well-known dot-product kernels on the hyperspheres with their universality. We take the example of kernel PCA in Section 6 to showcase the benefits of the proposed idea. We conclude the paper in Section 7 with discussions and some future research directions.

2. Geometric Limitations

In this section, we take a deeper look at the Aitchison geometry (Aitchison, 1994) on $\mathcal{S}^d$ 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

$$\mathrm{clr}(x) = \left(\log \frac{x_1}{x'}, \ldots, \log \frac{x_{d+1}}{x'}\right) \in \mathbb{R}^{d+1}$$

for all $x \in \mathcal{S}^d$, where $x' = (x_1 \cdots x_{d+1})^{1/(d+1)}$. It is a homeomorphism between $\mathcal{S}^d$ and a hyperplane in $\mathbb{R}^{d+1}$, so it transfers the linear structure and the inner product of the hyperplane to the open simplex $\mathcal{S}^d$. The geometry of clr-transformed data is called the Aitchison geometry. In this geometry, points close to the boundary of $\mathcal{S}^d$ 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, the Aitchison geometry views the data as *diverging to infinity*.

To be more specific, let us consider the following two sequences on $\mathcal{S}^2$: $p_n = (\frac{2}{3} - \frac{1}{n}, \frac{2}{n}, \frac{1}{3} - \frac{1}{n})$ and $q_n = (\frac{2}{3} - \frac{6}{n^{1.1}}, \frac{7}{n^{1.1}}, \frac{1}{3} - \frac{1}{n^{1.1}})$, where $n \geq 9$. Note that both sequences converge to the same point $(2/3, 0, 1/3)$, as displayed in the ternary plot in Figure 1(a). Thus *they are almost the same* for sufficiently large n. The isometric log-ratio (ilr) transformation, which maps $\mathcal{S}^2$ to $\mathbb{R}^2$, is applied to the sequences and the transformed points are shown in Figure 1(b). As the ilr transformation preserves the inner product, the figure exactly displays the Aitchison geometry of p_n and q_n. 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 $n \to \infty$. 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 1(c). We can see that the distance between the two sequences is clearly diverging toward infinity.

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.

2.1. Square-Root Transformation

The square-root transformation of compositional data have also been considered as a way of overcoming the simplex geometry. One sends the point $x = (x_1, \ldots, x_{d+1}) \in \Delta^d$ to $(\sqrt{x_1}, \ldots, \sqrt{x_{d+1}})$ 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 (Scealy & Welsh, 2011; 2014).

We point out one crucial disadvantage of the square-root transformation. Compositional data consist of relative ratios contained in the corresponding *radial vectors* (see Section 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 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.

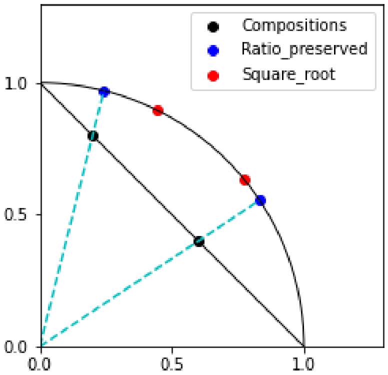


Figure 2. Comparison of the compositional radial vectors (dashed lines) and the square-root transformed points (red points) on the unit circle. The relative ratios of the red points are different from other points.

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.

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 paper. 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 $\mathcal{X}$ (Schölkopf et al., 2002). 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*

$$\langle f, \Phi(x) \rangle = f(x), \quad \forall f \in \mathcal{H}_K, \tag{1}$$

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 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 (2001) proved the consistency of SVM using universal kernels, together with examples of universal dot-product kernels on $\mathbb{R}^d$.

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*

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle \mu_{\mathbb{P}}, f \rangle \tag{2}$$

for all $f \in \mathcal{H}_K$ (Smola et al., 2007).

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 $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K} = 0$ 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. (2012) showed that universal kernels are characteristic, and thus we focus only on universal kernels in this work.

3.3. Pull-Back of RKHS

Let S 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 & Raghupathi (2016) 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

$$K \circ \varphi(s,t) = K(\varphi(s), \varphi(t)) \tag{3}$$

for all $s, t \in S$. 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 1. *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*

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

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

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

Theorem 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 S. The pull-back construction will be instrumental in formulating theoretical frameworks in Sections 4 and 5.

4. Radial Transformation and Equivalence of Function Spaces

As we showed inadequacies of traditional methods for compositional data with zeros in Section 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 2.1. 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.

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 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 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* (Scetbon & Harchaoui, 2021). 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.

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. *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.

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 (3). The pull-back map $\psi^* : \mathcal{H}_K \to \mathcal{H}_{K\circ\psi}$ defined in Section 3.3 establishes the following equivalence.

Theorem 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.*

$$\begin{tikzcd}
\Delta^d \arrow[r, "\Phi_{K\circ\psi}"] \arrow[d, "\psi"] & \mathcal{H}_{K\circ\psi} \\
\mathbb{S}^d_{\geq 0} \arrow[r, "\Phi_K"] & \mathcal{H}_K \arrow[u, "\psi^*"']
\end{tikzcd}$$

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.

We write the proof of Theorem 3 as follows. By the reproducing property (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. Then the commutativity of the diagram is readily seen from simple evaluations.

The kernel feature maps in Theorem 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 & Ruas (2007). Then the generalization of Theorem 3 is stated as follows.

Theorem 4. *The following diagram*

$$\begin{tikzcd}
\mathcal{P}(\Delta^d) \arrow[r] \arrow[d, "\psi_*"] & \mathcal{H}_{K\circ\psi} \\
\mathcal{P}(\mathbb{S}^d_{\geq 0}) \arrow[r] & \mathcal{H}_K \arrow[u, "\psi^*"']
\end{tikzcd}$$

is commutative where the horizontal maps are kernel mean embeddings.

The proof of Theorem 4 is straightforward from the definition of the push-forward map ψ_* and the generalized reproducing property (2).

We conclude from Theorems 3 and 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*.

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 3.3. We start with reviewing examples of kernels on hyperspheres.

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 (2013) provides an extensive survey of these kernels.

Isotropic kernels on spheres have been studied for a long time since Schoenberg (1942), and they are broadly used in directional data analysis. For a recent example, see Balasubramanian et al. (2021) 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.

5.2. Universal and Characteristic Kernels

As mentioned in Section 3.2, the universality or characteristicity of kernels is required to apply kernel mean embedding methods properly. Micchelli et al. (2006) 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 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.

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

The proof of Theorem 5(i) can be found in, for example, Lemma 4.55 of Steinwart & Christmann (2008). We state the characteristicity in Theorem 5(ii) for completeness, although universality is sufficient in practice. Since it is readily proved from the generalized reproducing property (2), the proof is omitted here.

We summarize some well-known and easily computable isotropic kernels on $\mathbb{S}^d$ in Table 1. Their universality properties are also marked for their use in mean embedding methods.

Table 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	$\langle x, y \rangle$	×
POLYNOMIAL	$(\gamma \langle x, y \rangle + 1)^p$	×
GAUSSIAN	$\exp\left(-\gamma \lVert x - y \rVert_2^2\right)$	○
VON-MISES	$\exp\left(\gamma \langle x, y \rangle\right)$	○
MATÉRN	$\frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\theta}{\gamma}\right) K_\nu \left(\frac{\theta}{\gamma}\right)$	○
RATIONAL QUADRATIC	$(\lVert x - y \rVert_2^2 + \gamma^2)^{-\beta}$	○

6. Empirical Examples

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 Appendix section A. The simulated data have about 40% of zero values. For illustration, Figure 3 displays the simulated data for $d = 2$ on a ternary plot. In the actual analysis, we use $d = 100$ and $d = 15$.

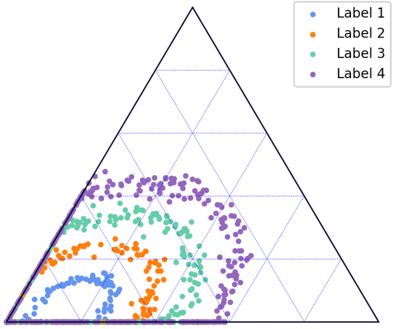
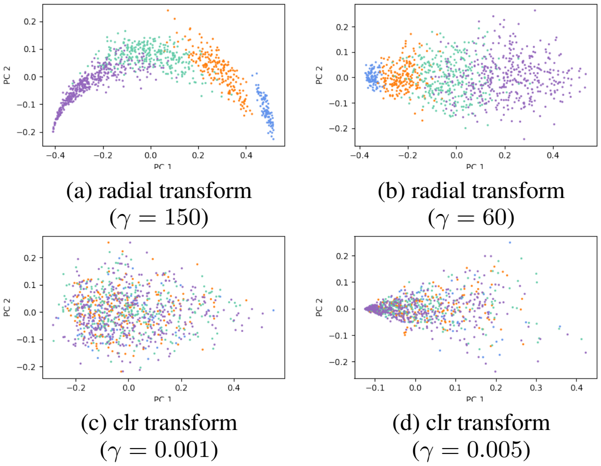


Figure 3. Ternary plot of the simulated data with $d = 2$.

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 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 $(1/2)x_{\text{min}}$ 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 (Ahn, 2010). 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 4(b). We refer to Appendix section B for use of other kernels and parameters.



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

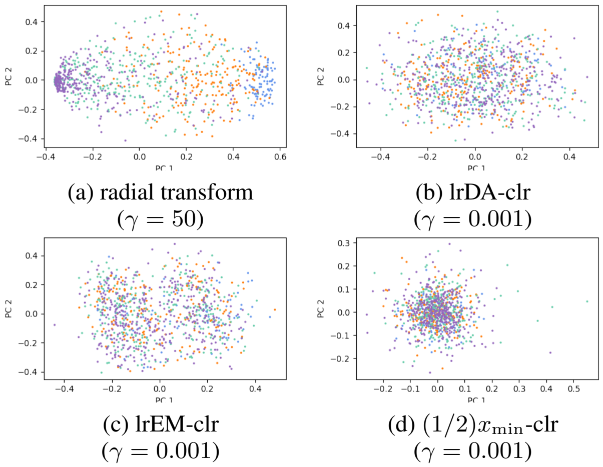
(b) radial transform ($\gamma = 60$)

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

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

Figure 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 3. Here, γ indicates the parameter for Gaussian kernel.

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 $(1/2)x_{\text{min}}$, and compared them with the radial transformed data in Figure 5. The results of lrDA and lrEM are produced by the R package `zCompositions` (Palarea-Albaladejo & Martin-Fernandez, 2015). 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 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.



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

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

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

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

Figure 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.

6.2. Quantitative Evaluation of the Proposed Method

We then provide a quantitative assessment of kPCA on new synthetic data and real-world data examples. The eigenvalues of the Gram matrix, denoted by $\lambda_1, \dots, \lambda_n$, are used to measure the effectiveness of kPCA. Note that as in linear PCA, eigenvalues of the Gram matrix can be interpreted as the amount of information that each 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 kPCA. We implement kPCA with the Gaussian kernel after the radial and the clr transformation, where we replace zeros with $(1/2)x_{\text{min}}$ before the clr transformation as in Section 6.1.

6.2.1. Synthetic Data

We simulate high-dimensional compositional data following Te Beest et al. (2021) 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

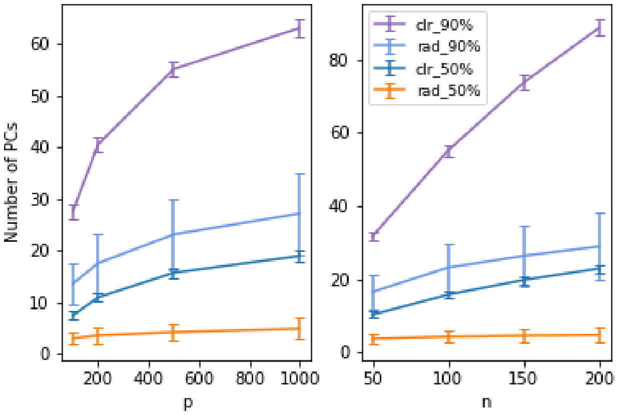


Figure 6. Number of PCs needed to capture the variability in synthetic data. The sample size is fixed at $n = 100$ for the left panel and the dimension is $p = 500$ on the right panel.

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. Figure 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, 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 6.1.

6.2.2. Real Data Examples

We also analyze real-world microbiome datasets, whose availability is listed in Appendix section D. Their attributes such as n, p, and the percentage of zeros are presented in Table 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, it is evidenced that the radial transform (r50 or r90) shows better performance than the clr transform (c50 or c90) with respect to the efficiency of dimension reduction.

Table 2. Number of PCs needed for real data examples.

Dataset	n	p	% 0	r50	r90	c50	c90
Hayden et al. (2020)	1279	643	93	5	35	22	128
Gimblet et al. (2017)	632	1860	95	1	8	8	142
Arumugam et al. (2011)	280	553	67	1	8	1	18
Carrieri et al. (2021)[1]	1200	186	58	4	15	24	111
Carrieri et al. (2021)[1]	278	186	69	4	16	19	82
Charlson et al. (2010)	60	856	89	4	17	9	39
Schiffer et al. (2019)	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).

7. Discussion and Future Works

In this work 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 (Lozupone et al., 2011), which is called the UniFrac distance. Principal coordinates analysis, equivalently multidimensional 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.

Acknowledgments

The authors would like to thank anonymous reviewers for their suggestions and feedback. The work of Junyoung Park and Cheolwoo Park were partially supported by the National Research Foundation of Korea (NRF-2021R1A2C1092925). The work of Changwon Yoon and Jeongyoun Ahn were partially supported by National Research Foundation of Korea (NRF-2021R1A2C1093526).

References

Ahn, J. 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.

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

Aitchison, J. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379, 1992.

Aitchison, J. Principles of compositional data analysis. *Lecture Notes-Monograph Series*, pp. 73–81, 1994.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Weissenbach, J., Ehrlich, S. D., and Bork, P. Enterotypes of the human gut microbiome. *Nature*, 473: 174–180, 2011.

Balasubramanian, K., Li, T., and Yuan, M. On the optimality of kernel-embedding based goodness-of-fit tests. *J. Mach. Learn. Res.*, 22:1–1, 2021.

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

Bogachev, V. I. and Ruas, M. A. S. *Measure Theory*, volume 1. Springer, 2007.

Butler, A. and Glasbey, C. A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(5): 505–520, 2008.

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.-J., Murphy, B., Mayes, A. E., Paterson, S., Grimshaw, S., Winn, M., Shand, C., Hadjidoukas, P., Rowe, W. P. M., Hawkins, S., MacGuire-Flanagan, A., Tazzioli, J., Kenny, J. G., Parida, L., Hoptroff, M., and Pyzer-Knapp, E. O. Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific Reports*, 11:4565, 2021.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., and Collman, R. G. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*, 5(12): e15216, 2010.

Chayes, F. On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193, 1960.

Gimblet, C., Meisel, J. S., Loesche, M. A., Cole, S. D., Horwinski, J., Novais, F. O., Misic, A. M., Bradley, C. W., Beiting, D. P., Rankin, S. C., Carvalho, L. P., Carvalho, E. M., Scott, P., and Grice, E. A. Cutaneous leishmaniasis induces a transmissible dysbiotic skin microbiota that promotes skin inflammation. *Cell Host & Microbe*, 22 (1):13–24.e4, 2017.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8: 2224, 2017.

Gneiting, T. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4):1327–1349, 2013.

Greenacre, M. Compositional data analysis. *Annual Review of Statistics and its Application*, 8:271–299, 2021.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hayden, H., Eng, A., Pope, C., Brittnacher, M., Vo, A., Weiss, E., Hager, K., Martin, B., Leung, D., Heltshe, S., Borenstein, E., Miller, S., and Hoffman, L. Fecal dysbiosis in infants with cystic fibrosis is associated with early linear growth failure. *Nature Medicine*, 26:215–221, 2020.

Li, B. and Ahn, J. Reproducing kernels and new approaches in compositional data analysis. *arXiv preprint arXiv:2205.01158*, 2022.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. Unifrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5 (2):169–172, 2011.

Lubbe, S., Filzmoser, P., and Templ, M. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, 210:104248, 2021.

Martín-Fernández, J. A., Palarea-Albaladejo, J., and Olea, R. A. Dealing with zeros. *Compositional Data Analysis: Theory and applications*, pp. 43–58, 2011.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688–2704, 2012.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–144, 2017.

Palarea-Albaladejo, J. and Martin-Fernandez, J. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

Paulsen, V. I. and Raghupathi, M. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge university press, 2016.

Pearson, K. 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.

Rasmussen, C. L., Palarea-Albaladejo, J., Johansson, M. S., Crowley, P., Stevens, M. L., Gupta, N., Karstad, K., and Holtermann, A. 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.

Ron, A. and Sun, X. Strictly positive definite functions on spheres in Euclidean spaces. *Mathematics of Computation*, 65(216):1513–1530, 1996.

Scealy, J. and Welsh, A. 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.

Scealy, J. and Welsh, A. Colours and cocktails: Compositional data analysis 2013 Lancaster lecture. *Australian & New Zealand Journal of Statistics*, 56(2):145–169, 2014.

Scetbon, M. and Harchaoui, Z. A spectral analysis of dot-product kernels. In *International Conference on Artificial Intelligence and Statistics*, pp. 3394–3402. PMLR, 2021.

Schiffer, L., Azhar, R., Shepherd, L., Ramos, M., Geistlinger, L., Huttenhower, C., Dowd, J. B., Segata, N., and Waldron, L. HMP16SData: Efficient access to the human microbiome project through bioconductor. *American Journal of Epidemiology*, 188(6):1023–1026, 2019.

Schoenberg, I. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1):96–108, 1942.

Schölkopf, B., Smola, A., and Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Science & Business Media, 2008.

Stephens, M. A. Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203, 1982.

Te Beest, D. E., Nijhuis, E. H., Möhlmann, T. W., and Ter Braak, C. J. Log-ratio analysis of microbiome data with many zeroes is library size dependent. *Molecular Ecology Resources*, 21(6):1866–1874, 2021.

Tsagris, M. and Stewart, C. A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412, 2018.

Wang, H., Liu, Q., Mok, H. M., Fu, L., and Tse, W. M. A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179(2):459–468, 2007.

Watson, D. and Philip, G. Measures of variability for geological data. *Mathematical Geology*, 21(2):233–254, 1989.

Zadora, G., Neocleous, T., and Aitken, C. A two-level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*, 55(2):371–384, 2010.

Zaremba, W., Gretton, A., and Blaschko, M. B-tests: Low variance kernel two-sample tests. *arXiv preprint arXiv:1307.1954*, 2013.

Appendix

A. Simulated data generation process

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

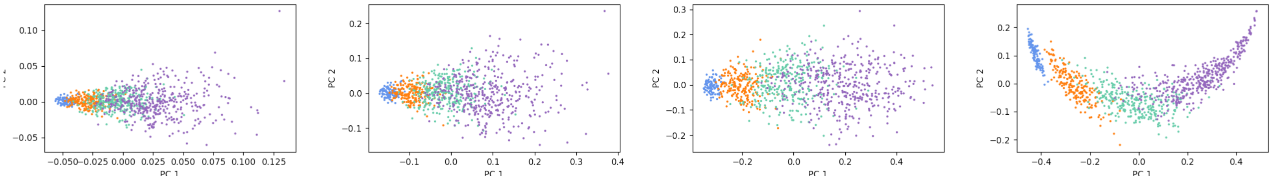
Step 2 (Scaling and shifting). The scale parameter $\dfrac{r}{10\sqrt{d/4+0.5}}$ is multiplied to each cluster. Then, we linearly shift the data by adding the d-dimensional vector $[0.15/(d-1), \cdots, 0.15/(d-1), 0.04/(d-1)]$.

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.

B. Kernel PCA with various kernel 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 4 in the paper. We use the same radial transformed and clr transformed data. 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.

B.1. Gaussian kernel



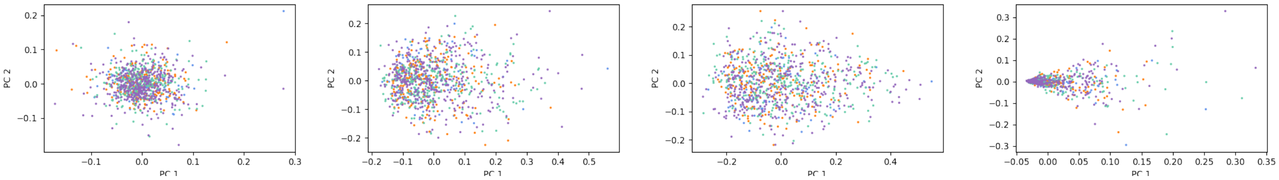
(a) radial transform ($\gamma = 1$)

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

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

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

Figure 7. Projection plots from kernel PCA with Gaussian kernel using the radial transformed data by various values of parameter.



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

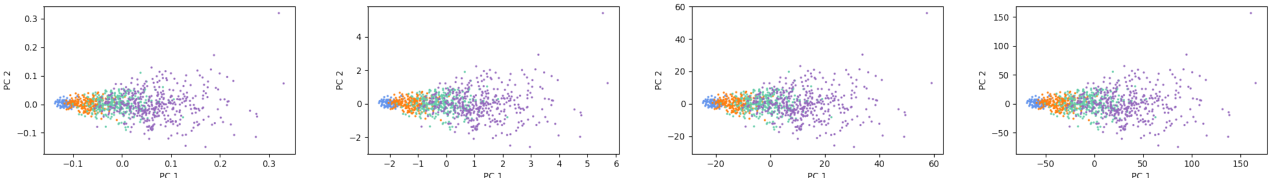
(b) clr transform ($\gamma = 0.0005$)

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

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

Figure 8. Projection plots from kernel PCA with Gaussian kernel using the clr transformed data by various values of parameter.

B.2. Polynomial kernel



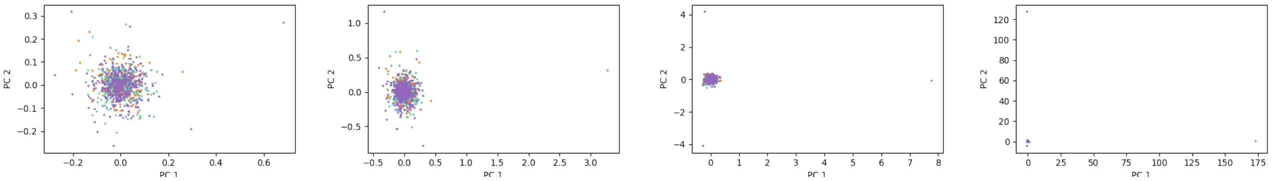
(a) radial transform ($\gamma = 1$)

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

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

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

Figure 9. Projection plots from kernel PCA with polynomial kernel using the radial transformed data by various values of parameter.



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

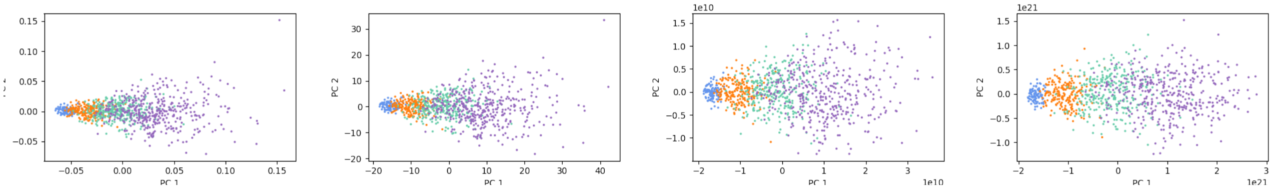
(b) clr transform ($\gamma = 0.0005$)

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

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

Figure 10. Projection plots from kernel PCA with polynomial kernel using the clr transformed data by various values of parameter.

B.3. von-Mises kernel



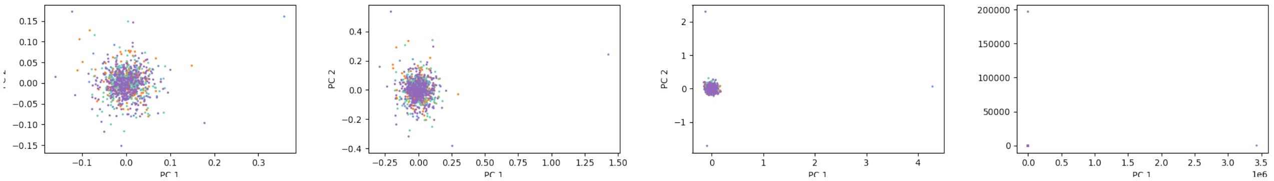
(a) radial transform ($\gamma = 1$)

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

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

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

Figure 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 ($\gamma = 0.001$)

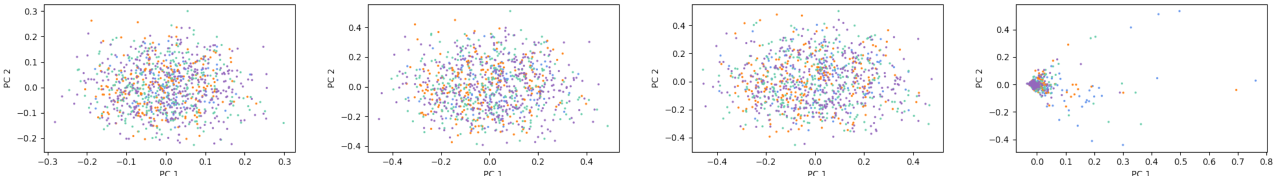
(d) clr transform ($\gamma = 0.01$)

Figure 12. Projection plots from kernel PCA with von-Mises kernel using the clr transformed data by various values of parameter.

C. Kernel PCA with various kernels and parameters using lrDA and lrEM zero replacement method

In this section, we present additional results from kernel PCA with various kernels and parameters regarding to Figure 5 in the paper. We use the same lrDA-clr and lrEM-clr transformed data. Again, the degree is $p = 3$ for the polynomial kernel.

C.1. Gaussian kernel



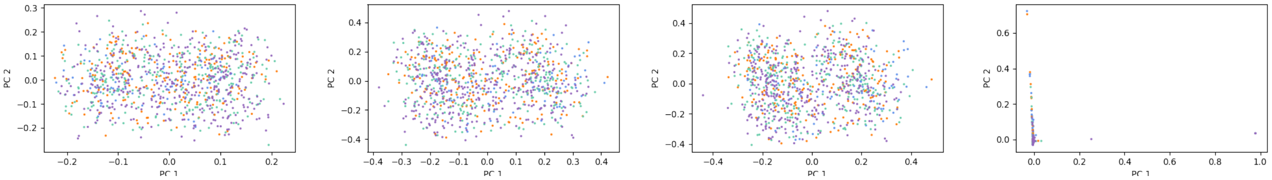
(a) lrDA-clr ($\gamma = 0.0001$)

(b) lrDA-clr ($\gamma = 0.0005$)

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

(d) lrDA-clr ($\gamma = 0.01$)

Figure 13. Projection plots from kernel PCA with Gaussian kernel using the lrDA-clr transformed data by various values of parameter.



(a) lrEM-clr ($\gamma = 0.0001$)

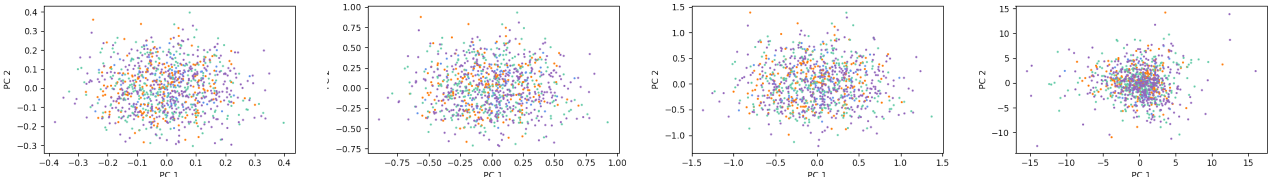
(b) lrEM-clr ($\gamma = 0.0005$)

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

(d) lrEM-clr ($\gamma = 0.01$)

Figure 14. Projection plots from kernel PCA with Gaussian kernel using the lrEM-clr transformed data by various values of parameter.

C.2. Polynomial kernel



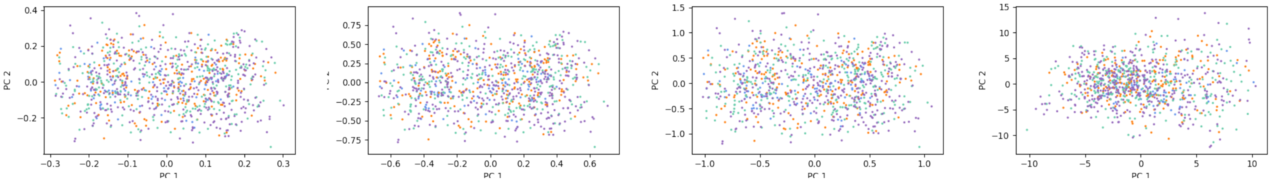
(a) lrDA-clr ($\gamma = 0.0001$)

(b) lrDA-clr ($\gamma = 0.0005$)

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

(d) lrDA-clr ($\gamma = 0.01$)

Figure 15. Projection plots from kernel PCA with polynomial kernel using the lrDA-clr transformed data by various values of parameter.



(a) lrEM-clr ($\gamma = 0.0001$)

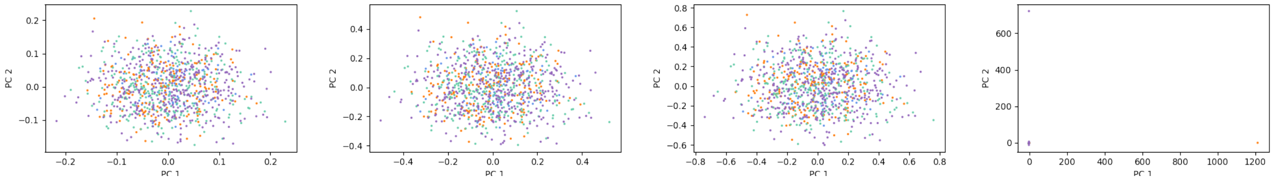
(b) lrEM-clr ($\gamma = 0.0005$)

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

(d) lrEM-clr ($\gamma = 0.01$)

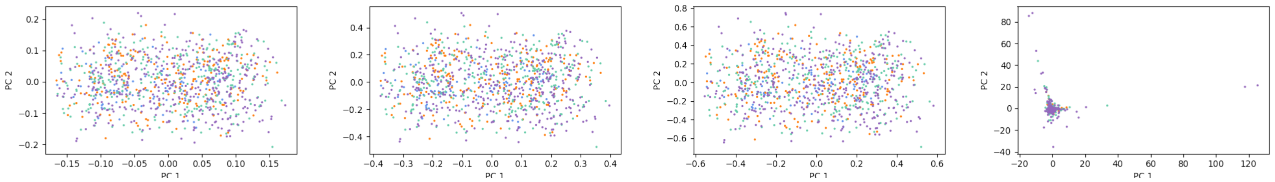
Figure 16. Projection plots from kernel PCA with polynomial kernel using the lrEM-clr transformed data by various values of parameter.

C.3. von-Mises kernel



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

D. Data availability

In this section, we present specific data availability for real data examples in section 6.2.2.

Table 3. Data availability for real data examples.

Dataset	Data Source
Hayden et al. (2020)	'BONUS-CF (WGS)' dataset from MicrobiomeDB.org
Gimblet et al. (2017)	'Experimental cutaneous leishmaniasis' dataset from MicrobiomeDB.org
Arumugam et al. (2011)	'enterotype' dataset in R package phyloseq
Carrieri et al. (2021)	Supplementary material of the referenced article
Charlson et al. (2010)	'throat.otu.tab' dataset in R package GUniFrac
Schiffer et al. (2019)	'vaginal.otu.tab' dataset in R package GUniFrac