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

Junyoung Park [1] **Jeongyoun Ahn** [2] **Cheolwoo Park** [1]

Abstract

Compositional data with a large number of components and an abundance of zeros are frequently observed in many fields recently. Analyzing such sparse high-dimensional compositional data naturally calls for dimension reduction or, more preferably, variable selection. Most existing approaches lack interpretability or cannot handle zeros properly, as they rely on a log-ratio transformation. We approach this problem with sufficient dimension reduction (SDR), one of the most studied dimension reduction frameworks in statistics. Characterized by the conditional independence of the data to the response on the found subspace, the SDR framework has been effective for both linear and nonlinear dimension reduction problems. This work proposes a compositional SDR that can handle zeros naturally while incorporating the nonlinear nature and spurious negative correlations among components rigorously. A critical consideration of sub-composition versus amalgamation for compositional variable selection is discussed. The proposed compositional SDR is shown to be statistically consistent in constructing a sub-simplex consisting of true signal variables. Simulation and real microbiome data are used to demonstrate the performance of the proposed SDR compared to existing state-of-art approaches.

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 (Huttenhower et al., 2012; Goodman & Gardner, 2018). 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 (Li, 2015). 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 (Lutz et al., 2022). 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 (Lee et al., 2022). 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. (2020), 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 (Pearson, 1897). 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 anal-

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

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

ysis) methods (Aitchison, 1982); we will introduce some of them in Section 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 (Lubbe et al., 2021). More importantly, Park et al. (2022) 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 (Nearing et al., 2022).

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. (2022) 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. (2009; 2004); Chen et al. (2017), which will be briefly reviewed in Section 3.

In Section 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. 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 (Li, 1991). 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. (2017).

Finally, we demonstrate the performance of the proposed method with synthetic and real microbiomes data in Section 5 and conclude the paper in Section 6.

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) (Gretton et al., 2005) is applied to obtain maximal dependence between variables and the response, the conditional covariance operator is used to obtain conditional independence for SDR (Chen et al., 2017), and the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) 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. (2012) to continuously relaxed algorithms with regularizations (Masaeli et al., 2010; Yamada et al., 2014).

Several studies have also been conducted to test the significance of the selected variables using kernels. These methods are based on Lee et al. (2016)'s pioneering work on *post-selection inference* (PSI), and kernel-based approaches have been successfully developed within this framework (Yamada et al., 2018; 2019; Lim et al., 2020; Freidling et al., 2021). 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 (Lin et al., 2014; Shi et al., 2016; Wang & Zhao, 2017; Lu et al., 2019). Rivera-Pinto et al. (2018) alternatively propose a forward selection process using the log-ratio balance (Egozcue et al., 2003).

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. (2021) 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 (2022) 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.

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* (Aitchison, 1982).

The sub-composition approach is simpler, in that it just *renormalizes* 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. (2015). 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

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

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 (Li, 2015). 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 (Pawlowsky-Glahn et al., 2015). In particular, amalgamations do not behave like linear projections in log-ratio geometry. Recent studies have attempted to reconcile amalgamation with log-ratio methods (Greenacre, 2020; Greenacre et al., 2021), 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. (2022) 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 $p_{\mathcal{S}} : \Delta^d \to \Delta^m$,

$$p_{\mathcal{S}}(x_0, \ldots, x_d) = \left(x_{s_1}, \ldots, x_{s_m}, \sum_{j \notin \mathcal{S}} x_j \right). \tag{1}$$

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.

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. (2017), who adopt the kernel dimension reduction (KDR) method of Fukumizu et al. (2009) for the variable selection purpose.

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

$$P_{Y|p(X)} = P_{Y|X}, \text{ or equivalently, } Y \perp\!\!\!\perp X \,|\, p(X) \tag{2}$$

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 (2018). Early studies on SDR tend to find the map p that achieves (2) among the orthogonal projections onto linear subspaces. However, because the orthogonal projections do not generally send simplex to simplex, the nonlinear SDR theory (Lee et al., 2013) 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 $p(X)$:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|p(X)] \Leftrightarrow Y \perp\!\!\!\perp \mathbb{E}[Y|X] \,|\, p(X). \tag{3}$$

This assumption is called *sufficient dimension reduction for conditional mean*, which is by definition a special case of SDR. Intuitively, (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* (Cook & Li, 2002), 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 (3) implies (2).

3.2. RKHS and Conditional Covariance Operator

While many SDR approaches are available, Fukumizu et al. (2004; 2009) 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 (1973).

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

$$\mathbb{E}_X[k_{\mathcal{X}}(X,X)] < \infty \;\text{ and }\; \mathbb{E}_Y[k_{\mathcal{Y}}(Y,Y)] < \infty. \tag{4}$$

Note that (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 (Muandet et al., 2017). 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

$$\begin{aligned}&\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \\ &\quad \mathbb{E}_{X,Y}\left[(f(X) - \mathbb{E}_X[f(X)])\,(g(Y) - \mathbb{E}_Y[g(Y)])\right]\end{aligned} \tag{5}$$

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

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2} \tag{6}$$

with $\|V_{YX}\| \le 1$ (Baker, 1973). This is called the *normalized cross-covariance operator* (NOCCO), which resembles the correlation in classical statistics (Fukumizu et al., 2007). It helps to define the following conditional covariance operator without worrying about the invertibility of Σ_{XX}:

Definition 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. (2009) 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 2. *For any $g \in \mathcal{H}_{\mathcal{Y}}$, we have*

$$\begin{aligned}&\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_{\mathcal{Y}}} = \\ &\inf_{f \in \mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X,Y}\left|(g(Y) - \mathbb{E}_Y[g(Y)]) - (f(X) - \mathbb{E}_X[f(X)])\right|^2.\end{aligned} \tag{7}$$

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

$$\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]]. \tag{8}$$

Note that the condition of (8) always holds when $k_{\mathcal{X}}$ is bounded and characteristic (Fukumizu et al., 2009). 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 (Gretton et al., 2012).

3.3. Kernel Feature Selection (KFS) via Minimization of Conditional Covariance

Motivated by (8), Fukumizu et al. (2009) and Chen et al. (2017) 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 $\mathcal{S} \subseteq \{1, 2, \dots, d\}$, 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

$$\operatorname*{argmin}_{\mathcal{S} \subseteq \{1,\dots,d\}} \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}), \tag{9}$$

where $\mathrm{Tr}(\cdot)$ 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 (Mika et al., 1999), 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 (Fukumizu et al., 2009; Chen et al., 2017) *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. Proposed Method

This section describes our kernel variable selection method for compositional data using the amalgamation in (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.

4.1. 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

$$X_{\mathcal{S}} = i_{\mathcal{S}} \circ p_{\mathcal{S}}(X), \quad X \in \mathcal{X}. \tag{10}$$

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

$$k_{\mathcal{S}}(z, w) = k_{\mathcal{X}}(i_{\mathcal{S}}(z), i_{\mathcal{S}}(w)). \tag{11}$$

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 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 appendix. 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 6.

4.2. SDR and Conditional Covariance Operator

From the discussion above, we can derive a theorem that parallels Theorem 2 in Chen et al. (2017) and Theorem 4 in Fukumizu et al. (2009):

Theorem 4. *Let $\Sigma_{YY|X_{\mathcal{S}}}$ denote the conditional covariance operator with the kernel k given in Lemma 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 $Y \perp\!\!\!\perp X \mid X_{\mathcal{S}}$. Here, the inequality $\preceq$ stands for the partial order of self-adjoint operators.

The inequality part follows immediately from Proposition 2. However, proving the equality condition needs exhaustive work due to the new projection $X_{\mathcal{S}}$ in (10). We give a full proof in the appendix. 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 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

$$\underset{\mathcal{S} \subseteq \{0, \ldots, d\}}{\mathrm{argmin}} \ \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}), \tag{12}$$

which is a compositional version of (9) with $X_{\mathcal{S}}$ defined in (10). 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. (2014) who investigate the performance of HSIC-Lasso under the two kernel choices.

For regression problems, Chen et al. (2017) 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 appendix.

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 cannot be applied to ensure the full SDR, which is claimed in Corollary 3 of Chen et al. (2017). 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 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 (12) achieves the *SDR for conditional mean*. However, Corollary 3 of Chen et al. (2017) inaccurately states that it achieves the full SDR. If we further assume the location regression model on the population (Section 3.1), we then obtain the full SDR:

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

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 (12) is equivalent to finding a subset $\mathcal{S}$ that minimizes this variance:

Corollary 6. *Under the assumptions of Proposition 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

$$Y = f(p_{\mathcal{S}}(X)) + \epsilon,\ X \perp\!\!\!\perp \epsilon, \text{ and } \mathbb{E}[\epsilon] = 0, \tag{13}$$

then Corollary 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. (2017), implicitly assuming (13).

4.3. Variable Selection Algorithm

The solution set of (12) 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

$$\underset{|\mathcal{S}| \le m}{\mathrm{argmin}}\, \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{S}}}). \tag{14}$$

Solving (14) 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. (2017) 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

$$\begin{aligned}\mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY|X_{\mathcal{S}}}) &= \mathrm{Tr}(\hat{\Sigma}^{(n)}_{YY} - \hat{\Sigma}^{(n)}_{YX_{\mathcal{S}}}(\hat{\Sigma}^{(n)}_{X_{\mathcal{S}}X_{\mathcal{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}^{(n)}_{X_{\mathcal{S}}Y}) \\ &= \epsilon_n \,\mathrm{Tr}(G_Y(G_{X_{\mathcal{S}}} + n\epsilon_n I_n)^{-1}),\end{aligned} \tag{15}$$

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 (15) is stated as

$$\min_{|\mathcal{S}| \le m} \mathrm{Tr}(\mathbf{Y}^T(G_{X_{\mathcal{S}}} + n\epsilon_n I_n)^{-1}\mathbf{Y}), \tag{16}$$

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 (16), justifying that minimizing the empirical estimate will asymptotically achieve the population minimum (12). See the appendix for proof.

Theorem 7. *Let $\hat{\mathcal{S}}^{(n)}$ be a global optimum that minimizes (16). If the regularization parameter ϵ_n satisfies*

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

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

A brute-force search of (16) 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. (2017). Note that we can express $X_{\mathcal{S}}$ as $(w \odot X, 1 - w^T X)$ where $w = (w_0, \ldots, w_d) \in \{0,1\}^d$ 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 $0 \le w_i \le 1$ and $\sum w_i \le m$, define

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

Then our relaxed objective is written as

$$\begin{aligned} \min_{w} \quad & \mathrm{Tr}(\mathbf{Y}^T (G_{X_w} + n\epsilon_n I_n)^{-1} \mathbf{Y}) \\ \text{subject to} \quad & \|w\|_1 \le m,\ 0 \le w_i \le 1,\ \forall i. \end{aligned} \tag{17}$$

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 5 (see also Ruan et al. (2021)). 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 (17) 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 (Rahimi & Recht, 2007).

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 (Lin et al., 2014; Lu et al., 2019) and selbal (Rivera-Pinto et al., 2018), chosen from the recent survey by Susin et al. (2020). 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 Appendix A; 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. (2020) 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.

5.1. Synthetic Data

We begin with simulations of microbiome count data proposed by Te Beest et al. (2021), 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

$$\log \mu_{ij} = a_i + t_j + e y_i, \tag{18}$$

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 1.

As illustrated in the figure, the proposed method clearly outperforms the log-ratio methods on average. The left panel

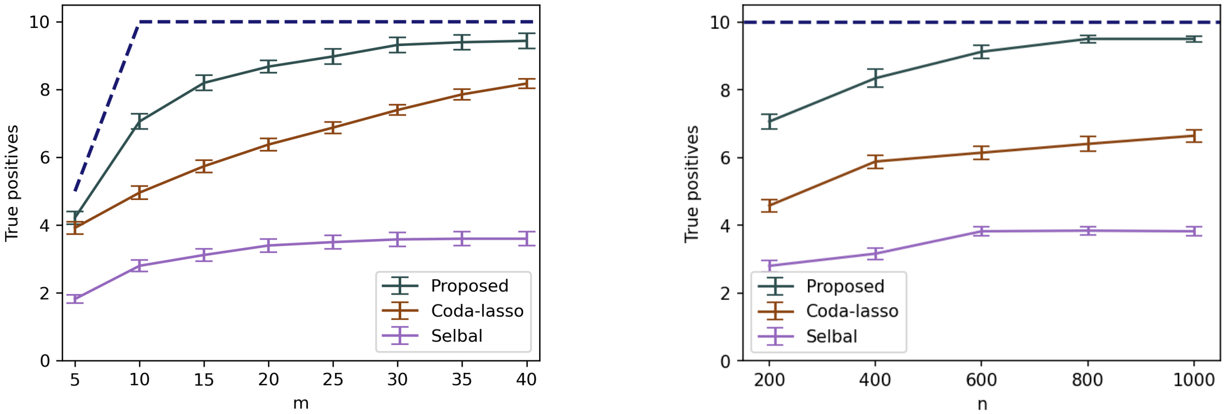

Figure 1. Variable selection results from 50 runs of synthetic data. The y-axis denotes the number of correctly selected features. The maximum number of true variables can be chosen by algorithms is indicated by the top dotted line. The x-axis of the left panel denotes the desired number $m \in \{5, 10, \ldots, 40\}$ 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 ± standard error are shown for each method. Note that the result of coda-lasso is displayed *in its favor*.

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 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 (18), 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 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.

Table 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

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 (Park et al., 2022) are more severe when the zero rates are moderately low. This issue is alleviated slightly if the zero proportion increases, as the model (18) 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.

5.2. BMI Microbiomes Data

We also evaluate our proposed method with the body mass index (BMI) dataset (Wu et al., 2011), which has been repeatedly analyzed with the constrained lasso approaches (Lin et al., 2014; Shi et al., 2016; Wang & Zhao, 2017). The dataset consists of 98 gut microbiome samples with BMI information and organized into 87 genera. For the purpose of comparison in Appendix A, 10 genera that appeared in only one sample are removed. As a result, the data have 77

Table 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)		

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 (Park et al., 2022). 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 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. (2014) and Shi et al. (2016). However, the selected genera from our method fairly differ from coda-lasso. Given the prediction accuracy and results presented in Section 5.1, our result should be considered more reasonable.

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. (2021) 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.

Acknowledgements

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

References

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

Baker, C. R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186: 273–289, 1973.

Brill, B., Amir, A., and Heller, R. Testing for differential abundance in compositional counts data, with application to microbiome studies. *The Annals of Applied Statistics*, 16(4):2648–2671, 2022.

Chen, J., Stern, M., Wainwright, M. J., and Jordan, M. I. Kernel feature selection via conditional covariance minimization. *Advances in Neural Information Processing Systems*, 30, 2017.

Cook, R. D. and Li, B. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

Freidling, T., Poignard, B., Climente-González, H., and Yamada, M. Post-selection inference with hsic-lasso. In *International Conference on Machine Learning*, pp. 3439–3448. PMLR, 2021.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

Fukumizu, K., Bach, F. R., and Gretton, A. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2), 2007.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Kernel dimension reduction in regression. *The Annals of Statistics*, 37 (4):1871–1905, 2009.

Goodman, B. and Gardner, H. The microbiome and cancer. *The Journal of pathology*, 244(5):667–676, 2018.

Greenacre, M. 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.

Greenacre, M., Grunsky, E., and Bacon-Shone, J. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Computers & Geosciences*, 148:104621, 2021.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer, 2005.

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

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., Fitzgerald, M. G., Fulton, R. S., et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Lee, K.-Y., Li, B., and Chiaromonte, F. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

Lee, S., Jung, S., Lourenco, J., Pringle, D., and Ahn, J. Resampling-based inferences for compositional regression with application to beef cattle microbiomes. *Statistical Methods in Medical Research*, 32(1):151–164, 2022.

Li, B. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018.

Li, H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

Li, K.-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414): 316–327, 1991.

Lim, J. N., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., and Shimodaira, H. More powerful selective kernel tests for feature selection. In *International Conference on Artificial Intelligence and Statistics*, pp. 820–830. PMLR, 2020.

Lin, W., Shi, P., Feng, R., and Li, H. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

Lu, J., Shi, P., and Li, H. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, 75(1):235–244, 2019.

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.

Lutz, K. C., Jiang, S., Neugent, M. L., De Nisco, N. J., Zhan, X., and Li, Q. A survey of statistical methods for microbiome data analysis. *Frontiers in Applied Mathematics and Statistics*, 8:884810, 2022.

Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

Masaeli, M., Dy, J. G., and Fung, G. M. From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th international conference on machine learning (ICML)*, pp. 751–758, 2010.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K.-R. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pp. 41–48. IEEE, 1999.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C., Wright, R. J., Dhanani, A. S., Comeau, A. M., et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):1–16, 2022.

Park, J., Yoon, C., Park, C., and Ahn, J. Kernel methods for radial transformed compositional data with many zeros. In *International Conference on Machine Learning*, pp. 17458–17472. PMLR, 2022.

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

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

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.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. Balances: a new perspective for microbiome analysis. *MSystems*, 3 (4):e00053–18, 2018.

Ruan, F., Liu, K., and Jordan, M. I. Taming nonconvexity in kernel feature selection—favorable properties of the laplace kernel. *arXiv preprint arXiv:2106.09387*, 2021.

Shi, P., Zhang, A., and Li, H. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Susin, A., Wang, Y., Lê Cao, K.-A., and Calle, M. L. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2):lqaa029, 2020.

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.

Tomassi, D., Forzani, L., Duarte, S., and Pfeiffer, R. M. Sufficient dimension reduction for compositional data. *Biostatistics*, 22(4):687–705, 2021.

Wang, S. Robust differential abundance test in compositional data. *Biometrika*, 110(1):169–185, 2022.

Wang, T. and Zhao, H. Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11(2):771–791, 2017.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052): 105–108, 2011.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, 26 (1):185–207, 2014.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics*, pp. 152–160. PMLR, 2018.

Yamada, M., Wu, D., Tsai, Y.-H. H., Takeuchi, I., Salakhutdinov, R., and Fukumizu, K. Post selection inference with incomplete maximum mean discrepancy estimator. *In International Conference on Learning Representations*, 2019.

Appendix

A. 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 5 produce different results depending on how the zeros are replaced (Lubbe et al., 2021). 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. (2022)) and the geometric Bayesian multiplicative (gbm) replacement (Martín-Fernández et al., 2015). 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.

A.1. Synthetic data

Table 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 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

A.2. BMI Microbiomes Data

Table 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)		

B. Proof of Results

B.1. Proof of Lemma 3

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

$$k(x, x') = k_{\mathcal{S}}(p_{\mathcal{S}}(x), p_{\mathcal{S}}(x')) = k_{\mathcal{X}}(x_{\mathcal{S}}, x'_{\mathcal{S}}) \tag{19}$$

with the corresponding RKHS $\mathcal{H}$ on $\mathcal{X}$. By pullback theorem of Paulsen & Raghupathi (2016), 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 (19) 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 & Christmann (2008).

B.2. Proof of Theorem 4

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

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

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

$$\mathrm{Var}_{Y|U}[g(Y)|U] = \mathbb{E}_{(U,V)|U}[\mathrm{Var}_{Y|U,V}[g(Y)|U, V]|U] + \mathrm{Var}_{(U,V)|U}[\mathbb{E}_{Y|U,V}[g(Y)|U, V]|U]. \tag{20}$$

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 2. Based on the one-to-one correspondence and the tower law, the first term of the right-hand side of (20) 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 (20) turns into

$$\langle g, (\Sigma_{YY|U} - \Sigma_{YY|X}) g\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_U[\mathrm{Var}_{X|U}[\mathbb{E}_{Y|X}[g(Y)|X]|U]]. \tag{21}$$

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 (21) 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 $Y \perp\!\!\!\perp X \,|\, U$ $(\because \sigma(U) \subseteq \sigma(X))$.

B.3. Proof of Proposition 5

Plugging $g = id_{\mathcal{Y}}$ into the equation (21), we have

$$\langle id_{\mathcal{Y}}, (\Sigma_{YY|X_{\mathcal{S}}} - \Sigma_{YY|X}) id_{\mathcal{Y}} \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{X_{\mathcal{S}}}[\mathrm{Var}_{X|X_{\mathcal{S}}}[\mathbb{E}_{Y|X}[Y|X]|X_{\mathcal{S}}]] = 0, \tag{22}$$

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. (2017) in case of *location regressions*.

B.4. Proof of Corollary 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 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

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

B.5. Proof of Theorem 7

We first state the following *uniform* convergence result:

Proposition 8. *If ϵ_n satisfies the asymptotic behavior given in Theorem 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 7 given Proposition 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 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 Fukumizu et al. (2009); their corresponding result of uniform convergence takes supremum over a continuous domain, not discrete. Proof of such a pointwise convergence can similarly be derived as guided in Appendix A.3 of Chen et al. (2017). Although our notation $X_{\mathcal{S}}$ of variable selection differs from $X_{\mathcal{T}}$ in the reference, the pointwise convergence can parallelly be followed by the law of large numbers.