Communications for Statistical Applications and Methods
2021, Vol. 28, No. 1, 89–97

https://doi.org/10.29220/CSAM.2021.28.1.089
Print ISSN 2287-7843 / Online ISSN 2383-4757

Binary classification on compositional data

Jae Yun Joo[a], Seokho Lee[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract

Due to boundedness and sum constraint, compositional data are often transformed by logratio transformation and their transformed data are put into traditional binary classification or discriminant analysis. However, it may be problematic to directly apply traditional multivariate approaches to the transformed data because class distributions are not Gaussian and Bayes decision boundary are not polynomial on the transformed space. In this study, we propose to use flexible classification approaches to transformed data for compositional data classification. Empirical studies using synthetic and real examples demonstrate that flexible approaches outperform traditional multivariate classification or discriminant analysis.

Keywords: Aitchison geometry, classification, compositional data, Gaussian mixture, isometric logratio transformation

1. Introduction

Compositional data is a special type of multivariate data with strictly positive real components under sum constraint. This kind of data often arises when we measure parts of a whole. A typical example of compositional data is a vector of proportions or percentages, where each component is positive and bounded, and the sum of components is 1 for proportions or 100% for percentages. Compositional data is commonly observed in various applications: ratio of components making up a rock, ratio of fine substances in atmosphere, to name a few (Pawlowsky-Glahn *et al.*, 2015). The sample space of compositional data having D components is the simplex embedded in D dimensional space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D)^T \in \mathbb{R}^D \,\middle|\, \sum_{j=1}^{D} x_j = \kappa,\ x_j > 0,\ j = 1, 2 \dots, D \right\}, \tag{1.1}$$

where κ is the sum constraint. Here we deal with compositional data of proportions, so that we set $\kappa = 1$. Since compositional vectors reside in $\mathcal{S}^D$, their geometry is different from the typical Euclidean geometry in real space. Addition (perturbation) and scalar multiplication (powering) operations, inner product, norm, and distance can be properly defined in the name of *Aitchison geometry* (Pawlowsky-Glahn and Egozcue, 2001).

Although compositional data can be described and understood under well-defined Aitchison geometry, statistical analysis related to compositional data is limited due to lack of statistical model on the simplex space. Dirichlet distribution is the only known statistical distribution that is well defined on simplex. Traditional multivariate data analysis techniques have been developed on Euclidean

This research was supported by Hankuk University of Foreign Studies Research Fund (of 2020).

[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin, Gyeonggi-do 17035, Korea. E-mail: lees@hufs.ac.kr

Published 31 January 2021 / journal homepage: http://csam.or.kr

space, where Gaussian distribution is the norm, so that they cannot be directly applied or generalized to compositional data analysis. Since main difficulty on compositional data comes from the bounded elements ($0 < x_j < \kappa,\ j = 1, 2, \ldots, D$) and the sum constraint ($\sum_{j=1}^{D} x_j = \kappa$), researchers often transform compositional vectors on $\mathcal{S}^D$ into $D - 1$ dimensional unrestricted vectors. Logratio transformations are sensible choices for compositional data because relative information is preserved. Isometric logratio (ilr) transformation is frequently used in compositional data analysis. Since $D - 1$ dimensional vector $\mathbf{z} \in \mathbb{R}^{D-1}$ from ilr transformation on $\mathbf{x} \in \mathbb{R}^D$ is free from constraint and boundedness, traditional multivariate analysis methods can be applied to $\mathbf{z}$ without any technical problem. These results and their interpretation seamlessly go back to the original compositional data because ilr transformation is an isometry (distance-preserving transformation) between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$.

For binary classification on compositional data, a common practice is to apply traditional classification/discriminant analysis methods, such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression to ilr transformed data. LDA and QDA are based on the assumption that class distributions are multivariate Gaussian. LDA and QDA may be inappropriate since the distribution on ilr transformed data from compositional data with Dirichlet distribution is not Gaussian. Logistic regression assumes a linear or polynomial classifier for classification. When two classes come from separate Dirichlet distributions, Bayes decision boundary on ilr transformed space is not of polynomial. Therefore, we expect that a more flexible classification rule is desirable.

In this article, we study the usage of flexible binary classification methods on compositional data. Support vector machine (SVM) is a popular one that produces a flexible decision boundary and does not require any distributional assumption for classific Therefore, if SVM is applied to ilr transformed data and the resulting classifier is reversely transformed back to the original simplex space, then the transformed classification rule outperforms the existing linear or polynomial classification rules in compositional data classification. We also consider Gaussian mixture for class distribution on ilr transformed data, which enables us to obtain posterior class probabilities while SVM does not have any probabilistic interpretation. We demonstrate two types of flexible classification, SVM and Gaussian mixture, are desirable candidates for binary classification on compositional data.

The remaining is organized as follows. We briefly review Aitchison geometry and ilr transformation in Section 2 to help understand the behavior of compositional data on the simplex space. In Section 3, we present why Gaussian-based discriminant analysis (LDA, QDA) and polynomial logistic regression are not appropriate on ilr transformed space although they are frequently used in practice, and explain why more flexible classification methods are appropriate on ilr transformed space to improve compositional data classification. Its empirical evidence is provided in Section 4 under synthetic and real examples. Finally, some concluding remarks are given in Section 5.

2. Brief review on Aitchison geometry and ilr transformation

We consider compositional vectors on $\mathcal{S}^D$ of (1.1) with $\kappa = 1$. The following definitions and properties on Aitchison geometry are well established and studied, for example, in Pawlowsky-Glahn *et al.* (2015).

Definition 1. *For* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *and* $\alpha \in \mathbb{R}$,

1. *(Perturbation)*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1y_1, x_2y_2, \ldots, x_Dy_D) \in \mathcal{S}^D.$$

2. *(Powering)*

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha) \in \mathcal{S}^D$$

with the closure operation $C(\mathbf{x}) = (x_1/\sum_{i=1}^D x_i, x_2/\sum_{i=1}^D x_i, \ldots, x_D/\sum_{i=1}^D x_i)^T$.

The simplex with perturbation and powering, $(\mathcal{S}^D, \otimes, \odot)$, is a vector space. Here, perturbation and powering are analogous to addition (or translation) and scalar multiplication, respectively, in real space. The following theorems show that perturbation and powering serve as basic operations required for a vector space structure of the simplex.

Theorem 1. $(\mathcal{S}^D, \oplus)$ *is a commutative group: for* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^D$*, it satisfies*

1. *(commutative)* $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$.

2. *(associative)* $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$.

3. *(identity)* $\mathbf{n} = C(1, 1, \ldots, 1) = (1/D, 1/D, \ldots, 1/D)$*, which is the unique barycenter of the simplex.*

4. *(inverse)* $\mathbf{x}^{-1} = C(1/x_1, 1/x_2, \ldots, 1/x_D)$*, leading to* $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$.

Often, $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}$ is defined and used for the perturbation difference.

Theorem 2. *For* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *and* $\alpha, \beta \in \mathbb{R}$*, the followings hold.*

1. *(associative)* $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha\beta) \odot \mathbf{x}$.

2. *(distributive 1)* $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$.

3. *(distributive 2)* $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$.

4. *(identity)* $1 \odot \mathbf{x} = \mathbf{x}$.

The following definitions of inner product, norm, and distance provide that $(\mathcal{S}^D, \oplus, \odot)$ is a finite dimensional Hilbert space.

Definition 2. *For* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$,

1. *(Aitchison inner product)*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}.$$

2. *(Aitchison norm)*

$$||\mathbf{x}||_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} = \sqrt{\frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\log \frac{x_i}{x_j} \right)^2}.$$

3. *(Aitchison distance)*

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2}.$$

As a Hilbert space, Cauchy-Schwartz inequality, Pythagoras theorem, and triangular inequality hold on $(\mathcal{S}^D, \oplus, \odot)$ as well. This geometry is called Aitchison geometry.

While compositional vectors behave well on simplex space, their sum constraint and bounded domain cause difficulties when they are put into traditional multivariate data analysis methods that are established in real space. To circumvent those difficulties, researchers introduced logratio transformation from the simplex onto real space without any constraint so that existing multivariate data analysis can be done without any trouble. Aitchison (1986) originally proposed two types of logratio transformations: additive logratio (alr) and centered logratio (clr) transformations. Logratio transformations on compositional data preserve its relative information on the simplex. However, alr transformation fails to preserve distance. (alr transformation is an isomorphism, but not an isometry) Contrast to alr transformation, clr transformation is an isometry from $\mathcal{S}^D$ to $\mathbb{R}^D$. However, clr transformation leads to degenerate distribution due to preserving dimensionality after transformation. Egozcue *et al.* (2003) propose a new logratio transformation, called isometric logratio (ilr) transformation, which is an isometry from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ associated with an orthogonal coordinate system in the simplex. Thus, traditional multivariate analysis applied to ilr transformed data can be seamlessly transmitted to the original simplex space through inverse ilr transformation. For $\mathbf{x} \in \mathcal{S}^D$, ilr transformed vector $\mathbf{z} = \text{ilr}(\mathbf{x}) \in \mathbb{R}^{D-1}$ is defined as:

$$z_j = \frac{1}{\sqrt{(D-j+1)(D-j)}} \left(\log \frac{x_j}{x_{j+1}} + \cdots + \log \frac{x_j}{x_D} \right), \quad j = 1, \ldots, D-1. \tag{2.1}$$

And, its inverse transformation becomes

$$\begin{aligned} x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1 \right), \\ x_j &= \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} z_k + \frac{\sqrt{D-j}}{\sqrt{D-j+1}} z_j \right), \quad j = 2, \ldots, D. \end{aligned} \tag{2.2}$$

Note that $z_j \in \mathbb{R}$ for $j = 1, \ldots, D-1$ and there is no constraint on z_j. And we can easily see that $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$, $\|\mathbf{x}\|_a = \|\mathbf{z}\|$, and $d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{z}_1, \mathbf{z}_2)$, where $\langle \cdot, \cdot \rangle$, $\| \cdot \|$, and $d(\cdot, \cdot)$ are inner product, norm, and distance defined on $D-1$ dimensional real space.

3. Classification on ilr transformed data

In the previous section, we explain that any multivariate data analysis method can be applied to the ilr transformed data for compositional data analysis because ilr transformation provides unconstrained coordinate system and still preserves distance and intrinsic dimensionality as well. However, there is still a concern that the distributions on ilr transformed data may be different from typical distributions that are commonly assumed in multivariate data analysis on real space. For example, multivariate Gaussian is the most popular distribution in multivariate data analysis under empirical reasons and/or theoretical considerations. However, it is not justified well to assume Gaussianity on ilr transformed data. In binary classification for compositional data, two classes are assumed to be distributed under their own class-specific distributions on the simplex space $\mathcal{S}^D$. Consider $\mathbf{x}_i^c \overset{iid}{\sim} f(\mathbf{x}|c)$ $(i = 1, \dots, n_c)$ to represent the observations belonging to the class c $(c = 0, 1)$ on $\mathcal{S}^D$. A common practice is to transform $\mathbf{x}_i^c$ into $\mathbf{z}_i^c$ by ilr transformation and, then, apply discriminant analysis or logistic regression on $\mathbf{z}_i^c$.

Discriminant analysis finds a decision rule that classifies $\mathbf{x}$ or its ilr-transformed $\mathbf{z}$, into $c = 1$ if $\delta(\mathbf{z}) > 1$, and $c = 0$ otherwise. Here, the classifier $\delta(\mathbf{z})$ is defined as

$$\delta(\mathbf{z}) = \frac{P(c = 1|\mathbf{z})}{P(c = 0|\mathbf{z})} \propto \frac{f(\mathbf{z}|c = 1)\pi_1}{f(\mathbf{z}|c = 0)\pi_0},$$

where $f(\mathbf{z}|c)$ are class specific distributions on ilr-transformed space and π_c are prior probabilities of class c. If $f(\mathbf{z}|c)$ is Gaussian, then the resulting decision boundary becomes linear (LDA) or quadratic (QDA) depending on covariance assumption. Therefore, discriminant analysis will perform reasonably well when Gaussianity is valid on ilr transformed space. Logistic regression assumes that $\log P(c = 1|\mathbf{z})/P(c = 0|\mathbf{z})$ is a polynomial of $\mathbf{z}$. If Bayes decision boundary on ilr-transformed space is not of polynomial, then logistic regression does not guarantee its performance.

Suppose that two classes follow separate Dirichlet distributions, which are most popularly assumed for compositional vector. From Dirichlet density function

$$f(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_D)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_D)} x_1^{\alpha_1 - 1} \cdots x_D^{\alpha_D - 1}, \quad 0 < x_j < 1, \ \sum_{j=1}^{D} x_j = 1, \ \alpha_j > 0,$$

ilr transformed vector $\mathbf{z}$ follows the distribution in the below theorem.

Theorem 3. *If $\mathbf{x} \sim$ Dirichlet$(\boldsymbol{\alpha})$ on $\mathcal{S}^D$ with $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^T$, then $\mathbf{z} = ilr(\mathbf{x})$ has the below density function on $\mathbb{R}^{D-1}$.*

$$\begin{aligned} f(\mathbf{z}|\boldsymbol{\alpha}) = & \frac{1}{\sqrt{D}B(\boldsymbol{\alpha})} \exp\left\{ \sum_{j=1}^{D-2} \left(\sqrt{\frac{D-j}{D-j+1}}\, \alpha_j - \frac{\sum_{k=j+1}^{D-1} \alpha_k}{\sqrt{(D-j+1)(D-j)}} \right) z_j + \frac{\alpha_{D-1}}{\sqrt{2}}\, z_{D-1} \right\} \\ & \times \left[1 - \exp\left(\sqrt{\frac{D-1}{D}}\, z_1 \right) - \sum_{j=2}^{D-1} \exp\left(\sqrt{\frac{D-j+1}{D-j}}\, z_j - \sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}}\, z_k \right) \right]^{\alpha_D - 1} \end{aligned}$$

with $B(\boldsymbol{\alpha}) = \{\Gamma(\alpha_1)\cdots\Gamma(\alpha_D)\}/\Gamma(\alpha_1 + \cdots + \alpha_D)$ is the beta function.

We omit the proof here because it can be easily obtained by applying the change-of-variable technique. Since $f(\mathbf{z}|\boldsymbol{\alpha})$ provided in Theorem 3 is far from a Gaussian density, discriminant analysis on ilr-transformed data is not appropriate for compositional data from a two-class Dirichlet population. In addition, with the parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ for two classes, $\log f(\mathbf{z}|\boldsymbol{\alpha}_1)/f(\mathbf{z}|\boldsymbol{\alpha}_0)$ is not of polynomial in $\mathbf{z}$, so that logistic regression is not a good candidate for compositional data classification. Instead of naively applying discriminant analysis or logistic regression, more flexible classification methods can be a better option for compositional data classification. Moreover, we have a limited knowledge on distributions established for the simplex. It may not be desirable to presume a specific one from the limited pool of distributions, such as Dirichlet, for class distribution.

Therefore, we expect that more flexible classification methods on the ilr-transformed data are desirable for compositional data classification. However, based on our limited knowledge, there are seldom literatures that consider flexible classification, such as SVM, for compositional data classification. In Section 4, we conduct comparison study for compositional data classification with several classification approaches. As existing methods, polynomial logistic regression and discriminant analysis under Gaussianity are considered as reference. Comparing to these frequently used methods, we consider SVM and Gaussian mixture discriminant analysis. SVM is a popular flexible binary classification that is free from distributional assumption. Discriminant analysis with Gaussian mixture as a class-specific distribution produces a non-polynomial decision boundary as well, whose functional form is determined flexibly according to the actual class distributions. Both methods aforementioned conduct classification procedure on ilr-transformed space after transformation. In addition, we consider discriminant analysis that is directly applied to the original simplex space under the assumption that class distributions are Dirichlet.

4. Numerical studies

In this section, we provide numerical comparison with polynomial logistic regression (Logistic), Gaussian-based discriminant analysis (LDA, QDA), SVM, Gaussian mixture discriminant analysis (GMDA) on the ilr space and Dirichlet discriminant analysis (DDA) on the simplex. For performance comparison, we use synthetic data generated from class distributions of Dirichlet and Dirichlet mixture. A real-world data example (Hydrochem data) is used for comparison as well, since the set of known distributions on simplex is too limited to represent various real-world compositional data.

4.1. Synthetic data examples

We consider two scenarios for two-class compositional data generation. n_0 and n_1 are class sizes in training data with $n = n_0 + n_1$ and $n_0 = n_1 = n/2$.

- (S1) Dirichlet class distribution

  D-dimensional vectors, $\mathbf{x}_i^c$ ($i = 1, \ldots, n_c; c = 0, 1$), are independently generated from Dirichlet($\boldsymbol{\alpha}_c$). We set $\boldsymbol{\alpha}_0 = (0.4, 0.4, 0.4, 1, \ldots, 1)^T$ and $\boldsymbol{\alpha}_1 = (1.5, 2, 1.5, 1, \ldots, 1)^T$.

- (S2) Dirichlet mixture class distribution

  D-dimensional vectors, $\mathbf{x}_i^c$ ($i = 1, \ldots, n_c;\ c = 0, 1$), are independently generated from Dirichlet mixture distribution, $\sum_{k=1}^{5} \pi_k$ Dirichlet($\boldsymbol{\alpha}_{c,k}$). We set $\boldsymbol{\alpha}_{0,1} = (2, 1, 9, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{0,2} = (1, 5, 10, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{0,3} = (8, 10, 10, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{0,4} = (1, 10, 5, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{0,5} = (2, 9, 1, 1, \ldots, 1)^T$ for the class 0, and $\boldsymbol{\alpha}_{1,1} = (6, 6, 18, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{1,2} = (9, 2, 9, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{1,3} = (9, 3, 3, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{1,4} = (9, 9, 2, 1, \ldots, 1)^T$, $\boldsymbol{\alpha}_{1,5} = (6, 18, 6, 1, \ldots, 1)^T$ for the class 1. And we consider the equal mixing probabilities, i.e., $\pi_k = 1/5$ for $k = 1, \ldots, 5$.

The last $D - 3$ elements in $\boldsymbol{\alpha}_c$ and $\boldsymbol{\alpha}_{c,k}$ are all set to 1, implying that only first 3 variables separates two classes and remaining variables have no discriminative power. After $\mathbf{x}_i^c$ were generated, we transformed it into $\mathbf{z}_i^c = \text{ilr}(\mathbf{x}_i^c)$ with ilr transformation in (2.1). Logistic, LDA, QDA, SVM, and GMDA use the transformed data $\mathbf{z}_i^c$ and DDA uses the original compositional data $\mathbf{x}_i^c$ in classification procedure. In the scenario of (S1), DDA is expected to outperform other candidates because it correctly assumes data generating process. Since each class consists of 5 different Dirichlet distributions in (S2), DDA is not better anymore and flexible classification methods will show better performance. Figure 1 depicts data distributions of $D = 3$ on the simplex and the ilr space. For (S1), Bayes decision boundary becomes nearly circular shape in the ilr-transformed real space as in the upper right panel of Figure 1. However, it takes a quite irregular shape in (S2) so that LDA or QDA seem not relevant

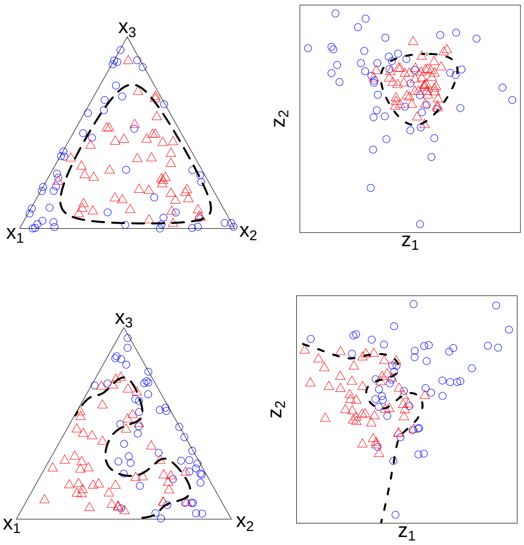

Figure 1: *Two-class compositional datasets of $D = 3$ are generated from the scenario of S1 (upper) and S2 (lower). Left panels depict the original compositional data on the simplex space and, in the right panels their ilr transformed data are displayed on the 2-dimensional real space.*

Table 1: Average of 100 test error rates and its standard deviation (in parenthesis) are presented for the case of $D = 3$

Scenario	n	Methods						
		Bayes	Logistic	LDA	QDA	SVM	GMDA	DDA
S1	100	0.2130 (0.0064)	0.2358 (0.0347)	0.4571 (0.0335)	0.2245 (0.0086)	0.2413 (0.0201)	0.2330 (0.0196)	**0.2184** (0.0085)
	250	0.2135 (0.0065)	0.2224 (0.0079)	0.4609 (0.0276)	0.2215 (0.0065)	0.2284 (0.0125)	0.2245 (0.0092)	**0.2150** (0.0066)
	500	0.2128 (0.0064)	0.2233 (0.0289)	0.4578 (0.0181)	0.2207 (0.0069)	0.2224 (0.0085)	0.2198 (0.0073)	**0.2136** (0.0065)
S2	100	0.1577 (0.0049)	0.2489 (0.0165)	0.3015 (0.0057)	0.2652 (0.0151)	**0.2054** (0.0234)	0.2098 (0.0222)	0.2535 (0.0126)
	250	0.1578 (0.0049)	0.2448 (0.0110)	0.3029 (0.0048)	0.2571 (0.0110)	**0.1781** (0.0110)	0.1813 (0.0109)	0.2458 (0.0075)
	500	0.1580 (0.0045)	0.2446 (0.0118)	0.3021 (0.0053)	0.2562 (0.0084)	**0.1708** (0.0082)	0.1715 (0.0066)	0.2449 (0.0057)

Best performer for each case is highlighted in bold face.

in this cases. To consider the shape of Bayes decision boundary, we fit logistic regression up to 5 degree polynomial models and choose the best among them for logistic regression. To quantify their performance, we additionally generated test data of size 5,000 from the same scenarios. Decision rules learned from the above methods are applied to test data and test error rates are computed. This procedure is repeated 100 times to reduce the effect from randomness in sampling, and their average and standard error are reported.

We summarize the averaged test error rates from the classifiers learned with training datasets of $D = 3$, $n = 100, 250, 500$ in Table 1 and $n = 250$, $D = 3, 6, 12$ in Table 2. Since each class of data is generated from a single Dirichlet distribution under the scenario (S1), DDA is expected to show the

Table 2: Average of 100 test error rates and its standard deviation (in parenthesis) are presented for the case of $n = 250$

Scenario	D	Methods						
		Bayes	Logistic	LDA	QDA	SVM	GMDA	DDA
S1	3	0.2135 (0.0065)	0.2224 (0.0079)	0.4609 (0.0276)	0.2215 (0.0065)	0.2284 (0.0125)	0.2245 (0.0092)	**0.2150** (0.0066)
	6	0.1070 (0.0039)	0.1254 (0.0091)	0.1368 (0.0073)	0.1312 (0.0062)	0.1247 (0.0086)	0.1236 (0.0082)	**0.1087** (0.0043)
	12	0.0842 (0.0039)	0.1080 (0.0110)	0.1178 (0.0090)	0.1186 (0.0072)	0.1030 (0.0088)	0.1112 (0.0084)	**0.0876** (0.0043)
S2	3	0.1578 (0.0049)	0.2448 (0.0110)	0.3029 (0.0048)	0.2571 (0.0110)	**0.1781** (0.0110)	0.1813 (0.0109)	0.2458 (0.0075)
	6	0.1311 (0.0046)	0.2283 (0.0105)	0.2438 (0.0083)	0.2322 (0.0010)	**0.2229** (0.0134)	0.2331 (0.0173)	0.2334 (0.0088)
	12	0.1090 (0.0039)	0.2237 (0.0125)	0.2156 (0.0083)	0.2408 (0.0106)	**0.2071** (0.0109)	0.2329 (0.0229)	0.2133 (0.0103)

Best performer for each case is highlighted in bold face.

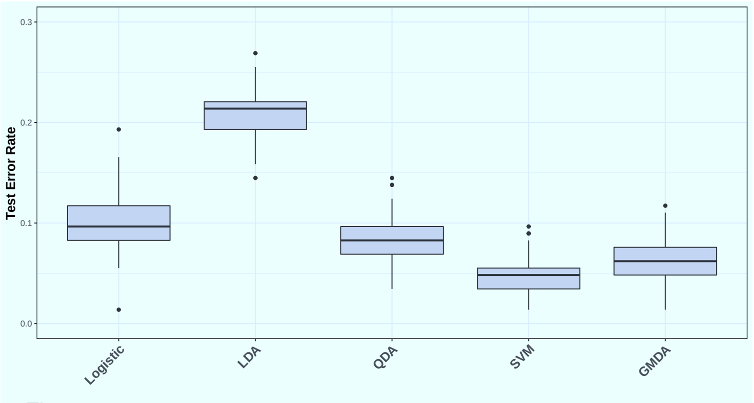

Figure 2: *Hydrochem data: boxplots of 100 test error rates are presented.*

best performance among others and this expectation appears in Table 1. Most classification methods, except LDA, performs reasonably well. This is because Bayes decision boundary on the real space is fitted well under a quadratic (QDA) or polynomial (Logistic) shape. The situation become changed in the scenario (S2), where each class comes from Dirichlet mixture. Flexible classification rules from SVM and GMDA clearly outperform other frequently-used classification methods. This simulation result demonstrates that flexible approach is necessary when the ilr-transformed decision boundary is not guaranteed to be of low-degree polynomial.

4.2. Hydrochem data examples

Next, we apply and compare all classification methods to Hydrochem data (Otero *et al.*, 2005). This data contains measurements of 14 components (H^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Sr^{2+}, Ba^{2+}, NH_4^+, Cl^-, HCO_3^-, NO_3^-, SO_4^{2-}, PO_4^{3-}, TOC) in water samples from the Llobregat river and its two main tributaries, Anoia and Cardener, in northeastern Spain. Total 485 water samples were obtained from 4 distinct water bodies (Anoia, Cardener, lower and upper Llobregat River), whose geological background and human activities in the vicinity vary greatly. For application for binary classification, we select only 2 groups, Anoia (143 samples) and lower Llobregat (135 samples), which are the two largest groups in the data.

We randomly split the 2-class data into training (70%) and test (30%) datasets and, then, applied all classification methods to the training data for learning their classifiers. Test error rate is evaluated by applying the fitted classifiers to the test dataset. This procedure is heavily influenced by random train/test splitting. Therefore, this procedure is repeated 100 times to prevent a generalization from a particular accidental splitting. Figure 2 presents boxplots of 100 test error rates from classifiers that are fitted by classification methods we consider. LDA is the worst performer, which indicates that a linear classifier is not appropriate for this classification problem. SVM turns out the best and GMDA also performs reasonably well comparing to Logistic and QDA. This real example, where its actual data generating process on the simplex space is unknown, indicates that flexible approaches, like SVM or GMDA, are competitive over the commonly used practices, such as LDA, QDA, and logistic regression.

5. Conclusion and remarks

In this work, we provide empirical evidence that flexible classification approaches, SVM and Gaussian mixture discriminant analysis, are promising options for compositional data classification, rather than traditional multivariate approaches which are commonly and currently used in practice. We use isometric logratio transformation for this application because ilr transformation is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$. Instead of ilr transformation, one may use a different logratio transformation, for example clr transformation. While clr is also an isometry between $\mathcal{S}^D$ and $\mathbb{R}^D$, a class distribution degenerates because the intrinsic dimensionality is $D - 1$. (for example, a sample covariance on $\mathbb{R}^D$ becomes singular.) A discriminant analysis (GMDA) is, therefore, not available for clr transformation, but SVM is still applicable. Unlike ilr transformation, the clr transformed z_j is obtained by logarithm of scaled x_j so that variable importance in the statistical analysis is preserved before/after transformation. Thus, variable selection or variable importance evaluation in classification can be implemented and evaluated under clr transformation. We leave this research direction as a future work.

References

Aitchison J (1986). *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

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

Otero N, Tolosana-Delgado R, Soler A, Pawlowsky-Glahn V, and Canals A (2005). Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river. *Water Research*, **39**, 1404–1414.

Pawlowsky-Glahn V and Egozcue JJ (2001). Geometric approach to statistical analysis on the simplex, *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15**, 384–398.

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

Received October 12, 2020; Revised November 27, 2020; Accepted December 22, 2020