Statistics and Its Interface Volume 12 (2019) 275–282

Two-sample test for compositional data with ball divergence

Jin Zhu, Kunsheng Lv, Aijun Zhang, Wenliang Pan*, and Xueqin Wang*

In this paper, we try to analyze whether the intestinal microbiota structures between gout patients and healthy individuals are different. The intestinal microbiota structures are usually measured by so-called compositional data, composed of multiple components whose value are typically non-negative and sum up to a constant. They are frequently collected and studied in many areas such as petrology, biology, and medicine nowadays. The difficulties to do statistical inference with compositional data arise from not only the constant restriction on the component sum, but also high dimensionality of the components with possible many zero measurements, which are frequently appeared in the 16S rRNA gene sequences. To overcome these difficulties, we first define the Bhattacharyya distance between two compositions such that the set of compositions is isometrically embedded in some spherical surfaces. And then we propose a two-sample test statistic for compositional data by Ball Divergence, a novel but powerful measure for the discrepancy between two probability measures in separable Banach spaces. Our test procedure demonstrates its excellent performance in Monte Carlo simulation studies even when the simulated data consist of thousand components with a high proportion of zero measurements. We also find that our method can distinguish two intestinal microbiota structures between gout patients and healthy individuals while the existing method does not.

AMS 2000 subject classifications: Primary 62P10; secondary 62G10.
Keywords and phrases: Ball divergence, Two-sample test, Compositional data.

1. INTRODUCTION

Gout is a common inflammatory disease directly caused by persistently elevating level of uric acid in the blood. The sudden attack of gout makes the patients suffer from swelling, extreme tenderness of the joint as well as alternating chills and fever. Gout is originally considered as a disease of older men, however, due to the increasing average age and intake of high protein food, women may also suffer from gout. Consequently, a lot of gout research which focuses on the female is conducted to obtain the insightful viewpoints [1, 8, 19]. Recently, Guo et al. proposed a novel and sensitive model to evaluate the risk of gout relying on the relative abundance of several bacterial markers among the intestinal microbiota [23]. Their proposed approach is superior to the traditional uric acid level based diagnostic method. Furthermore, their innovative work inspires scientists to explore more about gout via the intestinal microbiota. For example, one of the important questions is whether the intestinal microbiota structure of the health and gout are different.

To answer the mentioned question, in other words, to detect the distinction between two intestinal microbiota structures is a fundamental problem in statistics, which can be formalized as follows:

$$H_0 : \mu = \nu,$$

where μ, ν are two probability measures. Many classical methods have been developed to address this issue, including two-sample t-test, Wilcoxon test, Kolmogorov-Smirnov test for univariate random variables, and also Hotelling-T^2, multivariate version of the Kolmogorov-Smirnov test [4] and Generalized Cramér-von Mises [15] for multivariate random variables. However, it is worth noting that the intestinal microbiota data have two characteristics. First, the size of operational taxonomic units (OTUs) for each subject D is generally much larger than the sample size, such that the regularity conditions are generally not met [20]. Second, each subject with D OTUs satisfied that

$$x_d \geq 0, d = 1, \ldots, D \text{ and } \sum_{d=1}^{D} x_d = c,$$

where x_d is the count of the d-th OTUs after 16S rDNA pyrosequencing and c is the total number of OTUs which may vary for different subjects [23]. Since the component proportions are our primary concern, we should treat $(x_1, \ldots, x_D)$ as compositions in the context of compositional data analysis [13]. If we make a transformation for each subject so that

$$\sum_{d=1}^{D} y_d = 1, \ y_d = x_d/c,$$

*Corresponding author.

then $(y_1, \ldots, y_D)$ is the common D-part composition defined by Aitchison [7]. Nevertheless, due to the principle of subcompositional coherence [7], traditional multivariate test statistics can not be directly applied to this data because none of them obey the subcompositional coherence [13].

To tackle two-sample test problem for compositional data, Aitchison and Krzanowski [7, 16] first developed a classical parametric approach based on the likelihood ratio test theory, which assumes the compositions of data transformed by the log ratio function to be accord with multivariate Gaussian distribution. Henceforth, we denote this classical method as LR. Due to the difficulty of covariance estimation for high dimensional compositional data [21], the classical parametric approach is restricted. Later, Energy Distance (ED) [5], an influential evolution which was originally devoted to multivariate two-sample test problem, boosts the progress of distribution-free two-sample test method for high dimensional compositional data [13]. By calculating the Euclidean distance of isometric log-ratio transformed sample, ED is capable of separating two distinct compositional data with high dimensionality and is shown to be zero if and only if two distributions are identical [5]. Unfortunately, since numerous OTUs counts are zero, the log ratio transformation cannot be defined [18], and thus, ED based approach may not be applicable to the intestinal microbiota data directly. Alternatively, the Bhattacharyya distance, one of the commonest metrics for the compositional data [10, 11, 12], is still well-defined in the case of the large percentage of zero measurements. Specifically, the Bhattacharyya distance is equivalent to measuring the great circle distance between two sqrt-transformed samples whose component proportions sum up to one. However, the great circle distance defined in the unit sphere space is not of strong negative type, for which ED may not perform well or even totally lose its capacity [14].

A novel concept, Ball Divergence (BD), is recently developed to measure the difference between two probability measures in separable Banach spaces [17]. Taking non-negative value, BD possesses the property that it is equal to zero if and only if two probability measures are identical like ED. But BD is not restricted by the strong negative type, and thus, the Bhattacharyya distance or other distance (such as Angular distance or Aitchison distance [11]) for compositional data are all suitable for BD. With BD, a permutation based metric rank test is supplied to check the equality of distribution measures assumption. It is proven to possess robustness, consistency and well coping with imbalanced data [17]. The motivation of BD relies on the fact that two identical probability measures agree on all the balls in separable Banach spaces [2]. To construct an empirical BD statistic given two independent samples, any two points of each sample are utilized to draw a ball, and the differences of the proportions of two samples located in the ball are summed.

The remaining sections are organized in the following way. We will first revise BD and introduce the BD based two-sample test procedure for compositional data in Section 2. In Section 3, we carry out the Monte Carlo simulation studies to analyze the performances of LR, ED, and BD in the context of low-dimensional, high-dimensional, and high-proportional zero measurements scenarios. In Section 4, we employ BD test procedure on the intestinal microbiota data to analyze whether there exists a difference between the intestinal microbiota structures of the health and gout. Finally, Section 5 will give an overall summary of BD test procedure and provide some directions about future improvement.

2. METHODS

In this section, we will recap the definition of Ball Divergence in subsection 2.1, and propose a two-sample test procedure for compositional data with BD in subsection 2.2.

2.1 Ball divergence

Suppose $(V, \|\cdot\|)$ be a finite dimensional Banach space, where the norm $\|\cdot\|$ induces a metric $\rho(x, y) = \|x - y\|$ for two points $x, y \in V$. Let $\mathcal{B}$ be the Borel σ-algebra in V. μ, ν are two Borel probability measures defined on $\mathcal{B}$. Let B_r^u be a closed ball with the center u and the radius r. The Ball Divergence is defined as:

$$BD(\mu, \nu) = \iint_{V \times V} [\mu - \nu]^2 (B^x_{\rho(x,y)}) \times (\mu(dx)\mu(dy) + \nu(dx)\nu(dy)).$$

The crucial property of $BD(\mu, \nu)$, the homogeneity-zero property, is that $BD(\mu, \nu) = 0$ if and only if $\mu = \nu$.

Next, we introduce the sample version of BD which serves as the test statistic in the two-sample problem. For convenience, we define several notations. Let $\delta(x, y, z) = I(z \in B^x_{\rho(x,y)})$, where I is an indicator function, and hence, $\delta(x, y, z)$ takes value 1 if z is located in the closed ball $B^x_{\rho(x,y)}$, and 0 otherwise. Given two independent random samples $\mathcal{X}_{(n)} = \{X_1, \ldots, X_n\}$ associated with probability measures μ and $\mathcal{Y}_{(m)} = \{Y_1, \ldots, Y_m\}$ associated with ν, we define

$$P^{\mu\mu}_{ij} = \frac{1}{n}\sum_{u=1}^{n} \delta(X_i, X_j, X_u),\ P^{\mu\nu}_{ij} = \frac{1}{m}\sum_{v=1}^{m} \delta(X_i, X_j, Y_v),$$
$$P^{\nu\mu}_{kl} = \frac{1}{n}\sum_{u=1}^{n} \delta(Y_k, Y_l, X_u),\ P^{\nu\nu}_{kl} = \frac{1}{m}\sum_{v=1}^{m} \delta(Y_k, Y_l, Y_v).$$

From the definition, $P^{\mu\mu}_{ij}$ represents the proportion of sample $\mathcal{X}_{(n)}$ located in the closed ball $B^{X_i}_{\rho(X_i,X_j)}$ and $P^{\mu\nu}_{ij}$ calculates that of $\mathcal{Y}_{(m)}$. $P^{\nu\mu}_{kl}$ and $P^{\nu\nu}_{kl}$ have similar meaning with $P^{\mu\mu}_{ij}$ and $P^{\mu\nu}_{ij}$ except the closed ball is changed to $B^{Y_k}_{\rho(Y_k,Y_l)}$. The sample version of BD, $BD_{n,m}$, is a positive value defined by

$$BD_{n,m} = \frac{1}{n^2}\sum_{i,j=1}^{n} (P^{\mu\mu}_{ij} - P^{\mu\nu}_{ij})^2 + \frac{1}{m^2}\sum_{k,l=1}^{m} (P^{\nu\mu}_{kl} - P^{\nu\nu}_{kl})^2.$$

According to the definition of $BD_{n,m}$, the critical step of empirical BD computation is calculating the distance matrix $D^{XX} \in R^{n\times n}, D^{YY} \in R^{m\times m}$ as well as $D^{XY} \in R^{n\times m}$, where D^{XX}_{ij} measures the distance between X_i and X_j, D^{YY}_{kl} measures Y_k and Y_l, and D^{XY}_{ik} measures X_i and Y_k.

The sample version of BD possesses many desirable properties. For instance, it converges to $BD(\mu, \nu)$ and is consistent against any general alternative under mild assumptions [17].

2.2 Ball divergence based two-sample compositional data test

Suppose we collect two compositional data $\mathcal{X}_{(n)} = \{X_1, \ldots, X_n\}, \mathcal{Y}_{(m)} = \{Y_1, \ldots, Y_m\}$ and combine them to $\mathcal{Z}_{(n+m)} = \mathcal{X}_{(n)} \bigcup \mathcal{Y}_{(m)}$, for each subject $Z_t \in \mathcal{Z}_{(n+m)}$, Z_t consists of D component proportions z_j $(j = 1, \ldots, D)$, and satisfies that $z_j \geq 0$ $(j = 1, \ldots, D)$ and $\sum_{j=1}^{D} z_j = 1$. As we mentioned in subsection 2.1, the computation of $BD_{n,m}$ involves pairwise distance matrix, and hence, selecting a reasonable metric for compositional data is quite important.

Several metrics work for the compositional data, including the Aitchison distance, Mahalanobis distance, Bhattacharyya distance, etc. [11]. In terms of the gut microbiota data, the Bhattacharyya distance seems to be a more reasonable choice because the existence of zero counts makes the log-ratio transformation based distances such as Aitchison distance and Mahalanobis distance unavailable. To be specific, the Bhattacharyya distance between two compositions Z_1 and Z_2 of D components is defined as follows:

$$\text{Bhattacharyya}(Z_1, Z_2) = f\left(\sum_{j=1}^{D} \sqrt{z_{1j}}\sqrt{z_{2j}}\right),$$

where $z_{ij}, i = 1, 2, j = 1, \ldots, D$ are the component proportions of Z_1, Z_2, and f is the arc-cosine or negative logarithm function. In most cases, the arc-cosine function is a preferable choice because it is not only immune to the problem caused by zero proportion but also deserving of geometric interpretability. The Bhattacharyya distance with the arc-cosine function can be interpreted as the great-circle/geodesic distance between two points $\sqrt{Z_1}, \sqrt{Z_2}$ on the surface of a sphere.

In order to obtain the statistical inference result of BD based two-sample test, we use the non-parametric permutation technique to estimate the empirical null distribution and further derive the p-value.

3. SIMULATIONS

In this section, we conduct the Monte Carlo simulation studies for the Ball Divergence based two-sample test procedure for compositional data. We are interested in investigating the performance of BD test procedure when sample size n is larger or smaller than component number D. For comparison, the classical likelihood ratio (LR) test statistic based test procedure, as well as Energy distance (ED) based test procedure with the Bhattacharyya distance are also taken into consideration. The p-value of LR and ED based tests are derived by the permutation technique.

To create the D-part composition, we first generate a $D-1$ dimensional random value $(z_1, \ldots, z_{D-1})$ coming from commonly used distributions in Euclidean spaces. Next, following [6] and [3], we apply the additive logistic transform [13] to $(z_1, \ldots, z_{D-1})$ to obtain the compositional data. The additive logistic transform is defined as:

$$\phi : (z_1, \ldots, z_{D-1}) \to (1, e^{z_1}, \ldots, e^{z_{D-1}})/S,$$

where

$$S = 1 + \sum_{d=1}^{D-1} e^{z_d}.$$

For simplicity, we let $z_1, \ldots, z_{D-1}$ be $i.i.d$ univariate random variables. As a consequence, we can only describe the simulation settings of the common univariate distribution in Euclidean spaces, including normal, t, uniform, Cauchy as well as beta distributions. Because the location difference between two groups is almost canceled out especially when D is large, we mainly focus on the scale difference of distribution.

To describe these models, we give some notations first. We denote μ_1, μ_2 and σ_1, σ_2 as the location and scale parameters of normal distribution, df_1, df_2 as the degree of freedoms of t distribution, a_1, a_2 and b_1, b_2 as the minimum and maximum value of uniform distribution, α_1, α_2 and β_1, β_2 as two shape parameters of beta distribution, μ_1, μ_2 and γ_1, γ_2 as the location and scale parameters of Cauchy distribution. With respect to the sample size, setting two group with the same sample size, we let the sample size of each group n increase as $15, 50, 75, 150$. Fixing the significance level at 0.05, we replicate each model 1000 times to estimate the Type-I error and power.

We first pay attention to the case when $n > D$ and $D = 10$. To evaluate Type-I error, we consider the four models below.

- **Model 1**: The normal distribution. The location parameters are $\mu_1 = \mu_2 = 0$ and the scale parameters are $\sigma_1 = \sigma_2 = 1$.
- **Model 2**: The t distribution. The degree of freedoms are $df_1 = df_2 = 3$
- **Model 3**: The uniform distribution. The minimum value parameters are $a_1 = a_2 = 0$ and the maximum value parameters are $b_1 = b_2 = 1$.
- **Model 4**: The beta distribution. The shape parameters are $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 2$.

For the power evaluation, we consider the following four models.

Table 1. Type-I Error rates of the tests based on LR, ED and BD for Model 1 to 4.

	Model 1				Model 2			
size	15	50	75	150	15	50	75	150
LR	0.046	0.050	0.056	0.045	0.040	0.048	0.035	0.041
ED	0.041	0.046	0.036	0.052	0.051	0.049	0.034	0.041
BD	0.043	0.032	0.035	0.053	0.051	0.044	0.041	0.041
	Model 3				Model 4			
size	15	50	75	150	15	50	75	150
LR	0.055	0.040	0.043	0.047	0.053	0.041	0.042	0.047
ED	0.043	0.046	0.043	0.057	0.054	0.052	0.047	0.041
BD	0.041	0.045	0.035	0.054	0.048	0.048	0.038	0.052

Table 2. Empirical powers of the tests based on LR, ED and BD for Model 5 to 8.

	Model 5				Model 6			
size	15	50	75	150	15	50	75	150
LR	0.044	0.086	0.138	0.363	0.065	0.223	0.415	0.842
ED	0.047	0.063	0.071	0.157	0.077	0.133	0.177	0.463
BD	0.113	0.371	0.551	0.858	0.139	0.401	0.533	0.857
	Model 7				Model 8			
size	15	50	75	150	15	50	75	150
LR	0.057	0.085	0.129	0.364	0.069	0.692	0.955	1.000
ED	0.050	0.069	0.076	0.144	0.115	0.547	0.826	1.000
BD	0.138	0.623	0.832	0.995	0.722	0.998	1.000	1.000

- **Model 5**: The normal distribution. The location parameters are $\mu_1 = \mu_2 = 0$ and the scale parameters are $\sigma_1 = 0.8, \sigma_2 = 1$.
- **Model 6**: The truncated t distribution. The degree of freedoms are $df_1 = 2, df_2 = 3$, truncating at -10 and 10.
- **Model 7**: The uniform distribution. The minimum value parameters are $a_1 = 0, a_2 = 0.05$ and the maximum value parameters are $b_1 = 1, b_2 = 0.95$.
- **Model 8**: The beta distribution. The shape parameters are $\alpha_1 = 1, \alpha_2 = 2, \beta_1 = 3, \beta_2 = 6$.

The Type-I error rates of Model 1 to 4 and empirical powers of Model 5 to 8 are demonstrated in Table 1 and 2 respectively. Results of Table 1 show that LR, BD, and ED can control the Type-I errors well around the significance level. The Type-I error of LR is foreseeably reasonable since we utilize permutation technique to derive the null distribution. As regard to the empirical powers under different model settings, BD based two-sample compositional data test achieves satisfactory performance and is superior to test based on LR and ED while LR outperforms ED. It is also worth noting that BD retains the power in detecting the scale difference of the underlying distribution [17] though we have used the additive logistic transformation to obtain the simulated compositional data.

We shift our attention to another scenario that $n < D$ and $D = 3000$. Aside from increasing the component number, we truncate the tiny component proportions to zero when it is smaller than a threshold ϵ to imitate the gut microbiota data which contains multiple zero count OTUs. In the following models, we fix $\epsilon = 10^{-5}$. Since the estimation of the covariance matrix is infeasible in this situation, LR test procedure is not considered.

To evaluate Type-I error, we consider the four models below.

- **Model 9**: The normal distribution. The location parameters are $\mu_1 = \mu_2 = 0$ and the scale parameters are $\sigma_1 = \sigma_2 = 1$.
- **Model 10**: The t distribution. The degree of freedoms are $df_1 = df_2 = 10$.
- **Model 11**: The uniform distribution. The minimum value parameters are $a_1 = a_2 = 0$ and the maximum value parameters are $b_1 = b_2 = 1$.
- **Model 12**: The beta distribution. The shape parameters are $\alpha_1 = \alpha_2 = 8,\ \beta_1 = \beta_2 = 1$.

For the power evaluation, we consider the following four models. Among the 3000 components, we pick out 300 components randomly and let their distribution be distinct to another sample owing to the biological common sense that only a small part of OTUs exists significant difference.

- **Model 13**: The normal distribution. The 300 components of one sample have location parameter $\mu_2 = 0$ and scale parameter $\sigma_2 = 0.8$. The other components of two sample have location parameter $\mu_1 = 0$ and the scale parameter $\sigma_1 = 1$.
- **Model 14**: The truncated t distribution. The 300 components of one sample have degree of freedom $df_2 = 2$ while the other components of two sample have degree of freedom $df_1 = 3$. Each component is truncated at -10 and 10.
- **Model 15**: The uniform distribution. The 300 components of one sample have minimum value $a_2 = 0.05$ and maximum value $b_2 = 0.95$. The other components of two sample have minimum value $a_1 = 0$ and maximum value $b_1 = 1$.
- **Model 16**: The beta distribution. The 300 components of one sample have shape parameters $\alpha_2 = 1, \beta_2 = 3$. The other components of two sample have shape parameters $\alpha_1 = 2, \beta_1 = 6$.

We demonstrate the Type-I error rates and empirical powers of simulation studies of BD and ED test in Table 3 and 4, respectively. As we can see in Table 3, two methods can control the Type-I errors well around the significance level. With respect to the empirical powers, BD based two-sample test for compositional data shows satisfying performance and outperforms ED based test. Notice that the performance of ED based test in Model 15 is dissatisfactory. Compared with the other Models, in Model 15, the composition proportions are more averaged due to the random value before the additive logistic transformation uniformly ranges from 0 to 1, leading to the tiny difference composition proportions between two groups. In this case, BD is still able to

Table 3. Type-I Error rates of test procedures based on ED and BD for Model 9 to 12.

	Model 9				Model 10			
size	15	50	75	150	15	50	75	150
ED	0.057	0.047	0.046	0.042	0.056	0.050	0.041	0.056
BD	0.040	0.046	0.056	0.046	0.051	0.056	0.052	0.048
	Model 11				Model 12			
size	15	50	75	150	15	50	75	150
ED	0.051	0.041	0.047	0.045	0.048	0.047	0.038	0.052
BD	0.047	0.056	0.051	0.064	0.043	0.062	0.048	0.054

Table 4. Empirical powers of two test procedures based on ED and BD for Model 13 to 16.

	Model 13				Model 14			
size	15	50	75	150	15	50	75	150
ED	0.069	0.155	0.260	0.678	0.154	0.643	0.893	1.000
BD	0.628	0.993	0.999	1.000	0.308	0.779	0.936	0.998
	Model 15				Model 16			
size	15	50	75	150	15	50	75	150
ED	0.051	0.037	0.052	0.058	0.075	0.190	0.364	0.815
BD	0.631	0.999	1.000	1.000	1.000	1.000	1.000	1.000

discover the subtle distinction. Furthermore, compared with Model 5 to 8, the empirical powers of BD in Model 13 to 16 do not drop but improve. The reason for this improvement is that BD not only overcomes the barriers arisen from high dimensionality and zero measurements but also detects the more significant distribution distinctions resulted from the more distinct components.

To further evaluate the performance of the proposed test in the context of the zero measurements, we carried out simulations studies to examine the affection of the proportion of zero measurements. To control the percentage of zero measurements, for each observation, we randomly pick out a part of components then make their proportions become zero and conduct normalization procedure such that the proportions of all component sum up to 1. In this way, we can control the percentage of zero measurements varying from 30% to 90%.

In the following eight models, the sample sizes of two group is fixed at 15 and the number of compositions are fixed at 3000. For Type-I error evaluation, we consider the four models as follow.

- **Model 17**: The normal distribution. The location parameters are $\mu_1 = \mu_2 = 0$ and the scale parameters are $\sigma_1 = \sigma_2 = 1$.
- **Model 18**: The t distribution. The degree of freedoms are $df_1 = df_2 = 2$.
- **Model 19**: The uniform distribution. The minimum value parameters are $a_1 = a_2 = 0$ and the maximum value parameters are $b_1 = b_2 = 1$.
- **Model 20**: The Cauchy distribution. The location parameters are $\mu_1 = \mu_2 = 0$ and the scale parameters are $\gamma_1 = \gamma_2 = 0.1$.

Table 5. Type-I Error rates of two test procedures based on ED and BD for Model 17 to 20.

	Model 17				Model 18			
percentage (%)	30	50	70	90	30	50	70	90
ED	0.054	0.054	0.049	0.050	0.054	0.055	0.045	0.044
BD	0.038	0.051	0.051	0.048	0.052	0.041	0.045	0.058
	Model 19				Model 20			
percentage (%)	30	50	70	90	30	50	70	90
ED	0.064	0.052	0.056	0.052	0.055	0.039	0.045	0.049
BD	0.048	0.053	0.039	0.037	0.039	0.039	0.038	0.046

Table 6. Empirical powers of two test procedures based on ED and BD for Model 21 to 24.

	Model 21				Model 22			
percentage (%)	30	50	70	90	30	50	70	90
ED	1.000	1.000	0.883	0.172	1.000	1.000	0.989	0.277
BD	1.000	1.000	1.000	0.905	1.000	1.000	1.000	0.770
	Model 23				Model 24			
percentage (%)	30	50	70	90	30	50	70	90
ED	1.000	1.000	0.999	0.371	1.000	0.999	0.910	0.531
BD	1.000	1.000	0.999	0.508	1.000	1.000	0.962	0.680

For the power evaluation, we consider the four models in the following. Like Model 13 to 16, we let only 300 components be distinct to other components.

- **Model 21**: The normal distribution. The 300 components of one sample have location parameter $\mu_2 = 0$ and scale parameter $\sigma_2 = \sqrt{5}$. The other components of two sample have location parameter $\mu_1 = 0$ and the scale parameter $\sigma_1 = 1$.
- **Model 22**: The truncated t distribution. The 300 components of one sample have the degree of freedom $df_2 = 2$. The other components of two sample have the degree of freedom $df_2 = 10$. Each component is truncated at -10 and 10.
- **Model 23**: The uniform distribution. The 300 components of one sample have minimum value $a_2 = 0$ and maximum value $b_2 = 4$. The other components of two sample have minimum value $a_1 = 0$ and maximum value $b_1 = 6$.
- **Model 24**: The truncated Cauchy distribution. The 300 components of one sample have location parameter $\mu_2 = 0$ and scale parameter $\gamma_2 = 2$. The other components of two sample have location parameter $\mu_1 = 0$ and the scale parameter $\gamma_1 = 0.1$. Each component is truncated at -10 and 10.

The Type-I error rates and empirical powers of Model 17 to 24 are displayed in Table 5 and 6. As we can see in Table 5, both ED and BD can well control Type-I error rates around 0.05 regardless of highly proportional zero measurements. From Table 6, the increasing percentage of zero measurements have a negative impact on the empirical powers of two methods. However, if two distributions are

conspicuously different such as Model 21 and 22, Ball Divergence can temporarily maintain its performance and be very likely to reveal the heterogeneity of two datasets even though the percentage of zero measurements go up to 90% while ED may not.

4. REAL DATA ANALYSIS

In this section, we employ ED and BD based two sample test procedure for compositional data on the gut microbiota data to figure out whether the intestinal microbiota structure is different between health and gouty women. To solve the problem, we re-analyze a public dataset provided by Guo et al. [23] which is available on the Nature website (https://www.nature.com/articles/srep20602). The dataset consists of forty-nine males (24 gouty, 25 healthy) and thirty-four females (17 gouty, 17 healthy). For each subject, twenty-two body index, including age, BMI (body mass index), uric acid, total protein and so on, and the high-quality 16S rRNA gene sequence is collected. Bioinformatics pre-process procedure is applied to 16S rRNA gene sequences, leaving 3684 Operational Taxonomic Units (OTUs) for further analysis. We pick out the OTUs data of healthy and gouty female individuals and discard redundant 1299 OTUs whose value are zeros for all subjects to analyze their intestinal microbiota structure.

Denote the gut microbiota dataset of females as $\{(x_{i1}, \ldots, x_{i2385}), i = 1, \ldots, 34\}$ where x_{id} is the count of d-th OTU for the i-th subject. To apply our method, we first transform the data into a typical expression form of compositional data:

$$y_{id} = x_{id} / \sum_{d=1}^{2385} x_{id},\ d = 1, \ldots, 2385,\ i = 1, \ldots, 34$$

such that

$$y_{id} \in [0, 1], d = 1, \ldots, 2385, i = 1, \ldots, 34$$

and

$$\sum_{d=1}^{2385} y_{id} = 1.$$

Given the component proportions $\{(y_{i1}, \ldots, y_{i2385}),\ i = 1, \ldots, 34\}$, we perform the two-sample test procedure for compositional data based on ED and BD. The classical LR method is not considered because of the dimensional limitation. The p-value of two methods are presented in Table 7. Noticed that ED fails to detect the difference between healthy people and gouty patients, whereas BD is able to reveal significant distinctions between two groups. The reasonability of the result is supported by the result of Guo et al. [23].

Table 7. The p-value of two test procedures for the gut microbiota data

Method	ED	BD
p-value	0.120	0.020

To further confirm the result, we pick out 17 OTUs which show a significant difference between the dataset of gout patients and healthy individuals (p-value < 0.005, Wilcoxon-test), and later, visualize the centered and scaled component proportions of influential OTUs for each individual in Figure 1 on page 281. From Figure 1, compared with gout patients, the 17 influential OTUs tend to have larger proportion among healthy female. Therefore, the visualization result also supports our finding.

5. CONCLUSION AND DISCUSSIONS

The advanced modern techniques help scientists collect 16S rRNA gene sequences which can be utilized to study bacterial phylogeny, taxonomy [9], and further the organismal and functional structures of the intestinal microbiota. We are interested in investigating the organismal structural difference between gout and health female, a typical high dimensional two-sample compositional data problem. However, the high dimensionality, numerous zero counts, and compositions nature of 16S rRNA data makes the statistical inference be a tough issue. On the basis of Ball Divergence which is capable of distinguishing two probability distributions in the separable Banach spaces, we propose a Ball Divergence based two sample compositional data test procedure to resolve this problem. Thanks to the non-parametric metric rank statistic essence of Ball Divergence, we avoid estimating the covariance matrix in the context of high dimensionality. Moreover, by choosing a sensible distance, the Bhattacharyya distance, to measure the dissimilarities between two compositions, we successfully remove the obstacle resulted from hundreds of zero count OTUs. The simulation studies show that our proposed test procedure is superior to both ED and LR in the low-dimensional case and overcomes ED in the high-dimensional and many zero counts situation in which LR method is unavailable. Finally, our proposed test procedure discovers the significant difference of the gut microbiota between gouty and healthy women, and we can conclude the organismal structural difference really exists.

The BD based test procedure has supplied a solution for our question, and further, we can improve the performance of BD in the following aspects. First, since the perfect dissimilarity between two compositions is usually unknown in practice, the other metrics such as the Aitchison distance can substitute the Bhattacharyya distance in the BD based testing framework if the compositional data do not contain the zero value and the substitution might bring better performance. Second, from the definition of BD, BD treats the squared proportion difference of each closed ball

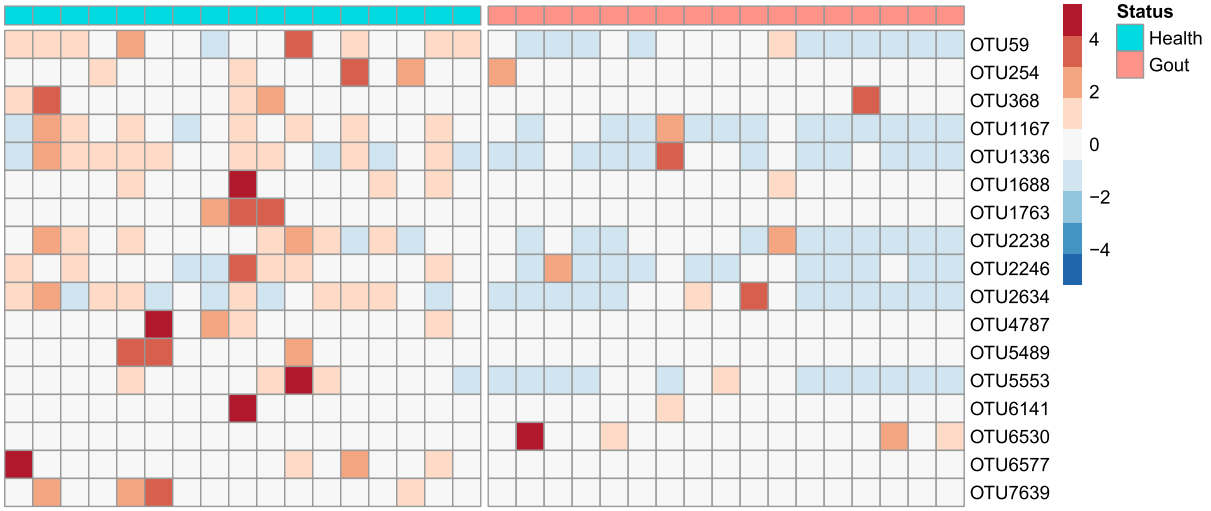

Figure 1. A heat map displays the centered and scaled component proportions of 17 influential OTUs for each individual. The left panel is the heat map for the healthy persons and the right for the gouty patients. For the heat map, each row corresponds to an OTU and each column corresponds to an individual, and more warm-toned grid means the corresponding OTU with a larger percentage in the individual. The different color distributions of two heat map indicate the intestinal microbiota structure of health and gout are distinct.

equivalently. Borrowing ideas of AdaBoost [22], it may be worthwhile to extend BD to a weighted version such that BD could pay adequate attention to those more meaningful closed balls. Third, since the high percentage of zero counts does harm to BD and ED test procedure, a potential approach may replace the zero measurements with a reasonably estimated value so as to relieve the setback.

ACKNOWLEDGMENT

Dr. Pan's research is partially supported by the National Natural Science Foundation of China (Grant No. 11701590), Natural Science Foundation of Guangdong Province of China (Grant No. 2017A030310053) and Young teacher program/Fundamental Research Funds for the Central Universities (Grant No. 17lgpy14). Dr. Wang's research is partially supported by the National Natural Science Foundation of China (Grant No. 11771462) and International Science & Technology Cooperation program of Guangdong (20163400042410001). Dr. Zhang's research is partially supported by the Basic Research Seed Fund (201611159250) of the University of Hong Kong.

Received 3 May 2018

REFERENCES

[1] De Souza, A. W., Fernandes, V. and Ferrari, A. J. (2005). Female gout: clinical and laboratory features. *The Journal of rheumatology* **32** 2186–2188.

[2] Preiss, D. and Tišer, J. (1991). Measures in Banach spaces are determined by their values on balls. *Mathematika* **38** 391–397. MR1147839

[3] Xia, F., Chen, J., Fung, W. K. and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063. MR3146800

[4] Fasano, G. and Franceschini, A. (1987). A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society* **225** 155–170.

[5] Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat* **5**.

[6] Fang, H., Huang, C., Zhao, H. and Deng, M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31** 3172–3180.

[7] Aitchison, J. (1986). The statistical analysis of compositional data. MR0865647

[8] Puig, J. G., Michán, A. D., Jiménez, M. L., de Ayala, C. P., Mateos, F. A., Capitán, C. F., de Miguel, E. and Gijón, J. B. (1991). Female gout: clinical spectrum and uric acid metabolism. *Archives of internal medicine* **151** 726–732.

[9] Janda, J. M. and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology* **45** 2761–2764.

[10] Martin-Fernández, J. and Bren, M. (2001). Some practical aspects on multidimensional scaling of compositional data. In *Proceedings of 2001 Annual Conference of the International Association for Mathematical Geology* 6–12.

[11] Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G. and Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. In *Proceedings of IAMG* **98** 526–531.

[12] Palarea-Albaladejo, J., Martín-Fernández, J. A. and Soto, J. A. (2012). Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *Journal of classification* **29** 144–169. MR2950940

[13] Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data with R* **122**. Springer. MR3099409

[14] Lyons, R. et al. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41** 3284–3305. MR3127883

[15] Chiu, S. N. and Liu, K. I. (2009). Generalized Cramér–von Mises goodness-of-fit tests for multivariate distributions. *Computational Statistics & Data Analysis* **53** 3817–3834. MR2749926

[16] Krzanowski, W. (1988). Principles of multivariate analysis: a user's perspective. Clarendon. MR0969370
[17] Pan, W., Tian, Y., Wang, X., Zhang, H. et al. (2018). Ball Divergence: Nonparametric two sample test. *The Annals of Statistics* **46** 1109–1137. MR3797998
[18] Dai, X. and Müller, H.-G. (2017). Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres. *arXiv preprint arXiv:1705.06226*. MR3852654
[19] Park, Y. B., Park, Y. S., Song, J., Lee, W. K., Suh, C. H. and Lee, S. K. (2000). Clinical manifestations of Korean female gouty patients. *Clinical rheumatology* **19** 142–146.
[20] Cao, Y., Lin, W. and Li, H. (2017). Two-sample tests of high-dimensional means for compositional data. *Biometrika*. MR3768869
[21] Cao, Y., Lin, W. and Li, H. (2018). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association* just-accepted 1–45.
[22] Freund, Y., Schapire, R. E. et al. (1996). Experiments with a new boosting algorithm. In *Icml* **96** 148–156. Bari, Italy. MR2920188
[23] Guo, Z., Zhang, J., Wang, Z., Ang, K. Y., Huang, S., Hou, Q., Su, X., Qiao, J., Zheng, Y., Wang, L. et al. (2016). Intestinal microbiota distinguish gout patients from healthy humans. *Scientific reports* **6** 20602.

Jin Zhu
Southern China Center for Statistical Science
School of Mathematics
Sun Yat-sen University
Guangzhou, GD 510275, China
E-mail address: zhu37@mail2.sysu.edu.cn

Kunsheng Lv
Southern China Center for Statistical Science
School of Mathematics
Sun Yat-sen University
Guangzhou, GD 510275, China
E-mail address: lvksh@mail2.sysu.edu.cn

Aijun Zhang
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong
E-mail address: ajzhang@hku.hk

Wenliang Pan
Southern China Center for Statistical Science
School of Mathematics
Sun Yat-sen University
Guangzhou, GD 510275, China
E-mail address: panwliang@mail.sysu.edu.cn

Xueqin Wang
Southern China Center for Statistical Science
School of Mathematics
Zhongshan School of Medicine
Sun Yat-sen University
Guangzhou, GD 510275, China
E-mail address: wangxq88@mail.sysu.edu.cn