Biometrika (2017), *pp.* 1–17

Two-sample tests of high-dimensional means for compositional data

BY YUANPEI CAO

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

yuanpeic@sas.upenn.edu

WEI LIN

Center for Statistical Science, Peking University, Beijing 100871, China

weilin@math.pku.edu.cn

AND HONGZHE LI

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

hongzhe@upenn.edu

SUMMARY

Compositional data are ubiquitous in many scientific endeavors. Motivated by microbiome and metagenomic research, we consider a two-sample testing problem for high-dimensional compositional data and formulate a testable hypothesis of compositional equivalence for the means of two latent log-basis vectors. We propose a test through the centered log-ratio transformation of the compositions. The asymptotic null distribution of the test statistic is derived and its power against sparse alternatives is investigated. A modified test for paired samples is also considered. Simulations show that the proposed tests can be significantly more powerful than tests that are applied to the raw and log-transformed compositions. Their usefulness is illustrated by applications to gut microbiome composition in obesity and Crohn's disease.

Some key words: Basis; Centered log-ratio transformation; Compositional equivalence; Extreme value distribution; Microbiome; Sparse alternative.

1. INTRODUCTION

Compositional data, which belong to a simplex, are ubiquitous in scientific disciplines such as geology, economics, and genomics. This paper is motivated by microbiome and metagenomic research, where the relative abundances of hundreds to thousands of bacterial taxa on a few tens to hundreds of individuals are available for analysis (The Human Microbiome Project Consortium, 2012). Due to varying amounts of DNA generating material across different samples, sequencing read counts are often normalized to relative abundances; the resulting data are therefore compositional (Li, 2015). One fundamental problem in microbiome data analysis is to test whether two populations have the same microbiome composition, which can be viewed as a two-sample testing problem for high-dimensional compositional data. Since the components of a composition must sum to one, applying standard multivariate statistical methods intended for unconstrained data directly to compositional data may result in inappropriate or misleading inferences (Aitchison, 2003).

Various methods for compositional data analysis have been developed since the seminal work of Aitchison (1982). Most existing methods for two-sample testing, however, deal only with the low-dimensional setting where the dimensionality is smaller than the sample size; see, e.g., the generalized likelihood ratio tests discussed in Aitchison (2003, §7.5). In this paper, we consider the two-sample testing problem for high-dimensional compositional data, where compositions in the $(p-1)$-dimensional simplex $\mathcal{S}^{p-1}$ are thought of as arising from latent basis vectors in the p-dimensional positive orthant $\mathbb{R}^p_+$. In microbiome studies, the basis components may represent the true abundances of bacterial taxa in a microbial community such as the gut of a healthy individual (Li, 2015). To circumvent the nonidentifiability issue associated with the basis vectors, we formulate a testable hypothesis of compositional equivalence for the means of two log-basis vectors. We then propose a test through the centered log-ratio transformation of the compositions. The proposed test thus is scale-invariant, which is crucial for compositional data analysis.

We emphasize here the extrinsic analysis point of view in compositional data analysis (Aitchison, 1982), which leads to biologically meaningful interpretations and is in contrast to intrinsic analysis, where interest lies solely in the composition. Classical extrinsic analysis, however, primarily concerns problems where the bases are observed, and thus differs radically from the focus of this paper.

Developing tests for the equality of two high-dimensional means has received much attention; see, e.g., Bai & Saranadasa (1996), Srivastava & Du (2008), Srivastava (2009), Chen & Qin (2010) and Cai et al. (2014). These existing tests, however, are not directly applicable to high-dimensional compositional data because the required regularity conditions are generally not met. For example, the covariance matrix of compositional data is singular, thereby violating the usual assumptions on the eigenvalues of the covariance matrix, such as those in Cai et al. (2014). Our assumptions are imposed on the latent log-basis vectors, which are free of the simplex constraint. We show that, under mild conditions, the centered log-ratio variables satisfy certain desired properties, which guarantee the validity of the proposed test. Then the asymptotic null distribution of the test statistic is derived and the power of the test against sparse alternatives is investigated. The proposed two-sample test is further modified to accommodate paired samples. All proofs are deferred to the Appendix.

2. A TESTABLE HYPOTHESIS OF COMPOSITIONAL EQUIVALENCE

Denote by $X^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})^{\mathrm{T}}$ the observed $n_k \times p$ data matrices for group k ($k = 1, 2$), where $X_i^{(k)}$ represent compositions that lie in the $(p-1)$-dimensional simplex $\mathcal{S}^{p-1} = \{(x_1, \ldots, x_p) : x_j > 0\,(j = 1, \ldots, p), \sum_{j=1}^p x_j = 1\}$. We assume that the compositional variables arise from a vector of latent variables, which we call the basis. For microbiome data, the basis components may refer to the true abundances of bacterial taxa in a microbial community. Denote by $W^{(k)} = (W_1^{(k)}, \ldots, W_{n_k}^{(k)})^{\mathrm{T}}$ the $n_k \times p$ matrices of unobserved bases, which generate the observed compositional data via the normalization

$$X_{ij}^{(k)} = W_{ij}^{(k)} \Big/ \sum_{\ell=1}^{p} W_{i\ell}^{(k)} \quad (i = 1, \ldots, n_k;\ j = 1, \ldots, p;\ k = 1, 2), \tag{1}$$

where $X_{ij}^{(k)}$ and $W_{ij}^{(k)} > 0$ are the jth components of $X_i^{(k)}$ and $W_i^{(k)}$, respectively.

Denote by $Z_i^{(k)} = (Z_{i1}^{(k)}, \dots, Z_{ip}^{(k)})^{\mathrm{T}}$ the log-basis vectors, where $Z_{ij}^{(k)} = \log W_{ij}^{(k)}$. Suppose that $Z_1^{(k)}, \dots, Z_{n_k}^{(k)}$ $(k = 1, 2)$ are two independent samples, each from a distribution with mean $\mu_k = (\mu_{k1}, \dots, \mu_{kp})^{\mathrm{T}}$ and common covariance matrix $\Omega = (\omega_{ij})$. One might attempt to test the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2. \tag{2}$$

These hypotheses, however, are not testable through the observed compositional data $X^{(k)}$ $(k = 1, 2)$. Clearly, a basis is determined by its composition only up to a multiplicative factor, and the set of bases giving rise to a composition $x \in \mathcal{S}^{p-1}$ forms the equivalence class $\mathcal{W}(x) = \{(tx_1, \dots, tx_p) : t > 0\}$ (Aitchison, 2003, p. 32). As an immediate consequence, a log-basis vector is determined by the resulting composition only up to an additive constant, and the set of log-basis vectors corresponding to x constitutes the equivalence class $\mathcal{Z}(x) = \{(\log x_1 + c, \dots, \log x_p + c) : c \in \mathbb{R}\}$. We therefore introduce the following definition.

DEFINITION 1. *Two log-basis vectors z_1 and z_2 are said to be compositionally equivalent if their components differ by a constant $c \in \mathbb{R}$, i.e., $z_1 = z_2 + c1_p$, where 1_p is the p-vector of 1s.*

Now, instead of testing the hypotheses in (2), we propose to test

$$H_0 : \mu_1 = \mu_2 + c1_p \text{ for some } c \in \mathbb{R} \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2 + c1_p \text{ for any } c \in \mathbb{R}, \tag{3}$$

which are testable using only the observed compositional data. Clearly, H_0 in (2) implies H_0 in (3), so that rejecting the latter would lead to rejection of the former. Note, however, that H_0 in (3) neither implies nor is implied by $E(X_1^{(1)}) = E(X_1^{(2)})$ or $E(\log X_1^{(1)}) = E(\log X_1^{(2)})$. We do not consider the latter two hypotheses because they are not scale-invariant, whereas we will derive in § 3·1 an equivalent form of H_0 in (3), from which its scale-invariance is obvious. Moreover, these two hypotheses do not allow us to obtain biological interpretations in terms of the true underlying abundances.

3. TESTS FOR COMPOSITIONAL EQUIVALENCE

3·1. *The centered log-ratio transformation and an equivalent hypothesis*

The unit-sum constraint entails that compositional variables must not vary independently, making many covariance-based multivariate analysis methods inapplicable. Aitchison (1982) proposed to relax the constraint by performing statistical analysis through log-ratios. Among various forms of log-ratio transformations, the centered log-ratio transformation has attractive features and has been widely used. For the observed compositional data $X^{(k)}$ $(k = 1, 2)$, the centered log-ratio matrices $Y^{(k)} = (Y_1^{(k)}, \dots, Y_{n_k}^{(k)})^{\mathrm{T}}$ are defined by

$$Y_{ij}^{(k)} = \log\{X_{ij}^{(k)}/g(X_i^{(k)})\} \quad (i = 1, \dots, n_k;\ j = 1, \dots, p;\ k = 1, 2), \tag{4}$$

where $g(x) = (\prod_{i=1}^p x_i)^{1/p}$ denotes the geometric mean of a vector $x = (x_1, \dots, x_p)^{\mathrm{T}}$. The relationship (4) can be expressed in the matrix form

$$Y_i^{(k)} = G \log X_i^{(k)}, \tag{5}$$

where $G = I_p - p^{-1}1_p1_p^{\mathrm{T}}$.

Let $\nu_k = E(Y_1^{(k)})$ $(k = 1, 2)$. In view of the scale invariance of the centered log-ratios, we can replace $X_i^{(k)}$ by $W_i^{(k)}$ in (5) and obtain

$$Y_i^{(k)} = GZ_i^{(k)}, \tag{6}$$

and hence

$$\nu_k = G\mu_k. \tag{7}$$

The matrix G has rank $p-1$ and hence a null space of dimension 1, $\mathcal{N}(G) \equiv \{x \in \mathbb{R}^p : Gx = 0\} = \{c1_p : c \in \mathbb{R}\}$. As a result, $\nu_1 = \nu_2$ if and only if $\mu_1 = \mu_2 + c1_p$ for some $c \in \mathbb{R}$. Therefore, testing the hypotheses in (3) is equivalent to testing

$$H_0 : \nu_1 = \nu_2 \quad \text{versus} \quad H_1 : \nu_1 \neq \nu_2. \tag{8}$$

Despite this equivalence, the hypotheses in (3) are meaningful only when the bases exist, which is the case in microbiome studies. On the other hand, the hypotheses in (8) concern only the compositions through the centered log-ratios, from which its scale invariance and testability using the observed compositional data are evident.

3·2. *A two-sample test for compositional equivalence*

A natural test statistic for testing H_0 in (8), and hence H_0 in (3), would be based on the differences $\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)}$, where $\bar{Y}_j^{(k)} = n_k^{-1}\sum_{i=1}^{n_k} Y_{ij}^{(k)}$ are the sample means of the centered log-ratios. Moreover, it is well-known that tests against sparse alternatives based on maximum type statistics are generally more powerful than those based on sum-of-squares type statistics (Cai et al., 2014). Since in microbiome studies we are mainly interested in the sparse setting where only a small number of taxa may have different mean abundances, we consider the test statistic

$$M_n = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leqslant j \leqslant p} \frac{(\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)})^2}{\hat{\gamma}_{jj}}, \tag{9}$$

where $\hat{\gamma}_{jj} = (n_1 + n_2)^{-1}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{ij}^{(k)} - \bar{Y}_j^{(k)})^2$ are the pooled sample centered log-ratio variances.

The asymptotic properties of M_n will be investigated in detail in § 4. Under suitable conditions on the log-basis variables $Z_{1j}^{(k)}$, we will show that the centered log-ratio variables $Y_{1j}^{(k)}$ are only weakly dependent and satisfy certain desired properties. As a result, $M_n - 2\log p + \log\log p$ converges in distribution to a type I extreme value or Gumbel variable; see Theorem 1. We can then define the asymptotic α-level test

$$\Phi_\alpha = I(M_n \geqslant q_\alpha + 2\log p - \log\log p), \tag{10}$$

where $q_\alpha = -\log \pi - 2\log\log(1-\alpha)^{-1}$ is the $(1-\alpha)$-quantile of the Gumbel distribution. The null hypothesis H_0 in (3) or equivalently (8) is rejected whenever $\Phi_\alpha = 1$.

Although M_n is similar to the test statistic M_I defined in Cai et al. (2014), the theoretical analysis is radically different, in that our assumptions are not imposed on the observed variables. Besides, the test statistic based on a linear transformation by the precision matrix proposed by Cai et al. (2014) is not considered here, because the covariance matrix of $Y_1^{(k)}$ is singular and its precision matrix is not well defined.

3·3. *A paired test for compositional equivalence*

So far we have been concerned with two independent samples. In practice, however, one may be interested in comparing compositions on the same sample before and after treatment. For such paired samples, the proposed test requires only slight modification. Now we observe a paired sample $(X_{ij}^{(1)}, X_{ij}^{(2)})$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$, which is generated by the basis $(W_{ij}^{(1)}, W_{ij}^{(2)})$; the log-basis $(Z_{ij}^{(1)}, Z_{ij}^{(2)})$ and the centered log-ratios $(Y_{ij}^{(1)}, Y_{ij}^{(2)})$ are the same as defined previously. Write $D_{ij} = Y_{ij}^{(1)} - Y_{ij}^{(2)}$ and $\bar{D}_j = n^{-1} \sum_{i=1}^{n} D_{ij}$. To test H_0 in (3) or equivalently (8), we propose the test statistic

$$\tilde{M}_n = n \max_{1 \leqslant j \leqslant p} \bar{D}_j^2 / \tilde{\gamma}_{jj},$$

where $\tilde{\gamma}_{jj} = n^{-1} \sum_{i=1}^{n} (D_{ij} - \bar{D}_j)^2$ are the sample variances of D_{ij}. Note that $\tilde{M}_n$ is different from M_n defined in (9) only in the variance estimates. Under appropriate assumptions on the log-basis differences $\Delta_j = Z_{1j}^{(1)} - Z_{1j}^{(2)}$ similar to Conditions 1–5 below, we can show that $\tilde{M}_n - 2 \log p + \log \log p$ converges in distribution to the same Gumbel variable as in Theorem 1. Hence, the test Φ_α defined in (10) is still valid with M_n replaced by $\tilde{M}_n$.

4. THEORETICAL RESULTS

4·1. *Assumptions and implications*

Since we are interested in testing the latent basis structures, we will impose conditions directly on the log-basis variables. Under the assumption of common basis covariance matrix, the two populations have a common centered log-ratio covariance matrix $\Gamma = \mathrm{cov}(Y_1^{(k)})$ $(k = 1, 2)$, which, in light of (6), is given by

$$\Gamma = G \Omega G^{\mathrm{T}}. \tag{11}$$

Denote the correlation matrices of $Z_1^{(k)}$ and $Y_1^{(k)}$ by $R = (\rho_{ij})$ and $T = (\tau_{ij})$, respectively.

We first impose the following conditions concerning the covariance structures of the log-basis variables:

Condition 1. $1/\kappa_1 \leqslant \omega_{jj} \leqslant \kappa_1$ for $j = 1, \ldots, p$ and some constant $\kappa_1 > 0$;

Condition 2. $\max_{1 \leqslant i < j \leqslant p} |\rho_{ij}| \leqslant r_1$ for some constant $0 < r_1 < 1$; and

Condition 3. $\max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} \rho_{ij}^2 \leqslant r_2$ for some constant $r_2 > 0$.

Conditions 1–3 are mild and standard in the high-dimensional testing literature. Condition 1 requires that the variances be bounded away from zero and infinity. Condition 2 is mild since $\max_{1 \leqslant i < j \leqslant p} |\rho_{ij}| = 1$ would imply that Ω is singular. Condition 3 is weaker than the usual assumption that the maximum eigenvalue of R is bounded.

Under Conditions 1–3, the following proposition shows that similar properties are satisfied by the centered log-ratio covariance matrix Γ and correlation matrix T.

PROPOSITION 1. *Assume that Conditions* 1–3 *hold. Then, for sufficiently large* p, *the centered log-ratio covariance matrix* Γ *and correlation matrix* T *satisfy the following properties:*

(i) $1/\kappa_2 \leqslant \gamma_{jj} \leqslant \kappa_2$ *for* $j = 1, \ldots, p$ *and some constant* $\kappa_2 > 0$*;*
(ii) $\max_{1 \leqslant i < j \leqslant p} |\tau_{ij}| \leqslant r_3$ *for some constant* $0 < r_3 < 1$*; and*
(iii) $\max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} \tau_{ij}^2 \leqslant r_4$ *for some constant* $r_4 > 0$.

We also need a moment condition on the log-basis variables and a restriction on the dimensionality.

Condition 4. There exist constants $\eta, K > 0$ such that

$$E[\exp\{\eta(Z_{1j}^{(k)} - \mu_{kj})^2/\omega_{jj}\}] \leqslant K \quad (j = 1, \ldots, p;\ k = 1, 2).$$

Condition 5. We have $n_1 \asymp n_2 \asymp n$ and $\log p = o(n^{1/3})$, where $n = n_1 n_2/(n_1 + n_2)$.

Condition 4 is a popular sub-Gaussian tail assumption that can easily be relaxed to the case of polynomial tails. It allows us to establish the following concentration properties for the centered log-ratio variables and the pooled sample variances.

PROPOSITION 2. *Under Conditions* 1 *and* 3–5*, the centered log-ratio variables satisfy*

$$\max_{i,j,k} |Y_{ij}^{(k)} - \nu_{kj}|/\gamma_{jj}^{1/2} = o_p(n^{1/2}/\log p), \tag{12}$$

and the pooled sample centered log-ratio variances $\hat{\gamma}_{jj}$ satisfy

$$\max_{j} |\hat{\gamma}_{jj} - \gamma_{jj}|/\gamma_{jj} = O_p\{(\log p/n)^{1/2}\}. \tag{13}$$

4·2. *Asymptotic properties of the two-sample test*

We are now in a position to state our main results concerning the asymptotic properties of the proposed two-sample test. The validity of the test relies on the fact that certain desired properties of the centered log-ratio variables can be related to those of the log-basis variables, which have been established in Propositions 1 and 2. The following theorem derives the asymptotic null distribution of M_n defined in (9).

THEOREM 1. *Under Conditions* 1–5*, we have, under H_0 in* (3) *or equivalently* (8)*,*

$$\text{pr}(M_n - 2\log p + \log\log p \leqslant t) \to \exp\{-\pi^{-1/2}\exp(-t/2)\}, \quad t \in \mathbb{R}, \quad n, p \to \infty.$$

Theorem 1 shows that the test Φ_α defined in (10) is indeed asymptotically of level α. To study the asymptotic power of the test, we consider the alternative

$$H_1 : \mu_{1j} \neq \mu_{2j} + c, \quad j \in S; \quad \mu_{1j} = \mu_{2j} + c, \quad j \in S^c, \tag{14}$$

for some $c \in \mathbb{R}$ and $S \subset \{1, \ldots, p\}$ with cardinality s, where S^c is the complement of S. This alternative, however, is difficult to analyze since c is unknown.

We now eliminate the constant c and connect H_1 in (14) to a more convenient form in terms of ν_1 and ν_2. Without loss of generality, define the signal vector $\delta = (\delta_1, \ldots, \delta_p)^T$ by

$$\mu_{1j} - \mu_{2j} - c = \delta_j \omega_{jj}^{1/2} \left(\frac{\log p}{n}\right)^{1/2} \quad (j = 1, \ldots, p), \tag{15}$$

where the scaling factor $\omega_{jj}^{1/2}(\log p/n)^{1/2}$ is introduced for technical reasons which will become clear in the proof of Theorem 2. Under H_1 in (14), we have $\delta_j \neq 0$ if and only if $j \in S$. Summing up the equations (15) and rearranging, we obtain

$$c = \bar{\mu}_1 - \bar{\mu}_2 - \frac{1}{p}\sum_{j=1}^{p} \delta_j \omega_{jj}^{1/2} \left(\frac{\log p}{n}\right)^{1/2} = \bar{\mu}_1 - \bar{\mu}_2 - O\left\{\frac{\|\delta\|_1}{p}\left(\frac{\log p}{n}\right)^{1/2}\right\},$$

where $\bar{\mu}_k = p^{-1}\sum_{j=1}^{p} \mu_{kj}$ $(k = 1, 2)$, $\|\delta\|_1 = \sum_{j=1}^{p} |\delta_j|$, and we have used the fact that $\max_j \omega_{jj} = O(1)$ by Condition 1. Since $\nu_{kj} = \mu_{kj} - \bar{\mu}_k$ $(k = 1, 2)$ by (7), we see that H_1 in

(14) implies

$$\begin{aligned} \nu_{1j} - \nu_{2j} &= \left\{ \delta_j \omega_{jj}^{1/2} + O\left(\frac{\|\delta\|_1}{p} \right) \right\} \left(\frac{\log p}{n} \right)^{1/2}, \quad j \in S, \\ \nu_{1j} - \nu_{2j} &= O\left\{ \frac{\|\delta\|_1}{p} \left(\frac{\log p}{n} \right)^{1/2} \right\}, \qquad\qquad j \in S^c. \end{aligned} \tag{16}$$

Compared with the usual sparse alternatives analyzed in the literature such as Cai et al. (2014), all components in the alternative (16) are shifted by a term of order $O\{\|\delta\|_1 p^{-1} (\log p/n)^{1/2}\}$. To prevent this term from interfering with signals at least of order $O\{(\log p/n)^{1/2}\}$, it suffices to assume that $\|\delta\|_1 = o(p)$. This key observation leads to the following theorem concerning the asymptotic power of Φ_α defined in (10).

THEOREM 2. *Assume that Conditions* 1 *and* 3–5 *hold. Under* H_1 *in* (14)*, if* $\|\delta\|_1 = o(p)$ *and* $\max_{j\in S} |\delta_j| \geqslant \surd 2 + \varepsilon$ *for some constant* $\varepsilon > 0$*, then* $\mathrm{pr}(\Phi_\alpha = 1) \to 1$ *as* $n, p \to \infty$.

Two remarks on Theorem 2 are in order. First, if the signals δ_i are bounded, then the condition $\|\delta\|_1 = o(p)$ holds provided the alternative (14) is sparse in the sense that $s = o(p)$. Second, by Theorem 3 of Cai et al. (2014), the condition $\max_{j\in S} |\delta_j| \geqslant \surd 2 + \varepsilon$ is minimax rate optimal for testing sparse alternatives in the classical two-sample testing problem. Thus, the proposed test achieves the best possible rate even though the bases are not observed.

5. SIMULATION STUDIES

We conducted simulation studies to evaluate the numerical performance of the proposed two-sample and paired tests. For comparison, we consider counterparts applied to the raw and log-transformed compositions, which are obtained by replacing $Y^{(k)}$ in the proposed tests with $X^{(k)}$ and $\log X^{(k)}$, respectively. The oracle tests based on the unobserved $W^{(k)}$, though impracticable, serve as the benchmarks for comparison.

We first examine the case of two independent samples. The simulated data were generated as follows. We first generated $Z^{(k)}$ from the following distributions:

(i) multivariate normal distribution, $Z_i^{(k)} \sim N_p(\tilde{\mu}_k, \Omega)$;

(ii) multivariate gamma distribution, $Z_i^{(k)} = \tilde{\mu}_k + F U_i^{(k)} / \surd 10$, where the $p \times p$ matrix $F = Q S^{1/2}$ with Q and S obtained from the singular value decomposition $\Omega = Q S Q^{\mathrm{T}}$, and the components of U_k were generated independently from the standard gamma distribution with shape parameter 10.

Then $W^{(k)}$ and $X^{(k)}$ were generated through the transformation $W_{ij}^{(k)} = \exp(Z_{ij}^{(k)})$ and (1). Note that $\mu_k = \tilde{\mu}_k$ in case (i) and $\mu_k = \tilde{\mu}_k + \surd 10 F 1_p$ in case (ii). The location parameters $\tilde{\mu}_k$ were set as follows. The components of $\tilde{\mu}_1$ were drawn from a uniform distribution $U(0, 10)$. Under H_0, we took $\tilde{\mu}_2 = \tilde{\mu}_1$; under H_1, we took

$$\tilde{\mu}_{2j} = \tilde{\mu}_{1j} - \delta_j \omega_{jj}^{1/2} \left(\frac{\log p}{n} \right)^{1/2},$$

where the signal vector δ has a support of size $s = \lfloor 0{\cdot}05p \rfloor$, $\lfloor 0{\cdot}1p \rfloor$, and $\lfloor 0{\cdot}5p \rfloor$, with the indices chosen uniformly from $\{1, \ldots, p\}$ and the nonzero δ_j drawn from $U[-2\surd 2, 2\surd 2]$. We considered the following covariance structures:

(i) banded covariance, $\Omega = D^{1/2}AD^{1/2}$, where A has non-zero entries $a_{jj} = 1$, $a_{j-1,j} = a_{j,j+1} = -0{\cdot}5$, and D is a diagonal matrix with entries drawn from $U(1,3)$;

(ii) sparse covariance, $\Omega = \mathrm{diag}(A_1, A_2)$, where $A_1 = B + \varepsilon I_q$, $A_2 = I_{p-q}$, $q = \lfloor 3p^{1/2} \rfloor$, B is a symmetric matrix with lower-triangular entries drawn from the uniform distribution on $[-1,-0{\cdot}5] \cup [0{\cdot}5,1]$ with probability $0{\cdot}5$ and set to 0 with probability $0{\cdot}5$, $\varepsilon = \max\{-\lambda_{\min}(B), 0\} + 0{\cdot}05$, and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix.

The simulation settings for the case of paired samples are similar, except that $Z_i^{(1)}$ and $Z_i^{(2)}$ are correlated and must be generated from a $2p$-dimensional joint distribution. The parameters $(\tilde{\mu}^*, \Omega^*)$ of the joint distribution were specified by $\tilde{\mu}^* = (\tilde{\mu}_1^{\mathrm{T}}, \tilde{\mu}_2^{\mathrm{T}})^{\mathrm{T}}$ and

$$\Omega^* = \begin{pmatrix} 1 & 0{\cdot}3 \\ 0{\cdot}3 & 1 \end{pmatrix} \otimes \Omega,$$

where $\tilde{\mu}_k$ and Ω were described above.

We set the sample sizes $n_1 = n_2 = 100$ for two independent samples and $n = 100$ for paired samples, with varying dimensions $p = 50, 100$ and 200. We repeated the simulation 1000 times for each setting and calculated the empirical sizes and powers of four tests with significance level $\alpha = 0{\cdot}05$. The results for two independent samples and paired samples are summarized in Table 1 and 2. The proposed test has higher power than those applied to the raw and log-transformed compositions, while controlling the size reasonably well around the nominal level 0·05, and closely mimics the performance of the oracle test. Its power gains over the tests based on log-transformed and raw compositions tend to be more pronounced in the more challenging scenarios with moderate dimensions and sparse signals. Its superiority does not seem to depend on the distributions or covariance structures.

To further examine the performance of the proposed test in very high-dimensional settings, we carried out simulations for two independent samples with dimension $p = 2000$ and sample sizes $n_1 = n_2 = 100$ and 200. The results are summarized in Table 3 and indicate that the proposed test still has approximately correct size and improved power over the two competing tests.

6. APPLICATIONS TO MICROBIOME DATA

6·1. *Analysis of obesity microbiome data*

We illustrate the proposed tests by applications to two microbiome datasets. We first consider a dataset from Wu et al. (2011), which was collected in a cross-sectional study of 98 subjects for investigating habitual diet effect on the human gut microbiome. The dataset was analyzed by regression in Lin et al. (2014) and was found to suggest an association between obesity and changes in gut microbiome composition. For each subject, DNA samples collected from stool samples were analysed by 454/Roche pyrosequencing of 16S rRNA gene segments from the V1–V2 region. An average of 9265 reads per sample were obtained, with a standard deviation of 3864, by denoising the pyrosequences prior to taxonomic assignment. The resulting 3068 operational taxonomic units were further merged into 87 genera that were observed in at least one sample. As suggested by Aitchison (2003) and Lin et al. (2014), zero counts were replaced by 0·5 before the count data were converted into compositional data by normalization. Demographic information including body mass index, BMI, was recorded on the subjects.

We are interested in testing whether lean and obese individuals have the same gut microbiome composition. To this end, we divided the subjects into a lean group (BMI < 25, $n_1 = 63$) and an obese group (BMI $\geqslant 25$, $n_2 = 35$), and performed various two-sample tests. The proposed

Table 1. *Empirical sizes and powers* (%) *of two-sample tests with* $\alpha = 0{\cdot}05$ *and* $n_1 = n_2 = 100$ *based on* 1000 *replications*

		Banded covariance			Sparse covariance		
	Method	$p = 50$	$p = 100$	$p = 200$	$p = 50$	$p = 100$	$p = 200$
Normal, H_0	Oracle	4·7	5·3	4·8	4·0	4·8	5·2
	Proposed	4·6	4·9	5·1	3·7	4·5	5·3
	Log	3·9	5·1	5·3	3·5	3·3	5·2
	Raw	0·9	1·0	0·3	1·5	1·0	1·3
Normal, H_1	Oracle	38·2	70·7	92·5	40·1	70·7	91·8
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	36·5	70·5	92·2	38·0	70·2	91·0
	Log	26·1	60·8	84·7	25·5	51·4	70·7
	Raw	4·0	5·5	8·2	7·0	16·8	23·7
Normal, H_1	Oracle	68·7	90·6	99·1	69·4	91·0	99·5
$s = \lfloor 0{\cdot}1p \rfloor$	Proposed	66·9	89·9	98·9	67·6	91·0	99·5
	Log	53·7	79·7	97·3	50·0	77·1	91·7
	Raw	9·1	10·1	14·1	16·6	31·7	49·2
Normal	Oracle	99·3	100·0	100·0	99·4	100·0	100·0
$s = \lfloor 0{\cdot}5p \rfloor$	Proposed	99·2	100·0	100·0	99·7	100·0	100·0
	Log	96·5	99·9	100·0	96·6	99·9	100·0
	Raw	39·2	37·1	61·5	55·0	84·0	96·0
Gamma, H_0	Oracle	5·6	4·4	4·7	5·9	4·8	4·8
	Proposed	5·3	4·9	4·8	5·0	4·9	5·1
	Log	4·7	3·6	3·7	5·0	4·7	4·5
	Raw	1·6	0·8	0·2	1·6	0·6	1·1
Gamma, H_1	Oracle	35·7	70·0	91·3	36·7	70·5	92·3
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	36·7	71·5	91·9	36·3	68·0	92·0
	Log	27·0	52·6	82·8	23·6	49·9	66·0
	Raw	5·1	4·4	10·2	4·2	6·0	9·4
Gamma, H_1	Oracle	68·5	91·8	99·6	69·0	91·7	99·5
$s = \lfloor 0{\cdot}1p \rfloor$	Proposed	66·8	91·5	99·5	66·9	90·8	99·7
	Log	52·4	78·4	96·2	50·9	75·6	90·7
	Raw	11·6	9·8	17·3	10·3	13·2	17·4
Gamma	Oracle	99·9	100·0	100·0	100·0	100·0	100·0
$s = \lfloor 0{\cdot}5p \rfloor$	Proposed	99·9	100·0	100·0	99·5	100·0	100·0
	Log	96·5	99·7	100·0	96·9	99·9	100·0
	Raw	42·7	53·1	61·9	40·4	50·7	62·9

test yielded a p-value of 0·001, indicating a marked difference between the two groups. In contrast, the tests based on the log-transformed and raw compositions gave p-values of 0·129 and 0·261, and hence failed to detect the difference at the 0·05 level. To assess the stability of our proposed test and perform power comparisons, we generated 5000 bootstrap subsamples within each group, with the subsampling proportion varying from 0·2 to 1. For each subsampling proportion, we obtained the empirical power as the proportion of subsamples where the null hypothesis was rejected at the 0·05 level. The empirical power curves based on the bootstrap subsamples, presented in Fig. 1(a), show that the proposed test greatly outperforms the competitors. We further conduct back-testing to check whether the signal disappears by breaking the association. We generated 1000 bootstrap samples from the pooled data and then randomly divided each sample into two groups with the same sizes as before. The histogram of p-values from our test based on the bootstrap samples is depicted in Fig. 1(b). The p-values are distributed quite evenly, indi-

Table 2. *Empirical sizes and powers* (%) *of paired tests with* $\alpha = 0{\cdot}05$ *and* $n = 100$ *based on* 1000 *replications*

		Banded covariance			Sparse covariance		
	Method	$p = 50$	$p = 100$	$p = 200$	$p = 50$	$p = 100$	$p = 200$
Normal, H_0	Oracle	4·8	5·6	6·3	5·9	6·6	6·0
	Proposed	4·9	5·5	6·8	5·9	6·1	6·5
	Log	5·4	4·4	7·1	5·2	3·7	4·1
	Raw	1·1	0·4	0·2	1·5	1·1	1·2
Normal, H_1	Oracle	55·3	86·8	98·3	54·4	85·8	99·0
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	52·9	86·5	98·4	54·0	84·8	98·9
	Log	39·6	75·0	95·6	35·7	68·7	87·3
	Raw	5·2	7·3	13·7	9·2	21·6	34·3
Normal, H_1	Oracle	85·1	98·6	99·9	83·9	98·6	99·9
$s = \lfloor 0{\cdot}1p \rfloor$	Proposed	82·8	98·4	99·9	82·8	98·3	99·9
	Log	71·2	94·5	99·6	65·4	90·0	98·3
	Raw	13·8	14·9	22·1	22·7	43·7	64·5
Normal	Oracle	100·0	100·0	100·0	100·0	100·0	100·0
$s = \lfloor 0{\cdot}5p \rfloor$	Proposed	100·0	100·0	100·0	100·0	100·0	100·0
	Log	99·5	100·0	100·0	99·7	100·0	100·0
	Raw	50·0	50·0	74·0	77·9	94·4	99·3
Gamma, H_0	Oracle	4·2	6·1	7·2	5·1	6·3	7·3
	Proposed	4·3	6·2	7·2	5·8	6·1	7·1
	Log	4·7	4·7	5·6	5·2	4·5	5·5
	Raw	1·2	0·5	0·4	1·4	1·7	2·0
Gamma, H_1	Oracle	55·5	86·1	98·4	51·3	84·2	98·3
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	53·9	86·2	98·4	50·1	84·0	98·0
	Log	42·4	77·3	94·8	35·9	67·2	88·1
	Raw	6·1	8·4	11·1	9·6	22·6	39·4
Gamma, H_1	Oracle	83·8	97·7	100·0	87·9	98·2	100·0
$s = \lfloor 0{\cdot}1p \rfloor$	Proposed	82·6	97·1	100·0	86·5	98·3	99·9
	Log	70·8	93·0	99·8	67·8	90·8	98·4
	Raw	12·0	13·2	20·7	24·0	44·0	61·5
Gamma	Oracle	100·0	100·0	100·0	100·0	100·0	100·0
$s = \lfloor 0{\cdot}5p \rfloor$	Proposed	100·0	100·0	100·0	100·0	100·0	100·0
	Log	99·4	100·0	100·0	99·8	100·0	100·0
	Raw	51·5	52·7	75·0	80·3	94·5	99·0

cating good accuracy of the asymptotics. Overall, our results confirm previous findings that lean and obese microbiomes differ at the taxonomic and functional levels (Turnbaugh et al., 2009).

To further assess the sensitivity of the results to zero replacements, we repeated the analysis with the zero counts replaced by 0·1 before normalization. The proposed test resulted in a p-value of 0·0001, while the tests based on the log-transformed and raw compositions gave p-values of 0·015 and 0·080, respectively. In this case only the proposed test rejects the null hypothesis at the 0·01 level and the inference does not seem sensitive to the zero replacement values.

6·2. *Analysis of Crohn's disease microbiome data*

Crohn's disease is a type of inflammatory bowel disease characterized by altered gut bacterial composition, whose etiology appears multifactorial and remains poorly understood. We analyze a dataset from a longitudinal study of 90 pediatric Crohn's disease patients reported by Lewis

Table 3. *Empirical sizes and powers* $(\%)$ *of two-sample tests with* $\alpha = 0{\cdot}05$ *and* $p = 2000$ *based on* 1000 *replications*

		Banded covariance		Sparse covariance	
	Method	$n_1 = n_2 = 100$	$n_1 = n_2 = 200$	$n_1 = n_2 = 100$	$n_1 = n_2 = 200$
Normal, H_0	Oracle	6·5	3·7	7·6	4·1
	Proposed	6·4	3·7	7·5	4·1
	Log	5·0	4·7	2·4	1·8
	Raw	0·2	0·1	0·5	0·9
Normal, H_1	Oracle	100·0	100·0	100·0	100·0
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	100·0	100·0	100·0	100·0
	Log	100·0	100·0	98·7	98·0
	Raw	48·4	55·0	60·7	68·6
Gamma, H_0	Oracle	7·0	6·1	6·4	6·7
	Proposed	6·9	6·1	6·6	7·1
	Log	4·9	4·5	2·7	2·2
	Raw	0·2	0·2	0·4	0·1
Gamma, H_1	Oracle	100·0	100·0	100·0	100·0
$s = \lfloor 0{\cdot}05p \rfloor$	Proposed	100·0	100·0	100·0	100·0
	Log	100·0	100·0	98·5	98·9
	Raw	36·1	45·3	22·1	22·0

et al. (2015). Among these patients, 26 were classified as responders to anti-tumor necrosis factor therapy, where response to therapy was defined as a reduction in fecal calprotectin, FCP, concentration to $\leqslant 250$ μg/g among those with baseline FCP > 250 μg/g. Twenty-four of the responders had stool samples collected at four time points: baseline, 1 week, 4 weeks, and 8 weeks into therapy. The bacterial composition was quantified using shotgun metagenomic sequencing and the MetaPhlAn package (Segata et al., 2012), yielding 43 genera that appeared in at least three samples across all time points. Since the read counts were not available, zero proportions were replaced by half or 10% of the minimum nonzero proportions in the dataset.

To determine the effect of the therapy among responders, we applied various paired tests to test for changes in gut microbiome composition between baseline and three later time points. As shown in Table 4, the p-values for the comparison between baseline and week 8 from all tests were significant or close to significant, with the strongest evidence provided by the proposed test. The comparisons at two earlier time points did not yield decisive conclusions. These inferences do not seem sensitive to the zero replacement strategies. The empirical power curves based on bootstrap subsamples in Fig. 1(c) exhibit more substantial power gains of the proposed test over the competitors with smaller sample sizes. Moreover, the histogram of p-values in Fig. 1(d) indicates that the proposed test survives the back-testing, where the observations at two time points were randomly interchanged for each subject in the bootstrap samples. Our results provide further support for the effect of the therapy on gut microbiome composition through reduced inflammation and suggest that it may take longer for the intestinal dysbiosis to be resolved.

7. Discussion

We have shown that it is possible to develop tests for high-dimensional parameters of the log-basis variables from which compositional data are derived, even though the bases are not observed. In this regard, our method substantially extends the scope of the log-ratio transformation methodology due to Aitchison (1982). The mild assumption that $\|\delta\|_1 = o(p)$ for the proposed

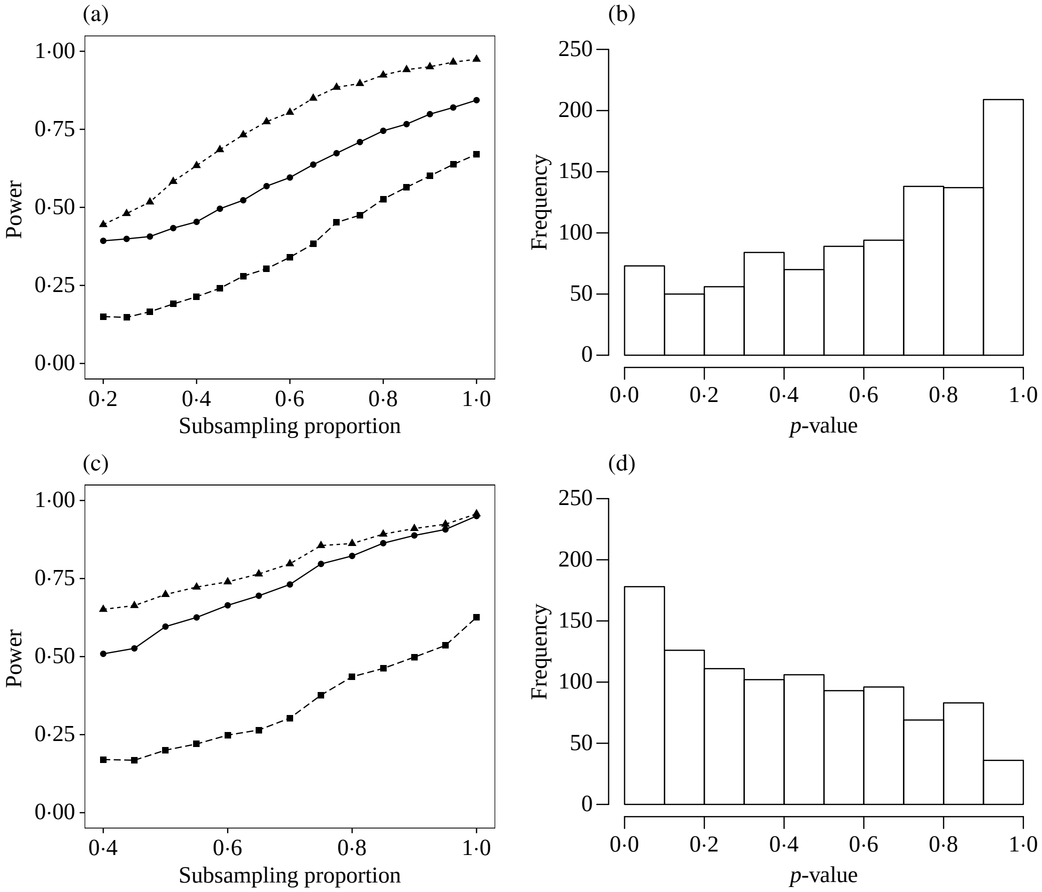

Fig. 1. Analysis of two microbiome datasets. Empirical power curves of the proposed test (triangles) and the tests based on log-transformed (dots) and raw (squares) compositions with $\alpha = 0{\cdot}05$ are shown in (a) for the obesity data and (c) for the Crohn's disease data. Histograms of p-values from the proposed test in the back-testing are shown in (b) for the obesity data and (d) for the Crohn's disease data, for 1000 replicates.

Table 4. *The p-values of paired tests applied to the Crohn's disease microbiome data with zeros replaced by half or 10% of the minimum nonzero proportions in the dataset*

	Zero replacement by half			Zero replacement by 10%		
	Proposed	Log	Raw	Proposed	Log	Raw
Baseline versus week 1	0·119	0·605	0·757	0·141	0·611	0·757
Baseline versus week 4	0·460	0·553	0·468	0·373	0·684	0·468
Baseline versus week 8	0·014	0·033	0·082	0·018	0·058	0·082

test to achieve the minimax optimal rate is due to the use of centered log-ratio variables as a proxy for the latent log-basis variables, which bears a striking resemblance to an approximate identifiability condition for large covariance estimation from compositional data considered in Cao et al. (arXiv:1601.04397).

Our testing framework may be extended in at least two directions. First, it would be worthwhile to exploit the covariance structure of compositional data for power enhancement, by borrowing ideas of Cai et al. (2014). Such an extension, however, seems nontrivial owing to the singularity of the centered log-ratio covariance matrix. Second, in addition to the global test developed in this paper, a multiple testing procedure with accurate error control would be helpful for identifying specific taxa that differ significantly between groups and contribute to the outcome of interest.

Acknowledgement

This research was supported in part by the U.S. National Institutes of Health and the National Natural Science Foundation of China. We thank the Associate Editor and three reviewers for their very helpful comments.

Appendix

We first introduce some notation. For a matrix $A = (a_{ij})_{p\times p}$, denote by $\|A\|_1$ and $\|A\|_{\max}$ the matrix 1-norm and entrywise ℓ_∞-norm, respectively, i.e., $\|A\|_1 = \max_{1\leqslant j\leqslant p}\sum_{i=1}^p |a_{ij}|$ and $\|A\|_{\max} = \max_{1\leqslant i,j\leqslant p}|a_{ij}|$. Write $a_{i\cdot} = p^{-1}\sum_{j=1}^p a_{ij}$ and $a_{\cdot\cdot} = p^{-2}\sum_{i=1}^p\sum_{j=1}^p a_{ij}$. We will use $C_1, C_2, \ldots > 0$ to denote generic constants, whose values may vary from line to line.

Proof of Proposition 1

By Condition 3, we have

$$\|R\|_1 \leqslant p^{1/2}\max_{1\leqslant j\leqslant p}\left(\sum_{i=1}^p \rho_{ij}^2\right)^{1/2} \leqslant p^{1/2}r_2^{1/2} = O(p^{1/2}). \tag{A1}$$

We write $\gamma_{ij} = \omega_{ij} - \omega_{i\cdot} - \omega_{j\cdot} + \omega_{\cdot\cdot}$. It follows from Condition 1 and (A1) that

$$|\omega_{i\cdot}| \leqslant \frac{1}{p}\sum_{j=1}^p|\omega_{ij}| \leqslant \frac{1}{p}\max_{1\leqslant j\leqslant p}|\omega_{jj}|\sum_{j=1}^p|\rho_{ij}| \leqslant \frac{1}{p}\kappa_1\|R\|_1 = O(p^{-1/2}), \tag{A2}$$

and similarly,

$$|\omega_{j\cdot}| = O(p^{-1/2}), \quad |\omega_{\cdot\cdot}| = O(p^{-1/2}). \tag{A3}$$

Hence

$$\|\Gamma - \Omega\|_{\max} \leqslant \max_{1\leqslant i,j\leqslant p}(|\omega_{i\cdot}| + |\omega_{j\cdot}| + |\omega_{\cdot\cdot}|) = O(p^{-1/2}). \tag{A4}$$

This and Condition 1 imply (i).

To show (ii), we write

$$\tau_{ij} = \frac{\gamma_{ij}}{(\gamma_{ii}\gamma_{jj})^{1/2}} = \frac{\omega_{ij} + \varepsilon_1}{\{(\omega_{ii} + \varepsilon_2)(\omega_{jj} + \varepsilon_3)\}^{1/2}},$$

where $\varepsilon_1 = -\omega_{i\cdot} - \omega_{j\cdot} + \omega_{\cdot\cdot}$, $\varepsilon_2 = -2\omega_{i\cdot} + \omega_{\cdot\cdot}$, and $\varepsilon_3 = -2\omega_{j\cdot} + \omega_{\cdot\cdot}$. By (A2) and (A3), we have $\varepsilon_i = O(p^{-1/2})$ $(i = 1, 2, 3)$. Therefore, by Condition 1,

$$\begin{aligned}\tau_{ij} &= \frac{\omega_{ij} + \varepsilon_1}{(\omega_{ii}\omega_{jj})^{1/2}}\left\{\frac{(\omega_{ii} + \varepsilon_2)(\omega_{jj} + \varepsilon_3)}{\omega_{ii}\omega_{jj}}\right\}^{-1/2} = \frac{\rho_{ij} + O(p^{-1/2})}{[\{1 + O(p^{-1/2})\}\{1 + O(p^{-1/2})\}]^{1/2}}\\ &= \rho_{ij} + O(p^{-1/2}),\end{aligned} \tag{A5}$$

which, together with Condition 2, implies (ii).

To show (iii), noting that $\tau_{ij}^2 - \rho_{ij}^2 = (\tau_{ij} - \rho_{ij})^2 + 2\rho_{ij}(\tau_{ij} - \rho_{ij})$ and using (A1) and (A5), we have

$$\sum_{i=1}^p(\tau_{ij}^2 - \rho_{ij}^2) = \sum_{i=1}^p(\tau_{ij} - \rho_{ij})^2 + 2\sum_{i=1}^p\rho_{ij}(\tau_{ij} - \rho_{ij}) = O(1) + 2\|R\|_1 O(p^{-1/2}) = O(1).$$

This and Condition 3 imply (iii) and complete the proof.

Proof of Proposition 2

We first write

$$Y_{ij}^{(k)} - \nu_{kj} = Z_{ij}^{(k)} - \mu_{kj} + \frac{1}{p}\sum_{j=1}^{p}(Z_{ij}^{(k)} - \mu_{kj}).$$

It follows from Condition 1 and Proposition 1 that

$$\frac{|Y_{ij}^{(k)} - \nu_{kj}|}{\gamma_{jj}^{1/2}} \leqslant \frac{\omega_{jj}^{1/2}}{\gamma_{jj}^{1/2}}\left(\frac{|Z_{ij}^{(k)} - \mu_{kj}|}{\omega_{jj}^{1/2}} + \frac{1}{p}\sum_{j=1}^{p}\frac{|Z_{ij}^{(k)} - \mu_{kj}|}{\omega_{jj}^{1/2}}\right) \leqslant 2(\kappa_1\kappa_2)^{1/2}\max_{i,j,k}\frac{|Z_{ij}^{(k)} - \mu_{kj}|}{\omega_{jj}^{1/2}}. \quad \text{(A6)}$$

Using Condition 4 and applying Markov's inequality and the union bound, we have

$$\text{pr}\left(\max_{i,j,k}|Z_{ij}^{(k)} - \mu_{kj}|/\omega_{jj}^{1/2} \geqslant t\right) \leqslant (n_1 + n_2)pK\exp(-\eta t^2)$$

for all $t > 0$. Hence, by Condition 5,

$$\max_{i,j,k}|Z_{ij}^{(k)} - \mu_{kj}|/\omega_{jj}^{1/2} = O_p\{(\log n + \log p)^{1/2}\} = o_p(n^{1/2}/\log p). \quad \text{(A7)}$$

Combining (A6) with (A7), we arrive at (12).

To prove (13), without loss of generality, we assume $\mu_1 = \mu_2 = 0$. Let $\hat{\gamma}_{jj}^{(k)}$ denote the sample centered log-ratio variances for population k ($k = 1, 2$). Note that

$$|\hat{\gamma}_{jj}^{(k)} - \gamma_{jj}| = |\hat{\omega}_{jj}^{(k)} - 2\hat{\omega}_{j\cdot}^{(k)} + \hat{\omega}_{\cdot\cdot}^{(k)} - (\omega_{jj} - 2\omega_{j\cdot} + \omega_{\cdot\cdot})| \leqslant 4\max_{i,j}|\hat{\omega}_{ij}^{(k)} - \omega_{ij}|.$$

It follows from Condition 1 and Proposition 1 that

$$\frac{|\hat{\gamma}_{jj}^{(k)} - \gamma_{jj}|}{\gamma_{jj}} \leqslant \frac{4}{\gamma_{jj}}\max_{i,j}\frac{|\hat{\omega}_{ij}^{(k)} - \omega_{ij}|}{(\omega_{ii}\omega_{jj})^{1/2}}(\omega_{ii}\omega_{jj})^{1/2} \leqslant 4\kappa_1\kappa_2\max_{i,j}\frac{|\hat{\omega}_{ij}^{(k)} - \omega_{ij}|}{(\omega_{ii}\omega_{jj})^{1/2}}.$$

The proof is completed by invoking the following lemma, which recaps a concentration result in Bickel & Levina (2008), and noting that $\hat{\gamma}_{jj} = n_1\hat{\gamma}_{jj}^{(1)}/(n_1 + n_2) + n_2\hat{\gamma}_{jj}^{(2)}/(n_1 + n_2)$.

LEMMA A1. *Under Condition* 4, *there exist constants* $C_1, C_2, C_3, C_4 > 0$ *such that*

$$\text{pr}\left\{\max_{i,j}\frac{|\hat{\omega}_{ij}^{(k)} - \omega_{ij}|}{(\omega_{ii}\omega_{jj})^{1/2}} \geqslant t\right\} \leqslant C_1 p\exp(-C_2 n_k t/2) + C_3 p^2\exp(-C_4 n_k t^2/4) \quad (t > 0;\ k = 1, 2).$$

Proof of Theorem 1

Let $t_p = t + 2\log p - \log\log p$ and

$$M_n^* = n\max_{1\leqslant j\leqslant p}\frac{(\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)})^2}{\gamma_{jj}}. \quad \text{(A8)}$$

We first show that under H_0 in (3) or equivalently (8), for any fixed $t \in \mathbb{R}$,

$$\text{pr}(M_n^* \leqslant t_p) \to \exp\{-\pi^{-1/2}\exp(-t/2)\} \quad \text{(A9)}$$

as $n, p \to \infty$. By the Bonferroni inequality, for any fixed integer m with $1 \leqslant m \leqslant p/2$,

$$\sum_{d=1}^{2m}(-1)^{d-1}\sum_{1\leqslant j_1<\cdots<j_d\leqslant p}\text{pr}\left(\bigcap_{k=1}^{d}E_{j_k}\right) \leqslant \text{pr}(M_n^* \geqslant t_p) \leqslant \sum_{d=1}^{2m-1}(-1)^{d-1}\sum_{1\leqslant j_1<\cdots<j_d\leqslant p}\text{pr}\left(\bigcap_{k=1}^{d}E_{j_k}\right),$$

where

$$E_j = \left\{n\frac{(\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)})^2}{\gamma_{jj}} \geqslant t_p\right\}.$$

Under H_0, we write

$$n^{1/2}\frac{\bar{Y}_j^{(1)}-\bar{Y}_j^{(2)}}{\gamma_{jj}^{1/2}}=\frac{n^{1/2}}{n_1}\sum_{i=1}^{n_1}\frac{Y_{ij}^{(1)}-\nu_{1j}}{\gamma_{jj}^{1/2}}-\frac{n^{1/2}}{n_2}\sum_{i=1}^{n_2}\frac{Y_{ij}^{(2)}-\nu_{2j}}{\gamma_{jj}^{1/2}}\equiv\sum_{i=1}^{n_1+n_2}\xi_{ij}.$$

By Proposition 2, it suffices to consider the event $\{\max_{i,j}|\xi_{ij}|\leqslant C_1(\log p)^{-1}\}$ for some constant $C_1>0$, which occurs with probability tending to 1. Let $N=(N_1,\ldots,N_p)^{\mathrm{T}}$ be multivariate normal with mean 0 and covariance matrix $nR/n_1+nR/n_2=R$. Applying Theorem 1.1 of Zaĭtsev (1987), for any sequence $\varepsilon_n=o(1)$, we have

$$\begin{aligned}\mathrm{pr}\left(\bigcap_{k=1}^{d}E_{j_k}\right)&=\mathrm{pr}\left(\min_{1\leqslant k\leqslant d}\left|\sum_{i=1}^{n_1+n_2}\xi_{ij_k}\right|\geqslant t_p^{1/2}\right)\\&\leqslant\mathrm{pr}\left(\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}-\varepsilon_n\right)+O\left\{d^{5/2}\exp\left(-C_2\frac{\log p}{d^3}\right)\right\}\\&\leqslant\mathrm{pr}\left(\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}-\varepsilon_n\right)+O(p^{-C_3}),\end{aligned}$$

and similarly,

$$\mathrm{pr}\left(\bigcap_{k=1}^{d}E_{j_k}\right)\geqslant\mathrm{pr}\left(\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}+\varepsilon_n\right)+O(p^{-C_3}).$$

Hence

$$\begin{aligned}\sum_{d=1}^{2m}(-1)^{d-1}\sum_{1\leqslant j_1<\cdots<j_d\leqslant p}&\mathrm{pr}\left\{\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}+(-1)^{d-1}\varepsilon_n\right\}+o(1)\\&\leqslant\mathrm{pr}(M_n^*\geqslant t_p)\\&\leqslant\sum_{d=1}^{2m-1}(-1)^{d-1}\sum_{1\leqslant j_1<\cdots<j_d\leqslant p}\mathrm{pr}\left\{\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}+(-1)^{d}\varepsilon_n\right\}+o(1).\end{aligned}$$

Then (A9) is proved by applying the following lemma, which follows from the same arguments as those for Lemma 6 of Cai et al. (2014), and letting $m\to\infty$.

LEMMA A2. *Under Conditions* 2 *and* 3, *we have*

$$\sum_{1\leqslant j_1<\cdots<j_d\leqslant p}\mathrm{pr}\left(\min_{1\leqslant k\leqslant d}|N_{j_k}|\geqslant t_p^{1/2}\pm\varepsilon_n\right)=\frac{1}{d!}\pi^{-d/2}\exp\left(-\frac{dt}{2}\right)\{1+o(1)\}.$$

Finally, consider the event $\{\max_j|\hat{\gamma}_{jj}-\gamma_{jj}|/\gamma_{jj}\leqslant C_4(\log p/n)^{1/2}\}$ for some constant $C_4>0$, which occurs with probability tending to 1 by Proposition 2. Then

$$|M_n-M_n^*|\leqslant n\max_{1\leqslant j\leqslant p}\frac{(\bar{Y}_j^{(1)}-\bar{Y}_j^{(2)})^2}{\hat{\gamma}_{jj}}\max_{1\leqslant j\leqslant p}\frac{|\hat{\gamma}_{jj}-\gamma_{jj}|}{\gamma_{jj}}\leqslant C_4M_n\left(\frac{\log p}{n}\right)^{1/2}=M_no\left(\frac{1}{\log p}\right)\tag{A10}$$

by Condition 5. This, together with (A9), completes the proof.

Proof of Theorem 2

In view of (A10) with M_n^* defined in (A8), it suffices to prove that, under H_1 in (14),

$$\mathrm{pr}(M_n^*\geqslant q_\alpha+2\log p-\log\log p)\to 1.\tag{A11}$$

By assumption, there exists some $j_0 \in S$ such that $|\delta_{j_0}| \geqslant \surd 2 + \varepsilon$. We write

$$n^{1/2}\frac{\bar{Y}^{(1)}_{j_0} - \bar{Y}^{(2)}_{j_0}}{\gamma^{1/2}_{j_0 j_0}} = n^{1/2}\frac{\bar{Y}^{(1)}_{j_0} - \nu_{1j_0}}{\gamma^{1/2}_{j_0 j_0}} - n^{1/2}\frac{\bar{Y}^{(2)}_{j_0} - \nu_{2j_0}}{\gamma^{1/2}_{j_0 j_0}} + n^{1/2}\frac{\nu_{1j_0} - \nu_{2j_0}}{\gamma^{1/2}_{j_0 j_0}} \equiv T_1 + T_2 + T_3.$$

Note that $T_1 = O_p(1)$ and $T_2 = O_p(1)$ by the central limit theorem. Define

$$T_4 = n^{1/2}\frac{\nu_{1j_0} - \nu_{2j_0}}{\omega^{1/2}_{j_0 j_0}}.$$

It follows from (A4) and Condition 1 that

$$|T_3 - T_4| = n^{1/2}\frac{|\nu_{1j_0} - \nu_{2j_0}|}{\gamma^{1/2}_{j_0 j_0}}\frac{|\gamma^{1/2}_{j_0 j_0} - \omega^{1/2}_{j_0 j_0}|}{\omega^{1/2}_{j_0 j_0}} = |T_3| O(p^{-1/2}).$$

Then, using (16) and the assumption $\|\delta\|_1 = o(p)$, we have

$$\begin{aligned} T_3 = T_4\{1 + O(p^{-1/2})\} &= n^{1/2}\frac{\delta_{j_0}\omega^{1/2}_{j_0 j_0} + o(1)}{\omega^{1/2}_{j_0 j_0}}\left(\frac{\log p}{n}\right)^{1/2}\{1 + O(p^{-1/2})\} \\ &= \{\delta_{j_0} + o(1)\}(\log p)^{1/2} \geqslant \left(\surd 2 + \frac{\varepsilon}{2}\right)(\log p)^{1/2} \end{aligned}$$

for sufficiently large p. Combining these bounds, we conclude that, with probability tending to 1,

$$n^{1/2}\frac{|\bar{Y}^{(1)}_{j_0} - \bar{Y}^{(2)}_{j_0}|}{\gamma^{1/2}_{j_0 j_0}} \geqslant (q_\alpha + 2\log p - \log\log p)^{1/2}.$$

This implies (A11) and completes the proof.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data (with Discussion). *J. R. Statist. Soc.* B **44**, 139–77.

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data.* Caldwell: Blackburn Press.

BAI, Z. & SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statst. Sinica* **6**, 311–29.

BICKEL, P. J. & LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.

CAI, T. T., LIU, W. & XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Statist. Soc.* B **76**, 349–72.

CHEN, S. X. & QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808–35.

LEWIS, J. D., CHEN, E. Z., BALDASSANO, R. N., OTLEY, A. R., GRIFFITHS, A. M., LEE, D., BITTINGER, K., BAILEY, A., FRIEDMAN, E. S., HOFFMANN, C., ALBENBERG, L., SINHA, R., COMPHER, C., GILROY, E., NESSEL, L., GRANT, A., CHEHOUD, C., LI, H., WU, G. D. & BUSHMAN, F. D. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host Microbe* **18**, 489–500.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Ann. Rev. Statist. Appl.* **2**, 73–94.

LIN, W., SHI, P., FENG, R. & LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–97.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. & HUTTENHOWER, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**, 811–814.

SRIVASTAVA, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Mult. Anal.* **100**, 518–32.

SRIVASTAVA, M. S. & DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Mult. Anal.* **99**, 386–402.

THE HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–14.

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A., AFFOURTIT, J. P., EGHOLM, M., HENRISSAT, B., HEATH, A. C., KNIGHT, R. & GORDON, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**, 480–4.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R., SINHA, R., GILROY, E., GUPTA, K., BALDASSANO, R., NESSEL, L., LI, H., BUSHMAN, F. D. & LEWIS, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–8.

ZAĬTSEV, A. YU. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Prob. Theory Rel. Fields* **74**, 535–66.