Supplementary File to "D-MANOVA: fast distance-based multivariate analysis of variance for large-scale microbiome association studies"

Jun Chen and Xianyang Zhang

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. Department of Statistics, Texas A&M University, College Station, TX, USA.

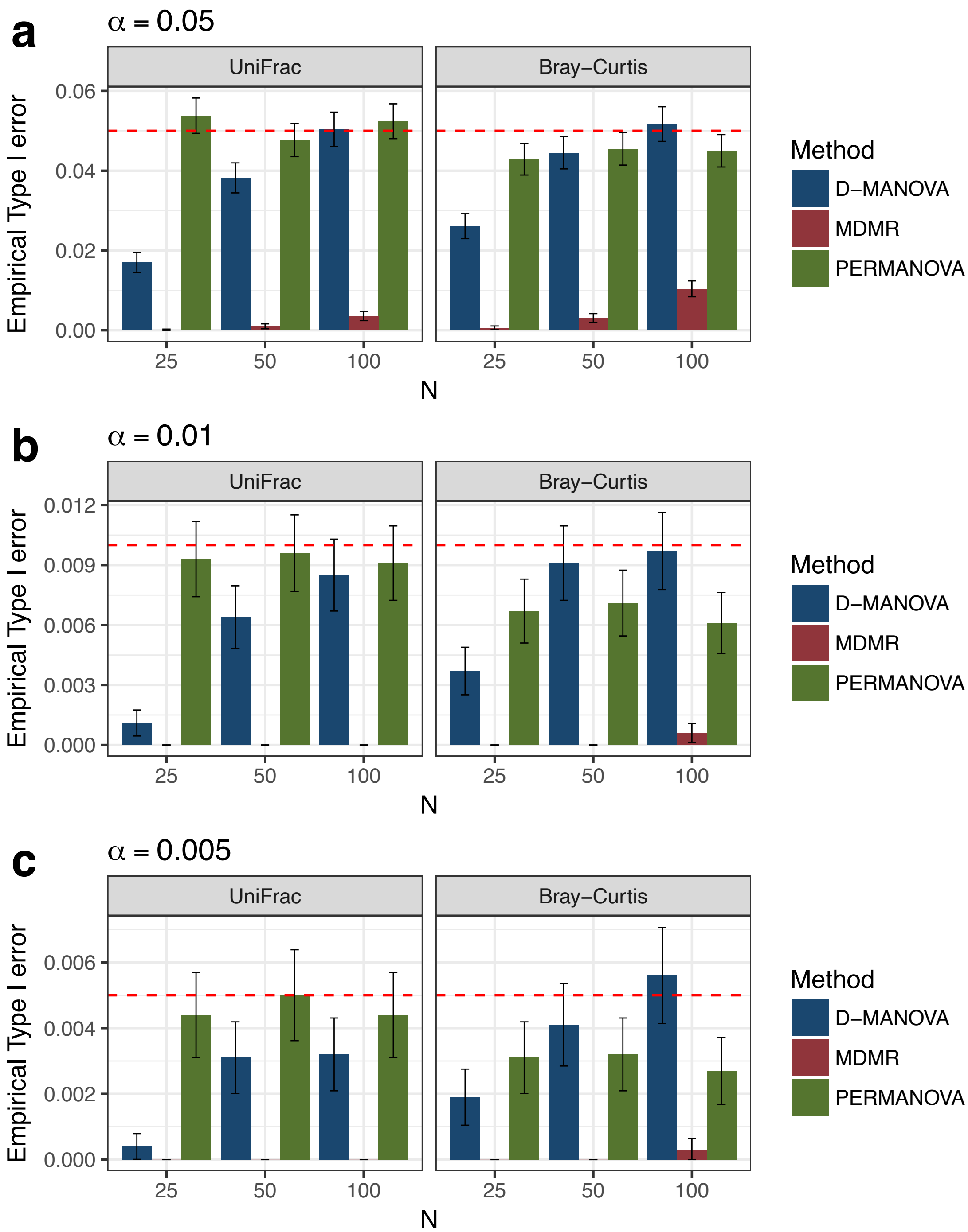

Figure S1: The empirical type I error rates of D-MANOVA, MDMR and PERMANOVA based on UniFrac and Bray-Curtis distances at different sample sizes (n=25, 50, 100) and varying α levels of 0.05 (**a**), 0.01 (**b**) and 0.005 (**c**). Simulation was repeated 10,000 times to calculate the empirical type I error. The error bar represents 95% confidence interval and the dashed line indicates the target α level.

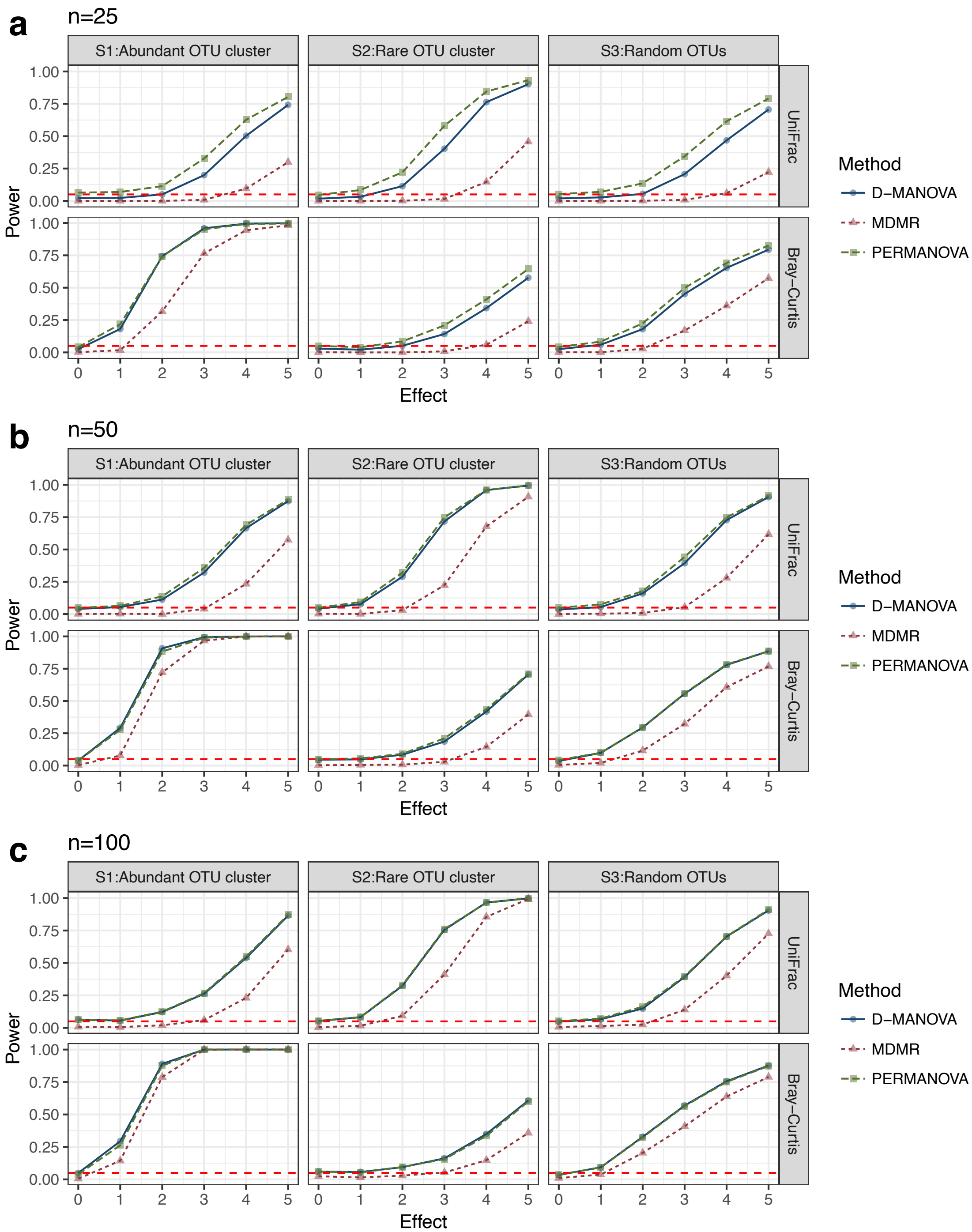

Figure S2: Power comparison of D-MANOVA, MDMR and PERMANOVA based on UniFrac and Bray-Curtis distances under different effect sizes (horizontal axis) and sample sizes (**a-c**). Three scenarios (Scene 1, Scene 2 and Scene 3), where the variable X affects an abundant OTU cluster, rare OTU cluster and random OTUs, respectively, were investigated. The power calculation was based on a nominal α level of 0.05 and a repetition of 1,000 simulation runs. The horizontal dashed line indicates the α level.

Table S1: P-values for testing the association of the gut microbiome with the demographic and lifestyle variables based on the American Gut dataset. Bray-Curtis distance was used. The runtime is expressed relative to the D-MANOVA. The computation was performed under R v3.3.2 on an iMAC (3.2 GHz Intel Core i5, 32 GB 1600 MHz DDR3, EI Capitan v10.11.5).

	R^{2*}	D-MANOVA	MDMR	PERMANOVA
Sex	0.29%	1.46E-112	0	<0.001
Age	0.27%	8.70E-100	0	<0.001
Race	0.21%	1.31E-45	1.89E-15	<0.001
Exercise frequency	0.17%	6.86E-58	0	<0.001
BMI	0.12%	1.28E-37	0	<0.001
Water source	0.11%	2.03E-18	3.72E-05	<0.001
Alcohol frequency	0.10%	5.73E-30	0	<0.001
Diet type	0.07%	5.51E-17	9.89E-13	<0.001
Tabacco frequency	0.04%	1.90E-09	1.15E-07	<0.001
Sleep duration	0.03%	1.72E-06	1.01E-05	<0.001
C-section	0.03%	1.33E-06	1.23E-05	<0.001
Dog as pet	0.03%	2.26E-05	9.65E-05	<0.001
Handness	0.02%	0.646	0.841	0.644
Runtime	-	1	$\times$12.7	$\times$567.4

*R^2 is the percent of variation explained by a variable, where the variability is summarized by pairwise distances.

Supplementary Note 1. Proof of Theorem 2.1

Let $\mathcal{H}$ be a Hilbert space equipped with the inner product $< \cdot, \cdot >$ and the inner product induced norm $\| \cdot \|$. Assume that

$$d_{ij}^2 = \|\phi(Y_i) - \phi(Y_j)\|^2, \tag{1}$$

where $\phi(\cdot) : \mathcal{Y} \to \mathcal{H}$ is an embedding from $\mathcal{Y}$ to $\mathcal{H}$. Define $\Phi = (\phi(Y_1), \dots, \phi(Y_n))^\top \in \mathcal{H}^{\otimes n}$ with $\mu = E\phi(Y_1)$ and $\mathcal{H}^{\otimes n}$ being the n-ary Cartesian power of $\mathcal{H}$. For $f = (f_1, \dots, f_n)^\top, g = (g_1, \dots, g_n)^\top \in \mathcal{H}^{\otimes n}$, let $< f, g >= \sum_{i=1}^n < f_i, g_i >$ and $\|f\|^2 = \sum_{i=1}^n \|f_i\|^2$. Define

$$f \circ g^\top = \begin{pmatrix} < f_1, g_1 > & < f_1, g_2 > & \cdots & < f_1, g_n > \\ \cdots & \cdots & \cdots & \cdots \\ < f_n, g_1 > & < f_n, g_2 > & \cdots & < f_n, g_n > \end{pmatrix},$$

and we have $G = D\Phi \circ \Phi^\top D$. We assume that $\mathbf{1}$ is contained in the column space of Z, which implies that $H^{X|Z} D = H^{X|Z}$ and $H^{I|X,Z} D = H^{I|X,Z}$. Consider the linear model,

$$\Phi = XB + ZA + E,$$

where $B \in \mathcal{H}^{\otimes p_1}$, $A \in \mathcal{H}^{\otimes p_2}$ and $E = (e_1, \dots, e_n)^\top \in \mathcal{H}^{\otimes n}$. Here $e_1, \dots, e_n$ are independent mean-zero random variables in $\mathcal{H}$, which are independent of X and Z. Note that

$$H^{X|Z}\Phi = H^{X|Z} XB + H^{X|Z} E.$$

Under the null $B = 0$, we have $H^{X|Z}\Phi = H^{X|Z} E$. In this case, we get

$$\mathrm{tr}(H^{X|Z} G H^{X|Z}) = \mathrm{tr}(H^{X|Z}\Phi \circ \Phi^\top H^{X|Z}) = \mathrm{tr}(H^{X|Z} E \circ E^\top H^{X|Z}) = \sum_{j,k=1}^n h_{jk} K(e_j, e_k), \tag{2}$$

where $K(e_j, e_k) =< e_j, e_k >$. By Mercer's theorem, K is semi-positive definite and thus admits the spectral decomposition of the form

$$K(e_j, e_k) = \sum_{l=1}^{+\infty} \lambda_l \psi_l(e_j)\psi_l(e_k), \tag{3}$$

where $\mathbb{E}[\psi_s(e_i)\psi_l(e_i)] = \mathbf{1}\{s = l\}$ and $\mathbb{E}[\psi_l(e_i)] = 0$. Based on the setup above, we have the following theorem.

Theorem 0.1. *Assume that* $\mathbb{E}\|e_1\|^4 < \infty$ *and*

$$\|H^{X|Z}\|_{2,4} = \sup_{a:\|a\|_2=1} \|H^{X|Z} a\|_4 \to 0. \tag{4}$$

Then under the null,

$$\frac{tr(H^{X|Z} G H^{X|Z})/m_1}{tr(H^{I|X,Z} G H^{I|X,Z})/(n - m_2)} \to^d T_0 = \frac{\sum_{l=1}^{+\infty} \lambda_l \chi^2_{m_1,l}/m_1}{\sum_{l=1}^{+\infty} \lambda_l},$$

where $\{\chi^2_{m_1,l}\}_{l=1}^{+\infty}$ *are independent chi-square random variables with* m_1 *degrees of freedom.*

Proof. Suppose $H^{X|Z} = (\zeta_{ij})$ admits the spectral decomposition $H^{X|Z} = U^\top U$ with $U =$

$(u_1, \ldots, u_{m_1})^\top = (u_{ij}) \in \mathbb{R}^{m_1 \times n}$ whose rows (i.e., u_is) are the eigenvectors of $H^{X|Z}$. Here U is only defined up to an $m_1 \times m_1$ orthonormal transformation. Condition (4) implies that

$$\|U\|_4 := (\sum_{i=1}^{m_1} \sum_{j=1}^{n} u_{ij}^4)^{1/4} \to 0, \tag{5}$$

which does not depend on the choice of eigenvectors. To see this, let $L = (L_{ij}) \in \mathbb{R}^{m_1 \times m_1}$ be an orthonormal matrix. Note that for any $1 \le i \le m$,

$$\|\sum_{i=1}^{m} L_{ji} u_i\|_4 \le \sum_{i=1}^{m} |L_{ji}| \|u_i\|_4 \to 0,$$

which implies that $\|LU\|_4 \to 0$.

In view of (2) and (3), we have

$$\mathrm{tr}(H^{X|Z} G H^{X|Z}) = \sum_{l=1}^{+\infty} \lambda_l \sum_{i=1}^{m_1} V_{l,i,n}^2$$

where $V_{l,i,n} = \sum_{j=1}^{n} u_{ij} \psi_l(e_j)$. Note that

$$\begin{aligned}
\lim_n \mathrm{cov}(V_{l,i,n}, V_{l',i',n}) =& \lim_n \sum_{j,j'=1}^{n} u_{ij} u_{i'j'} \mathbb{E} \psi_l(e_j) \psi_{l'}(e_{j'}) \\
=& \lim_n \sum_{j=1}^{n} u_{ij} u_{i'j} \mathbb{E} \psi_l(e_j) \psi_{l'}(e_j) \\
=& \mathbf{1}\{l = l', i = i'\}.
\end{aligned}$$

Under the assumption $\mathbb{E}\|\varphi(e_1)\|^4 < \infty$, we have

$$\mathbb{E} K(e_1, e_1)^2 = \mathbb{E} \left(\sum_l \lambda_l \psi_l(e_1)^2 \right)^2 < \infty,$$

which implies $\mathbb{E}[\psi_l(e_1)^4] < \infty$ for any l with $\lambda_l \neq 0$. Together with (5), the Lyapunov condition is satisfied and thus $(V_{l,i,n})_{1 \le l \le K, 1 \le i \le m_1}$ for any finite K converges to a multivariate normal distribution say $(V_{l,i})_{1 \le l \le K, 1 \le i \le m_1}$ by the Cramér-Wold device, where $\mathrm{cov}(V_{l,i}, V_{l',i'}) = \mathbf{1}\{l = l', i = i'\}$.

Denote $V_n(K) = \sum_{l=1}^{K} \lambda_l \sum_{i=1}^{m_1} V_{l,i,n}^2$ and define $V(K)$ in the same way by replacing $V_{l,i,n}$ with $V_{l,i}$. We aim to show that

$$V_n(\infty) \to^d V(\infty). \tag{6}$$

In view of Theorem 8.6.2 of Resnick (1999), we only need to show

(A) $V_n(K) \to^d V(K)$ for any K;

(B) $\mathbb{E}|V(\infty) - V(K)|^2 \to 0$ as $K \to +\infty$;

(C) $\lim_{K \to +\infty} \lim_{n \to +\infty} \mathbb{E}|V_n(\infty) - V_n(K)|^2 = 0$.

(A) follows from the finite dimensional convergence and the continuous mapping theorem. To show

(B), we note that

$$\mathbb{E}|V(\infty)-V(K)|^2=\mathbb{E}\left(\sum_{l=K+1}^{+\infty}\lambda_l\chi^2_{m_1,l}\right)^2=m_1^2\left(\sum_{l=K+1}^{+\infty}\lambda_l\right)^2+2m_1\sum_{l=K+1}^{+\infty}\lambda_l^2\to 0,$$

where we have used the fact that $\sum_{l=1}^{+\infty}\lambda_l<\infty$. Some algebra yields that

$$\begin{aligned}
&\sum_{i,i'=1}^{m_1}\mathrm{cov}(V^2_{l,i,n},V^2_{l',i',n})\\
=&\mathrm{cov}(\psi_l(e_1)^2,\psi_{l'}(e_1)^2)\sum_{j=1}^{n}\sum_{i,i'=1}^{m_1}u_{ij}^2u_{i'j}^2+2\mathrm{cov}(\psi_l(e_1)\psi_l(e_2),\psi_{l'}(e_1)\psi_{l'}(e_2))\sum_{j\neq j'}\sum_{i,i'=1}^{m_1}u_{ij}u_{i'j}u_{ij'}u_{i'j'}\\
=&\mathrm{cov}(\psi_l(e_1)^2,\psi_{l'}(e_1)^2)\sum_{j=1}^{n}\zeta_{jj}^2+2\mathrm{cov}(\psi_l(e_1)\psi_l(e_2),\psi_{l'}(e_1)\psi_{l'}(e_2))\sum_{j\neq j'}\zeta_{jj'}^2\\
\le&C_1\sum_{i,j}\zeta_{ij}^2=C_1m_1,
\end{aligned}$$

for some constant $C_1>0$. Using this result, we have

$$\begin{aligned}
\mathbb{E}|V_n(K)-V_n(\infty)|^2=&\mathbb{E}\left(\sum_{l=K+1}^{+\infty}\lambda_l\sum_{i=1}^{m_1}V^2_{l,i,n}\right)^2\\
\le&2m_1^2(\mathbb{E}V^2_{l,i,n})^2\left(\sum_{l=K+1}^{+\infty}\lambda_l\right)^2+2\mathbb{E}\left\{\sum_{l=K+1}^{+\infty}\lambda_l\sum_{i=1}^{m_1}(V^2_{l,i,n}-\mathbb{E}V^2_{l,i,n})\right\}^2\\
\le&2m_1^2(\mathbb{E}V^2_{l,i,n})^2\left(\sum_{l=K+1}^{+\infty}\lambda_l\right)^2+2\sum_{l,l'=K+1}^{+\infty}\lambda_l\lambda_{l'}\sum_{i,i'=1}^{m_1}\mathrm{cov}(V^2_{l,i,n},V^2_{l',i',n})\\
\le&2m_1^2(\mathbb{E}V^2_{l,i,n})^2\left(\sum_{l=K+1}^{+\infty}\lambda_l\right)^2+2C_1m_1\left(\sum_{l=K+1}^{+\infty}\lambda_l\right)^2\to 0.
\end{aligned}$$

Thus (C) holds as well.

To deal with the denominator of the statistic, we note that

$$\mathrm{tr}(H^{I|X,Z}GH^{I|X,Z})=\sum_{i=1}^{n-m_2}\left\|\sum_{j=1}^{n}r_{ij}\varphi(e_j)\right\|^2=\sum_{i=1}^{n-m_2}\sum_{j,k=1}^{n}r_{ij}r_{ik}K(e_j,e_k),$$

where we assume $H^{I|X,Z}=(h_{ij})$ has the spectral decomposition $R'R$ with $R=(r_{ij})\in\mathbb{R}^{(n-m_2)\times n}$. Note that

$$\frac{1}{n-m_2}\mathbb{E}\mathrm{tr}(H^{I|X,Z}GH^{I|X,Z})=\mathbb{E}K(e_1,e_1),$$

and

$$
\begin{aligned}
&\frac{1}{(n-m_2)^2}\mathrm{var}\left(\mathrm{tr}(H^{I|X,Z}GH^{I|X,Z})\right)\\
=&\frac{1}{(n-m_2)^2}\sum_{i,i'=1}^{n-m_2}\sum_{j,k,j',k'=1}^{n} r_{ij}r_{ik}r_{i'j'}r_{i'k'}\mathrm{cov}(K(e_j,e_k),K(e_{j'},e_{k'}))\\
=&\frac{\mathrm{var}(K(e_1,e_1))}{(n-m_2)^2}\sum_{i,i'=1}^{n-m_2}\sum_{j=1}^{n} r_{ij}^2 r_{i'j}^2+\frac{2\mathrm{var}(K(e_1,e_2))}{(n-m_2)^2}\sum_{i,i'=1}^{n-m_2}\sum_{j\neq k} r_{ij}r_{ik}r_{i'j}r_{i'k}\\
=&\frac{\mathrm{var}(K(e_1,e_1))}{(n-m_2)^2}\sum_{j=1}^{n} h_{jj}^2+\frac{2\mathrm{var}(K(e_1,e_2))}{(n-m_2)^2}\sum_{j\neq k} h_{jk}^2\\
\leq&\frac{C'}{(n-m_2)^2}\sum_{j,k} h_{j,k}^2=\frac{C'}{n-m_2}\to 0,
\end{aligned}
$$

where $C'>0$. Thus by the law of large numbers,

$$
\frac{1}{n-m_2}\mathrm{tr}(H^{I|X,Z}GH^{I|X,Z})\to^p \mathbb{E}K(e_1,e_1)=\sum_{l=1}^{+\infty}\lambda_l. \tag{7}
$$

The conclusion thus follows from (6), (7), and the Slutsky's theorem. $\square$

Supplementary Note 2: derivation of the chi-square approximation

The idea of the chi-square approximation is to match the first two moments of the chi-square distribution with those of T_0. To this end, we note that $\mathbb{E}[T_0] = 1$ and the variance of $m_1 T_0$ is equal to

$$\mathrm{var}(m_1 T_0) = \frac{2m_1 \sum_{l=1}^{+\infty} \lambda_l^2}{(\sum_{l=1}^{+\infty} \lambda_l)^2} = \frac{2m_1 \mathbb{E}K(e_1, e_2)^2}{(\mathbb{E}K(e_1, e_1))^2} = \frac{2m_1}{p}$$

with $p = (\mathbb{E}K(e_1, e_1))^2 / \mathbb{E}K(e_1, e_2)^2$. Therefore,

$$\mathbb{E}(pm_1 T_0) = pm_1 \quad \text{and} \quad \mathrm{var}(pm_1 T_0) = 2m_1 p.$$

Note that

$$H^{I|X,Z}\Phi = H^{I|X,Z}E.$$

Suppose $\widetilde{G} = (\tilde{g}_{ij}) = H^{I|X,Z} G H^{I|X,Z}$ with $H^{I|X,Z} = (h_{ij})$. Then we have

$$\widetilde{G} = H^{I|X,Z} E \circ E^\top H^{I|X,Z}.$$

We can estimate $\mathbb{E}K(e_1, e_1)$ by

$$\widehat{\mu}_1 = \frac{1}{n - m_2} \mathrm{tr}(\widetilde{G}).$$

To estimate $\mathbb{E}K(e_1, e_2)^2$, we note that

$$\begin{aligned}
\sum_{i \neq k} \mathbb{E}\tilde{g}_{ik}^2 =& \sum_{i \neq k} \mathbb{E} \left(\sum_{j_1, j_2} h_{i,j_1} h_{k,j_2} K(e_{j_1}, e_{j_2}) \right)^2 \\
=& \sum_{i \neq k} \mathbb{E} \sum_{j_1, j_2, j_3, j_4} h_{i,j_1} h_{k,j_2} h_{i,j_3} h_{k,j_4} K(e_{j_1}, e_{j_2}) K(e_{j_3}, e_{j_4}) \\
=& \mathbb{E}K(e_1, e_2)^2 \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1}^2 h_{k,j_2}^2 + \mathbb{E}K(e_1, e_2)^2 \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1} h_{k,j_1} h_{i,j_2} h_{k,j_2} \\
&+ \{\mathbb{E}K(e_1, e_1)\}^2 \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1} h_{k,j_1} h_{i,j_2} h_{k,j_2} + \mathbb{E}K(e_1, e_1)^2 \sum_{i \neq k} \sum_{j_1} h_{i,j_1}^2 h_{k,j_1}^2 \\
=& \mathbb{E}K(e_1, e_2)^2 \left\{ \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1}^2 h_{k,j_2}^2 + \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1} h_{k,j_1} h_{i,j_2} h_{k,j_2} \right\} \\
&+ \{\mathbb{E}K(e_1, e_1)\}^2 \sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1} h_{k,j_1} h_{i,j_2} h_{k,j_2} + \mathbb{E}K(e_1, e_1)^2 (\sum_j h_{jj}^2 - \sum_{i,j} h_{ij}^4).
\end{aligned}$$

where the last three terms are of smaller order $O(n)$. Thus a natural estimator for $\mathbb{E}K(e_1, e_2)^2$ would be

$$\widehat{\mu}_2 = \frac{\sum_{i \neq k} \tilde{g}_{ik}^2}{\sum_{i \neq k} \sum_{j_1 \neq j_2} h_{i,j_1}^2 h_{k,j_2}^2} = \frac{\sum_{i \neq k} \tilde{g}_{ik}^2}{(n - m_2)^2 + \sum_{i,j} h_{i,j}^4 - 2 \sum_i h_{ii}^2}.$$

We then estimate p by

$$\widehat{p} = \frac{\widehat{\mu}_1^2}{\widehat{\mu}_2}.$$

Therefore, we can approximate the distribution of pm_1T_0 by $\chi^2_{\hat{p}m_1}$.

Supplementary Note 3: Simulation setup

We study the type I error control and power (i.e., the probability of rejecting the null hypothesis under the alternative) using simulations. We simulate a covariate of interest (X) and a confounder (Z), which are bivariate normally distributed with mean 0, sd 1 and correlation 0.5. We use the Dirichlet distribution to simulate the baseline microbiome composition, following the same strategy as described in [2]. The parameters of the Dirichlet distribution were estimated based on a human upper respiratory microbiome dataset (60 subjects, 856 OTUs) [1], which can be accessed in the R *GUniFrac* package. Next, we let X and Z affect the abundances of a subset of OTUs. Depending on how the affected OTUs are distributed on the phylogenetic tree, we study three scenarios: Scene 1. X and Z affect a cluster of abundant OTUs (38 OTUs, 11.9% of total abundance), Scene 2. X and Z affect a cluster of rare OTUs (42 OTUs, 2.6% of total abundance), and Scene 3. X and Z affect 39 OTUs randomly distributed on the tree. The OTU clusters are formed by applying the Partitioning Around Medoid algorithm (20 clusters) based on the patristic distances among OTUs. For those affected OTUs, we apply a fold change of $e^{aX+0.5Z}$ to their proportions. We vary the coefficient a to create different levels of signal strength. The null situation is simulated by setting $a = 0$. Finally, we normalize the proportion data to sum one and generate the counts using the multinomial distribution with a sequencing depth of 10,000. We calculate the UniFrac and Bray-Curtis (BC) distances, two most widely used distance metrics, based on the OTU count data and the phylogenetic tree. We compare the proposed method (D-MANOVA, *dmanaova* function in R *GUniFrac* package) to PERMANOVA (999 permutations, *adonis* function in R *vegan* package) and MDMR (*mdmr* function, R *MDMR* package) based on these distance matrices.

References

[1] Charlson, E.S., Chen, J., Custers-Allen, R., et al. (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers, *PloS one*, **5**, e15216.

[2] Chen, J., Bittinger, K., Charlson, E.S., et al. (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances, *Bioinformatics*, **28**, 2106-2113.