Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab498
Advance Access Publication Date: 13 July 2021
Applications Note

Genetic and population analysis

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

Jun Chen [1,*] **and Xianyang Zhang**[2,*]

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55901, USA and [2]Department of Statistics, Texas A&M University, College Station, TX 77840, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on May 3, 2021; revised on June 20, 2021; editorial decision on June 30, 2021; accepted on July 6, 2021

Abstract

Summary: PERMANOVA (permutational multivariate analysis of variance based on distances) has been widely used for testing the association between the microbiome and a covariate of interest. Statistical significance is established by permutation, which is computationally intensive for large sample sizes. As large-scale microbiome studies, such as American Gut Project (AGP), become increasingly popular, a computationally efficient version of PERMANOVA is much needed. To achieve this end, we derive the asymptotic distribution of the PERMANOVA pseudo-F statistic and provide analytical *P*-value calculation based on chi-square approximation. We show that the asymptotic *P*-value is close to the PERMANOVA *P*-value even under a moderate sample size. Moreover, it is more accurate and an order-of-magnitude faster than the permutation-free method MDMR. We demonstrated the use of our procedure D-MANOVA on the AGP dataset.

Availability and implementation: D-MANOVA is implemented by the *dmanova* function in the CRAN package *GUniFrac*.

Contact: chen.jun2@mayo.edu or zhangxiany@stat.tamu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the past decade, numerous microbiome studies have been conducted to elucidate the role of the human microbiome in health and disease, generating an enormous amount of microbiome sequencing data (Kashyap *et al.*, 2017). Microbiome data have complex structures including zero-inflation, skewed abundance distribution and phylogenetic relatedness among features. To address these statistical challenges, one popular approach summarizes the microbiome data in the form of pairwise distances and statistical analyses are then performed based on the distance matrices (Chen *et al.*, 2012). One widely used distance-based method is PERMANOVA (permutational multivariate analysis of variance based on distances), which aims to identify covariates that could significantly explain the inter-subject variability captured by the pairwise distances (McArdle and Anderson, 2001). As a key component in microbiome data analysis, PERMANOVA has been routinely used in establishing an overall association between the microbiome and a covariate of interest. PERMANOVA uses permutation to assess the statistical significance and could be extremely slow at a large sample size. For example, running a single-threaded instance with 1000 permutations on a sample size of 5000 takes ~1 h on a desktop computer. In practice, many hypotheses may be tested and more permutations are needed to assess a lower Type I error level, further exacerbating the computational burden. Although methods exist for estimating the tail probability of permutation tests (Knijnenburg *et al.*, 2009), an analytical method, an analytical method, which accurately approximates the PERMANOVA *P*-value without permutation, is highly desirable. Recently, McArtor *et al.* (2017) proposed the MDMR method for analytical *P*-value calculation based on the asymptotic distribution of the PERMANOVA pseudo-F statistic. However, no rigorous proof was given. In addition, we found that MDMR could be conservative under many settings. Here, we rigorously derive the asymptotic distribution of the pseudo-F statistic, which is different from the one used in MDMR, and provide an accurate chi-square approximation. We show that our approach, D-MANOVA, provides more accurate approximation than MDMR and is also an order-of-magnitude faster.

2 Materials and methods

Suppose we have n subjects, p_1 variables of interest and p_2 covariates we want to adjust. Let $X \in \mathbb{R}^{n\times p_1}$ and $Z \in \mathbb{R}^{n\times p_2}$ be the design matrices for the variables of interest and the covariates, respectively. Define $H^{X,Z}$ and H^{Z} as the projection matrices onto the corresponding column spaces. Further let $H^{X|Z} = H^{X,Z} - H^{Z}$ and $H^{I|X,Z} = I_n - H^{X,Z}$ with $I_n \in \mathbb{R}^{n\times n}$ being the $n \times n$ identity matrix, $\text{rank}(H^{X|Z}) = m_1$ and

rank$(H^{I|X,Z}) = n - m_2$. Let $\{Y_i\}_{i=1}^n$ be the responses, which belong to a metric space denoted by $(\mathcal{Y}, d)$, and $d_{ij} = d(Y_i, Y_j)$ be the pairwise distance. Denote $A = (-d_{ij}^2/2) \in R^{n\times n}$. We define G as the Gower's centered matrix

$$G = \left(I_n - \frac{11^\top}{n}\right) A \left(I_n - \frac{11^\top}{n}\right) = DAD,$$

where $1 \in \mathbb{R}^{n\times 1}$ is the vector of all 1 s and $D = I_n - 11^\top/n$. The distance-based pseudo-F statistic is defined as

$$T = \frac{\text{tr}(H^{X|Z} G H^{X|Z})/m_1}{\text{tr}(H^{I|X,Z} G H^{I|X,Z})/(n-m_2)}, \quad (1)$$

where tr$(\cdot)$ denotes the trace of a matrix. The pseudo-F statistic is the basis for distance-based multivariate analysis of variance and quantifies the association between the multivariate Y, whose variability is encoded in the distance matrix, and the covariate of interest X while adjusting other covariates Z. Compared to the classic F-statistic for linear models, the distribution of the distance-based pseudo-F statistic is unknown and permutation, as implemented in PERMANOVA, is usually employed to obtain the P-value. To obtain an analytical P-value without permutation, McArtor *et al.* (2017) proposed an asymptotic null distribution for the pseudo-F statistic. However, no rigorous theoretical proof for their asymptotic null distribution was given. Here, we fill this gap and derive a more accurate asymptotic null distribution. Let $\mathcal{H}$ be a Hilbert space equipped with the inner product $< \cdot, \cdot >$ and the inner product induced norm $||\cdot||$. Assume that

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

where $\phi(\cdot) : \mathcal{Y} \to \mathcal{H}$ is an embedding from $\mathcal{Y}$ to $\mathcal{H}$. Define $\Phi = (\phi(Y_1), \ldots, \phi(Y_n))^\top \in \mathcal{H}^{\otimes n}$ with $\mathcal{H}^{\otimes n}$ being the n-ary Cartesian power of $\mathcal{H}$. Then, the distance-based multivariate analysis of variance can be re-formulated in the linear model

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

where $B \in \mathcal{H}^{\otimes p_1}$, $A \in \mathcal{H}^{\otimes p_2}$ and $E = (e_1, \ldots, e_n)^\top \in \mathcal{H}^{\otimes n}$. Here, $e_1, \ldots, e_n$ are independent mean-zero random variables in $\mathcal{H}$, which are independent of X and Z. Let $K(e_j, e_k) = < e_j, e_k >$. By Mercer's theorem, K is semi-positive definite and thus admits the spectral decomposition of the form $K(e_j, e_k) = \sum_{l=1}^{+\infty} \lambda_l \psi_l(e_j)\psi_l(e_k)$, where $\mathbb{E}[\psi_s(e_i)\psi_l(e_i)] = 1\{s = l\}$ and $\mathbb{E}[\psi_l(e_i)] = 0$. Based on this setup, we have the following theorem, whose proof is given in Supplementary Note S1.

Theorem 2.1 *Assume that* $\mathbb{E}||e_1||^4 < \infty$ *and*

$$||H^{X|Z}||_{2,4} = \sup_{a:||a||_2=1} ||H^{X|Z} a||_4 \to 0. \quad (3)$$

Then under the null,

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

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

Theorem 2.1 shows that as $n \to +\infty$, the distance-based pseudo-F statistic converges to a weighted sum of independent chi-squared random variables. As the weights are unknown, the limiting distribution is non-pivotal. Here, we develop a chi-square approximation, which has a computational complexity $O(n^2)$ and also provides accurate enough approximation. The idea is to match the first two moments of the chi-square distribution with those of T_0. Suppose $p = (\mathbb{E}K(e_1,e_1))^2/\mathbb{E}K(e_1,e_2)^2$, $\tilde{G} = (\tilde{g}_{ij}) = H^{I|X,Z} G H^{I|X,Z}$ with $H^{I|X,Z} = (h_{ij})$. Based on the derivation detailed in Supplementary Note S2, the distribution of T_0 can be approximated by $\frac{1}{pm_1}\chi^2_{pm_1}$, where

$$\hat{p} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2}, \hat{\mu}_1 = \frac{1}{n-m_2}\text{tr}(\tilde{G}), \hat{\mu}_2 = \frac{\sum_{i\neq k}\tilde{g}_{ik}^2}{(n-m_2)^2 + \sum_{i,j} h_{i,j}^4 - 2\sum_i h_{ii}^2}.$$

We implemented D-MANOVA by the *dmanova* function inour *GUniFrac* package (Chen *et al.*, 2012). To facilitate its use, the interface and the output are similar to those of the *adonis* function in the CRAN *vegan* package.

3 Results

We conduct simulations (Supplementary Note S3) to study the performance of D-MANOVA, comparing to PERMANOVA and MDMR. Figure 1a compares the P-values of D-MANOVA and PERMANOVA on the log scale [$n = 100$, Bray–Curtis (BC) distance, Scenario 3 in Supplementary Note S3] based on 1000 simulation runs under the null (H_0, left) and the alternative (H_1, right). We can see that D-MANOVA and PERMANOVA, P-values are highly correlated under both H_0 and H_1. Since the lowest P-value is 0.001 for PERMANOVA with 999 permutations, we see a large number of 0.001 under H_1 while D-MANOVA has no such restriction. Figure 1b compares the performance of the three competing methods under sample sizes of 100, 200 and 500 based on the BC distance. Under H_0 (first point of the power curve), all the methods control the Type I error under the nominal level with MDMR being more conservative. In terms of statistical power, D-MANOVA almost achieves the same power as PERMANOVA, while MDMR is less powerful under $n = 100$ and 200. The conservativeness has also been noted by the MDMR authors, and they do not recommend to run MDMR on sample sizes <200. However, even under $n = 500$, we still observe some power loss, indicating the approximation of D-MANOVA is more accurate. It is interesting to study the performance of D-MANOVA under small sample sizes. We thus repeat the simulations at $n = 25$ and 50. Supplementary Figure S1 shows that the Type I error of D-MANOVA is well controlled at different α

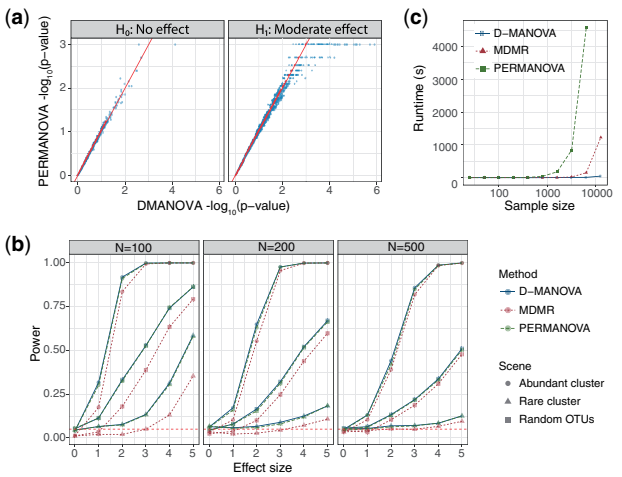

Fig. 1. Performance comparison of D-MANOVA, MDMR and PERMANOVA (999 permutations) based on simulations. Bray-Curtis distance was used. (**a**) Scatter plots comparing the P-values of D-MANOVA and PERMANOVA on the log scale under the null (H_0) and alternative (H_1). (**b**) Power comparison at sample sizes 100, 200 and 500. Simulation was averaged over 1000 runs. (**c**) Runtime comparison at varying sample sizes ($n = 50, 100, \ldots, 6400, 12800$). Runtimes were averaged over three repetitions. 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)

levels and the size is closer to the nominal level as the sample size increases. Supplementary Figure S2 shows that the power of D-MANOVA is close to that of PERMANOVA even at $n = 25$. MDMR, on the other hand, is substantially less powerful under small sample sizes. We also compare the average computation time of the three methods at different sample sizes (Fig. 1c). At $n = 12\ 800$, PERMANOVA could not complete the analysis in hours while MDMR takes around 20 min. In contrast, D-MANOVA uses less than one minute. Therefore, D-MANOVA significantly improves over MDMR in terms of both accuracy and computational efficiency.

We finally demonstrate the use of D-MANOVA using the publicly available dataset (figshare doi:10.6084/m9.figshare.6137198) from the American Gut Project (AGP) (McDonald *et al.*, 2018). We aim to test the association of the demographic and lifestyle variables with the gut microbiome composition based on the BC distance. We focus the analysis on the American and European populations with an age range between 18 and 80. A total of 7730 subjects were included in the analysis. The country residence was adjusted when testing the associations. Supplementary Table S1 shows the D-MANOVA, MDMR and PERMANOVA association *P*-values for these demographic/lifestyle variables ordered by effect sizes as measured by the distance-based R^2. Due to the large sample size, all the variables except the 'handness' are found to be significantly associated with the gut microbiome composition. For those significant variables, PERMANOVA *P*-values are all <0.001, so more permutations are needed to produce accurate p-values. For the 'handness' variable, D-MANOVA achieves a similar *P*-value as PERMANOVA. In contrast, MDMR tends to produce larger *P*-values, consistent with the conservativeness noted in the simulations. In terms of computational speed, D-MANOVA is about 13 times faster than MDMR and 567 times faster than PERMANOVA.

Simulations demonstrated that D-MANOVA had good type I error control at the 0.005 level, which should suffice for most community-level analyses since the number of tests is usually limited. However, when an extremely small type I error rate is needed to account for testing thousands or even millions of hypotheses, we recommend using our procedure to filter out most insignificant hypotheses and those hypotheses with extremely small *P*-values can be further validated by permutation. As the sample size increases, the detectable effect sizes become much smaller and statistical significance from community-level analyses may have limited practical utility. In such case, lower-level analyses (e.g. species- or genus-level) may be more meaningful. D-MANOVA could be possibly applied to those lower-level analyses by defining a relevant distance metric on the lower-level units.

Funding

This work was supported by the Center for Individualized Medicine at Mayo Clinic (J.C.), National Science Foundation [DMS-1830392 and DMS-1811747, X.Z.] and National Institutes of Health [R21 HG011662, J.C. and X.Z.].

Conflict of Interest: none decalred.

References

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, **28**, 2106–2113.

Kashyap,P.C. *et al.* (2017) Microbiome at the frontier of personalized medicine. *Mayo Clin. Proc.*, **92**, 1855–1864.

Knijnenburg,T.A. *et al.* (2009) Fewer permutations, more accurate P-values. *Bioinformatics*, **25**, i161–i168.

McArdle,B.H. and Anderson,M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

McArtor,D.B. *et al.* (2017) Extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic. *Psychometrika*, **82**, 1052–1077.

McDonald,D. *et al.*; The American Gut Consortium (2018) American gut: an open platform for citizen science microbiome research. *Msystems*, **3**, e00031-18.