Lecture 6: Generalized multivariate analysis of variance
Beta-diversity; ordination analysis

ISME J. 2016 Mar 25. doi: 10.1038/ismej.2016.37
Measuring association of the ‘entire’ microbiome with other variables

• Distance matrices capture some aspects of the data (e.g. microbiome composition, relative abundance, phylogenetic relationships).

• Euclidean distance (square-root of sums of square differences between components of the centered data) captures the covariances of the variables.

• Can these characteristics be used to draw association of the entire microbiome with other variables of interest (e.g. treatment group, locus of sampling, etc.)?
A general strategy for multivariate analysis

• Apply a normalization to the data (e.g. relative abundance);
• Calculate a distance metric between the observations (e.g. Unifrac, Jensen-Shannon, Chi-Square);
• Perform ordination and/or clustering analysis to visualize relationships between observations;
• Test for differences between predefined groups (e.g. treatment levels, phenotypes)
ANOVA

• Idea: $SS_{\text{total}} = SS_{\text{error}} + SS_{\text{treatments}}$
• F test: $F = \frac{SS_{\text{treatments}}/(I - 1)}{SS_{\text{error}}/(n_T - I)}$
• $F = \frac{\text{(variance between)}}{\text{(variance within treatments)}}$
• $I$ – number of treatments
• $n_T$ – total number of cases
ANOVA example

1. Within group means
   - $Y_1 = (6+8+4+5+3+4)/6 = 5$
   - $Y_2 = ... = 9$
   - $Y_3 = ... = 10$

2. Overall mean $Y = 8$

3. Between group sum of squares
   - $SS_{treatments} = n_1(Y_1-Y)^2 + n_2(Y_2-Y)^2 + n_3(Y_3-Y)^2 = 84$
   - $(k - 1) = 3 - 1 = 2$

4. Within group sum of squares
   - $SS_{error} = 68$
   - $(nT - k) = 18 - 3 = 15$

5. $F = (84/2) / (68/15) = 42/4.5 = 9.3$

6. $F_{critical} (2, 15) = 3.68$

7. Conclusion: The group effects are statistically significantly different.

8. Next: perform post-hoc pairwise tests to detect the pairs that are different

### ANOVA Example Table

<table>
<thead>
<tr>
<th>a1</th>
<th>a2</th>
<th>a3</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SS</th>
<th>SS1</th>
<th>SS2</th>
<th>SS3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6 - $Y_1)^2$ = (6 - 5)^2 = 1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>(8 - 5)^2 = 9</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(4 - 5)^2 = 1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(5 - 5)^2 = 0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(3 - 5)^2 = 4</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>(4 - 5)^2 = 1</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Euclidean MANOVA

• A direct extension of the univariate ANOVA to multiple variables.
• \( SS = \sum (Y_i - \bar{Y})^T(Y_i - \bar{Y}) \)
• \( SS = \sum d^2 \), where \( d \) is the Euclidean distance from the center.
Geometric representation of MANOVA (Anderson, 2001)

\[ F = \frac{SS_A/(a - 1)}{SS_W/(N - a)} \]

*\( SS_A \) – between group sums of squares

*\( SS_W \) – within group sums of squares

*\( SS_T \) – total sum of squares

\( SS_T = SS_W + SS_A \)

Key: Mean within group squared distance is equal to sum of squared distances to the centroid.

**Fig. 1.** A geometric representation of MANOVA for two groups in two dimensions where the groups differ in location. The within-group sum of squares is the sum of squared distances from individual replicates to their group centroid. The among-group sum of squares is the sum of squared distances from group centroids to the overall centroid. (---) Distances from points to group centroids; (-----) distances from group centroids to overall centroid; (○), overall centroid; (□), group centroid; (●), individual observation.
Calculating F-statistic from arbitrary distance matrices

\[ F = \frac{SS_A/(a-1)}{SS_W/(N-a)} \]

**Fig. 3.** Schematic diagram for the calculation of (a) a distance matrix from a raw data matrix and (b) a non-parametric MANOVA statistic for a one-way design (two groups) directly from the distance matrix. \( SS_T \), sum of squared distances in the half matrix (■) divided by \( N \) (total number of observations); \( SS_W \), sum of squared distances within groups (□) divided by \( n \) (number of observations per group). \( SS_A = SS_T - SS_W \) and \( F = [SS_A/(a-1)]/[SS_W/(N-a)] \), where \( a \) = the number of groups.
Obtaining p-values

- The F-statistic does not follow Fisher’s F-ratio under null, therefore we need to evaluate its distribution under null.
- Null hypothesis: there is no difference between groups; therefore, we can compute null distribution empirically by shuffling the group labels.
- For each reshuffling of labels compute F statistic, the p-value is then

\[
P = \frac{(\text{No. of } F^\pi \geq F)}{(\text{Total no. of } F^\pi)}
\]
Post-hoc tests for multi-level factors

• When a factor has more than 2 levels, it is not immediately clear which pair of groups are different from each other.
• To figure this out a post-hoc pairwise tests need to be carried out.
• The t-statistic is computed as square root of the F statistic for the pair only.
• Pairwise p-values are calculated with additional permutations.
• Multiple comparison correction may be necessary.
More sophisticated designs

• Two-way MANOVA
  • Straightforward extension with all interactions considered.

• Stratification/block design
  • When an effect is to be determined within the levels of another factor
  • E.g. Location of sampling vs. treatment
More sophisticated regression scenarios

• Based on Zapala & Schork, PNAS 2006.
• Suppose we have $M$ predictor variables
• We treat the multivariate $(N \times P)$ data (microbiome abundance, gene expression, etc.) as the response variable $Y$
• The basic multivariate regression model is $Y = X\beta + \varepsilon$,
• where $\beta$ is the coefficient matrix, and $\varepsilon$ is an error term.
• Define the hat matrix as usual $H = (X'X)^{-1}X'$. 
Regression scenario (continued)

- \( G = -\frac{1}{2} \left( I - \frac{1}{n} 11' \right) D^{(2)} \left( I - \frac{1}{n} 11' \right); \)
- Then \( F = \frac{\text{tr}(HGH)/(M-1)}{\text{tr}[(I-H)G(I-H)]/(N-M)}. \)
- This is how PERMANOVA is implemented in R/vegan package, function adonis().
Assumptions of PERMANOVA

• PERMANOVA is defined for balanced sample sizes, but can be rewritten for \( n_x \neq n_y \).
• Homoscedasticity is an underlying assumption.
• Do violations of these assumptions lead to undesired behaviors?

• Simulation to test these assumptions:
  • Let X be 1,000 dimensional uncorrelated standard normal
  • Let Y be 1,000 dimensional uncorrelated multivariate normal with each component
    • mean = 1/sqrt(1000)\*e
    • S.D. = 0.8
  • Simulate data with \( n_x, n_y \in \{5,10,15,20\} \)
  • Compute Euclidean distances, PERMANOVA p-values
Empirical robustness of PERMANOVA to heteroscedasticity and unbalanced sample sizes

Empirical type I error and power ($\alpha = 0.05$)
Robustness of PERMANOVA

• When both homoscedasticity and balanced sample sizes are violated adverse statistical behavior can be observed.

• If X is the more dispersed sample then
  • $n_x < n_y$ leads to type I error inflation,
  • $n_x > n_y$ leads to loss of power,
  • where $n_x$ is the number of observations in the more dispersed sample.
Idea: Univariate approach to heteroscedasticity issues

- Consider the square of Welch t-statistic $T_W^2 = \frac{(\bar{x} - \bar{y})^2}{s_x^2/n_x + s_y^2/n_y}$.
- If we can write $T_W^2$ in terms of pairwise distances, we can generalize it to multivariate data.
- We can use permutation testing to assess the significance.

Multivariate Welch t-test on distances

Alexander V. Alekseyenko

Departments of Public Health Sciences and Oral Health Sciences, Program for Human Microbiome Research, The Biomedical Informatics Center Medical University of South Carolina, 135 Cannon Street, MSC 200, Charleston, SC 29466, USA
Key equations for $T^2_W$ derivation

• $s_x^2 = \frac{1}{n_x(n_x-1)} \sum_{i<j}^n (x_i - x_j)^2 = \frac{1}{n_x(n_x-1)} \sum_{i<j}^n d_{ij}^2$,

• Where $\sum_{i<j}^n$ denotes double summation $\sum_{i=1}^n \sum_{j=i+1}^n$.

• Let $Z = (z_1, \ldots, z_{n_x+n_y}) = (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y})$,

  • $(\bar{x} - \bar{y})^2 = \frac{n_x+n_y}{n_x n_y} \left[ \frac{1}{n_x+n_y} \sum_{i<j}^n (z_i - z_j)^2 - \frac{1}{n_x} \sum_{i<j}^n (x_i - x_j)^2 - \frac{1}{n_y} \sum_{i<j}^n (y_i - y_j)^2 \right]$. 
Pseudo-F vs $T^2_W$

$$F = \frac{1}{n_x + n_y} \sum_{i,j=1}^{n_x+n_y} d_{ij}^2 - \frac{1}{n_x} \sum_{i,j=1}^{n_x} d_{ij}^2 - \frac{1}{n_y} \sum_{i,j=n_x+1}^{n_x+n_y} d_{ij}^2 \left(\frac{1}{n_x} \sum_{i,j=1}^{n_x} d_{ij}^2 + \frac{1}{n_y} \sum_{i,j=n_x+1}^{n_x+n_y} d_{ij}^2\right) / (n_x - n_y - 2)$$

$$T^2_W = \frac{n_x + n_y}{n_x n_y} \times \frac{1}{n_x} \sum_{i,j=1}^{n_x} d_{ij}^2 - \frac{1}{n_x} \sum_{i,j=1}^{n_x} d_{ij}^2 - \frac{1}{n_y} \sum_{i,j=n_x+1}^{n_x+n_y} d_{ij}^2 \left(\frac{1}{n_x} \sum_{i,j=1}^{n_x} d_{ij}^2 + \frac{1}{n_y} \sum_{i,j=n_x+1}^{n_x+n_y} d_{ij}^2\right)$$

How do these compare when $n_x = n_y$ or $\frac{1}{n_x(n_x-1)} \sum_{i,j} d_{ij}^2 - \frac{1}{n_y(n_y-1)} \sum_{i,j} d_{ij}^2$?
Empirical performance of $T^2_W$ vs PERMANOVA

Empirical type I error and power ($\alpha = 0.05$)

Type I error

Homoscedastic
Typical experimental scenarios at $n \approx 10$ or $n \approx 50$
Performance in a real dataset

Table 2. Comparison of PERMANOVA and $T_{W}^{2}$ on mouse gut microbiome dataset.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N obs.</th>
<th>$H$</th>
<th>$\omega^2$</th>
<th>d</th>
<th>P-values</th>
<th>N obs.</th>
<th>$H$</th>
<th>$\omega^2$</th>
<th>d</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. vs. All Abx.</td>
<td>10 vs. 40</td>
<td>1.4</td>
<td>0.22</td>
<td>1.21</td>
<td>0.040</td>
<td>0.0001</td>
<td>10 vs. 36</td>
<td>1.4</td>
<td>0.29</td>
<td>1.34</td>
</tr>
<tr>
<td>C. vs. Penicillin</td>
<td>10 vs. 10</td>
<td>0.85</td>
<td>0.12</td>
<td>1.90</td>
<td>0.00001</td>
<td>0.0002</td>
<td>10 vs. 9</td>
<td>1.1</td>
<td>0.07</td>
<td>1.94</td>
</tr>
<tr>
<td>C. vs. Vancomycin</td>
<td>10 vs. 10</td>
<td>1.8</td>
<td>0.08</td>
<td>2.26</td>
<td>0.00009</td>
<td>0.0001</td>
<td>10 vs. 9</td>
<td>1.6</td>
<td>0.21</td>
<td>2.70</td>
</tr>
<tr>
<td>C. vs. Tetracycline</td>
<td>10 vs. 10</td>
<td>1.2</td>
<td>0.12</td>
<td>2.05</td>
<td>0.00005</td>
<td>0.0005</td>
<td>10 vs. 10</td>
<td>1.0</td>
<td>0.07</td>
<td>1.89</td>
</tr>
<tr>
<td>C. vs. Van. + Tetr.</td>
<td>10 vs. 10</td>
<td>1.1</td>
<td>0.10</td>
<td>1.97</td>
<td>0.002</td>
<td>0.002</td>
<td>10 vs. 8</td>
<td>1.4</td>
<td>0.11</td>
<td>2.24</td>
</tr>
</tbody>
</table>

Table 3. Comparison of PERMANOVA and $T_{W}^{2}$ on human skin microbiome dataset.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N obs.</th>
<th>$H$</th>
<th>$\omega^2$</th>
<th>d</th>
<th>P-values</th>
<th>N obs.</th>
<th>$H$</th>
<th>$\omega^2$</th>
<th>d</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control vs. Lesion</td>
<td>49 vs. 51</td>
<td>1.07</td>
<td>0.014</td>
<td>0.77</td>
<td>0.0003</td>
<td>0.0002</td>
<td>10 vs. 8</td>
<td>1.6</td>
<td>0.21</td>
<td>2.70</td>
</tr>
<tr>
<td>Control vs. Unaffected</td>
<td>49 vs. 51</td>
<td>1.04</td>
<td>-0.0006</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>10 vs. 8</td>
<td>1.6</td>
<td>0.21</td>
<td>2.70</td>
</tr>
<tr>
<td>C. vs. Van. + Tetr.</td>
<td>10 vs. 10</td>
<td>1.1</td>
<td>0.10</td>
<td>1.97</td>
<td>0.002</td>
<td>0.002</td>
<td>10 vs. 8</td>
<td>1.4</td>
<td>0.11</td>
<td>2.24</td>
</tr>
<tr>
<td>C. vs. Vancomycin</td>
<td>10 vs. 10</td>
<td>1.8</td>
<td>0.08</td>
<td>2.26</td>
<td>0.00009</td>
<td>0.0001</td>
<td>10 vs. 9</td>
<td>1.6</td>
<td>0.21</td>
<td>2.70</td>
</tr>
<tr>
<td>C. vs. Penicillin</td>
<td>10 vs. 10</td>
<td>0.85</td>
<td>0.12</td>
<td>1.90</td>
<td>0.00001</td>
<td>0.0002</td>
<td>10 vs. 9</td>
<td>1.1</td>
<td>0.07</td>
<td>1.94</td>
</tr>
</tbody>
</table>

In this work, we have examined the performance of PERMANOVA and $T_{W}^{2}$ in the context of comparing microbial communities in different conditions. The results indicate that PERMANOVA and $T_{W}^{2}$ are comparable in terms of their ability to detect significant differences between groups. The use of these tests is supported by the availability of software implementations and their ease of use in statistical analysis. Further research is needed to explore the optimal conditions and parameters for these tests in different experimental settings.
Welch distance MANOVA with $W_d^*$ (to be published)

- In univariate case $W_d^*$ is equivalent to $W^*$.
- Let $x_i \sim N(0, s_i)$,
  1. $s_i^2 = 1$;
  2. $s_1^2 = 1, s_2^2 = 0.8, s_3^2 = 0.8^2$;
  3. $s_1^2 = 1, s_2^2 = 0.2, s_3^2 = 0.2^2$.
- $n_i \in \{5, 10, 20, 40\}$.
- Note that univariate $W^*$ test based on F distribution applies here.

![Graph showing the fraction of rejected null hypotheses for different test cases and variance levels.](image-url)
PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances

Zheng-Zheng Tang\textsuperscript{1,}*,†, Guanhua Chen\textsuperscript{1,}*,† and Alexander V. Alekseyenko\textsuperscript{2,3,4}

\textsuperscript{1}Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA, \textsuperscript{2}Biomedical Informatics Center, \textsuperscript{3}Department of Public Health Sciences and \textsuperscript{4}Department of Oral Health Sciences, Medical University of South Carolina, Charleston, SC 29403, USA

\*To whom correspondence should be addressed.
\†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Motivation: Recent advances in sequencing technology have made it possible to obtain high-throughput data on the composition of microbial communities and to study the effects of dysbiosis on the human host. Analysis of pairwise intersample distances quantifies the association between the microbiome diversity and covariates of interest (e.g. environmental factors, clinical outcomes, treatment groups). In the design of these analyses, multiple choices for distance metrics are available. Most distance-based methods, however, use a single distance and are underpowered if the distance is poorly chosen. In addition, distance-based tests cannot flexibly handle confounding variables, which can result in excessive false-positive findings.

Results: We derive presence-weighted UniFrac to complement the existing UniFrac distances for more powerful detection of the variation in species richness. We develop PERMANOVA-S, a new distance-based method that tests the association of microbiome composition with any covariates of interest. PERMANOVA-S improves the commonly-used Permutation Multivariate Analysis of Variance (PERMANOVA) test by allowing flexible confounder adjustments and ensembling multiple distances. We conducted extensive simulation studies to evaluate the performance of different distances under various patterns of association. Our simulation studies demonstrate that the power of the test relies on how well the selected distance captures the nature of the association. The PERMANOVA-S unified test combines multiple distances and achieves good power regardless of the patterns of the underlying association. We demonstrate the usefulness of our approach by reanalyzing several real microbiome datasets.

Availability and Implementation: miProfile software is freely available at https://medschool.vanderbilt.edu/tang-lab/software/miProfile.

Contact: z.tang@vanderbilt.edu or g.chen@vanderbilt.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Copyright The Author 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics, 32(17), 2016, 2618–2625
doi: 10.1093/bioinformatics/btw311
Advance Access Publication Date: 19 May 2016
Original Paper

• Based on Tang et al. Bioinformatics 2016.
• Suppose we want to consider K distances simultaneously, $D_1, \ldots, D_K$.
• We would like to know the significance of the entire ensemble
• Determine which individual distance performs best
PERMANOVA-S: Ensembling algorithm

1. For each $D_k$, compute the observed pseudo-F statistic $F_k$;
2. Obtain B permutations and compute $F_k^{(1)}, ..., F_k^{(B)}$;
3. Compute p-value for each $k$, $p_k$, and $p_{\text{min}} = \min(p_1, ..., p_K)$;
4. For each $k$, compute the permutation p-value as $p_k^{(b)} = (B - \text{rank}(F_k^{(b)}))/B$;
5. For each permutation $b$, obtain minimal permutation p-value $p_{\text{min}}^{(b)} = \min(p_1^{(b)}, ..., p_K^{(b)})$.
6. The final (unified) p-value is the proportion of $p_{\text{min}}^{(1)}, ..., p_{\text{min}}^{(b)}$ smaller than $p_{\text{min}}$. 
Summary

- PERMANOVA is useful for omnibus hypothesis testing;
- PERMANOVA has undesirable behavior with unbalanced heteroscedastic data;
- $T^2_W$ corrects that behavior in two sample case;
- PERMANOVA testing can be done with ensembling multiple distances.