Lecture 6: Distance-based 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: $SS_{total} = SS_{error} + SS_{treatments}$
• F test: $F = [SS_{treatments}/(I - 1)]/[SS_{error}/(n_T - I)]$
• $F = (\text{variance between})/(\text{variance within treatments})$
• $I$ – number of treatments
• $n_T$ – total number of cases

ANOVA example

<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>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>
</tr>
<tr>
<td>$(8 - 5)^2 = 9$</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>$(4 - 5)^2 = 1$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$(5 - 5)^2 = 0$</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$(3 - 5)^2 = 4$</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>$(4 - 5)^2 = 1$</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

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
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. A. 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. $SS_A$ sum of squared distances in the half matrix I divided by $N$ (total number of observations); $SS_W$ sum of squared distances within groups I divided by $a$ (number of observations per group). $SS_A = SS_W$, $SS_T$ and $F = ([SS_A/(a-1)]/[SS_T/(N-a)])$, 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 it’s 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

$$P = \frac{(\text{No. of } F^n \geq F)}{(\text{Total no. of } F^n)}$$
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 = $1/\sqrt{1000} \times e$
  - 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
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: Borrow from univariate approach to heteroscedastic data

- 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.
Key equations for $T_W^2$ derivation

• $s^2 = \frac{1}{n_x(n_x-1)} \sum_{i<j} (x_i - x_j)^2 = \frac{1}{n_x(n_x-1)} \sum_{i<j} d_{ij}^2$, 
• Where $\sum_{i<j}^n$ denotes double summation $\sum_{i=1}^n \sum_{j=i+1}^n$.
• Let $Z = (z_1, ..., z_{n_x+n_y}) = (x_1, ..., x_{n_x}, y_1, ..., y_{n_y})$,
  \begin{align*}
  \cdot \ (\bar{x} - \bar{y})^2 &= \frac{n_x+n_y}{n_x n_y} \left[ \frac{1}{n_x} \sum_{i<j} (z_i - z_j)^2 - \frac{1}{n_y} \sum_{i<j} (x_i - x_j)^2 - \frac{1}{n_x} \sum_{i<j} (y_i - y_j)^2 \right].
  \end{align*}

Pseudo-F vs $T_W^2$

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

\[
T_W^2 = \frac{n_x + n_y}{n_x n_y} \times \frac{1}{n_x(n_x-1)} \sum_{i,j=1}^{n_x} d_{ij}^2 - \frac{1}{n_y(n_y-1)} \sum_{i,j=n_x+1}^{n_x+n_y} d_{ij}^2
\]

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 $T_{W}^{2}$ vs PERMANOVA

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

Table 2. Comparison of PERMANOVA and $T^2_{Wd}$ 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>PERMANOVA</th>
<th>$T^2_{Wd}$</th>
<th>N obs.</th>
<th>$H$</th>
<th>$\omega^2$</th>
<th>d</th>
<th>PERMANOVA</th>
<th>$T^2_{Wd}$</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>
<td>0.015</td>
<td>0.0014</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>0.00002</td>
<td>10 vs. 9</td>
<td>1.1</td>
<td>0.07</td>
<td>1.94</td>
<td>0.015</td>
<td>0.015</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>0.00001</td>
<td>10 vs. 9</td>
<td>1.6</td>
<td>0.21</td>
<td>2.70</td>
<td>0.00001</td>
<td>0.00082</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>0.00005</td>
<td>10 vs. 10</td>
<td>1.0</td>
<td>0.07</td>
<td>1.89</td>
<td>0.007</td>
<td>0.006</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>
<td>0.001</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Welch distance MANOVA with $W_d^*$

- 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.

**$W_d^*$ and PERMANOVA on heteroscedastic univariate data**

![Graph A](image1.png)

**Key derivations for $W_d^*$**

\[
W^* = \frac{\sum w_j (\bar{x}_j - \hat{\mu})^2 / (k-1)}{1 + \left[\frac{2(k-2)}{k^2 - 1}\right] \sum h_j},
\]

\[
w_j = n_j / s_j^2,
\]

\[
\hat{\mu} = \frac{\sum w_j \bar{x}_j / W}{W},
\]

\[
W = \sum w_j,
\]

\[
h_j = (1 - w_j / W)^2 / (n_j - 1).
\]

\[
w_j = n_j / s_j^2 = (n_j - 1) n_j^2 \left( \sum_{p<q} d_{pq}^2 \right)^{-1}
\]

\[
\sum_{j=1}^k w_j (\bar{x}_j - \hat{\mu})^2 = \frac{1}{W} \sum_{i<j} w_i w_j (\bar{x}_i - \bar{x}_j)^2.
\]
Application example: racial disparity in colorectal cancer microbiome

![Image of PCoA plot](image.png)

Table 2: Significance of the primary and interaction effects by PERMANOVA and \( \chi^2 \) tests

<table>
<thead>
<tr>
<th>Covariate</th>
<th>PERMANOVA ( p ) value</th>
<th>( \chi^2 ) ( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>0.064</td>
<td>0.907</td>
</tr>
<tr>
<td>Location</td>
<td>0.907</td>
<td>0.907</td>
</tr>
<tr>
<td>Race and location</td>
<td>0.282</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Table 3: One versus all post hoc comparisons of the interaction terms

<table>
<thead>
<tr>
<th>Group</th>
<th>( t ) statistic</th>
<th>( \chi^2 ) ( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA distal</td>
<td>8.88</td>
<td>0.039</td>
</tr>
<tr>
<td>CA distal</td>
<td>1.93</td>
<td>0.075</td>
</tr>
<tr>
<td>AA proximal</td>
<td>0.36</td>
<td>0.936</td>
</tr>
<tr>
<td>CA proximal</td>
<td>0.70</td>
<td>0.665</td>
</tr>
</tbody>
</table>

PERMANOVA-S: accommodates multiple distances

- 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

Genome analysis

PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances

Zheng-Zheng Tang\(^{3,4,\ast} \), Guanhua Chen\(^{3,4,\ast} \) and Alexander V. Alekseyenko\(^{2,3,4} \)

\(^{1}\)Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA, \(^{2}\)Biomedical Informatics Center, \(^{3}\)Department of Public Health Sciences and \(^{4}\)Department of Oral Health Sciences, Medical University of South Carolina, Charleston, SC 29425, USA