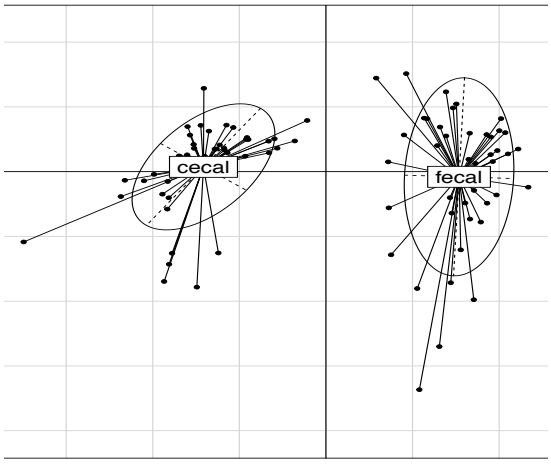
Lecture 6: Generalized multivariate analysis of variance

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.)?



2

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.)?

3

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)**

4

ANOVA

- Idea: $SS_{total} = SS_{error} + SS_{treatments}$
- F test: $F = [SS_{treatments}/(I - 1)]/[SS_{error}/(n_T - I)]$
- F = (variance between)/(variance within treatments)
- I – number of treatments
- n_T – total number of cases

5

ANOVA example

a1	a2	a3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

SS_1	SS_2	SS_3
$(6 - Y_1)^2$=(6 – 5)^2 = 1	1	9
$(8 - 5)^2 = 9$	9	1
$(4 - 5)^2 = 1$	0	1
$(5 - 5)^2 = 0$	4	4
$(3 - 5)^2 = 4$	9	9
$(4 - 5)^2 = 1$	1	4

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

6

Euclidean MANOVA

- A direct extension of the univariate ANOVA to multiple variables.
- $SS = \Sigma(\mathbf{Y_i} - \mathbf{Y})^T(\mathbf{Y_i} - \mathbf{Y})$
- $SS = \Sigma\, d^2$, where d is the Euclidean distance from the center.

7

Geometric representation of MANOVA (Anderson, 2001)

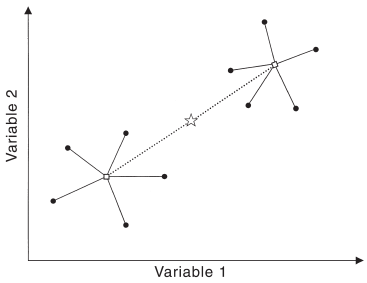


Fig. 1. A geometric representation of MANOVA for two groups in two dimensions where the groups differ in location. The within-group sum of squares is the sum of squared distances from individual replicates to their group centroid. The among-group sum of squares is the sum of squared distances from group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

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

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.

8

Calculating F-statistic from arbitrary distance matrices

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

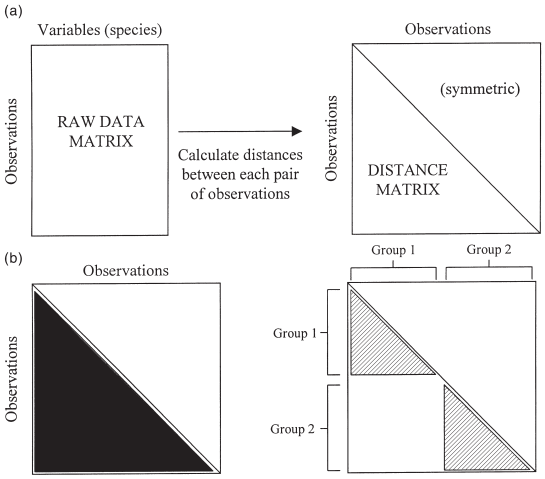


Fig. 3. Schematic diagram for the calculation of (a) a distance matrix from a raw data matrix and (b) a non-parametric MANOVA statistic for a one-way design (two groups) directly from the distance matrix. SS_T, sum of squared distances in the half matrix (■) divided by N (total number of observations); 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.

9

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

$$P = \frac{(\text{No. of } F^{\pi} \geq F)}{(\text{Total no. of } F^{\pi})}$$

10

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.
- Pairwise p-values are calculated with additional permutations.
- Multiple comparison correction may be necessary.

11

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

12

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 $\boldsymbol{Y}$
- The basic multivariate regression model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \varepsilon$,
- where $\boldsymbol{\beta}$ is the coefficient matrix, and ε is an error term.
- Define the hat matrix as usual $\boldsymbol{H} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

13

Regression scenario (continued)

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

14

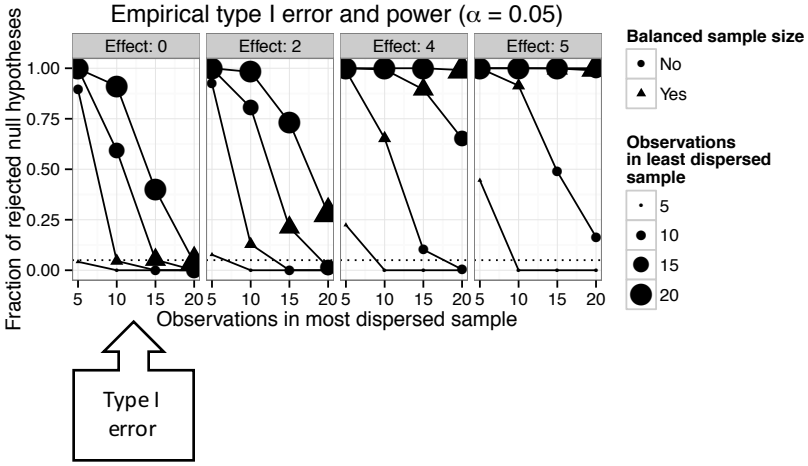
Assumptions of PERMANOVA

- 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 asumptions:
  - 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

15

Empirical robustness of PERMANOVA to heteroscedasticity and unbalanced sample sizes



16

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.

17

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.

18

Key equations for T_W^2 derivation

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

19

Pseudo-F vs T_W^2

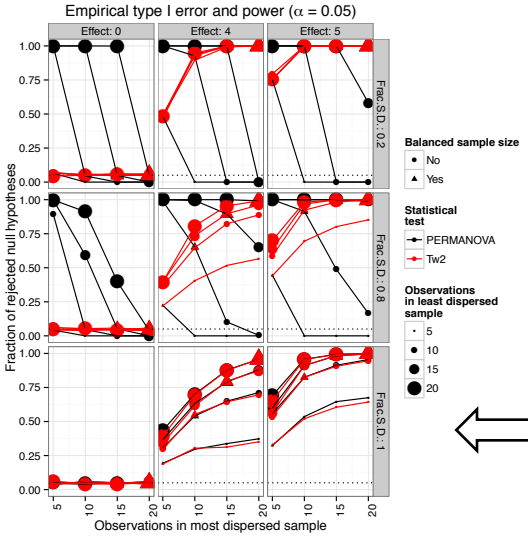
$$F = \frac{\frac{1}{n_x+n_y}\sum_{\substack{i<j\\ i,j=1}}^{n_x+n_y} d_{ij}^2 - \frac{1}{n_x}\sum_{\substack{i<j\\ i,j=1}}^{n_x} d_{ij}^2 - \frac{1}{n_y}\sum_{\substack{i<j\\ i,j=n_x+1}}^{n_x+n_y} d_{ij}^2}{\left(\frac{1}{n_x}\sum_{\substack{i<j\\ i,j=1}}^{n_x} d_{ij}^2 + \frac{1}{n_y}\sum_{\substack{i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

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

How do these compare when $n_x = n_y$ or $\frac{1}{n_x(n_x-1)}\sum_{i<j} d_{ij}^2 - \frac{1}{n_y(n_y-1)}\sum_{i<j} d_{ij}^2$?

20

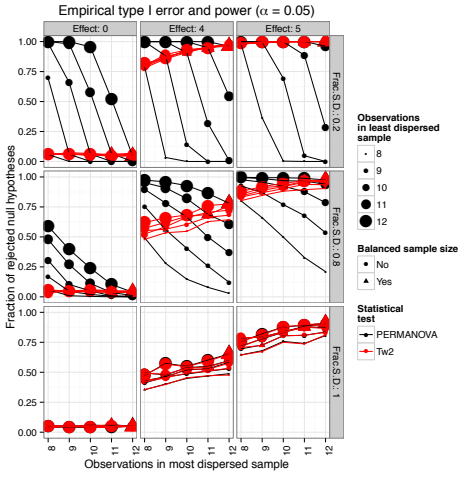
Type I error

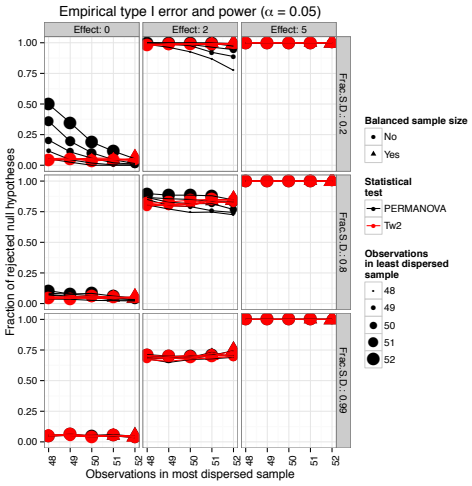


Empirical performance of T_W^2 vs PERMANOVA

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.

	Cecal microbiome				P-values		Fecal microbiome				P-values	
Comparison	N obs.	$\mathcal{H}$	ω^2	d	PERMANOVA	T_W^2	N obs.	$\mathcal{H}$	ω^2	d	PERMANOVA	T_W^2
C. vs. All Abx.	10 vs. 40	1.4	0.22	1.21	0.040	0.0001	10 vs. 36	1.4	0.29	1.34	0.015	0.0014
C. vs. Penicillin	10 vs. 10	0.85	0.12	1.90	0.00001	0.00002	10 vs. 9	1.1	0.07	1.94	0.015	0.015
C. vs. Vancomycin	10 vs. 10	1.8	0.08	2.26	0.00009	0.0001	10 vs. 9	1.6	0.21	2.70	0.00001	0.00002
C. vs. Tetracycline	10 vs. 10	1.2	0.12	2.05	0.00005	0.00005	10 vs. 10	1.0	0.07	1.89	0.007	0.006
C. vs. Van. + Tetr.	10 vs. 10	1.1	0.10	1.97	0.002	0.002	10 vs. 8	1.4	0.11	2.24	0.001	0.002

23

PERMANOVA-S: accommodates multiple distances

- Based on Tang et al. *Bioinformatics* 2016.
- Suppose we want to consider K distances simultaneously, $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K$.
- We would like to know the significance of the entire ensemble
- Determine which individual distance performs best

24

PERMANOVA-S: Ensembling algorithm

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

25

Summary

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

26