

### Assignment

- Use the spaeth.csv data from earlier:
- Eliminate species with 2 or fewer occurrences, then log transform the data
- Calculate Bray-Curtis similarity with the log transformed data
  - Perform PCoA with k=3
    - Report total variance explained, and the percent variance explained by each axis
  - Repeat PCoA with k=4
    - Report total variance explained, and the percent variance explained by each axis
- Plot the first two axes of PCoA with different symbols or colors for creek.
- What species have the highest correlation with axis 1 and 2? How does that help interpret your plot?

### PCoA Example

Site	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8	sp9	sp10	sp11	sp12	sp13	sp14	sp15	sp16	sp17	sp18	sp19	sp20
sample1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sample30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

### PCoA Example

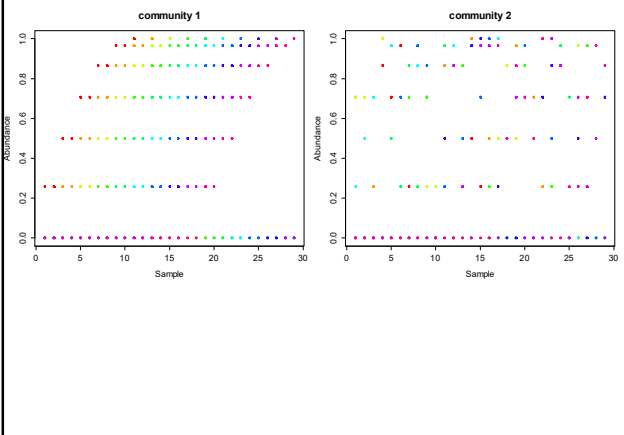
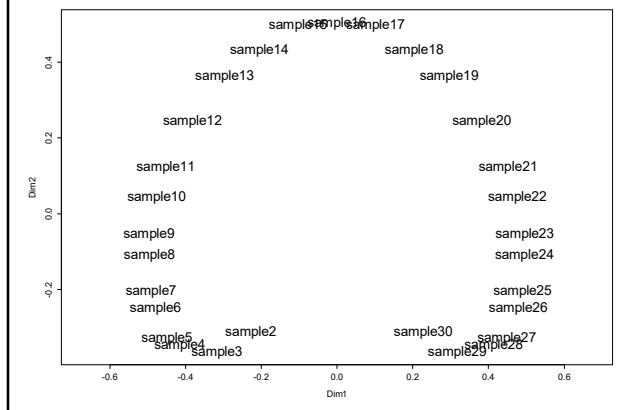


Fig: 4.844 3.047 1.394

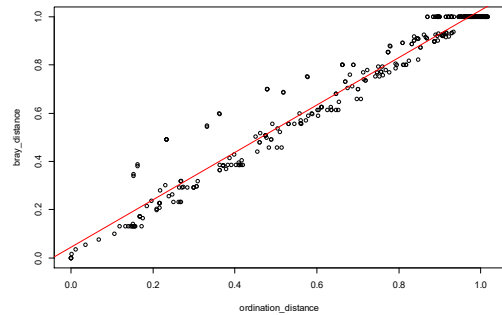


### Ordination object

- `cmdscale` function returns
  - points – k axes scores for each sample
  - eigenvalues
  - Goodness of Fit measure
- How do we assess how well the analysis did?
  - Goodness of Fit (GOF) ranges from 0-1, and can be interpreted similarly to  $r^2$
  - We can also calculate  $r^2$  directly from the scores.

Fig: 4.42 2.78 1.28 ...  
GOF: 0.8579 0.8919

Total Variance Explained for 3 axes = 96.63%



```
ordination_distance<-vegdist(prin_coord$points,method="euclidean")  
100*cor(ordination_distance,bray_distance)^2
```

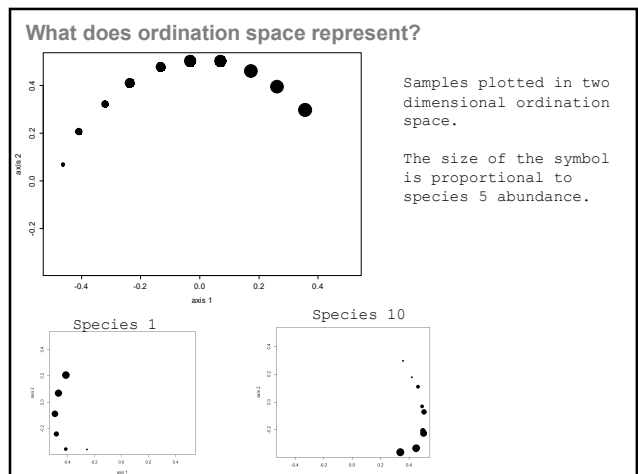
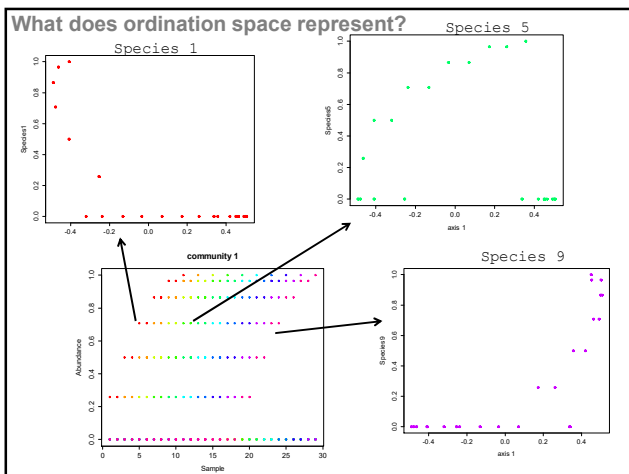
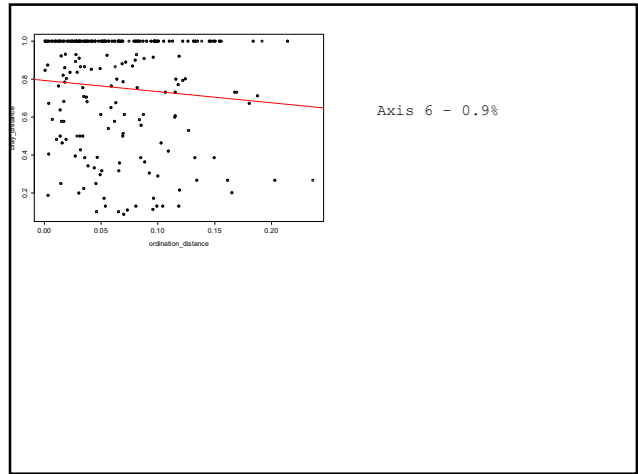
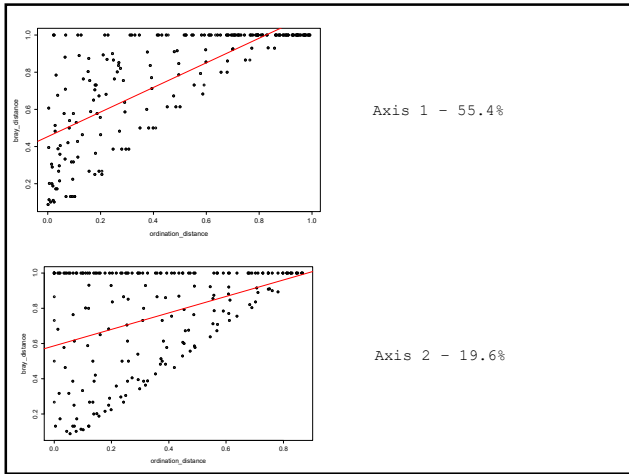
### Partitioning variation

- Total of all eigenvalues = 9.14
- Relative importance of each axis given by dividing each eigenvalue by the total...

Eigenvalue	Proportion of total variance
• 4.43	0.48
• 2.78	0.30
• 1.28	0.14
• 0.654	0.07
• 0.19	0.02
• 0.076	0.008
• 0.052	0.006
• 0.029	0.003

- Variance explained by each axis is different.
- For each axis, assess  $r^2$  for the euclidean distance among all samples on that axis to the original distance matrix
- This loop does this for all k axes

```
* pct_var<-0  
* for(k in 1:2)  
* {  
*   pct_var[k]<-100*cor(vegdist(prin_coord$points[,k],method="euclidean"),bray_distance)^2  
* }  
* pct_var
```



Raw data, Second sample file

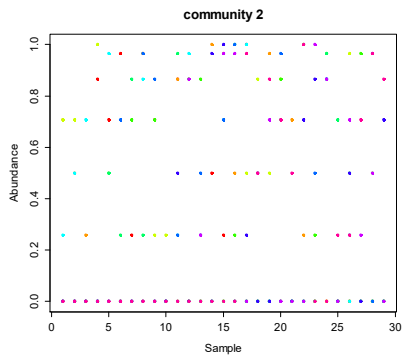


Fig: 2.23 1.94 1.39

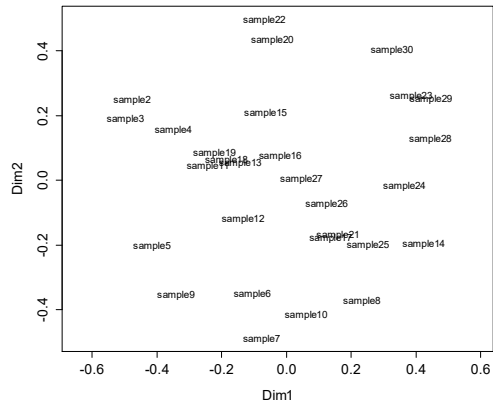
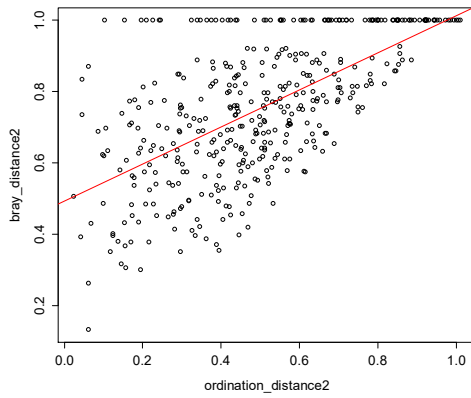
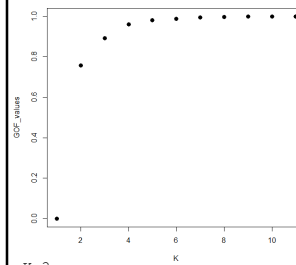


Fig: 2.23 1.94 1.39  
Total Variance Explained = 52%



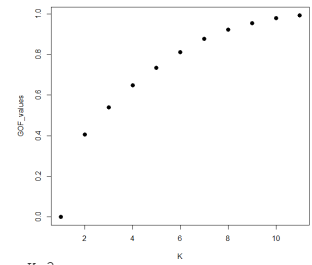
GOF values

Highly Structure Dataset



K=3  
Eigenvalues: 90% in first 3  
GOF 0.85-0.89  
Correlations between axes and  
bray Curtis matrix 55.5 19.5  
and 12.7 (sum 87.7)

Randomized Dataset

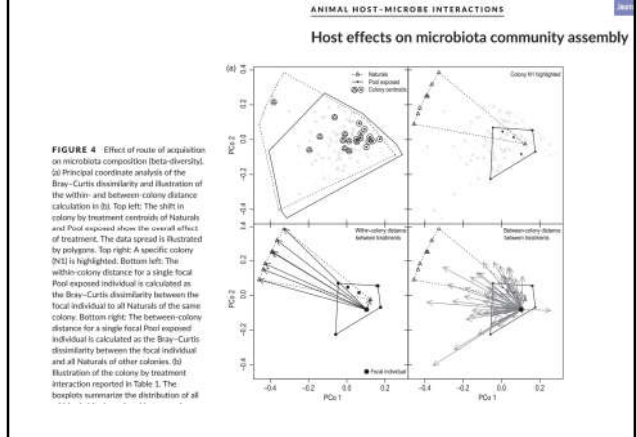


K=3  
Eigenvalues: 54% in first 3  
GOF 0.45-0.54  
Correlations between axes and  
bray Curtis matrix 22.1 11.9  
and 3.9 (sum 12.5)

## Reporting PCoA Results

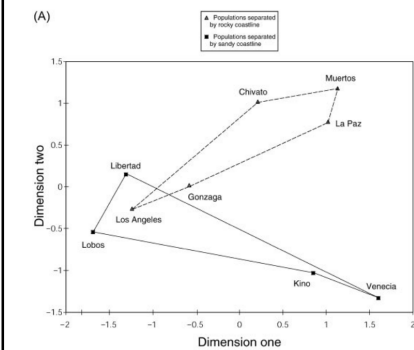
- Dimensions of original data matrix
  - In our example - 10 species and 29 samples
- Transformations or handling of missing/rare data
  - None in our example
- Type of distance matrix
  - Bray Curtis
- Number of dimensions retained
  - 3
- Total variance explained
- Relative importance of axes

## Other Examples of PCoA this semester



## Population subdivision in marine environments: the contributions of biogeography, geographical distance and discontinuous habitat to genetic differentiation in a blennioid fish, *Axoclinus nigricaudus*

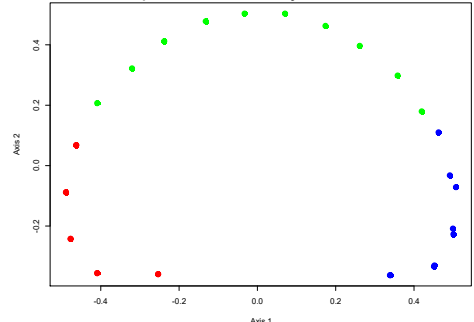
C. BUGHINO<sup>1</sup> and M. W. NACHREIN<sup>1</sup>  
 Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA



**Fig. 4** (A) Multidimensional scaling plot of populations based on genetic distances ( $D_{ij}$ ) between pairs of populations. Populations from the western Gulf that are connected by rocky habitat are shown by triangles and populations from the eastern Gulf that are separated by sandy habitat are shown by squares. The area of the polygon defined by the western Gulf populations equals 0.89 and the area of the polygon defined by the eastern Gulf populations equals 1.17. Greater genetic distances among eastern Gulf populations relative to western Gulf populations are consistent with discontinuous habitat (sand) reducing genetic interchange. (B) Scatterplot of residual genetic distances (controlling for biogeography) on geographical distance for pairs of populations separated by discontinuous habitat (sand and open water) and continuous habitat (rock). A modified Mantel permutation test shows that population pairs separated by open water are more genetically distant overall than populations pairs connected by rocky coastline (Table 3:  $t = 2.94$ ,  $P < 0.05$ ). In addition, as geographical distance increases, populations separated by sandy

## Sample Grouping

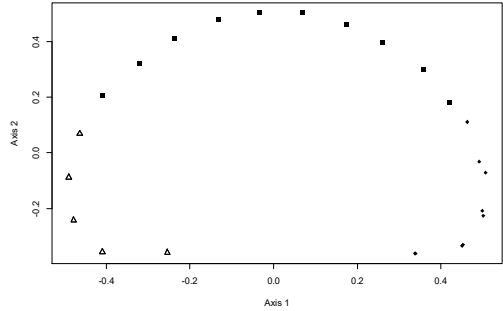
- Very often you want to show groups of samples. This can be done with point colors or symbols:



```
sample_groups = factor assigning each sample to A, B or C
sample_colors = c("red", "green", "blue")
plot(prin_coordspoints[,1:2], pch=19, col=sample_colors[sample_groups])
```

○ △ × ◇ ▽ ✖ ✗ ✘ ✙ ✚ ✛ ✜ ✝ ✞ ✟ ✠ ✡ ✢ ✣ ✤ ✥ ✦ ✧ ✨ ✩ ✪ ✫ ✬ ✭ ✮ ✯ ✰ ✱ ✲ ✳ ✴ ✵ ✶ ✷ ✸ ✹ ✺ ✻ ✼ ✽ ✾ ✿  
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

```
sample_symbols<-c(2, 15, 18)
```



```
plot(prin_coord$points[,1:2],pch=sample_symbols[sample_groups],col="black")
```