Lecture 5: Ecological distance metrics; Principal Coordinates Analysis

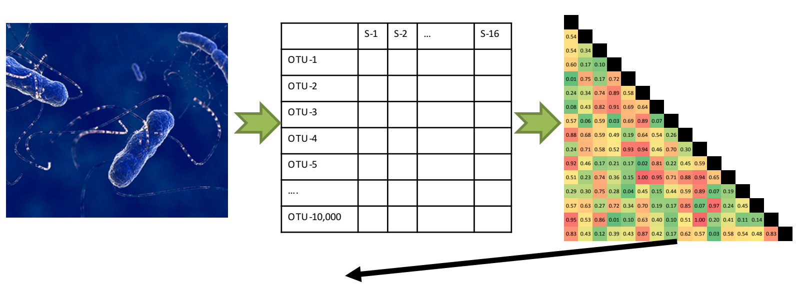
Univariate testing vs. community analysis

- Univariate testing deals with hypotheses concerning individual taxa
  - Is this taxon differentially present/abundant in different samples?
  - Is this taxon correlated with a given continuous variable?
- What if we would like to draw conclusions about the community as a whole?

Useful ideas from modern statistics

- Distances between anything (abundances, presence-absence, graphs, trees)
- Direct hypotheses based on distances.
- Decompositions through iterative structuration.
- Projections.
- Randomization tests, probabilistic simulations.

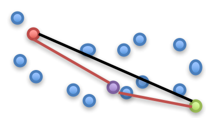
Data → Distances → Statistics



Visualization (Principal coordinates analysis)
Statistical hypothesis testing (PERMANOVA)

What is a distance metric?

- Scalar function d(.,.) of two arguments
- $d(x, y) >= 0$, always nonnegative;
- $d(x, x) = 0$, distance to self is 0;
- $d(x, y) = d(y, x)$, distance is symmetric;
- $d(x, y) < d(x, z) + d(z, y)$, triangle inequality.



Using distances to capture multidimensional heterogeneous information

- A “good” distance will enable us to analyze any type of data usefully
- We can build specialized distances that incorporate different types of information (abundance, trees, geographical locations, etc.)
- We can visualize complex data as long as we know the distances between objects (observations, variables)
- We can compute distances (correlations) between distances to compare them
- We can decompose the sources of variability contributing to distances in ANOVA-like fashion

Distance and similarity

- Sometimes it is conceptually easier to talk about similarities rather than distances
  - E.g. sequence similarity
- Any similarity measure can be converted into a distance metric, e.g.
  - S
  - If S is (0, 1), D=1-S
  - If S>0, D = 1/S or D = exp(-S)

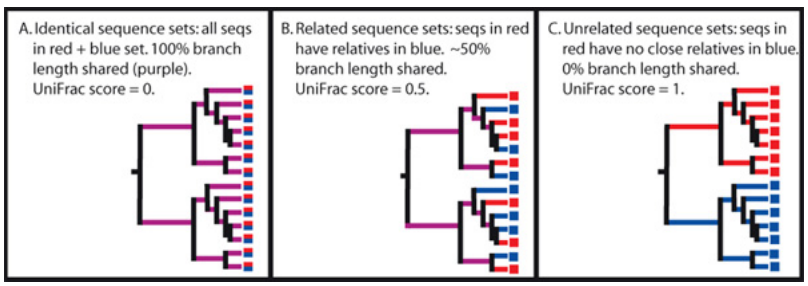
A few useful distances and similarity indices

- Distances:
  - Euclidean: (remember Pythagoras theorem) $\Sigma(x_i-y_i)^2$
  - Weighted Euclidean: $\chi^2 = \Sigma(e_i - o_i)^2/e_i$
  - Hamming/L1, Bray Curtis = $\Sigma\ \mathbf{1}_{\{xi=yi\}}$
  - Unifrac (later)
  - Jensen-Shannon: $(D(\mathbf{X}|\mathbf{M}) + D(\mathbf{Y}|\mathbf{M}))/2$, where
    - $\mathbf{M} = (\mathbf{X} + \mathbf{Y})/2$
    - Kullback-Leibler divergence: $D(\mathbf{X}|\mathbf{Y}) = \Sigma \ln[x_i/y_i]x_i$
- Similarity:
  - Correlation coefficient
  - Matching coefficient: $(f_{11}+f_{00})/(f_{11} + f_{10} + f_{01} + f_{00})$
  - Jaccard Similarity Index: $f_{11}/(f_{11} + f_{10} + f_{01})$

x\y	1	0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

Unifrac distance (Lozupone and Knight, 2005)

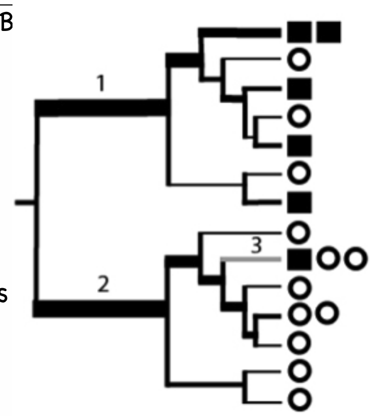
- Is a distance between groups of organisms related by a tree
- Definition: Ratio of the sum of the length of the branches leading to sample X or Y, but not both, to the sum of all branch lengths of the tree.



Weighted Unifrac (Lozupone et al., 2007)

$$\sum_{i=1}^{n} b_i \times |\frac{A_i}{A_T} - \frac{B_i}{B}$$

- n = number of branches in the
- b_i = length of the ith branch
- A_i = number of descendants of ith branch in group A
- A_T = total number of sequences in group A



A note of warning!

- “Garbage in, garbage out”
- Wrong normalization => wrong distance => wrong answer
- However, given the many choices there isn’t much beyond prior knowledge, experience and intuition to guide in selection of the distance.

Distance matrix

- It is convenient to organize distances as a matrix, $A=(a_{ij})$
- Distance matrices are:
  - Symmetric: $a_{ij} = a_{ji}$
  - Diagonals are 0: $a_{ii} = 0$.
- A distance matrix is Euclidean if it is possible to generate these distances from a set of n points in Euclidean space.
- Distance matrix is commonly represented by just lower (or upper) diagonal entries.

Some uses of distances

- Suppose D is a distance matrix for n objects. The objects are of several kinds indicated by a factor variable F;
- Intra-group distances are the distances between objects of the same kind;
- Inter-group distances are the distances between objects of different kinds;
- Mean distance between an object and a group of other objects is equal to the distance between the object and the center of the group.

PRINCIPAL COORDINATE ANALYSIS

Vector

A **vector**, **v**, of dimension n is an $n \times 1$ rectangular array of elements

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

vectors will be column vectors.

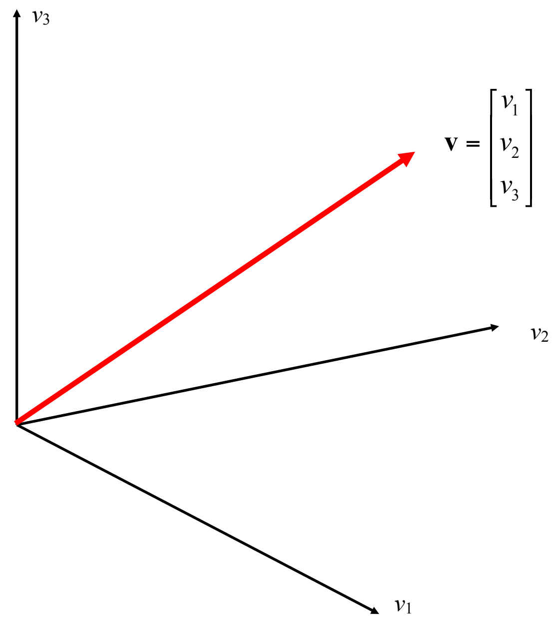
They may also be row vectors, when **transposed** $\mathbf{v}^T = [v_1, v_2, \ldots, v_n]$.

A **vector**, **v**, of dimension n

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

can be thought a point in n dimensional space

Every multivariate sample can be represented as a vector in some vector space



Vector Basis

- A basis is a set of linearly independent (dot product is zero) vectors that **span** the vector space.
- **Spanning** the vector space: Any vector in this vector space may be represented as a linear combination of the basis vectors.
- The vectors forming a basis are orthogonal to each other. If all the vectors are of length 1, then the basis is called orthonormal.

Matrix

An $n \times m$ matrix, A, is a rectangular array of elements

$$A = \left(a_{ij}\right) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}$$

n = # of rows

m = # of columns

dimensions = $n \times m$

Note: Let A and B be two matrices whose inverse exists. Let $C = AB$. Then the inverse of the matrix C exists and $C^{-1} = B^{-1}A^{-1}$.

Proof

$$C[B^{-1}A^{-1}] = [AB][B^{-1}A^{-1}] = A[B\,B^{-1}]A^{-1} = A[I]A^{-1}$$
$$= AA^{-1} = I$$

Diagonalization

Thereom If the matrix A is symmetric with distinct eigenvalues, $\lambda_1, \ldots, \lambda_n$, with corresponding eigenvectors $\vec{x}_1, \ldots, \vec{x}_n$

Assume $\vec{x}_i' \vec{x}_i = 1$

then $A = \lambda_1 \vec{x}_1 \vec{x}_1' + \ldots + \lambda_n \vec{x}_n \vec{x}_n'$

$$= \left[\vec{x}_1, \ldots, \vec{x}_n\right] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{x}_1' \\ \\ \vec{x}_n' \end{bmatrix} = PDP'$$

Basic idea for analysis of multidimensional data

- Compute distances
- Reduce dimensions
- Embed in Euclidean space
- The general framework behind this process is called **Duality diagram**: ($\mathbf{X}_{nxp}$, $\mathbf{Q}_{pxp}$, $\mathbf{D}_{nxn}$)
  - $\mathbf{X}_{nxp}$ (centered) data matrix
  - $\mathbf{Q}_{pxp}$ column weights (weights on variables)
  - $\mathbf{D}_{nxn}$ row weights (weights on observations)

Duality diagram defines the geometry of multivariate analysis

$$\begin{array}{ccc} \mathbb{R}^{p*} & \xrightarrow[X]{} & \mathbb{R}^{n} \\ Q \uparrow \;\; \downarrow V & & D \downarrow \;\; \uparrow W \\ \mathbb{R}^{p} & \xleftarrow[X^t]{} & \mathbb{R}^{n*} \end{array}$$

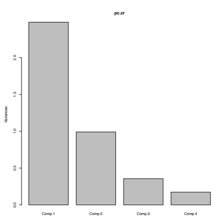
- $V = X^T DX$
- $W = XQX^T$
- Duality:
  - The eigen decomposition of VQ leads to eigen-decomposition of WD
- *Inertia* is equal to trace (sum of the diagonal elements) of VQ or WD.

Principal Component Analysis (PCA)

- Let Q = I and D = 1/n I and let X be centered.
- $VQ = X^T DXQ = 1/n\ X^T X$.
- The inertia Tr(VQ) = sum of the variances.
- PCA decomposes the variance of X into independent components.
- To decompose the inertia means to find the eigen-system of VQ or equivalently WD matrices.
- Eigenvalues give the amount of inertia explained in corresponding dimension.
- Eigenvectors give the dimensions of variability.

Example PCA

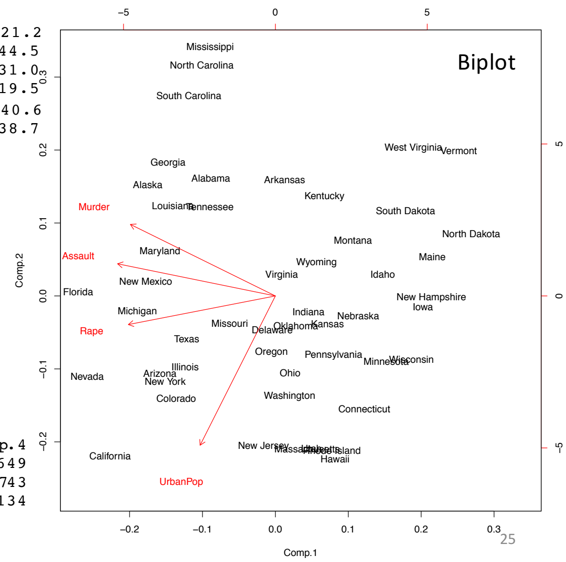
	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7



Screeplot: plot of inertia

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
Murder	-0.536	0.418	-0.341	0.649
Assault	-0.583	0.188	-0.268	-0.743
UrbanPop	-0.278	-0.873	-0.378	0.134
Rape	-0.543	-0.167	0.818	



Centering

- Let Y be not centered data matrix with n observations (rows) and p variables (columns)
- Let **H** = (**I** – 1/n **1**x**1’**)
- Then X = HY is centered

From Euclidean distances to PCA to PCoA

- Note that if **D** is a Euclidean distance, then
- **X X**' = 1/n **H D**$^{(2)}$ **H**.
- PCoA is a generalization of PCA in that knowledge of **X** is not required, all you need to represent the points is **D**, the inter-point distance matrix.

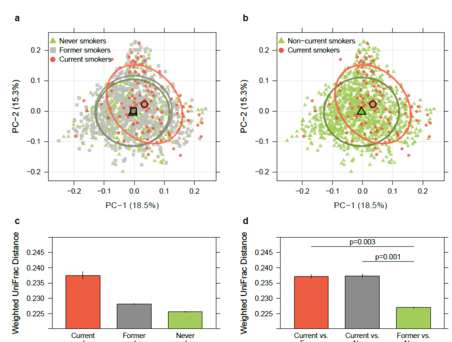
Representation of (arbitrary) distances in Euclidean space

- The idea is to use singular value decomposition (SVD) on the centered interpoint distance matrix to extract Euclidean dimensions
- SVD: X = U S V, where S is diagonal matrix with diagonal elements s_1, s_2, …, s_n, and U and V are unit matrices (i.e. their determinant is 1 and they span their corresponding spaces)

PCoA details

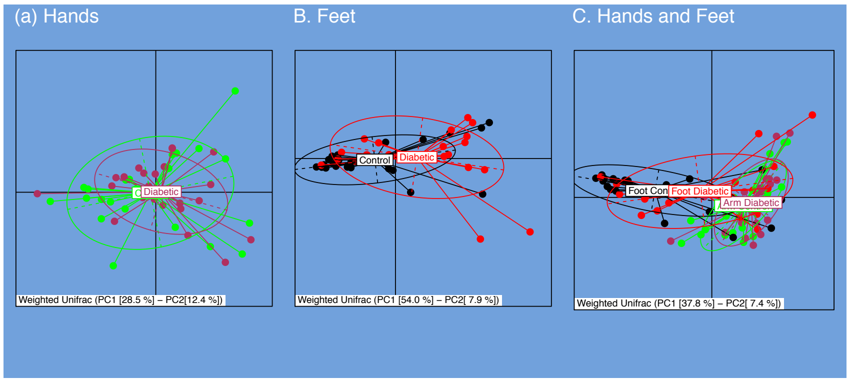
- Algorithm starting from **D** inter-point distances:
  - Center the rows and columns of the matrix of square (element-by-element) distances: $\mathbf{S} = -1/2\mathbf{H}\,\mathbf{D}^{(2)}\mathbf{H}$
  - Compute SVD by diagonalizing **S**, $\mathbf{S} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^T$
  - Extract Euclidean representations: $\underline{X} = \mathbf{U}\,\mathbf{\Lambda}^{1/2}$
- The relative values of diagonal elements of **Λ** gives the proportion of variability explained by each of the axes.
- The values of **Λ** should always be looked at in deciding how many dimensions to retain.

Beta-diversity; ordination analysis



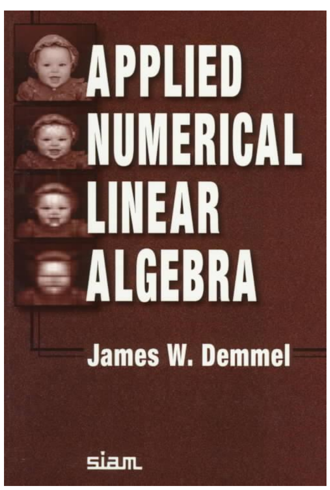
ISME J. 2016 Mar 25. doi: 10.1038/ismej.2016.37

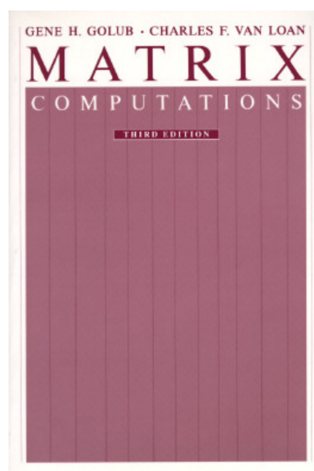
Differentiation of microbiota between diabetic and non-diabetic subjects and across body sites



Redel et al. J Infect Dis. 2013 207(7):1105-14

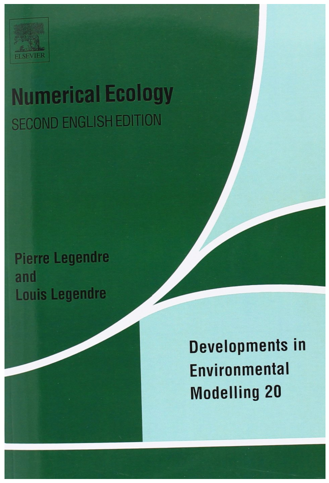
Suggested reading/references





+ any proof-based linear algebra text book.

Suggested reading



- Susan Holmes "Multivariate Data Analysis: The French Way", IMS Lecture Notes–Monograph Series, 2006.