Introduction to multivariate analysis

Outline for today

- Why do a multivariate analysis
- Ordination, classification, model fitting
- Principal component analysis
- Discriminant analysis, quickly
- Species presence/absence data
- Distance data

Data are usually multivariate

Typically we measure many variables on the populations, species, and ecosystems that we study.

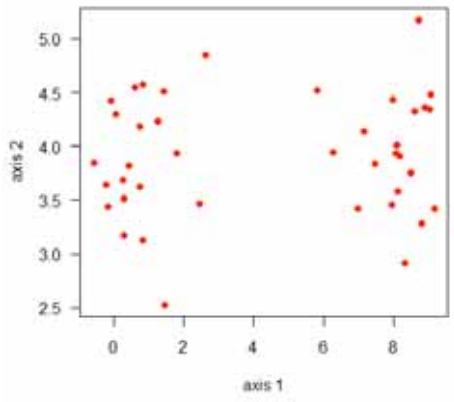
This creates a challenge: how to display and analyze measurements on all those variables.

Need ways to make it easier to find the important patterns and relationships among the many variables.

Some of the goals of multivariate methods

- To visualize complex data in few dimensions
- To find the most relevant combinations of variables (e.g., “size”)
- To reduce the number of comparisons and tests
- To reduce noise in the data

Large number of variables → multivariate analysis →



Ordination, Classification, Model fitting

Multivariate methods are used for

- Model fitting (multivariate analysis of variance; multiple regression)
- Ordination (scaling): arrange sampling units along gradients or according to combinations of variables
- Classification: place sampling units into groups

Example of where multivariate model fitting is useful

Example: "Subjects by trials" repeated measures design (from previous lecture on models with random effects)

Fixed effects: rodent treatment
Random effect: plot

Time is a fixed factor but sphericity assumption violated (nearby points in time more highly correlated than distant points in time). MANOVA often used to analyze these types of experiments, although requires more replication than available in this study.

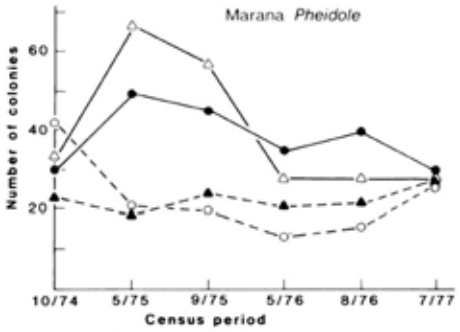
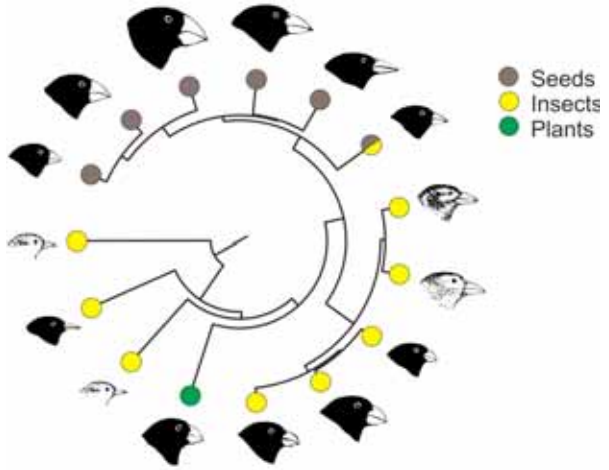


FIG. 3. Changes in density of *Pheidole* spp. (including *P. xerophila tucsonica, P. sitarches,* and *P. gilvescens*) on two rodent removal plots (——) and two control plots (– –) at Marana, Arizona over a 2¾-yr period.

Ordination

This is what we'll mainly focus on here.
Start with Principal Components Analysis because it is the most straightforward multivariate method.
Example 1: Differences in beak and body dimensions of Darwin's finches



Example 1: Darwin's Finches

Data: Means of measurements (5 variables) on 14 species

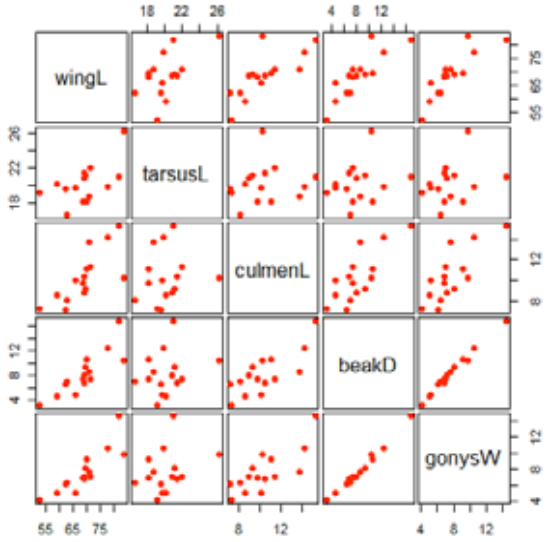
		VARIABLES (traits)				
	species	wingL	tarsusL	culmenL	beakD	gonysW
	C.heliobates	68.79	21.35	10.45	6.75	6.78
	C.pallida	71.2	21.96	11.36	7.51	7.02
	C.parvulus	62.28	19.55	7.2	6.51	6.13
	C.pauper	68.89	20.82	8.91	7.95	7.11
	C.psittacula	69.34	21.11	9.25	9.34	8.06
UNITS	*Certhidea.fusca*	59.02	20.04	8.57	4.56	5.04
	Certhidea.olivacea	52.48	19.1	7.3	3.17	4.11
	G.conirostris	77.47	19.77	14.22	12.35	10.59
	G.difficilis	68.31	18.15	9.75	7.47	6.89
	G.fortis	69.69	18.08	11.1	10.62	9.22
	G.fuliginosa	62.36	16.55	8.13	6.97	6.33
	G.magnirostris	81.79	20.88	15.25	16.84	14.53
	G.scandens	70.9	18.71	13.76	8.54	7.67
	Pinaroloxias	65.93	19.69	10.09	4.7	5.1
	Platyspiza	83.43	26.24	10.26	10.37	9.81

This is a matrix (rectangular array of numbers)

Example 1: Darwin's Finches

The data have only 5 variables but visualizing them still represents a challenge.

```
pairs(mydata)
```



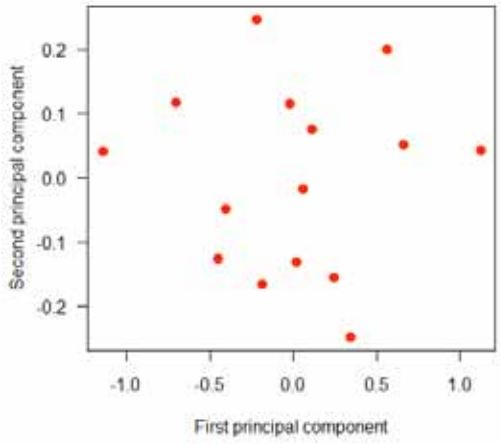
Example 1: Darwin's Finches

How to visualize multivariate data --- a continuing challenge

These are "Chernoff faces", which display multivariate data in the shape of a human face. The individual parts of the face represent values of the variables by their shape, size, placement and orientation. Humans are good at distinguishing faces.

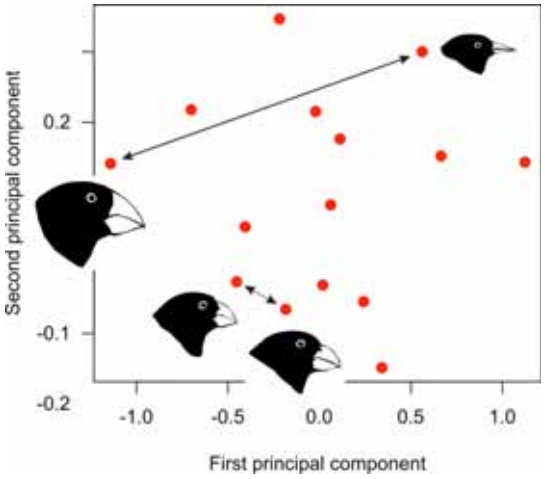
Principal components analysis

96% of all the variation among the Darwin's finch species is in just 2 dimensions. These dimensions are linear combinations of variables that co-vary among the species.



Principal components analysis

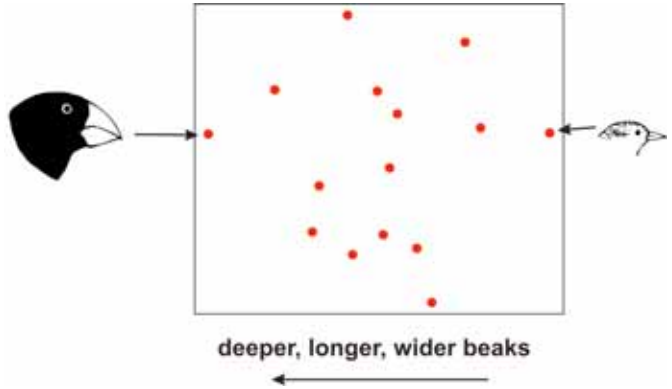
Even though we've gone from 5 dimensions to 2, distances between species are largely preserved. Points close together indicate species that are similar. Points far apart indicate species that are different.



Principal components analysis

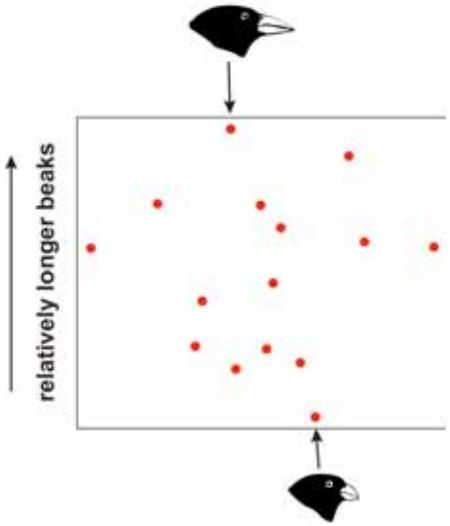
Even though they are composite variables, the axes in this case are interpretable.

The first axis arranged the species according to differences in overall beak size.



Principal components analysis

The second axis arranged the species according to differences in beak length (relative to overall beak size). It represents an axis of beak shape.



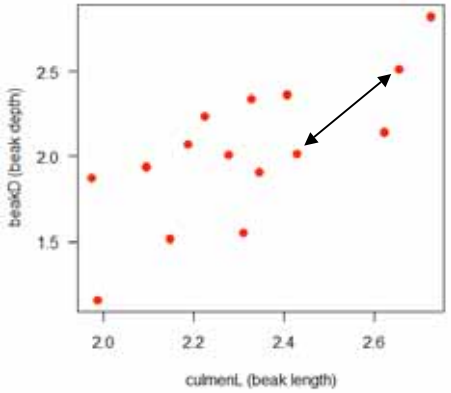
Principal components analysis – 2D illustration

To see what the analysis actually does, let's focus on two traits (everything is the same when there are more than two traits). I've log-transformed all the variables to help put on a similar scale.

The data on 15 species and 2 variables can be represented in either of two ways.

The first is the distance between pairs of points (species) (arrow on right). "Euclidean distance" is the straight line distance between two points

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



Principal components analysis – 2D illustration

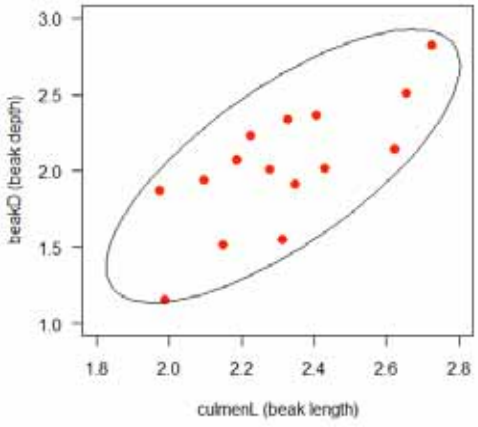
The other way to represent the data is by measuring the association between the variables (as illustrated by the ellipse)

The following is a covariance matrix between the variables:

	culmenL	beakD
culmenL	0.0518	0.0698
beakD	0.0698	0.1741

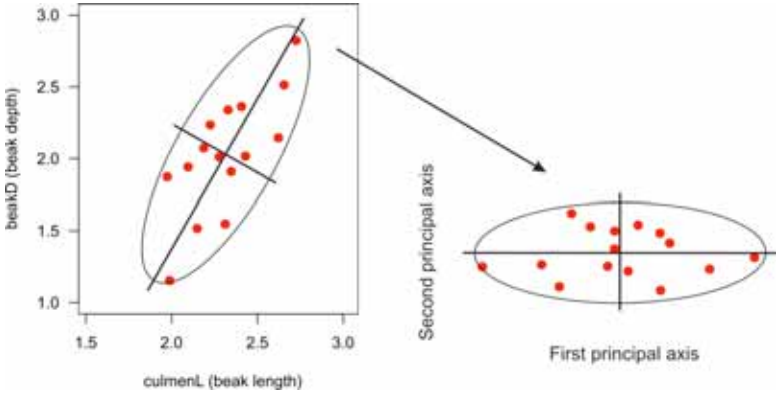
The elements of the covariance matrix are:

$var(x_1)$	$cov(x_1,x_2)$
$cov(x_1,x_2)$	$var(x_2)$



Principal components analysis – 2D illustration

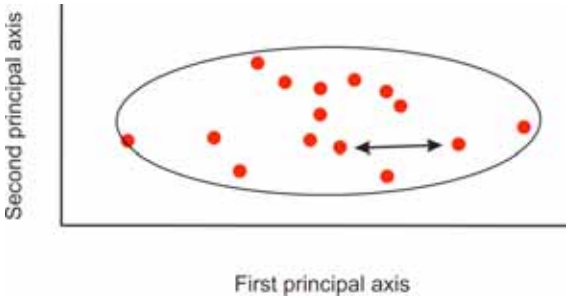
Principal components analysis finds a linear transformation of the data to create a composite variable with maximum possible variance. This is the first major axis, or principal component. A second linear transformation creates a variable with the next largest variance perpendicular ("orthogonal") to the first. (And so on, when there are more than 2 variables.)



Principal components analysis

The procedure amounts to nothing more than a rotation of the axes.

The Euclidean distances between pairs of species are unaltered by the transformation*.



*warning: in some non-R programs the default procedure is to standardize the variables ("correlation matrix") before carrying out the analysis, which WILL change the distances. Turn this off except as a last resort (when variables can't otherwise be put on a common scale)

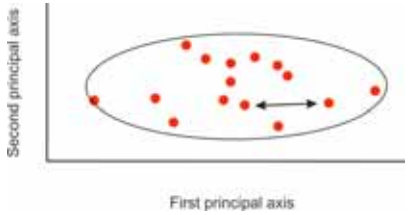
Principal components analysis

The covariance matrix of the new composite variables has variances down the diagonal and zeros off the diagonal (principal component axes are always uncorrelated)

	pc1	**pc2**
pc1	0.192	0
pc2	0	0.019

These variances are called the **eigenvalues**

They sum to the same total as the variances of the original traits.



Principal components analysis

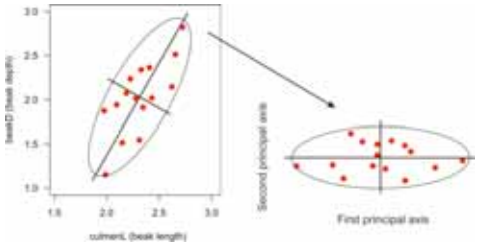
The vectors that contain the constants for transforming the original variables into the principal components are called the **eigenvectors**.

	eigen1	**eigen2**
culmenL	0.413	-0.911
beakD	0.911	0.413

pc1= 0.413*culmenL+ 0.911*beakD

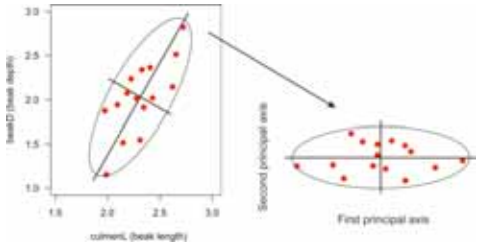
pc2= −0.911*culmenL+ 0.413*beakD

The constants are called **loadings.**



Principal components analysis

	eigen1	**eigen2**
culmenL	0.413	-0.911
beakD	0.911	0.413



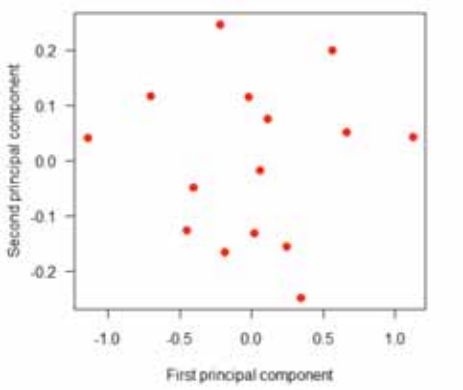
The constants are called **loadings** because they indicate the importance of each variable to the principal component.

pc1 reflects beak size because both traits make positive contributions (depth contributes more than length). The axis separates big-beaked birds at one extreme from small-beaked birds at the other.

pc2 reflects beak shape because beak depth loads positively but beak length negatively. It separates short deep beaks at one extreme from long shallow beaks at the other.

Principal components analysis

The idea is the same with 3, 4, 5, or any number of variables. The plot below is from the analysis of all 5 variables for the 15 Darwin's finch species. The only difference is that when we look at only the first two principal components we aren't seeing all the differences among the species. The eigenvalues tell you how much of the total variation is captured by the first two principal components.

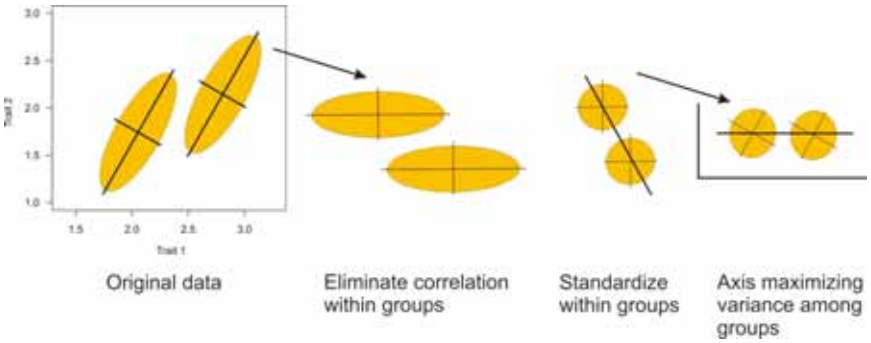


Discriminant function analysis, quickly

Discriminant analysis shares some features with principal component analysis, but is for grouped data.

The procedure finds axes that maximize variation among groups relative to variation between groups.

This procedure DOES alter the distances between pairs of data points



[break]

Correspondence Analysis for species presence/absence data

…is an analogous method for visualizing associations between species in presence/absence or abundance at sites (or, equivalently, differences among sites in the presence/absence or abundance of species).

Data: presence/absence of 16 ant species in 4 geographic regions (Gotelli and Ellison 2004). In this matrix, the sites are the "units", and the species are the "variables".

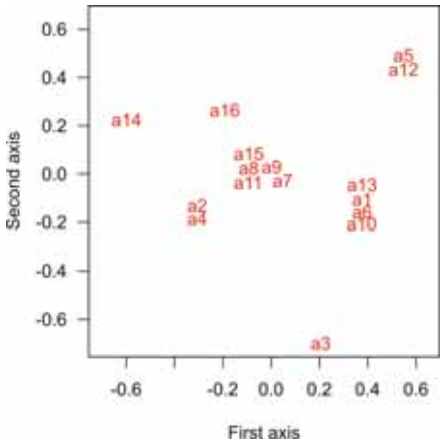
		VARIABLES (ant species)															
	Site	**a1**	**a2**	**a3**	**a4**	**a5**	**a6**	**a7**	**a8**	**a9**	**a10**	**a11**	**a12**	**a13**	**a14**	**a15**	**a16**
	1. CT	0	1	0	1	0	0	0	1	1	0	1	0	0	0	1	1
UNITS	**2. MA.mainland**	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1	0
	3. MA.islands	1	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1
	4. VT	0	1	0	1	0	0	1	1	1	0	1	0	0	1	1	1

The first goal is to visualize the relationships between the different ant species among sites.

Correspondence Analysis for species presence/absence data

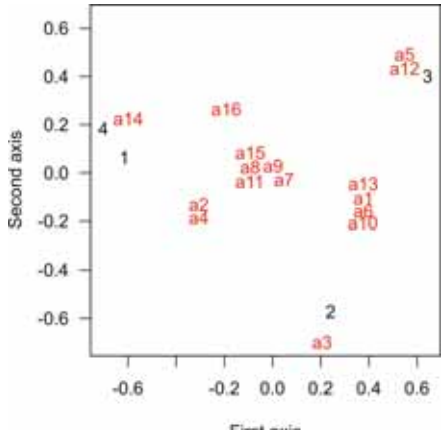
Correspondence Analysis uses a measure of species associations based on differences between observed cell counts and expected counts under independence in a contingency table of species presence/absence or abundances.

Applied to the ants it produces a plot in which points close together on the axes indicate species that tend to occur together at sites. Points far apart indicate species that occur together infrequently.



Correspondence Analysis for species presence/absence data

Likewise, the sites can be compared by the observed and expected numbers of species shared. The method can thus ordinate sites and species simultaneously on the same axes. Species next to sites in the plot indicate species that occur predominantly there, whereas species falling between site points are shared among sites.



Correspondence Analysis for species presence/absence data

By rearranging the site and ant species by their order along the first axis, we can see the "correspondence" between sites and species.

	Site	a14	a2	a4	a16	a15	a8	a11	a9	a7	a3	a1	a13	a6	a10	a5	a12
	4. VT	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
UNITS	**1. CT**	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	2. MA.mainland	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0
	3. MA.islands	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1

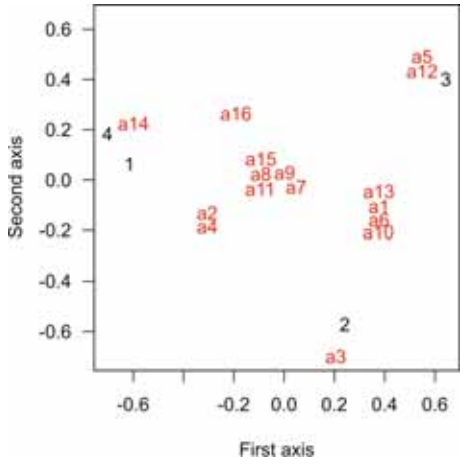
This first axis behaves like an ecological gradient, with species found at all sites clustered in the middle, and species with non-overlapping distributions at either end.

Correspondence Analysis for species presence/absence data

The first axis explains most of the variation in the data. It also maximizes the association between units (sites, rows) and variables (species, columns).

Eigenvalues corresponding to each axis indicate the correlation between species and site scores.

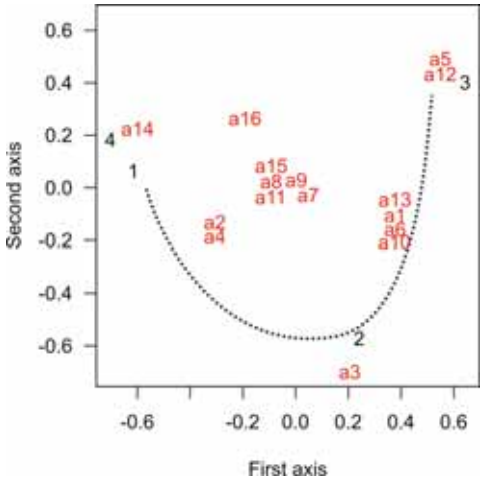
For the ant data:
Axis 1: 0.56
Axis 2: 0.39



Correspondence Analysis for species presence/absence data

A worry with CA in general is the "arch" or, more seriously, a "horseshoe" in the ordination. When sites at the ends of ecological gradients have few species in common, they may be arranged such that they appear similar by the fact that they are missing many of the same species.

It arises because the measure of distance between assemblages is not linear and doesn't increase with increasing "ecological distance". The problem occurs mostly when beta diversity is high.



Distance data

Can be analyzed similarly.

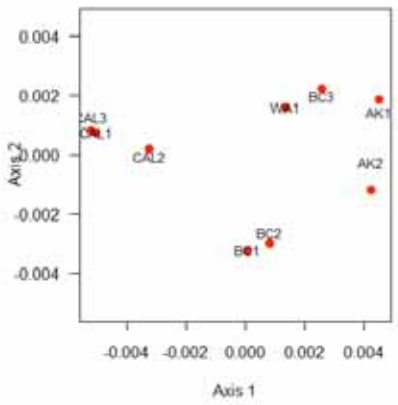
Example: Percent sequence divergence between Pacific Ocean basin freshwater stickleback populations in over 1000 bp of the Eda gene.

	BC1	BC2	WA1	CAL1	AK1	AK2	BC3	CAL2	CAL3
BC1	0.00000	0.00076	0.00454	0.00683	0.00680	0.00454	0.00605	0.00531	0.00681
BC2	0.00076	0.00000	0.00530	0.00758	0.00605	0.00530	0.00529	0.00607	0.00757
WA1	0.00454	0.00530	0.00000	0.00683	0.00302	0.00378	0.00151	0.00530	0.00681
CAL1	0.00683	0.00758	0.00683	0.00000	0.00988	0.00988	0.00834	0.00531	0.00379
AK1	0.00680	0.00605	0.00302	0.00988	0.00000	0.00302	0.00151	0.00833	0.00983
AK2	0.00454	0.00530	0.00378	0.00988	0.00302	0.00000	0.00530	0.00835	0.00985
BC3	0.00605	0.00529	0.00151	0.00834	0.00151	0.00530	0.00000	0.00682	0.00832
CAL2	0.00531	0.00607	0.00530	0.00531	0.00833	0.00835	0.00682	0.00000	0.00378
CAL3	0.00681	0.00757	0.00681	0.00379	0.00983	0.00985	0.00832	0.00378	0.00000

Distance data

Various forms of "multidimensional scaling" using distance measurements. The overall goal is to produce a plot in which data points that are similar are close to one another in the plot, whereas points apart from one another are more similar.

Distances may not be preserved, eg in non-metric scaling only the rank order of distances between points is maintained.



Discussion paper next week:

Harvey and Rambaut, 2000. Comparative analyses for adaptive radiations.

Download from "**assignments**" tab on course web site.

Presenters: Yasha P. and Gyan H.