Introduction to multivariate analysis

Outline

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

Data are usually multivariate

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

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

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

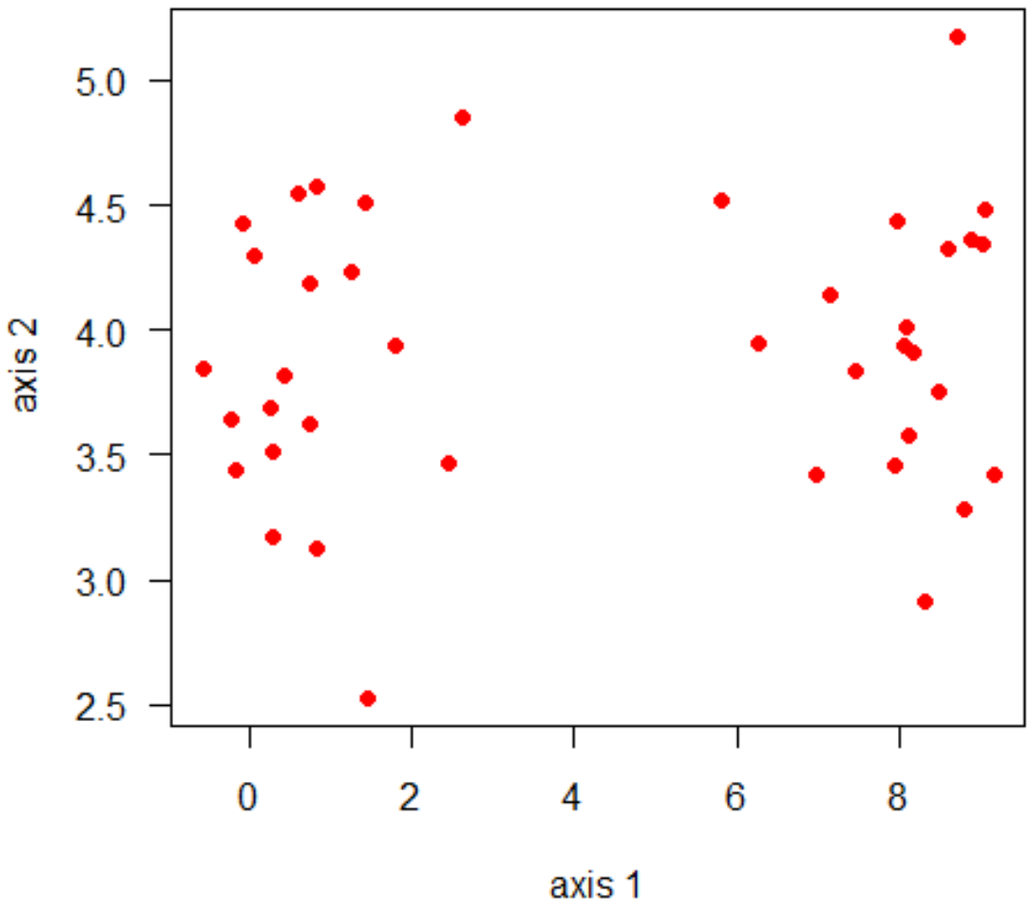
Ordination is usually what we want:

Arrange sampling units along gradients or according to combinations of variables

- To visualize complex data in few dimensions
- To find meaningful combinations of the original variables that can be used in subsequent analyses

Large number of variables

multivariate analysis



Ordination, Classification, Model fitting

Multivariate methods are used for

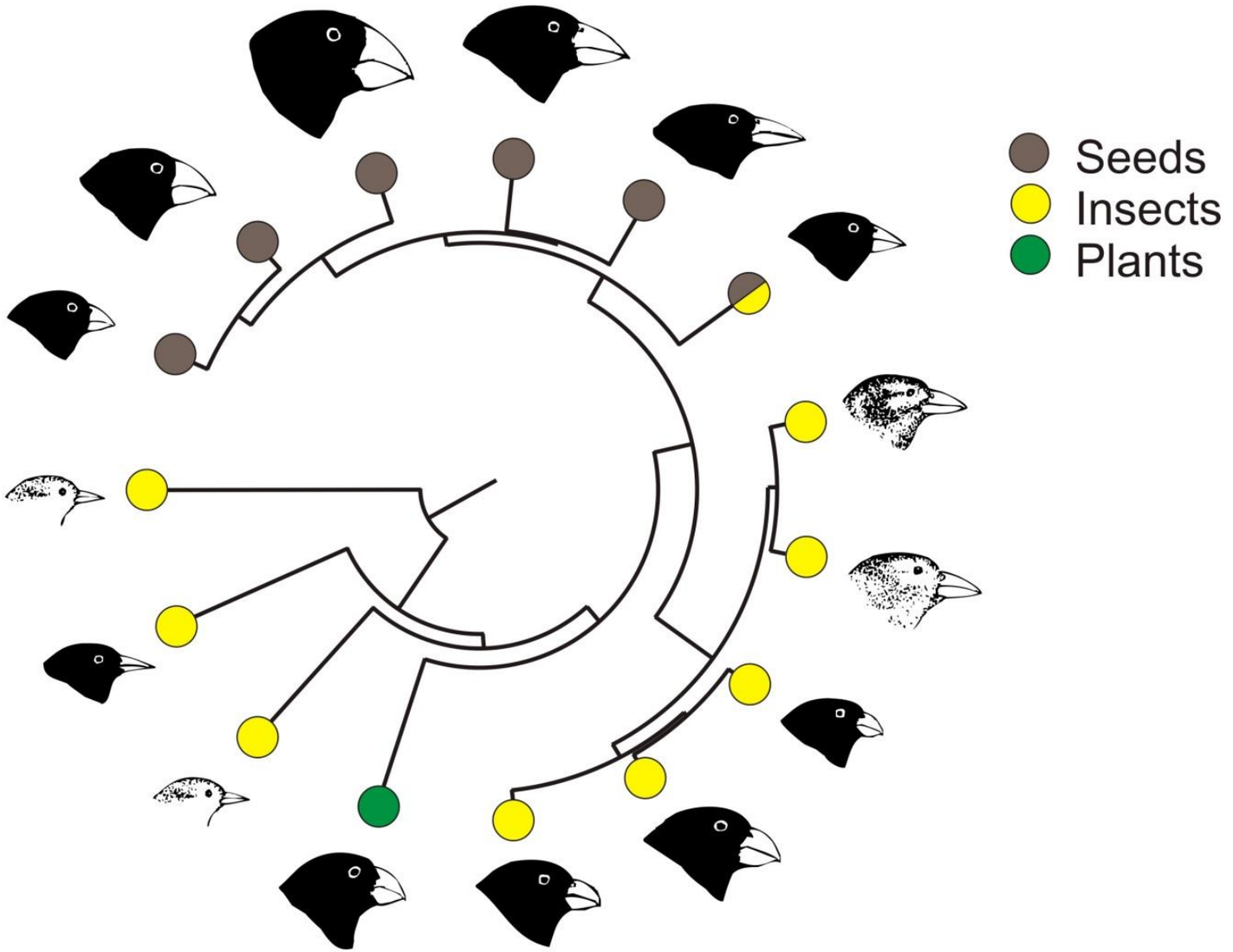
- Ordination: arrange sampling units along composite variables
- Classification: place sampling units into groups
- Model fitting: multivariate analysis of variance; multiple regression

Today, I'll focus mainly on ordination.

We'll start with Principal Components Analysis because it is the most straightforward multivariate method.

Principal Component Analysis

Example 1: Differences in beak and body dimensions of Darwin's finches



Principal Component Analysis

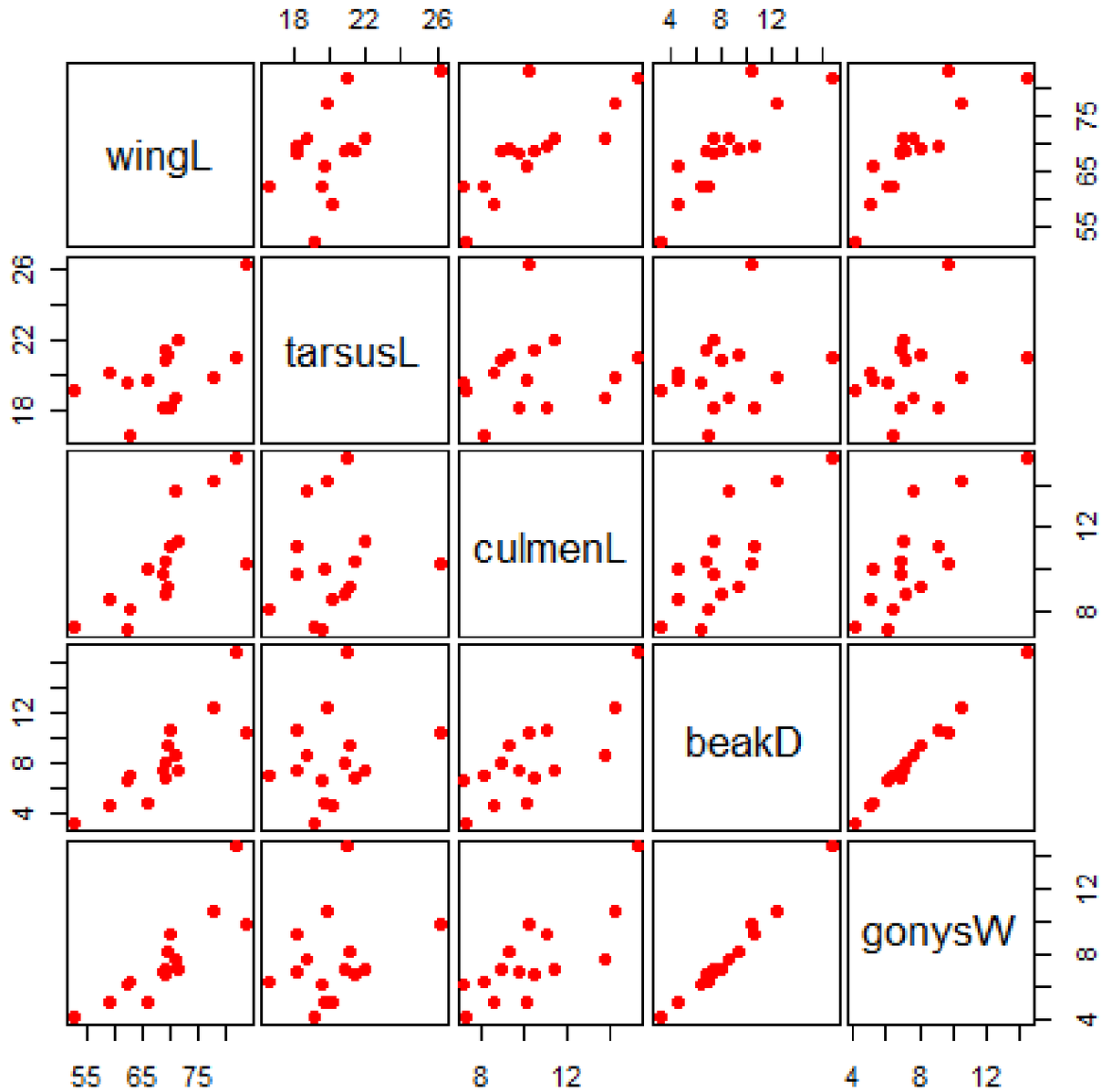
mydata: Data frame of means for 14 species of 5 traits

		VARIABLES (traits)				
		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 (species)	***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**

Principal Component Analysis

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

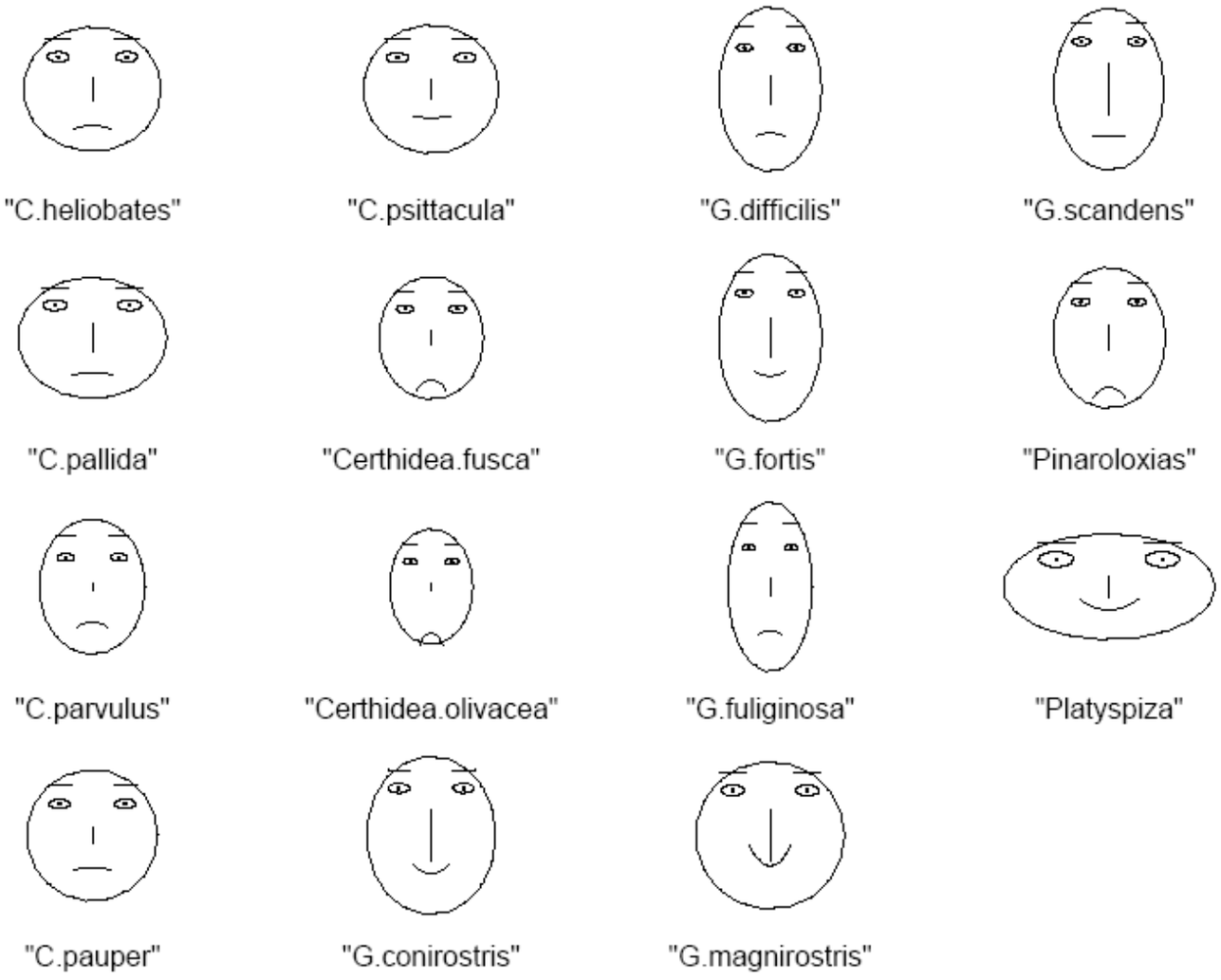
```
pairs(mydata)
```



Principal Component Analysis

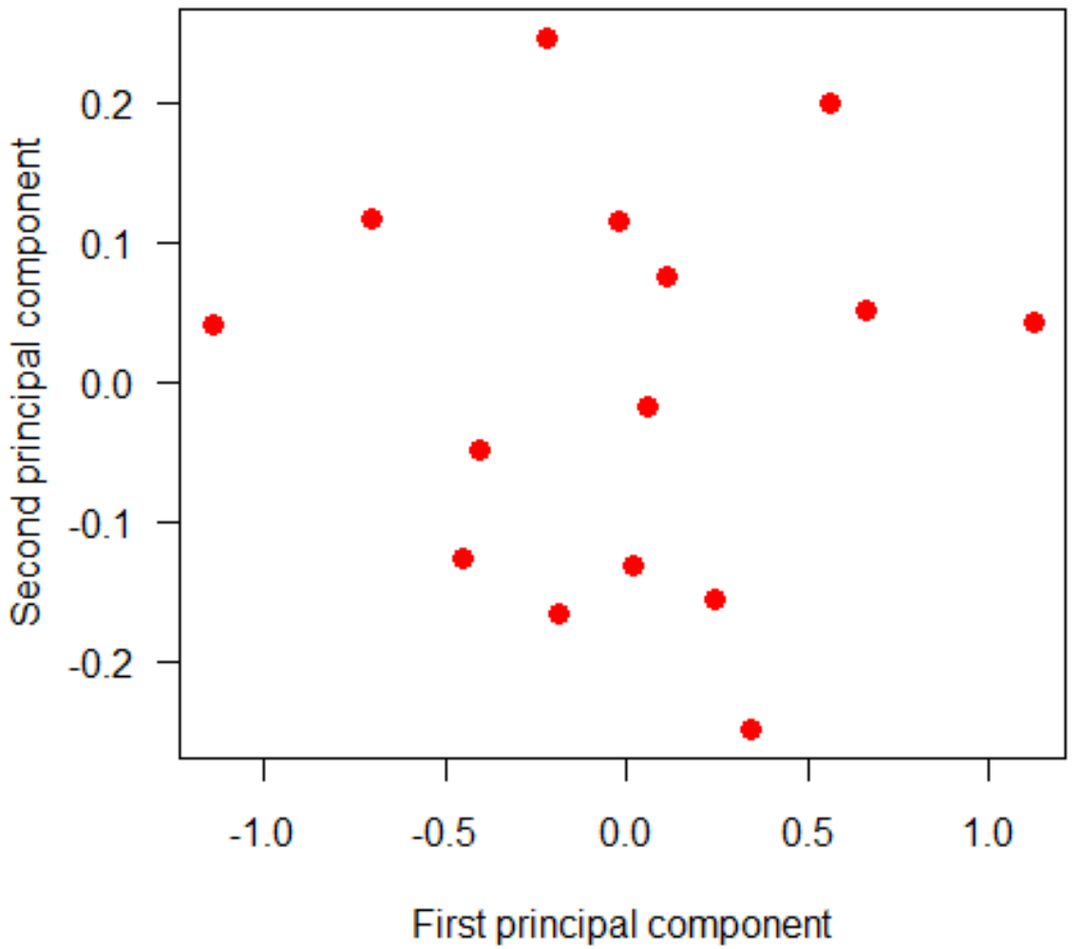
How to visualize multivariate data?

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.



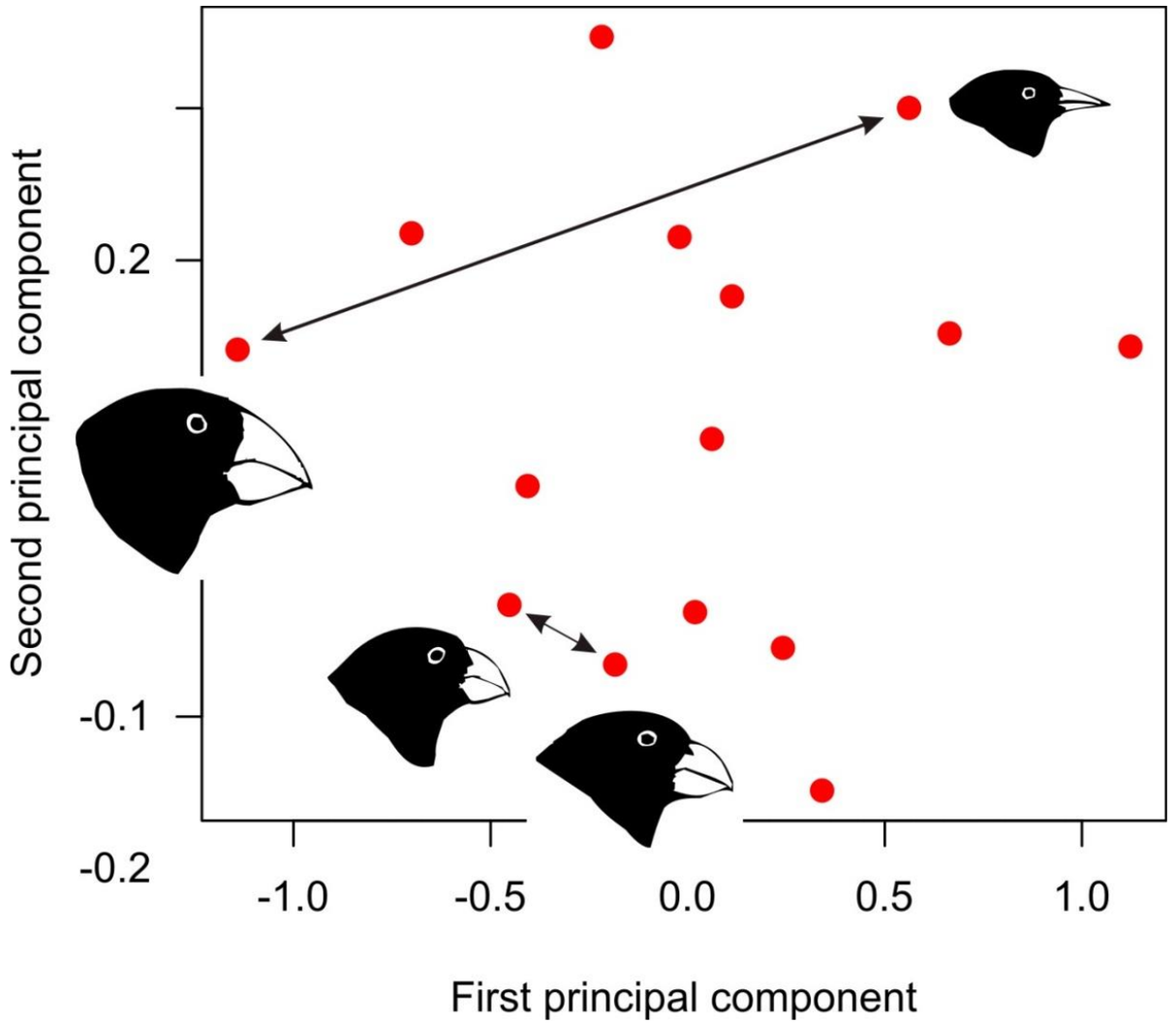
Principal Component Analysis

PC's are linear combinations of the original variables. 96% of all the variation among the Darwin's finch species is described by just 2 dimensions or principal components. This is because the traits strongly co-vary among species.



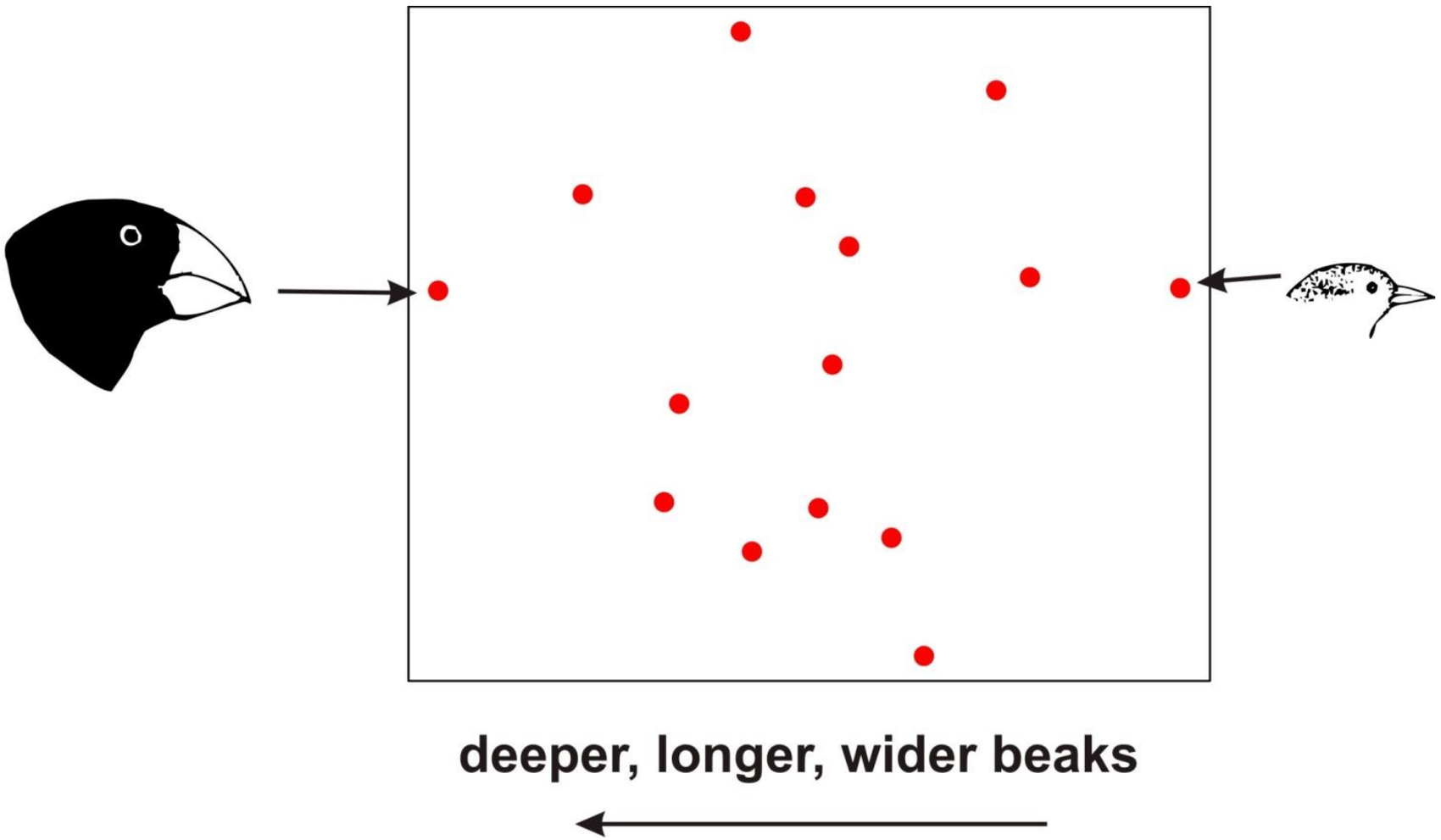
Principal component analysis

Even though we're visualizing 5 dimensions of data in only 2 dimensions, distances between species are approximately preserved. Points close together indicate species that are similar. Points far apart indicate species that are more different.



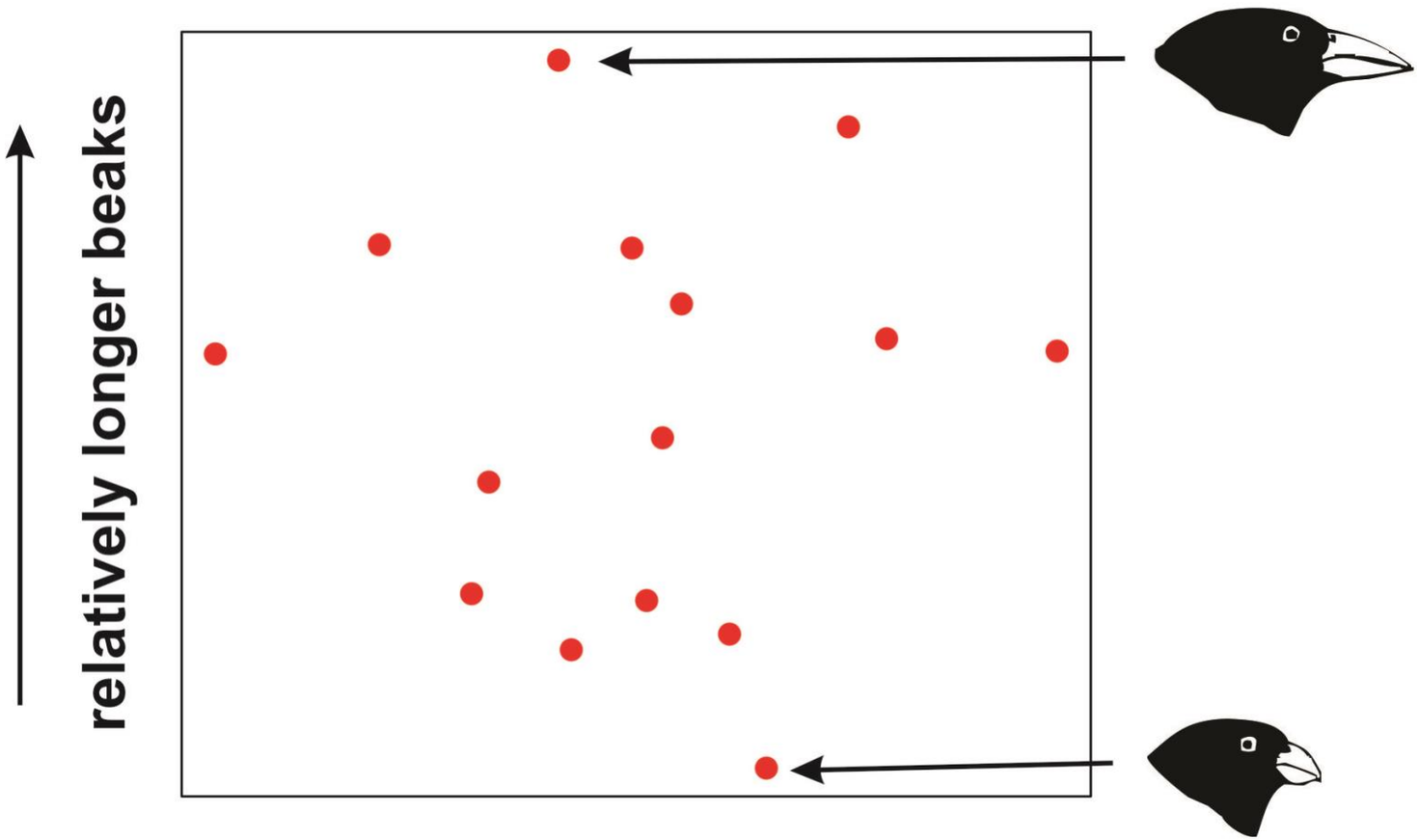
Principal component analysis

Even though they are composite variables, the axes in this case are interpretable. The first axis arranges the species according to differences in overall beak size.



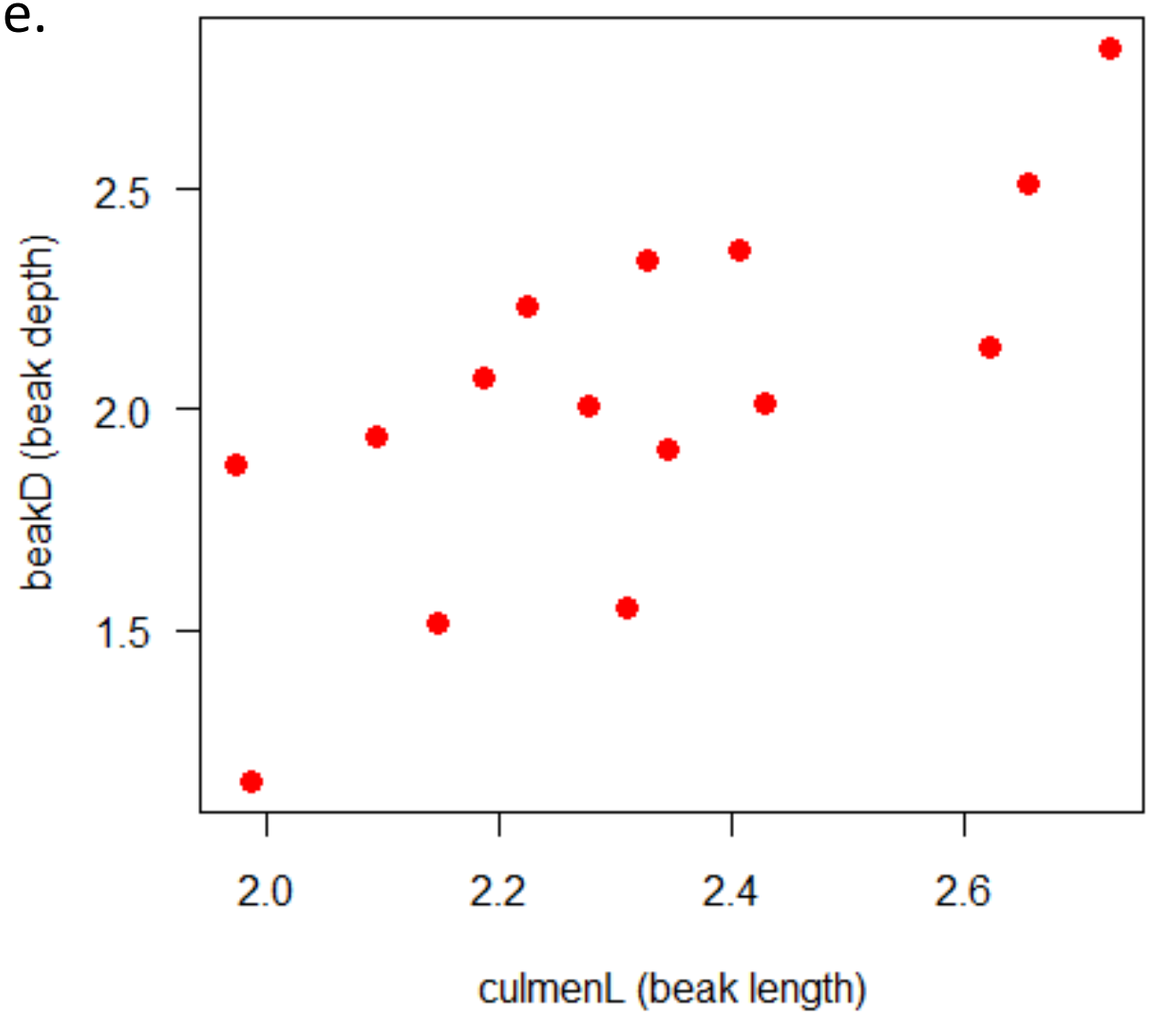
Principal component analysis

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



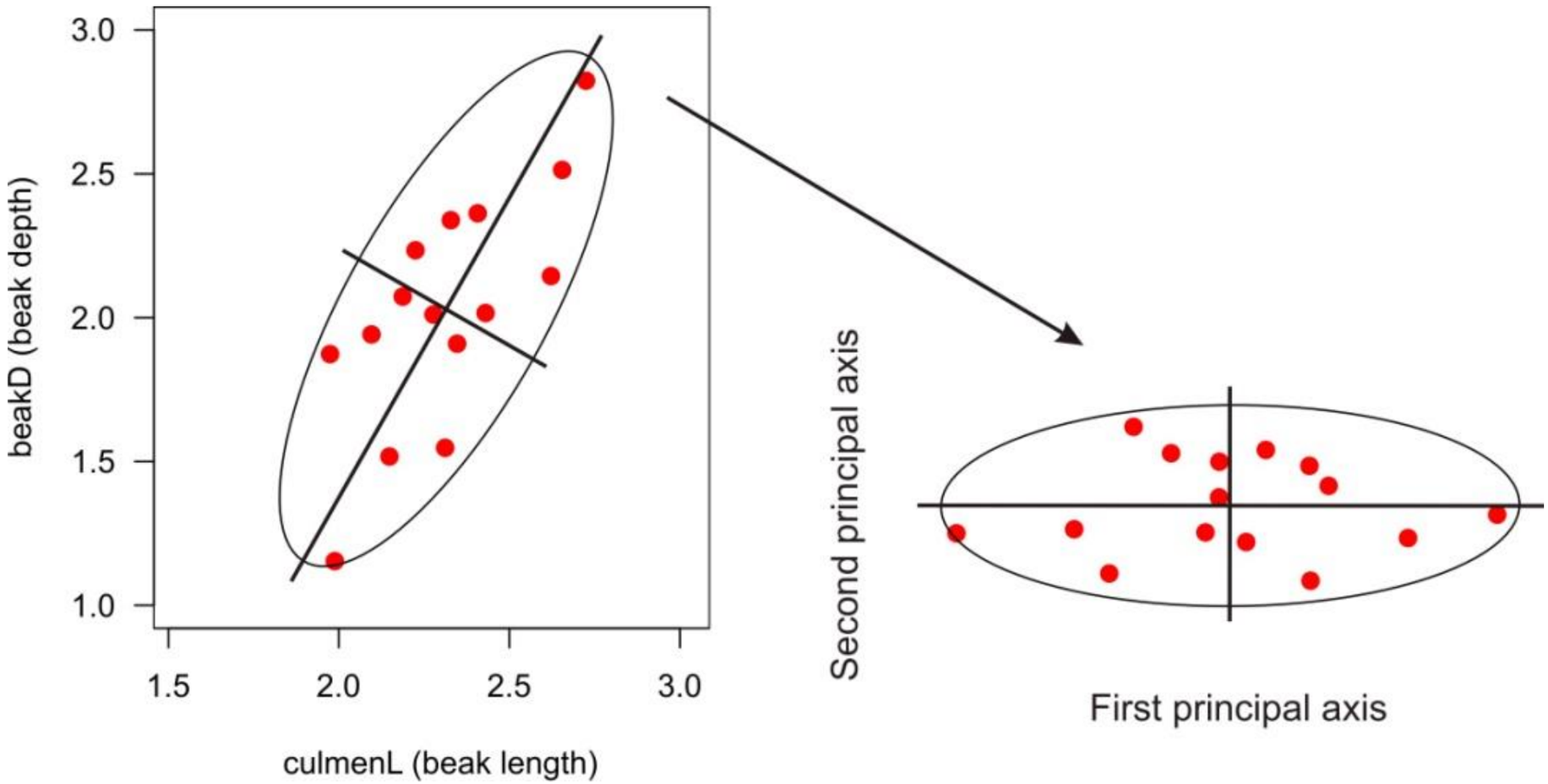
Principal component analysis – how it works

What does PCA do? The method amounts to nothing more than a rotation of the axes, allowing you to view much of the data in a small number of dimensions. To see this, imagine just two traits were measured instead of 5. I've log-transformed all the variables to put on a similar scale.



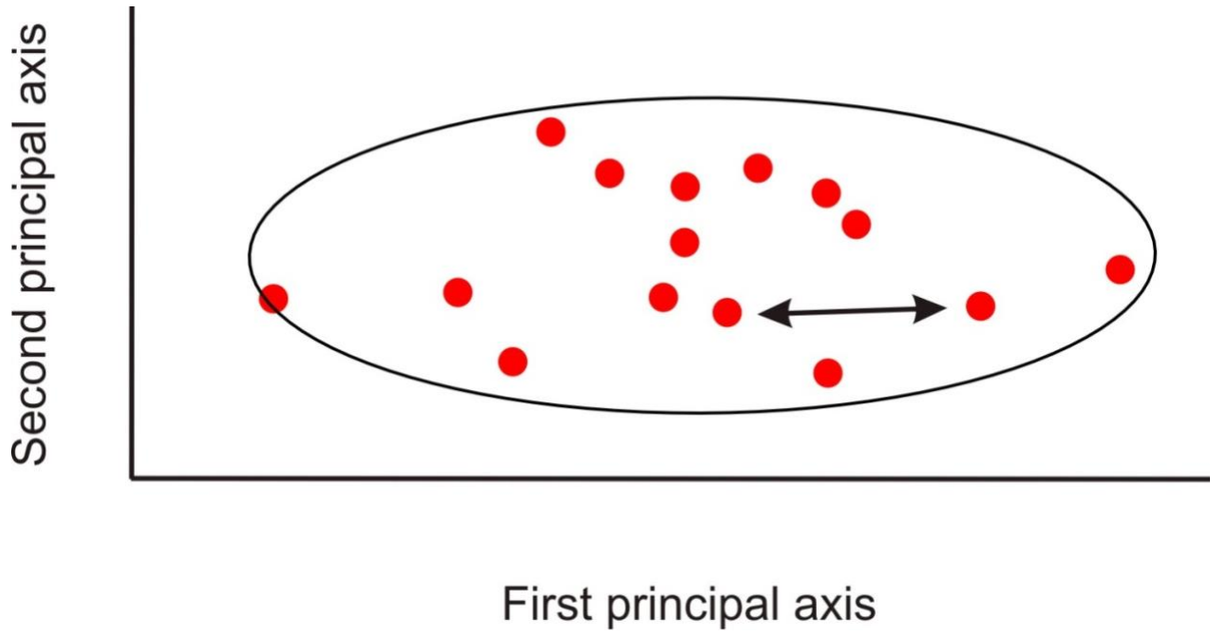
Principal component analysis – how it works

Principal components analysis just rotates the original scatter of points so that new axes are uncorrelated.



Principal component analysis – how it works

Because it is nothing more than a rotation of the axes, the distances between pairs of points (species) are unchanged by the transformation, provided that all the PC axes are retained*.



*Warning: in some stats programs the default procedure is to standardize the variables (using “correlation matrix” instead of the covariance matrix) before carrying out the analysis. Use the correlation matrix only if variables lack a common scale. Euclidean distances will then be based on standardized data, not the original measurements.

Principal component analysis – how it works

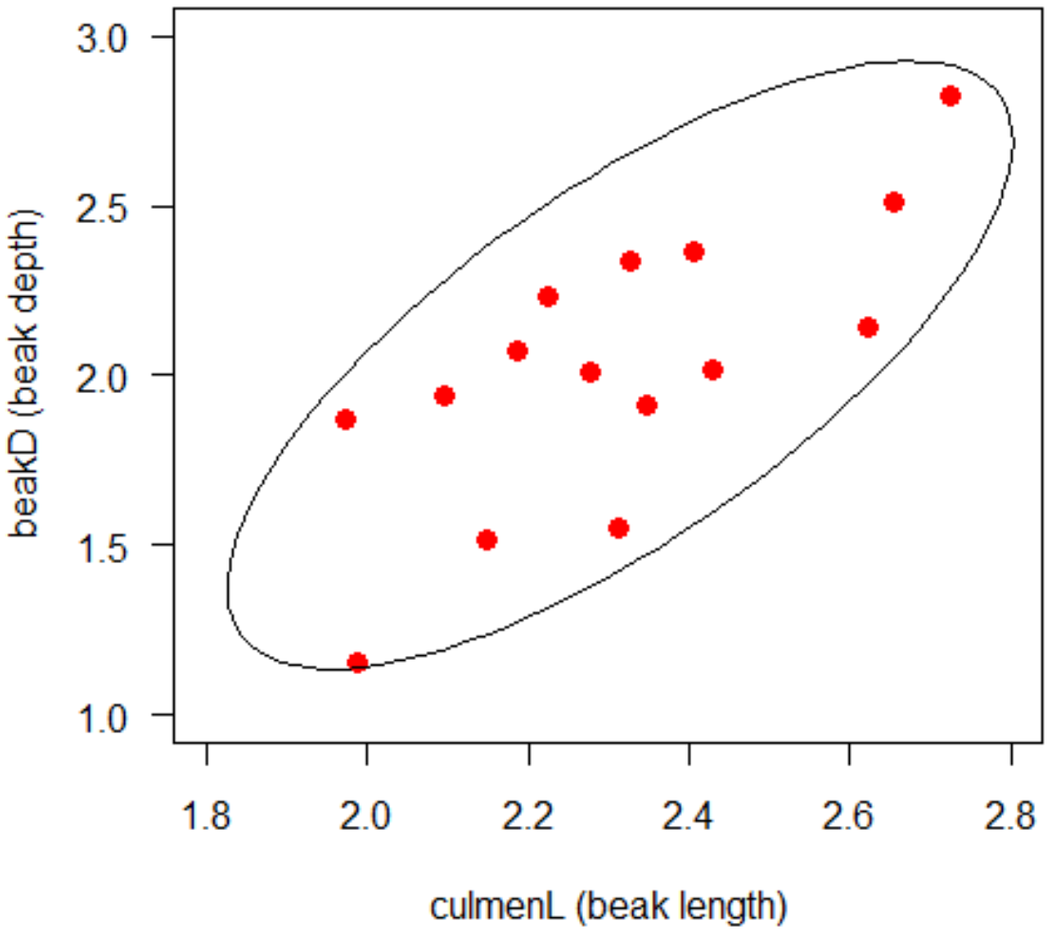
The computations behind the scenes involve describing the data by the associations between the variables (as illustrated by the ellipse)

The elements of the covariance matrix:

	x_1	x_2
x_1	$\text{var}(x_1)$	$\text{cov}(x_1,x_2)$
x_2	$\text{cov}(x_1,x_2)$	$\text{var}(x_2)$

For the two finch variables:

	culmenL	**beakD**
culmenL	0.0518	0.0698
beakD	0.0698	0.1741



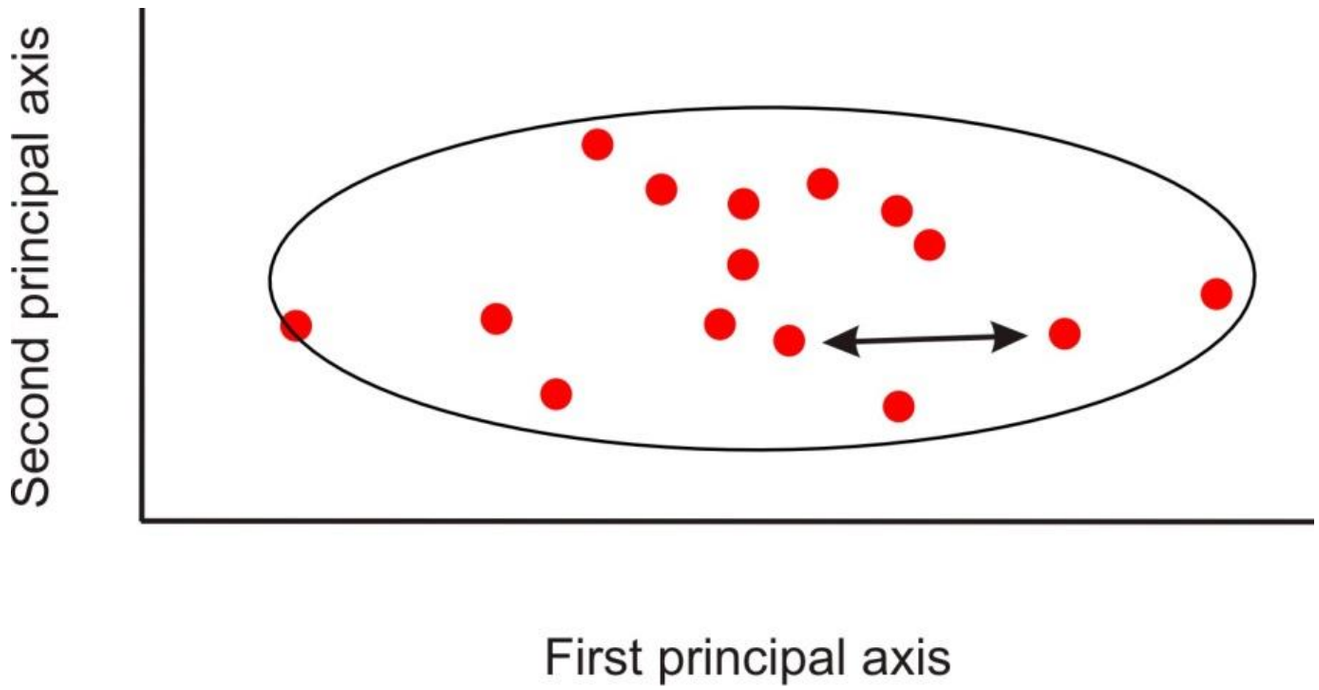
Principal component analysis

The covariance matrix of the new, composite variables has variances on the diagonal and zeros off the diagonal. The component axes have zero covariance.

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

These variances are called the **eigenvalues**

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

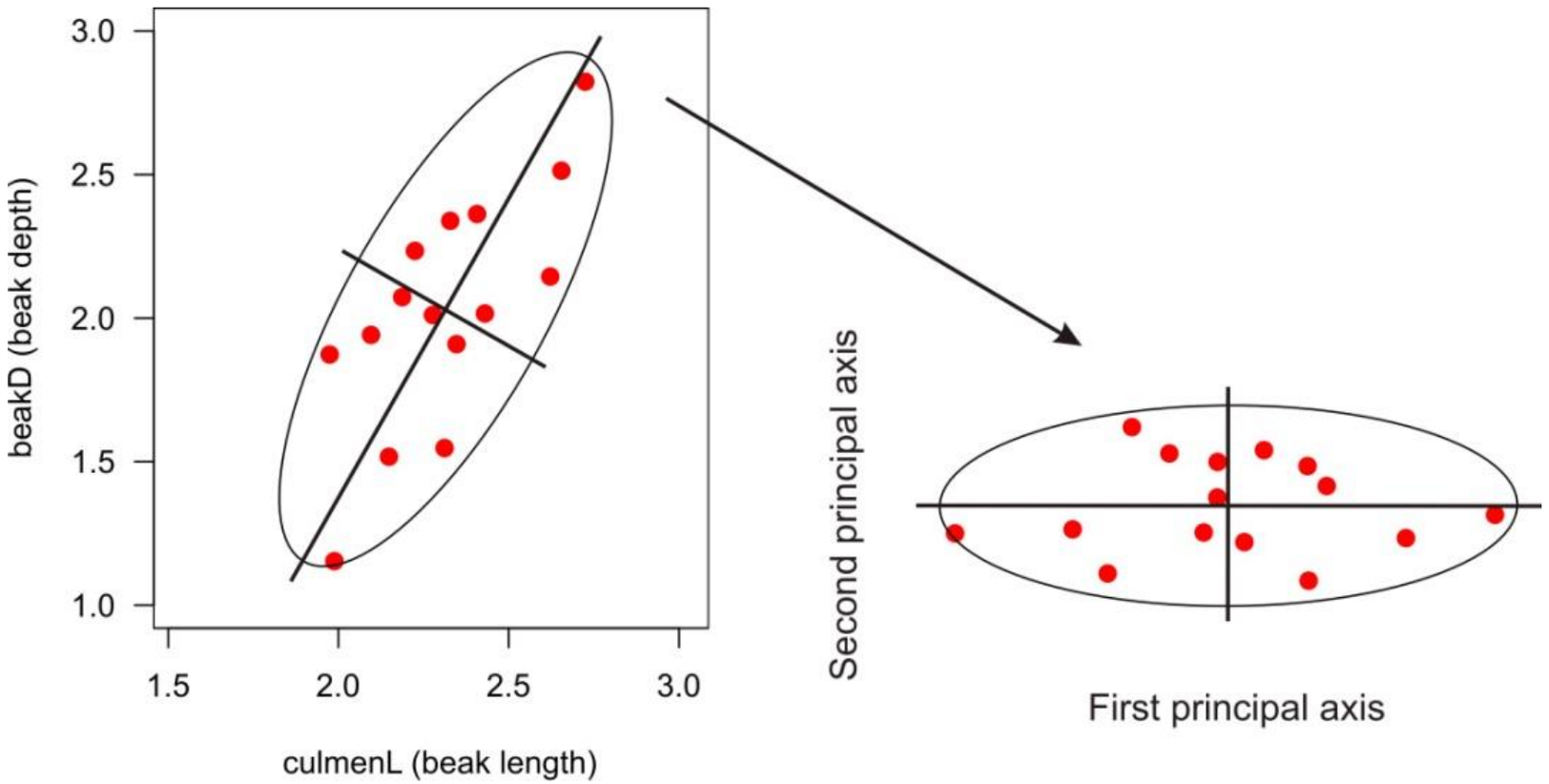


Principal component analysis

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

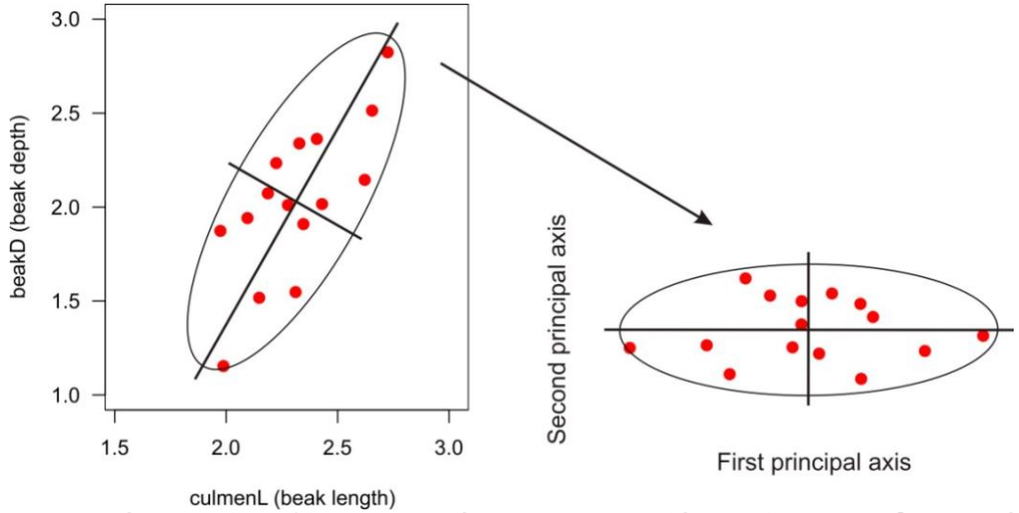
	eigenvec1	**eigenvec2**
culmenL	0.413	-0.911
beakD	0.911	0.413

These constants are called **loadings.**



Principal component analysis

	eigenvec1	eigenvec2
culmenL	0.413	-0.911
beakD	0.911	0.413



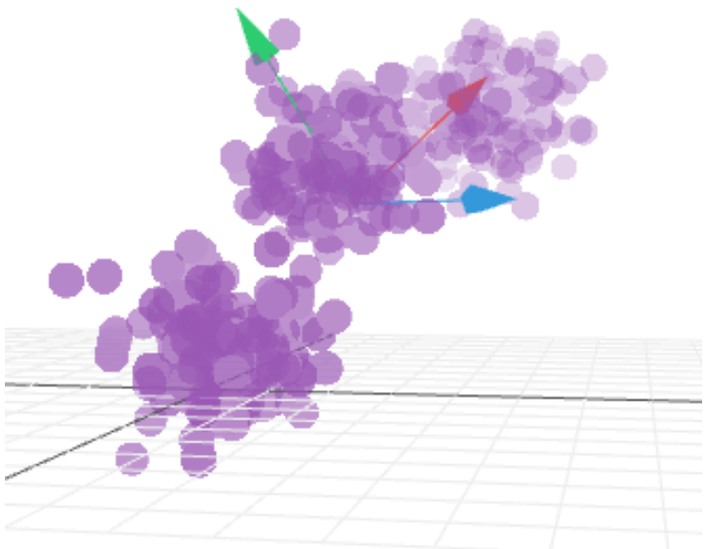
The constants are called **loadings** because they indicate the contribution of each variable to the principal component.
pc1 can be interpreted as “beak size” because both beak traits make contributions having the same sign (depth contributes more than length). The axis separates big-beaked birds at one extreme from small-beaked birds at the other.

pc2 is interpreted as “beak shape” because beak depth loads positively but beak length negatively. It separates short deep beaks at one end from long shallow beaks at the other.

Principal component analysis

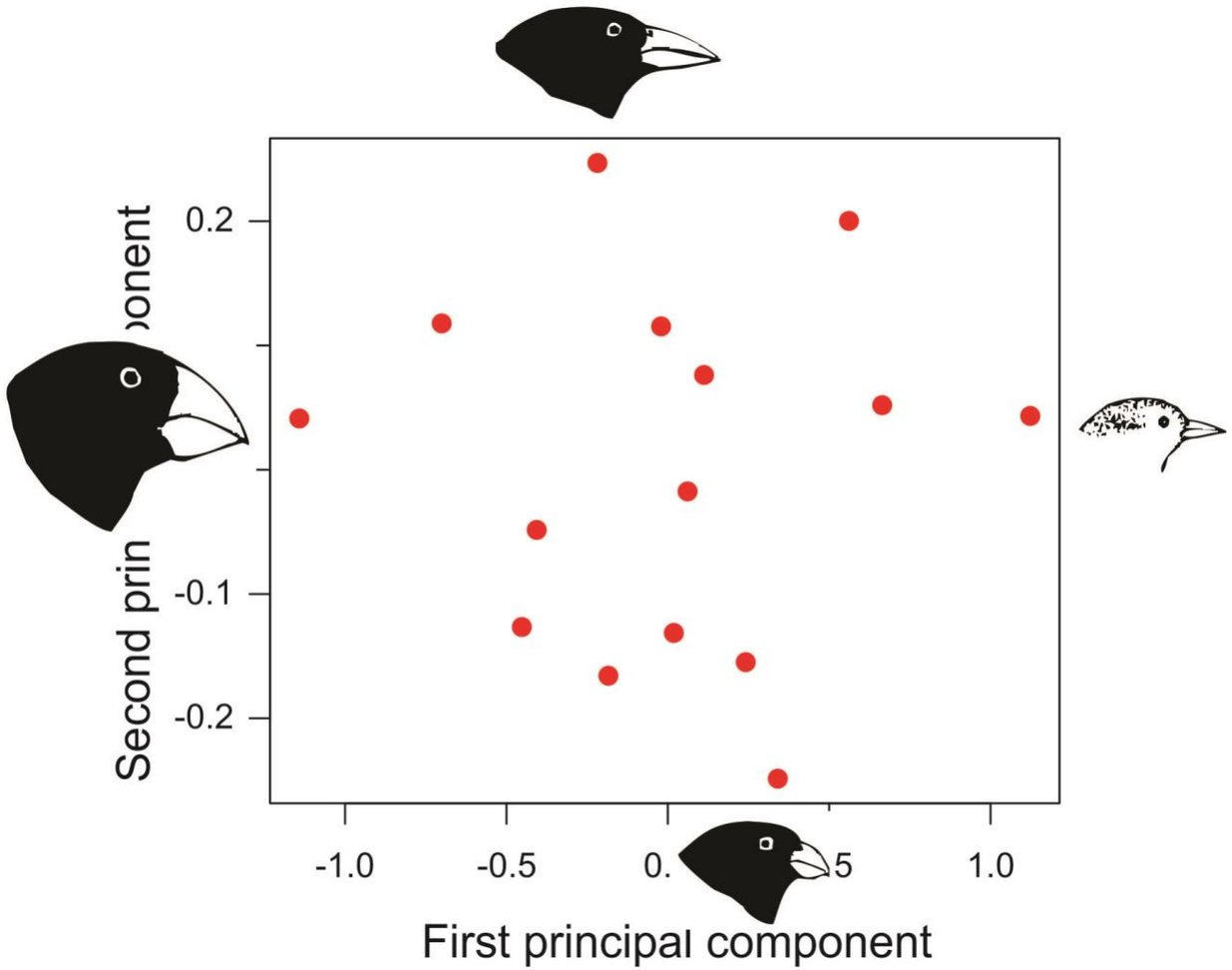
We can still visualize this process with 3 variables instead of 2 (it gets hard to visualize with more than 3 variables)

3D example



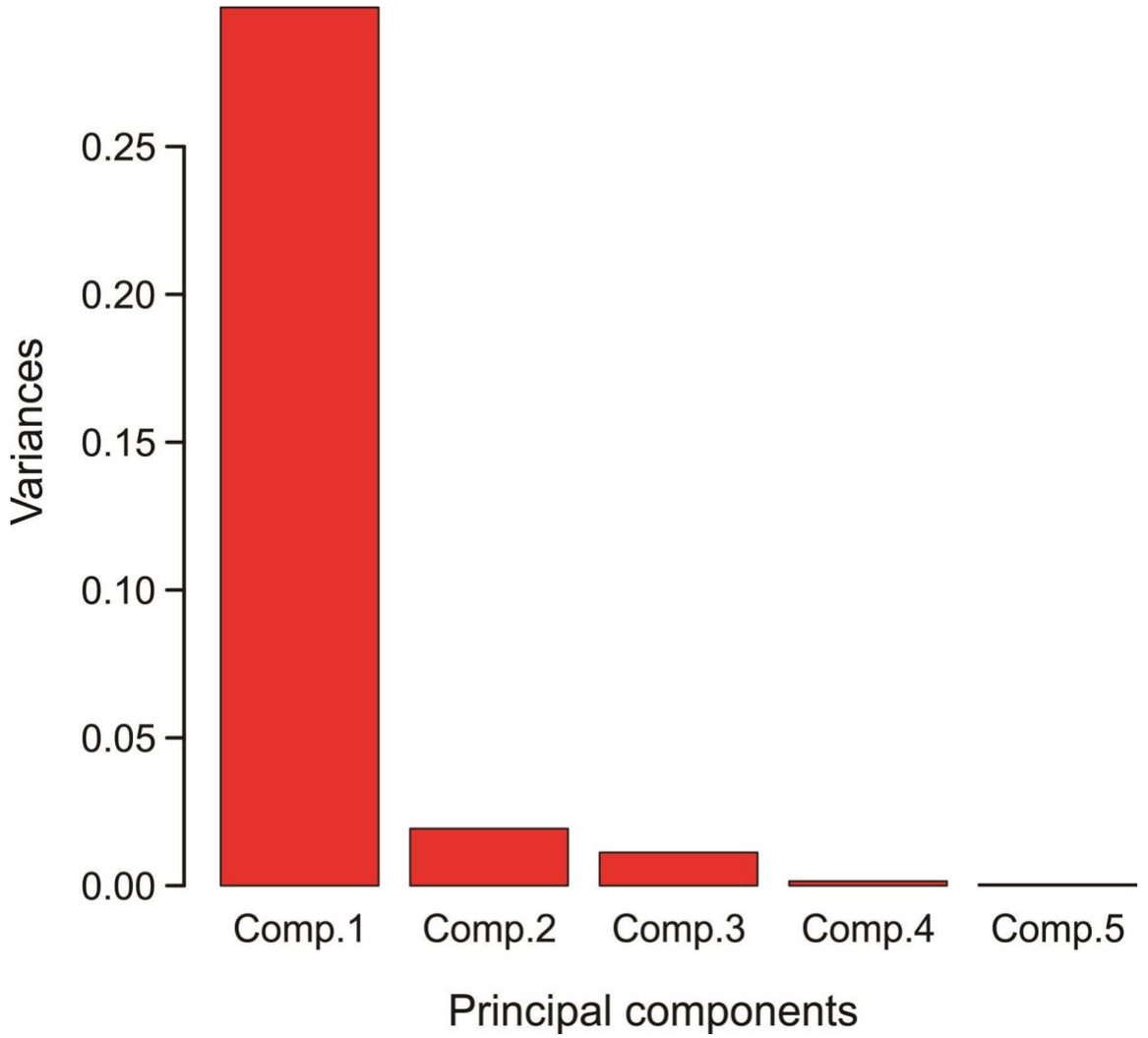
Principal component analysis

The idea is the same with 5, or any number of variables. The plot below is of the first two PC axes from the PC 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 <u>all</u> the differences among the species.



Principal component analysis

The eigenvalues tell you how much of the total variation is captured by the first two principal components. These can be visualized in a “scree plot.”



Principal component analysis

The eigenvectors indicate the loadings.

For the Darwin's finch data:
PC1 can be interpreted as "beak size".
PC2 can be interpreted as "beak shape"
PC3 is mainly one trait, tarsus length
Etc…

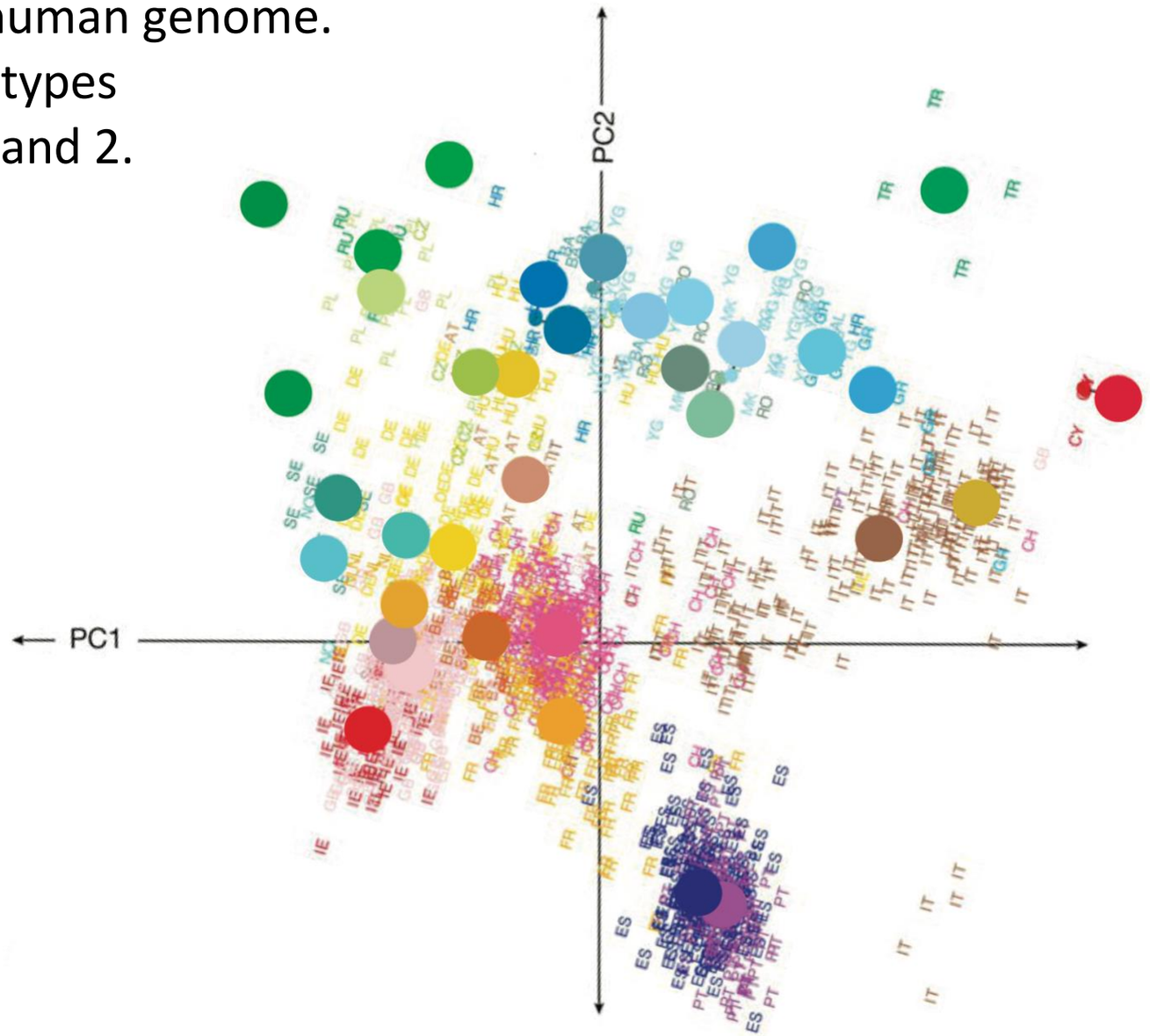
	PC1	PC2	PC3	PC4	PC5
wingL	-0.195	0.062	-0.335	0.577	0.716
tarsusL	-0.052	-0.043	-0.919	-0.071	-0.383
culmenL	-0.326	0.932	0.039	0.002	-0.153
beakD	-0.733	-0.330	0.193	0.395	-0.400
gonysW	-0.562	-0.127	-0.073	-0.711	0.397

Principal component analysis

Example 2: 197146 traits in 1387 individual humans sampled from Europe.
Data are loci (nucleotides) in the human genome.
At every locus, individuals of genotypes
AA, Aa and aa are scored as 0, 1 and 2.
The covariance matrix is
197146 × 197146 in size.

PC1 and PC2 are shown.
Points are individuals.
Colors are countries.
Circles are country means.

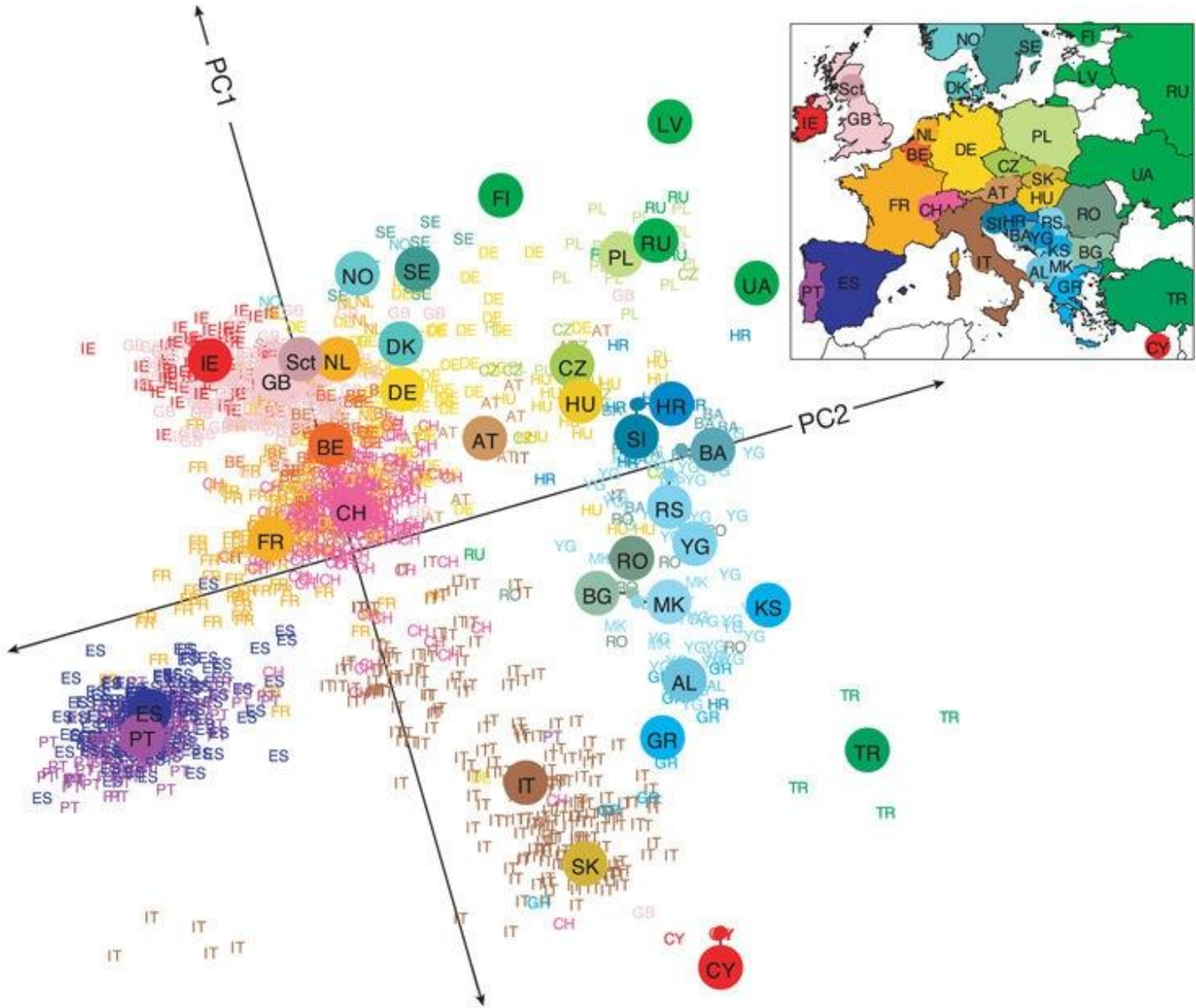
Individuals cluster by country.



Novembre et al (2008) *Nature*

Principal component analysis

Example 2: 197146 traits in 1387 individual humans sampled from Europe. The authors rotated the above figure to the configuration shown, and placed it next to a map of Europe. A strong correspondence was observed between position of individuals along PC1 and PC2 and geographical map position, revealing how land configuration influenced gene flow between countries.



Novembre et al (2008) *Nature*

Correspondence Analysis

An ordination method to visualize spatial associations between species based on their presence/absence (or abundance) at sites. Equivalently, the method visualizes differences among sites in the extent to which they share the same species.

The method is like principal component analysis.

Instead of using a covariance matrix describing association between pairs of traits (as in PCA), correspondence analysis uses a contingency table to measure association between all pairs of species based on their frequency of co-ocurrence (measured relative to the co-occurrence expected under independence).

Equivalently, it measured similarity between all pairs of sites based on the number of shared species (relative to the random expectation).

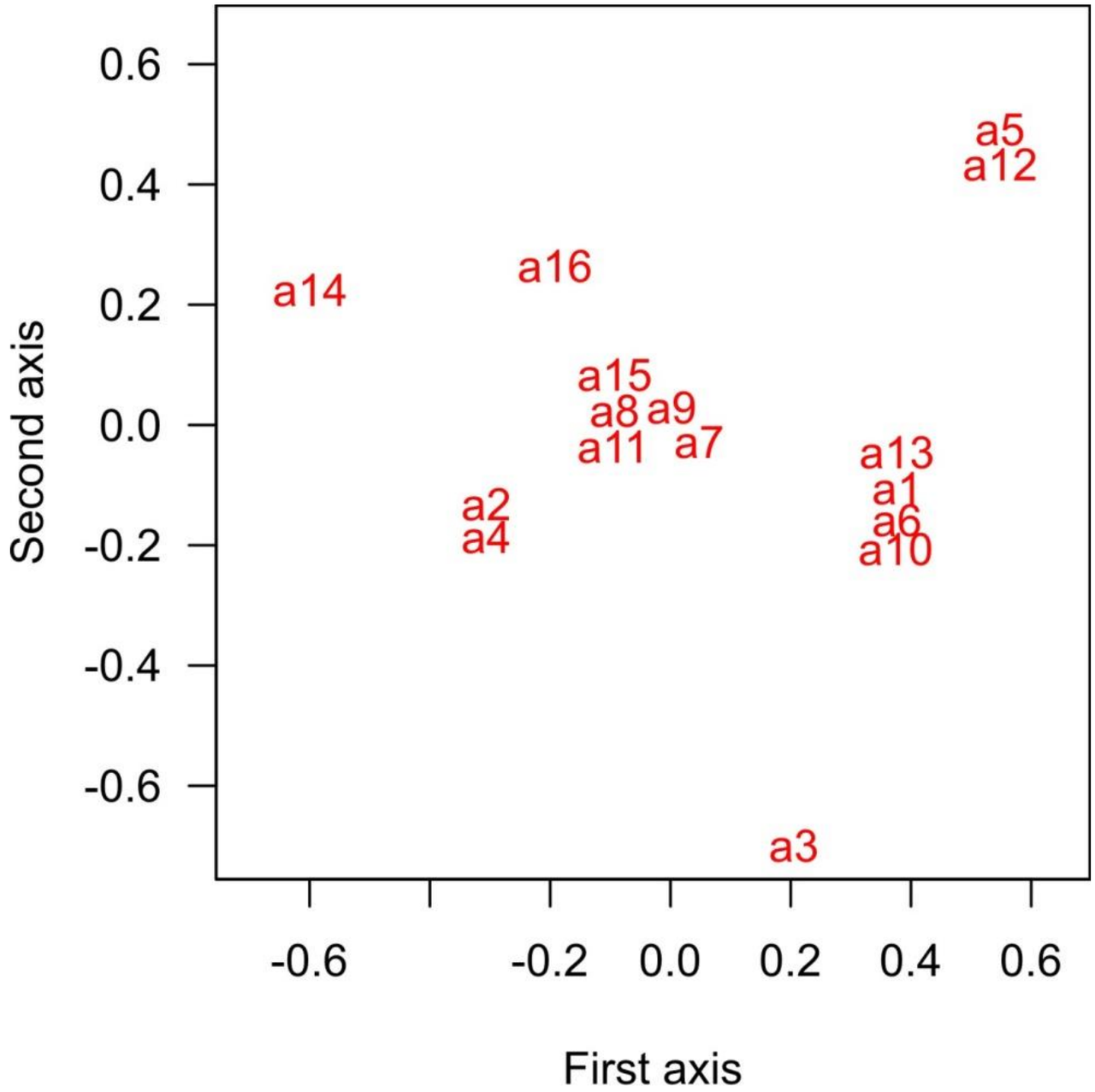
Correspondence Analysis

Example: Presence/absence of 16 ant species in 4 geographic regions (Gotelli and Ellison 2004). In this matrix, the sites are the spatial units, and the species are the variables. The data indicate a species is present at a site (1) or not (0).

		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

Correspondence Analysis

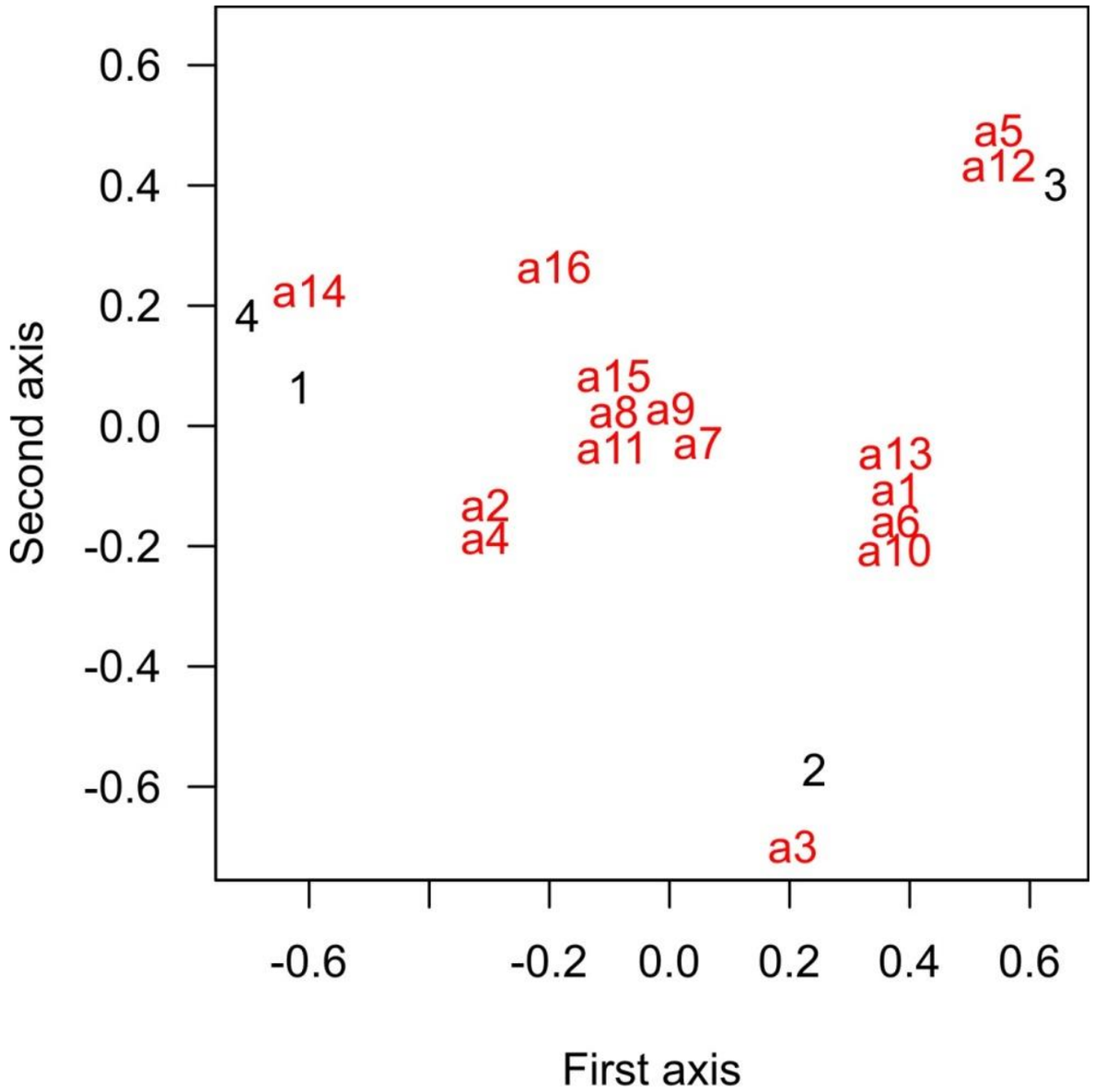
Correspondence analysis produces a plot in which points close together on the axes indicate species that tend to co- occur at sites. Points far apart indicate species that occur infrequently at the same sites.



Correspondence Analysis

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.

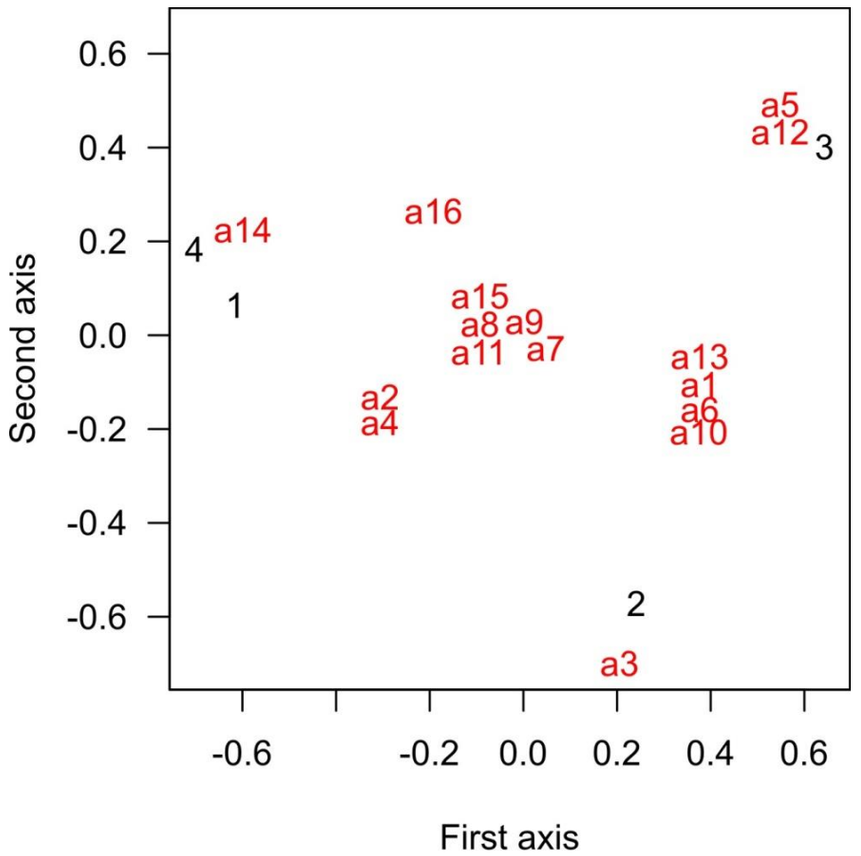
Species next to sites in the plot indicate species that occur predominantly there, whereas species falling between site points are shared among sites.



Correspondence Analysis

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

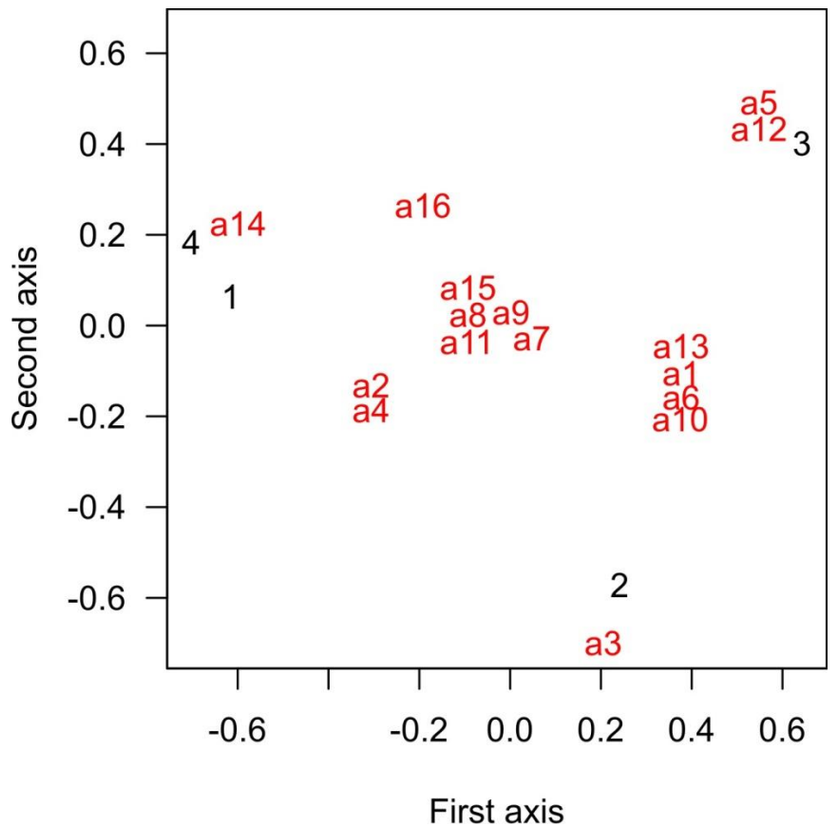
	Site	a14	a2	a4	a16	a15	a8	a11	a9	a7	a3	a1	a13	a6	a10	a5	a12
	4. VT	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
UNITS	**1. CT**	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	2. MA.mainland	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0
	3. MA.islands	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1



Correspondence Analysis

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

	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

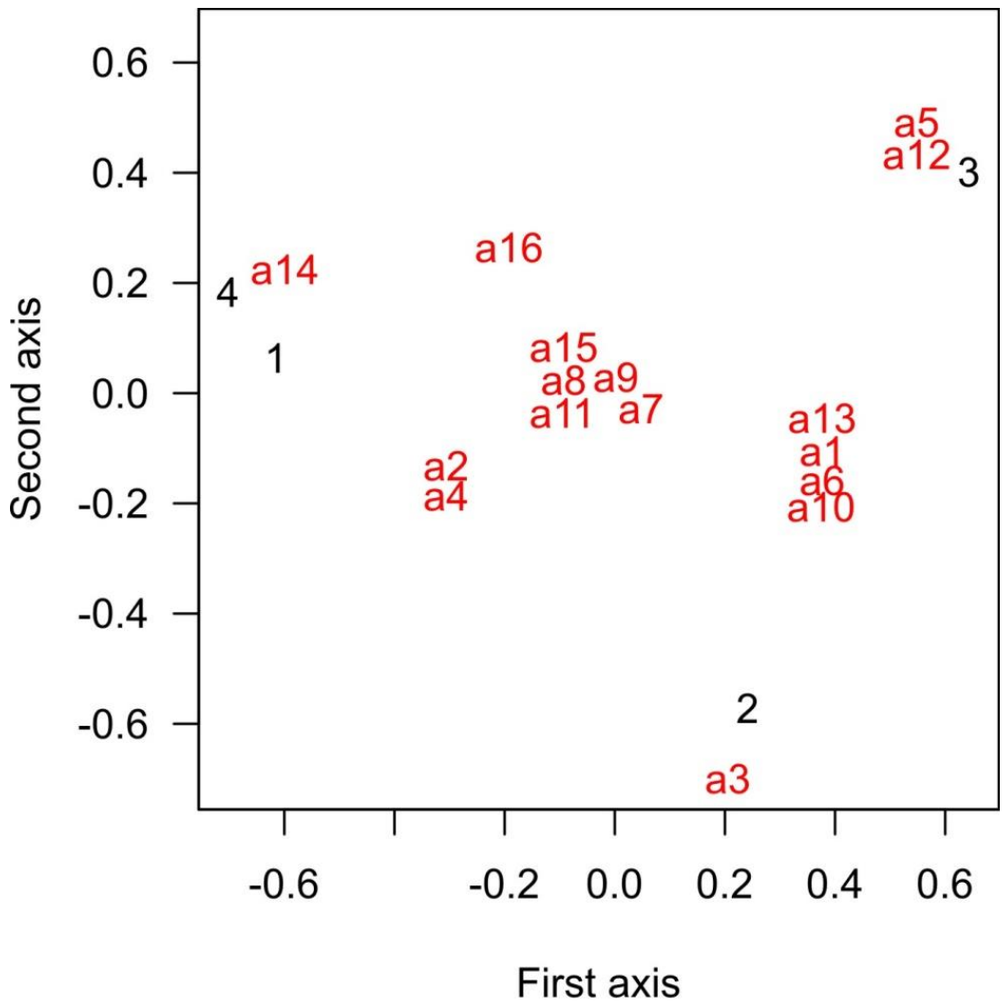


Correspondence Analysis

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 for each axis indicate the correlation between species and site scores.

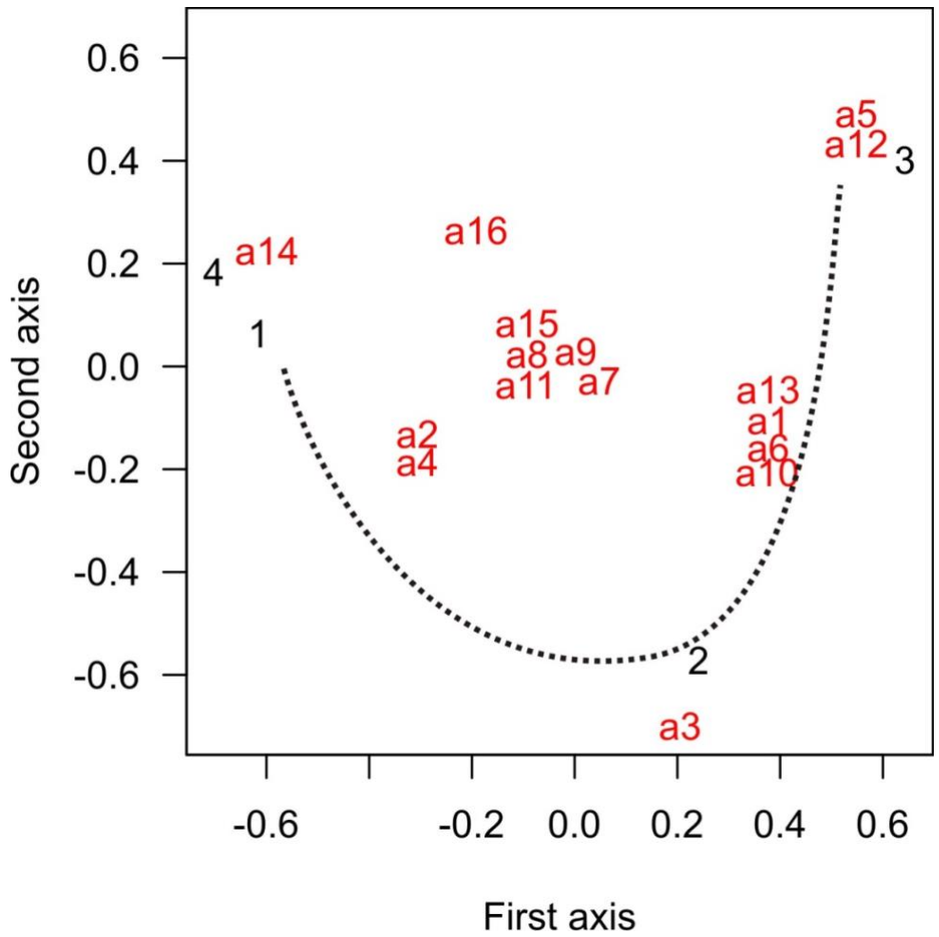
For the ant data:
Axis 1: 0.56
Axis 2: 0.39



Correspondence Analysis

The second axis is often less interpretable, and is sometimes characterized by an "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" on axis 2 only because they are missing some of the same species.

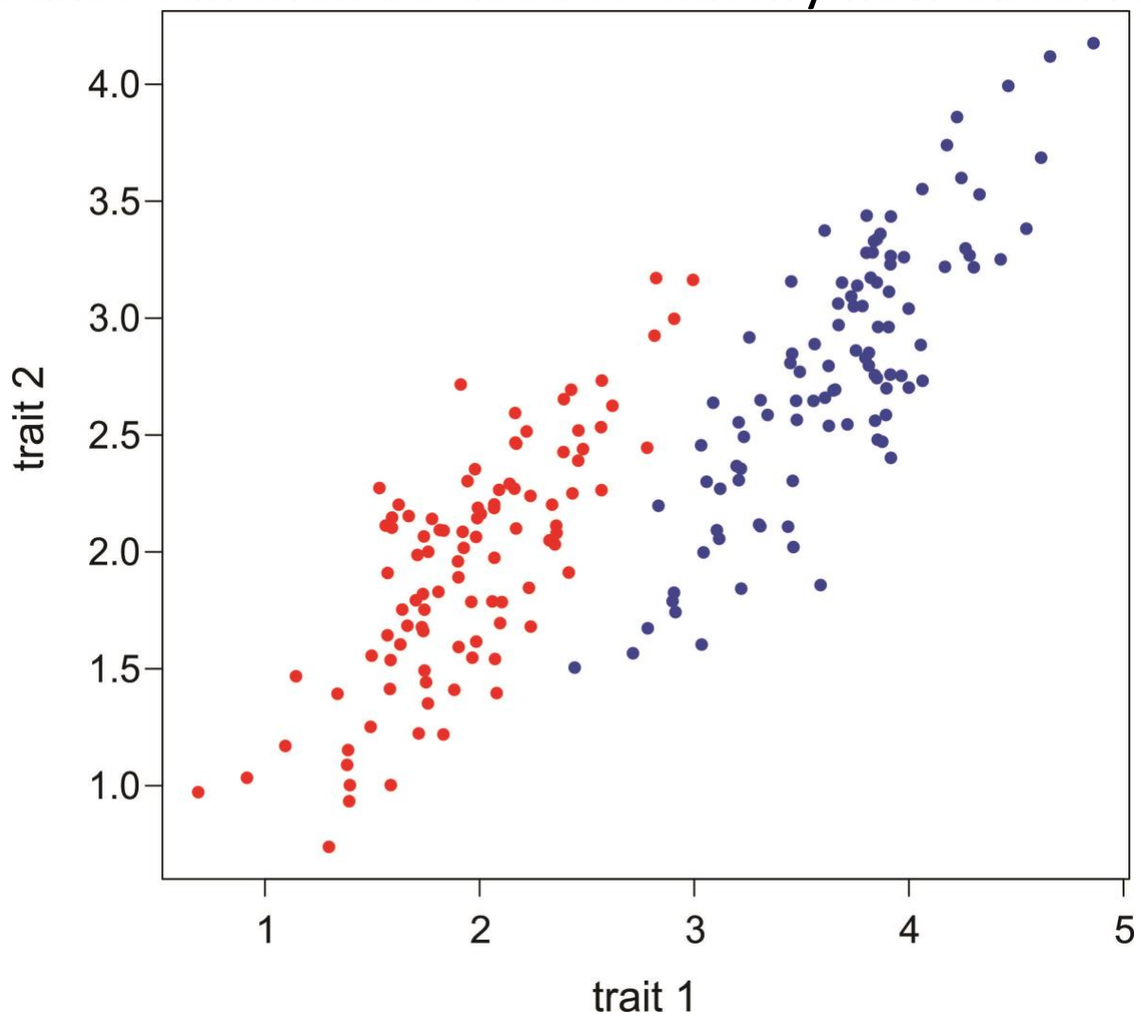
It arises because the measure of distance between assemblages is not linear and doesn't increase with increasing "ecological distance". The problem is most acute when spatial turnover of species (beta diversity) is high.



Discriminant function analysis

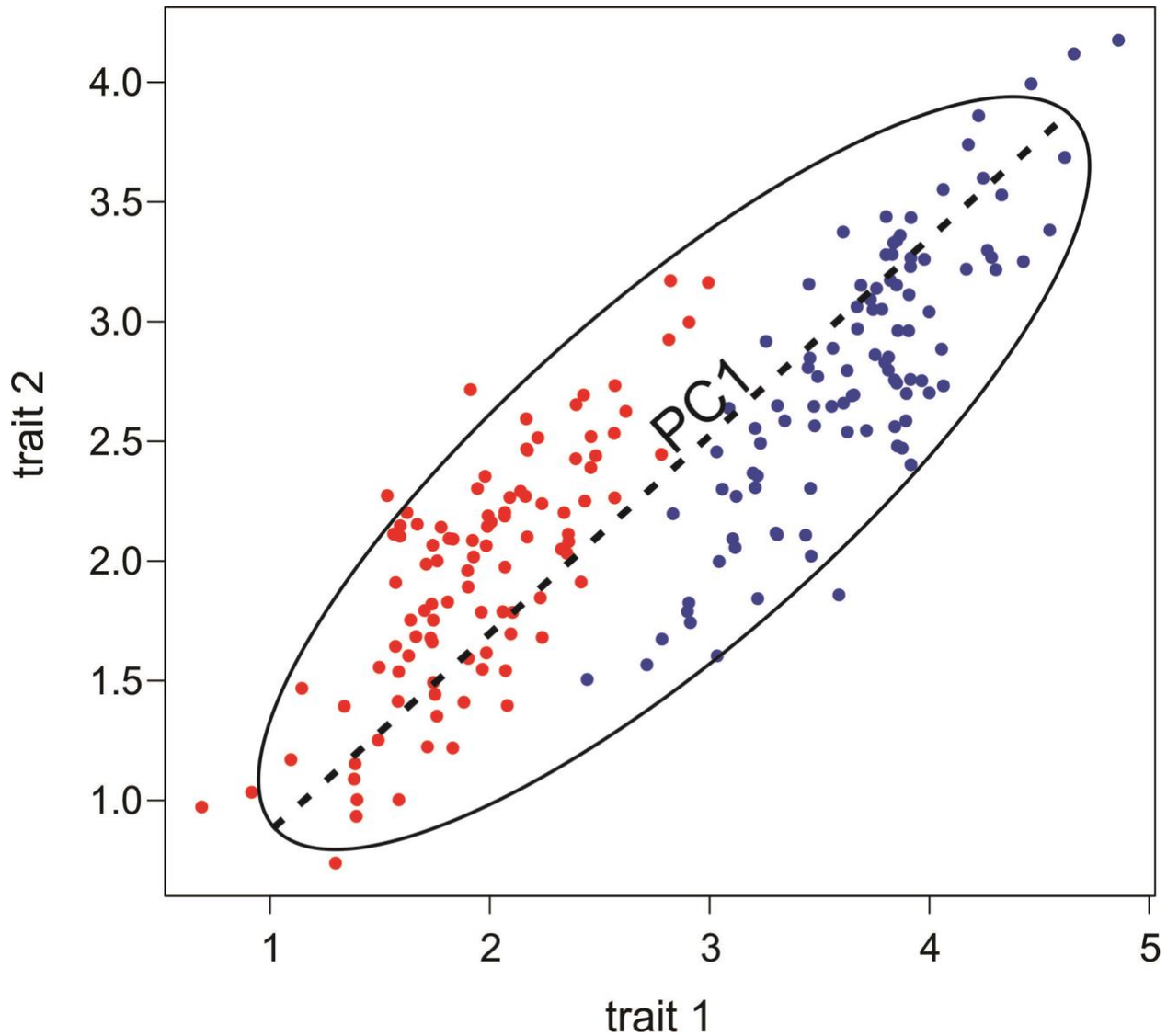
Discriminant function analysis is for classification rather than ordination.

It finds axes that maximally separate two or more previously identified groups. It finds axes that maximize variation among groups relative to variation between groups. These axes can then be used to classify new observations into the same groups.



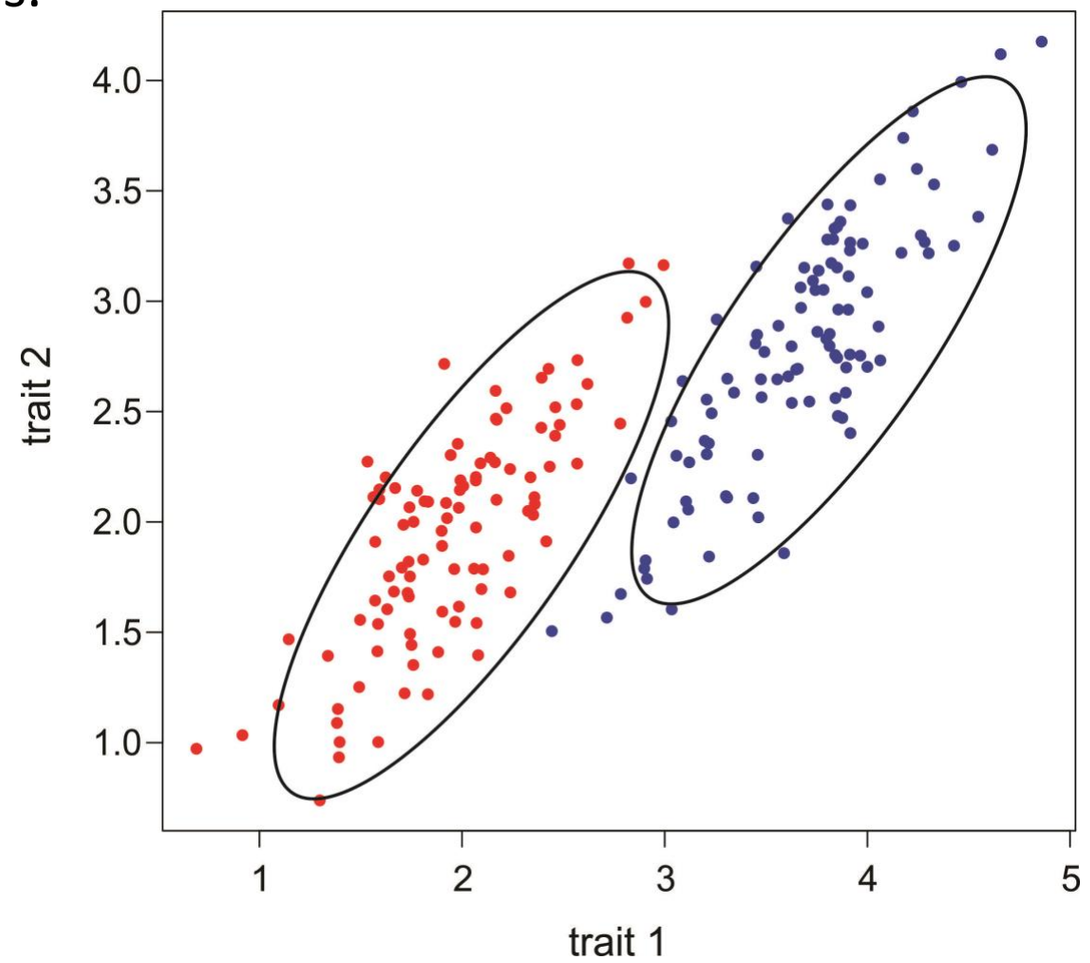
Discriminant function analysis

Principal component analysis is blind to groups, and only finds the directions of maximum total variance,



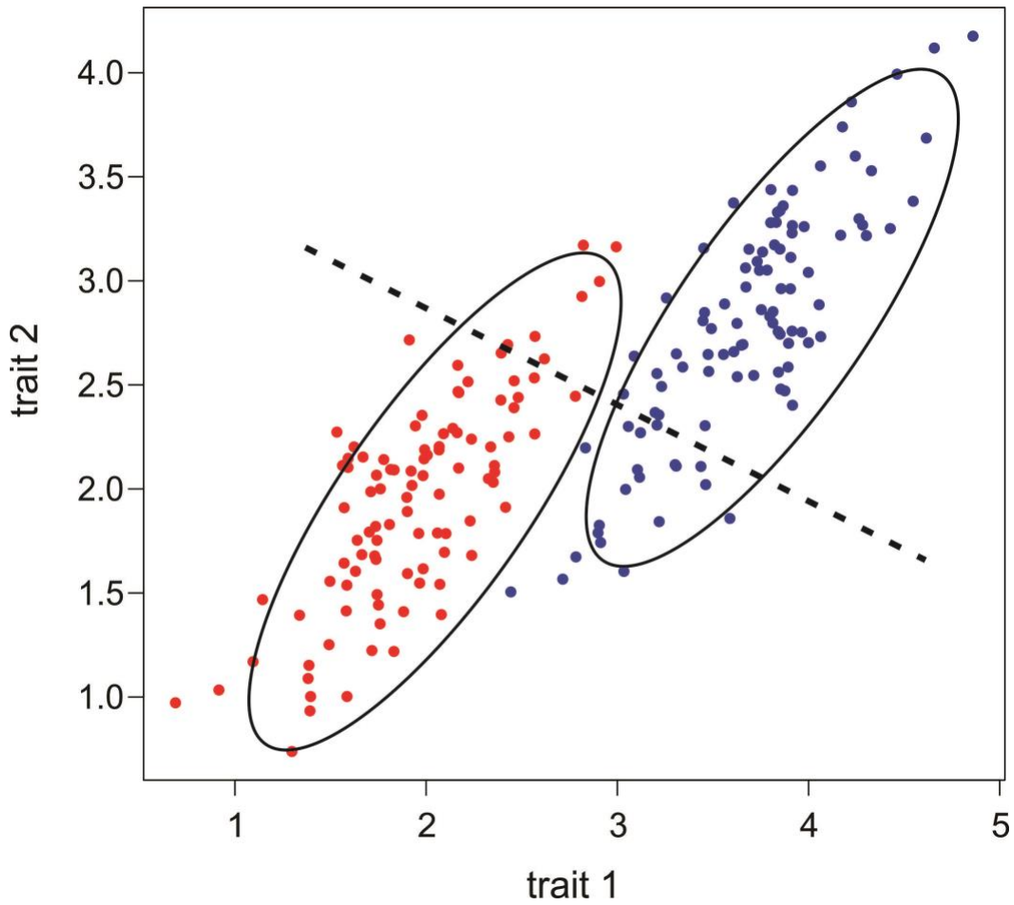
Discriminant function analysis

Discriminant function analysis explicitly finds the axes that best separate the groups. It doesn't do this simply by finding the direction of the biggest difference between means.



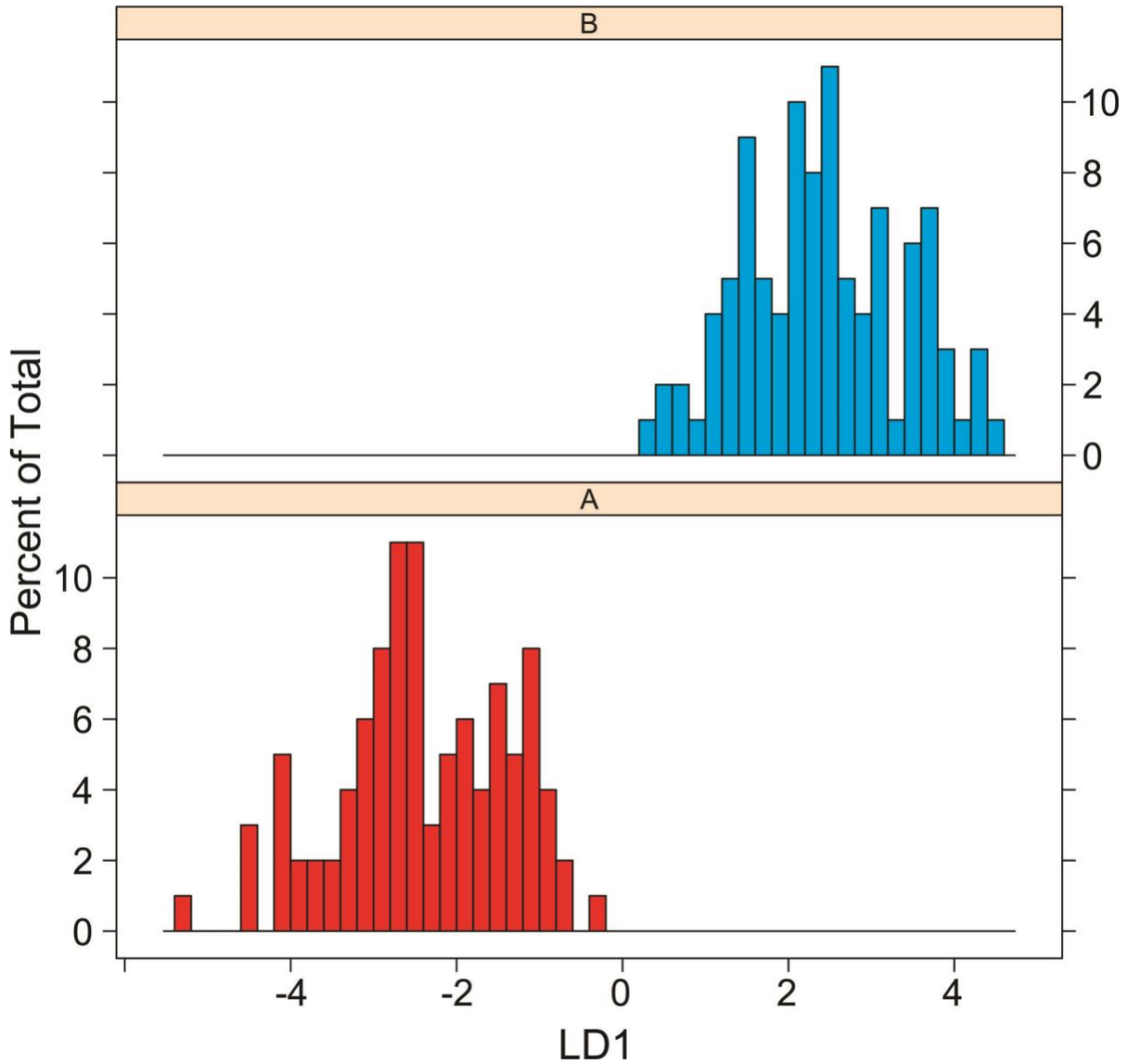
Discriminant function analysis

Instead, it finds the axes that best separate groups relative to within-group variation.



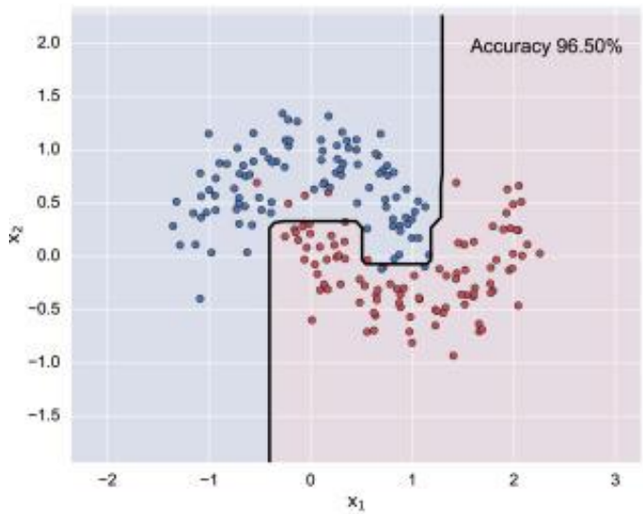
Discriminant function analysis

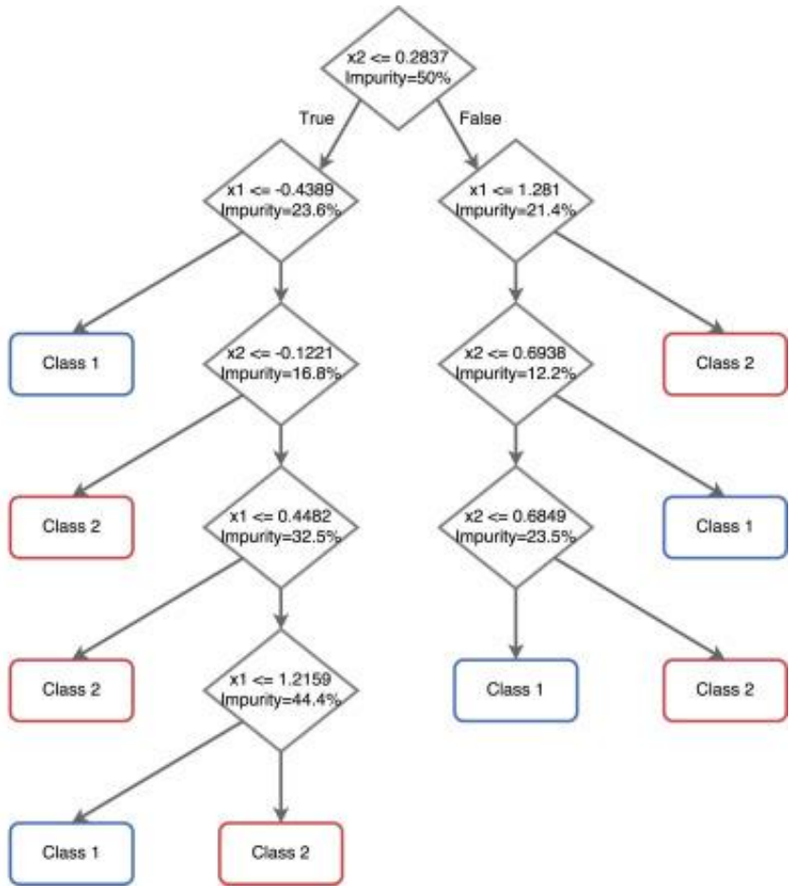
The resulting axes best separate the groups in the data, and can be used for classification of new observations.



Decision trees

Nonlinear classification method. The data are recursively partitioned into non-overlapping rectangular regions each associated with a decision rule. A fitted decision tree can be visualized as a series of IF-THEN decision rules to classify observations.





Fratello & Tagliaferri (2019) *Encycl Bioinf Comp BIol*

Random Forest

An ensemble machine learning method based on bootstrap resamples of the data. A decision tree is fitted to each bootstrap sample (the algorithm “learns” the boundaries of the decisions from the data). The algorithm then “learns” from the many decision trees which features matter most, and the results are then combined to make a final prediction. The method provides insights into which features are most important for making predictions.

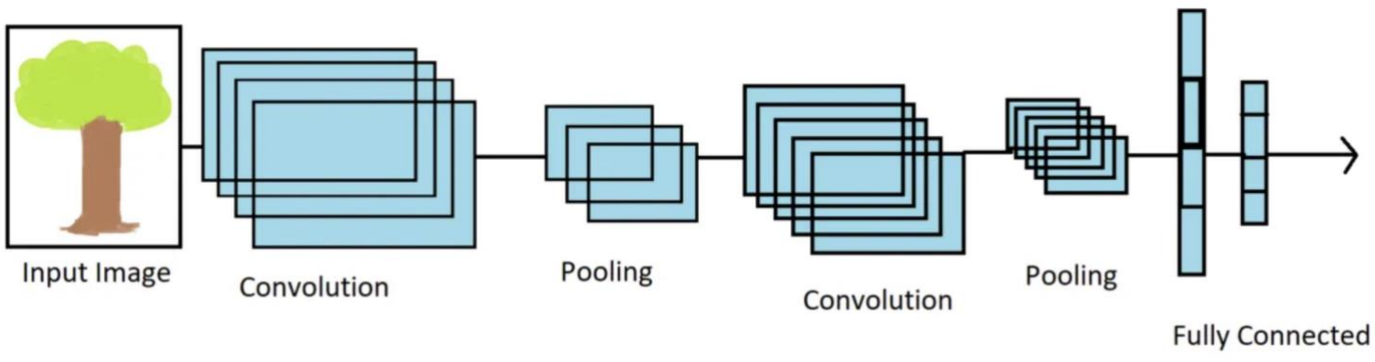
Involves **training** (decision trees are trained independently on bootstrap samples) and **prediction** (each tree makes a prediction, and then the Random Forest combines these predictions, e.g., by majority voting for classification, to produce the final prediction.

Convolutional Neural Net (CNN)

A deep learning approach to classification based on structured data having spatial or temporal relationships, such as images and text.

Don’t have time to go into details, but youtube has some excellent video introductions. I added a couple of links to R tips AI page.

A good way to motivate your learning is to try it yourself!



medium.com

Workshop this week

Multivariate methods**:**

Principal components analysis
Correspondence analysis
Image classification using a CNN (optional)

Optional CNN might take all two hours, so plan your own workshop.

Train a CNN to classify low-resolution grayscale images of dogs and cats.
Practice on example of handwritten digit classification before lab time.
Bring a digital image of your pet if you want to predict what it is by the end.
I've added image processing tips to the R tips Data page.

We'll use Google Colabs (unless you succeed in installing and testing keras on your laptop BEFORE coming to the workshop – it's tricky and there's not enough time in the workshop to help with this).

Discussion paper next week: (the last!):

Phylogeny and conservation. Download from “Handouts” tab on course web site.

Presenters: Jennifer & Louis

Moderators: Eric & Ralitza