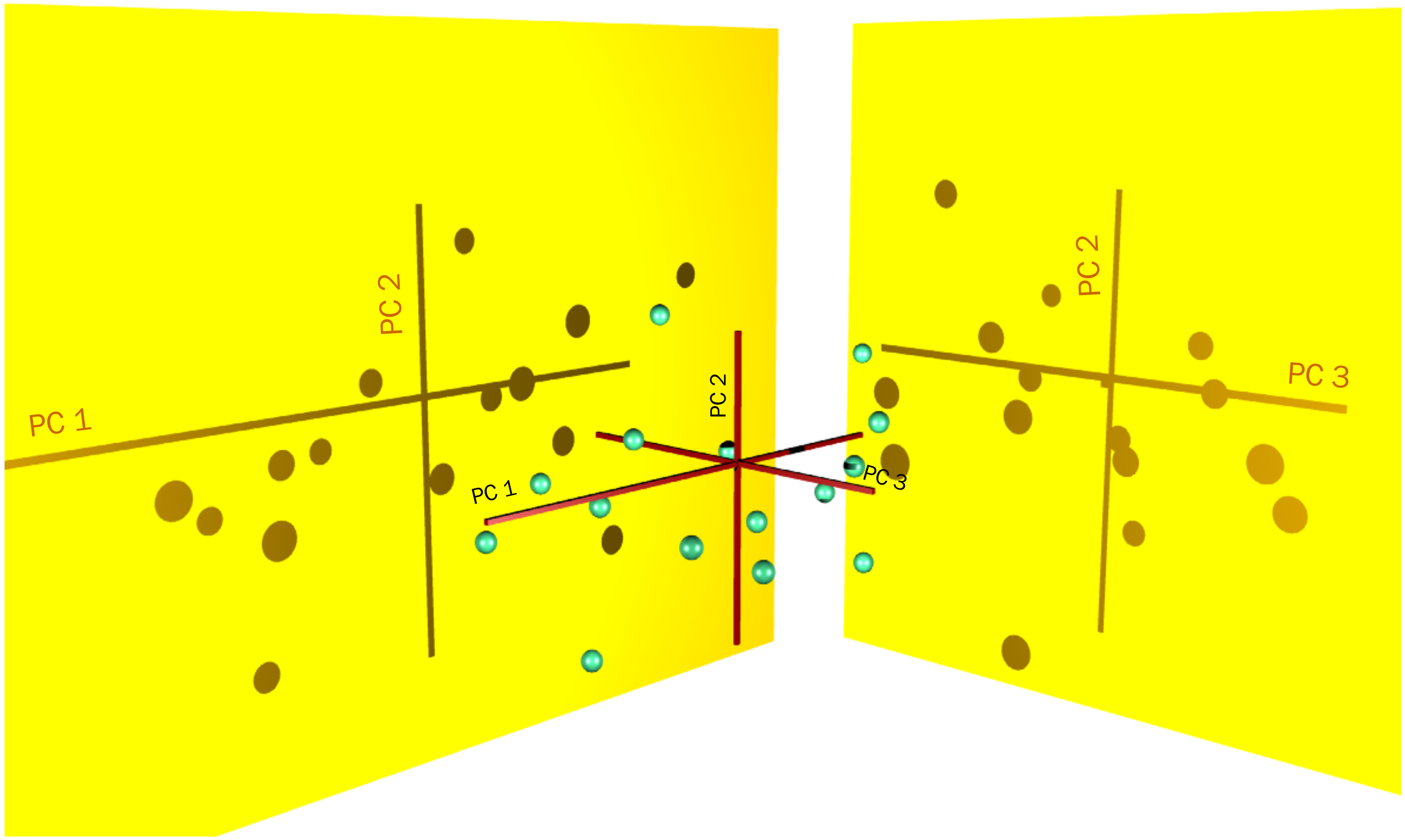
Procrustes, PCA, and 3D coordinates



Principal components space - multidimensional

Module Overview

Day 1 – Introduction to R (aka, the R boot camp)
Thursday, 20 June, 2013

Day 2 – Introduction to Geometric Morphometrics in R
Friday, 21 June, 2013

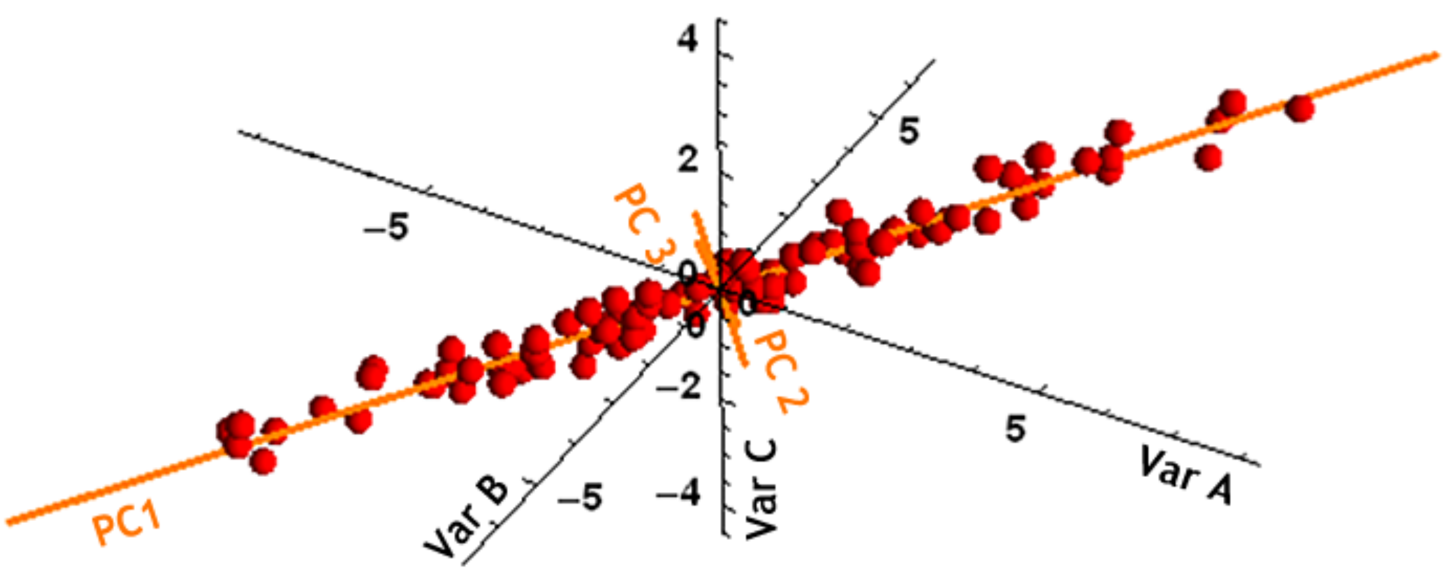
Day 3 – Procrustes Analysis, Shape Space, and Statistical Testing
Saturday, 22 June, 2013

Day 4 – Morphological Evolution and Shape Modeling
Sunday, 23 June, 2013

Day 5 – Phylogenetics of shape and review
Monday, 24 June 2013

Ordination and Principal Components Analysis

1. Introduction to Ordination
2. Why PCA is an important part of Geometric Morphometrics
3. Technical explanation of what PCA does
4. Eigenvalues, Eigenvectors and Scores
5. Morphological meaning of principal component axes
6. Modeling in shape space



Ordination

Ordering specimens along new variables

Principal Components Analysis (PCA)

Arranges data by major axes based on measured variables

Principal Coordinates Analysis (PCO)

Arranges data by major axes based on distance measures

Canonical Variates Analysis (CVA)
(or Discriminant Function Analysis, DFA)

Finds best separation between groups

Multidimensional Scaling (MDS)

Arranges data so the distances on 2D plot are as similar as possible to original multivariate distances

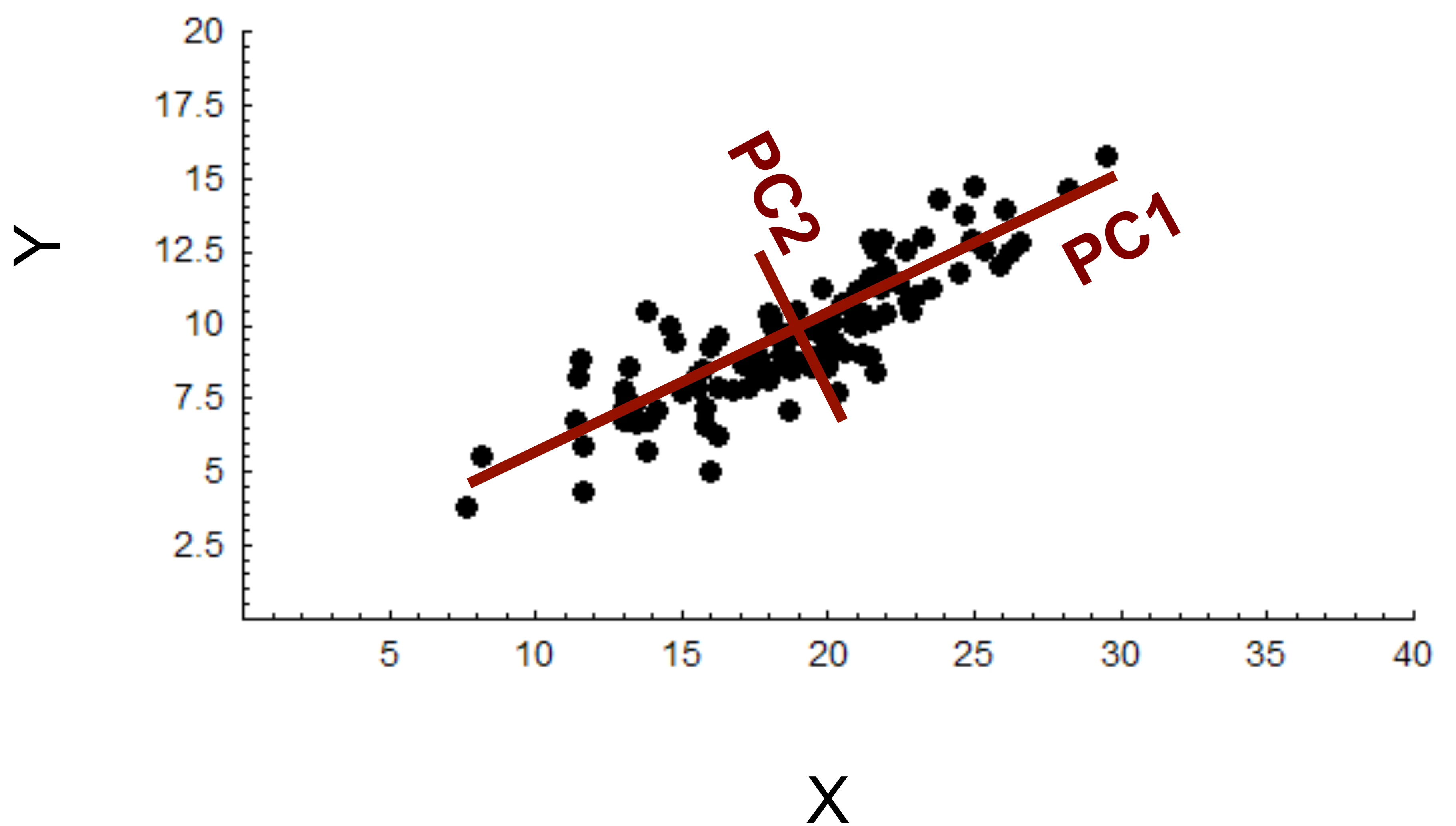
What does PCA do?

1. Rotates data to its major axes for better visualization

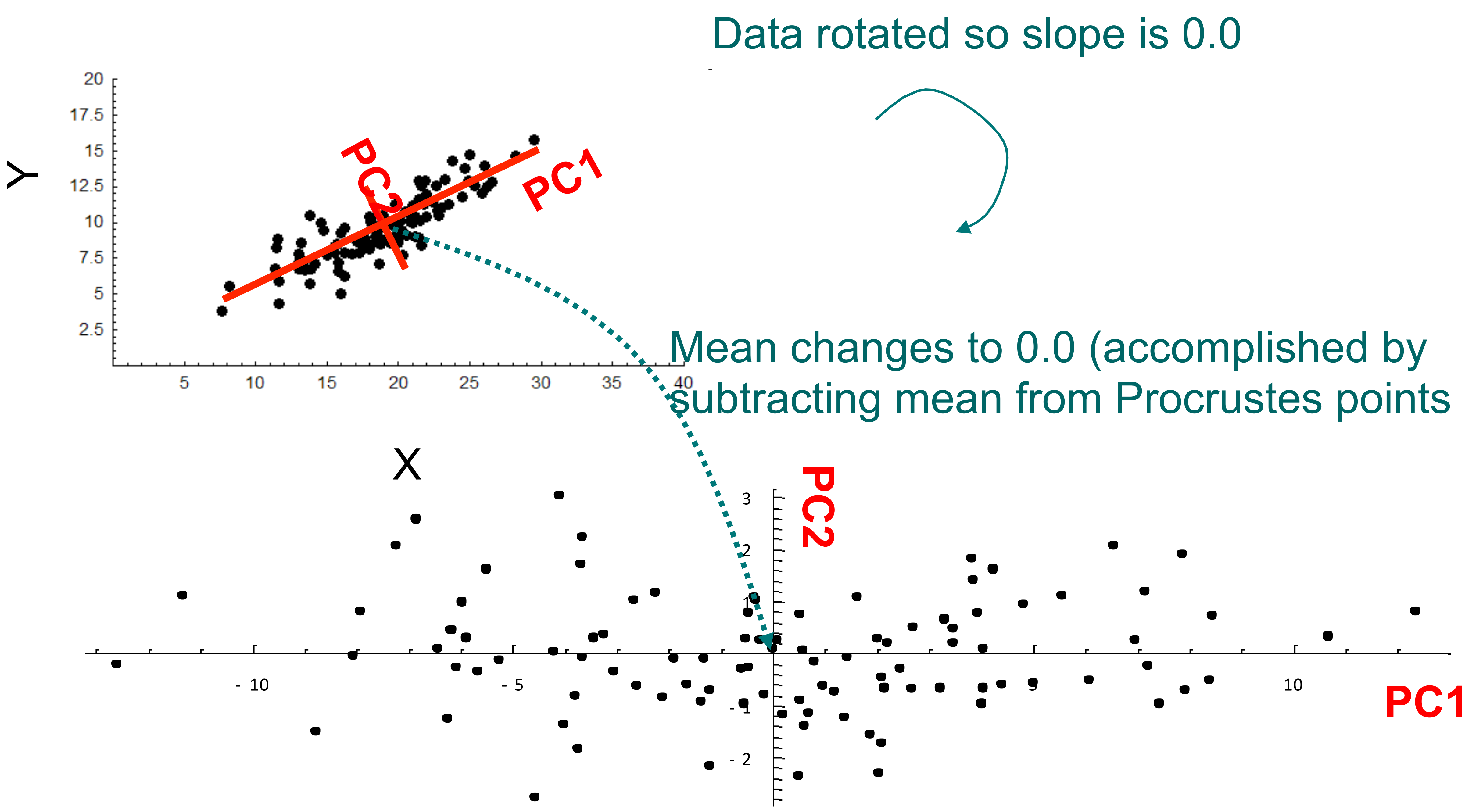
2. Preserves original distances between data points
   in other words, PCA does not distort the variation data
   (iff the covariance method is used, standard practice in geometric morphometrics)

3. Removes correlations between variables to make further statistical analysis simpler

The principal components (PCs) of a data set are its major axes



Principal components are a 'rigid rotation' of the original data



Note that variance increases along horizontal axis, but decreases along vertical axis.

Important points: the "meaning" of PCA

1. Principal components analysis finds the axes of greatest variation in a data set

2. PCA removes correlations from the data

3. Principal components scores are "shape variables" that are the basis for further analysis

4. <u>But PCA is nothing more than a rotation of the data!</u>

Behind the scenes in PCA of landmarks

Procrustes

This aligns shapes and minimizes differences between them to ensure that only real shape differences are measured.

1. **Subtract mean (consensus) from each shape to produce "residuals"**
   This centers the PC axes on the mean (consensus) shape.
2. **Calculate covariance matrix of residuals**
   Estimates variance and covariance among the original variables
3. **Calculate eigenvalues and eigenvectors of covariance matrix**
   Finds the major axes of the data and the variation along them.
4. **Multiply residuals times eigenvectors to produce scores**
   Rotates the original data onto the major axes and gives the coordinates for their new position.

Output of PCA

Eigenvalues

variance on each PC axis
(In R: *svd(cov(residuals))$d, or plotTangentSpace(coords)$summary$stdev^2*)

Eigenvectors

loading of each original variable on each PC axis
(In R: *svd(cov(residuals))$u*)

Scores (=shape variables)

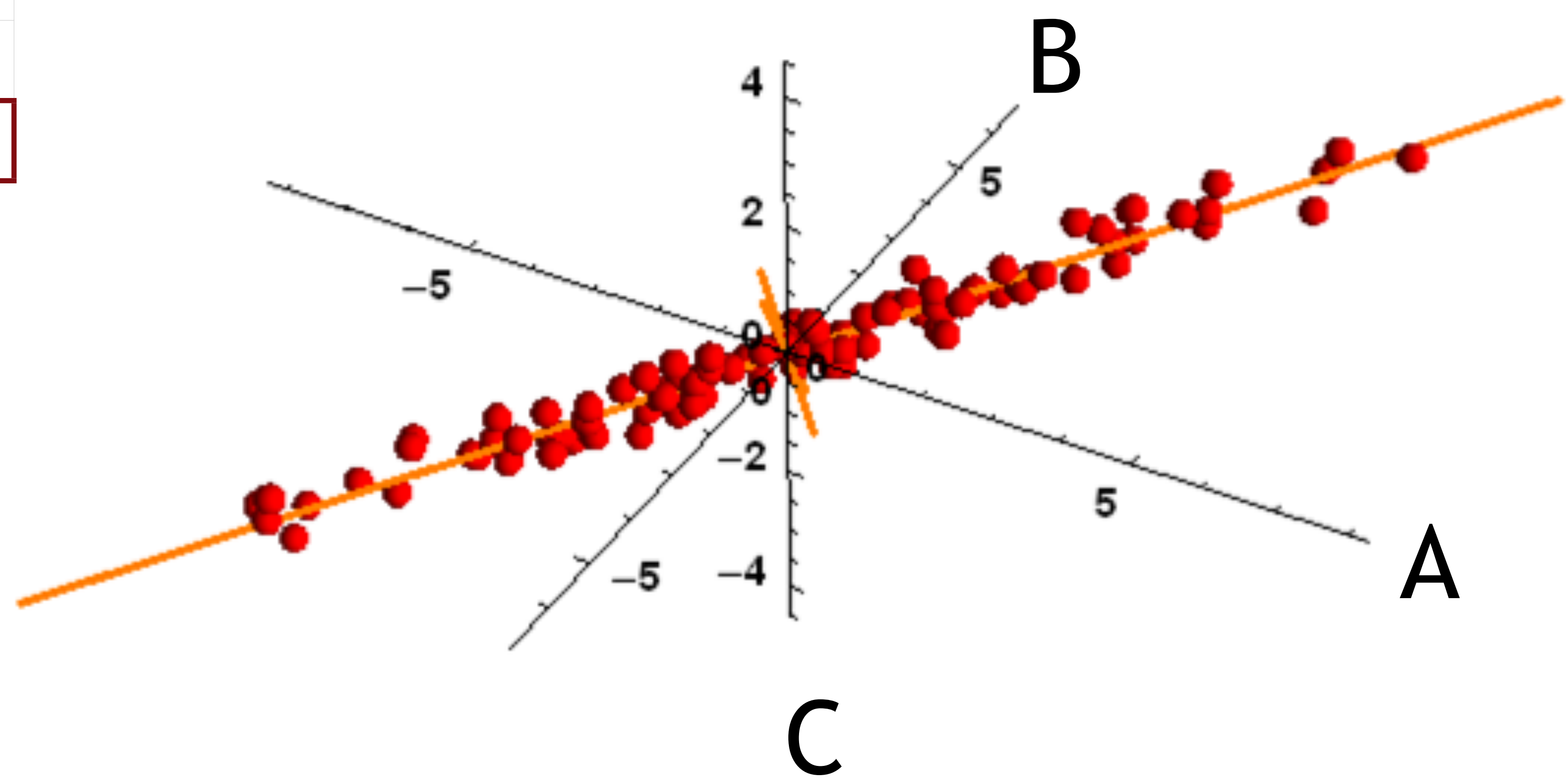
location of each data point on each PC axis
(In R: *resids%*%svd(cov(residuals))$u*)

resids are the residuals of the Procrustes coordinates (coords - consensus)
cov is the covariance matrix of the residuals

PCA is based on the covariance matrix

Diagonal elements are variances, off-diagonal are covariances (slopes)

	A	B	C
A	6.56	4.69	2.59
B	4.69	4.21	1.38
C	2.59	1.38	1.36

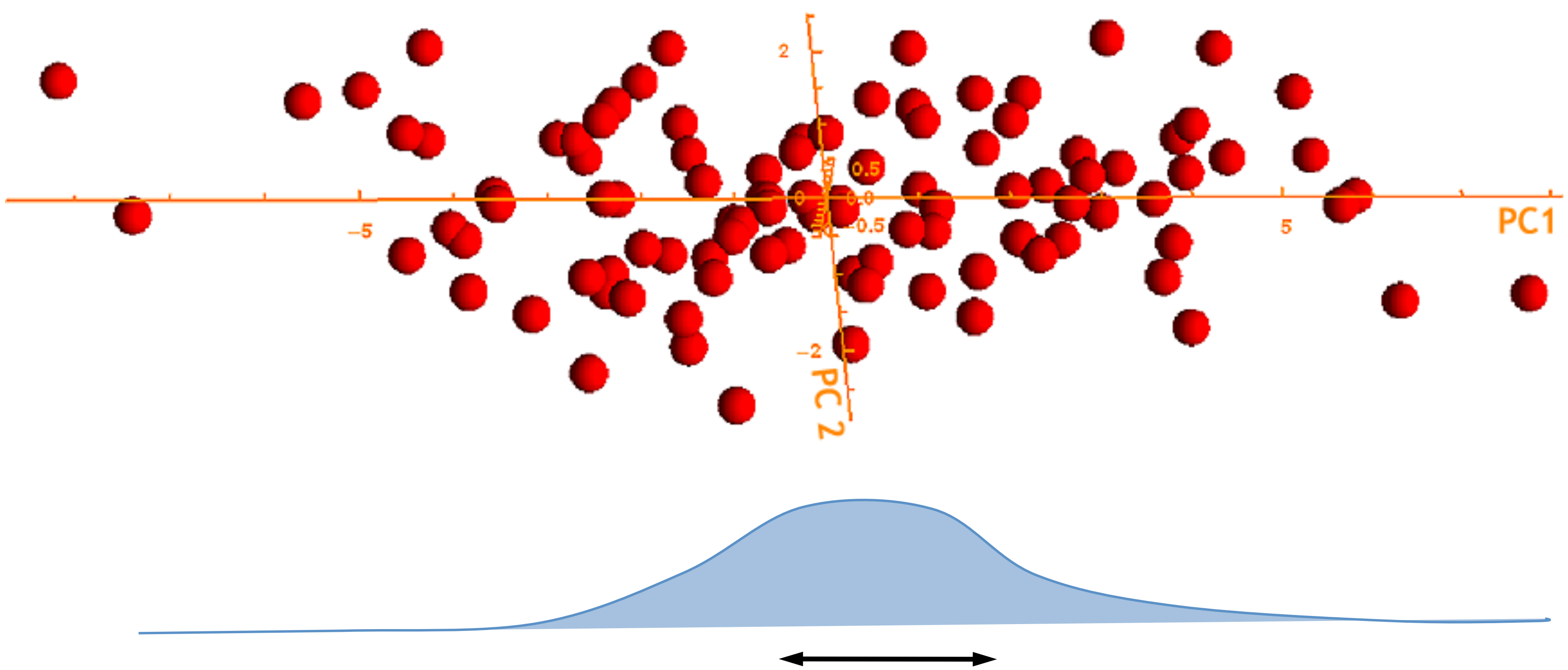


Eigenvalues

Variance of data along each PC axis

PC 1 = 11.08
PC 2 = 1.01
PC 3 = 0.04

	PC1	PC2	PC3
PC1	11.08	0	0
PC2	0	1.01	0
PC3	0	0	0.04



Important point: the meaning of eigenvalues

Between 95% and 99% of data lie within 2.0 SDs of the mean

1. If you know the variance, you know the standard deviation is its square-root;

2. You know that nearly all the data have a range of 4 * SD;

3. If the mean is 0.0, then nearly all the data lie
   between -2 * SD and +2 * SD;

4. The eigenvalues (or singular values) of a PC are variances, therefore the
   range of data on that PC can be calculated from them.

Important point: the meaning of eigenvalues (cont.)

Total variance of morphometric data set is the total amount of shape variation, which can be calculated three ways:

1. Summing squared distances between landmark points and the consensus (sample mean) for all the objects and dividing by (*n*);
2. Summing the eigenvalues that are returned by the PCA;
3. Summing squared PC scores (have a mean of zero so no subtraction is required) and dividing by (*n*);

If these three calculations don't give the same number, <u>something is wrong</u>

Useful variants on Eigenvalues

Eigenvalues

PC 1 = 11.08
PC 2 = 1.01
PC 3 = 0.04

12.13

Percent explained

100 * 11.08 / 12.13 = 91.3
100 * 1.01 / 12.13 = 8.3
100 * 0.04 / 12.13 = 0.3

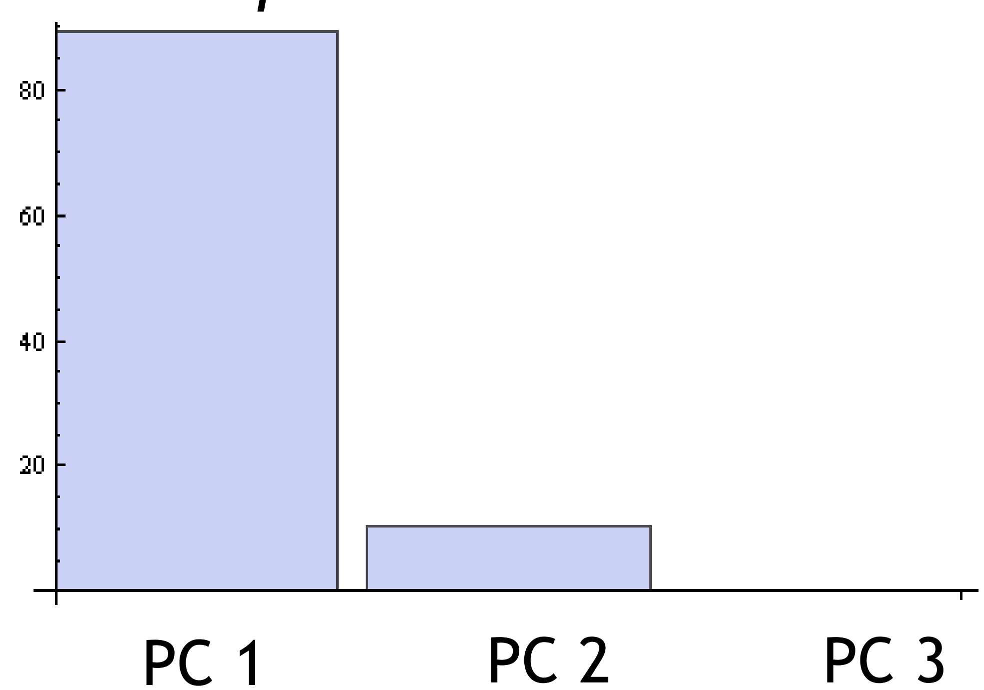
100.0

Standard Deviation

(plotTangentSpace reports this)

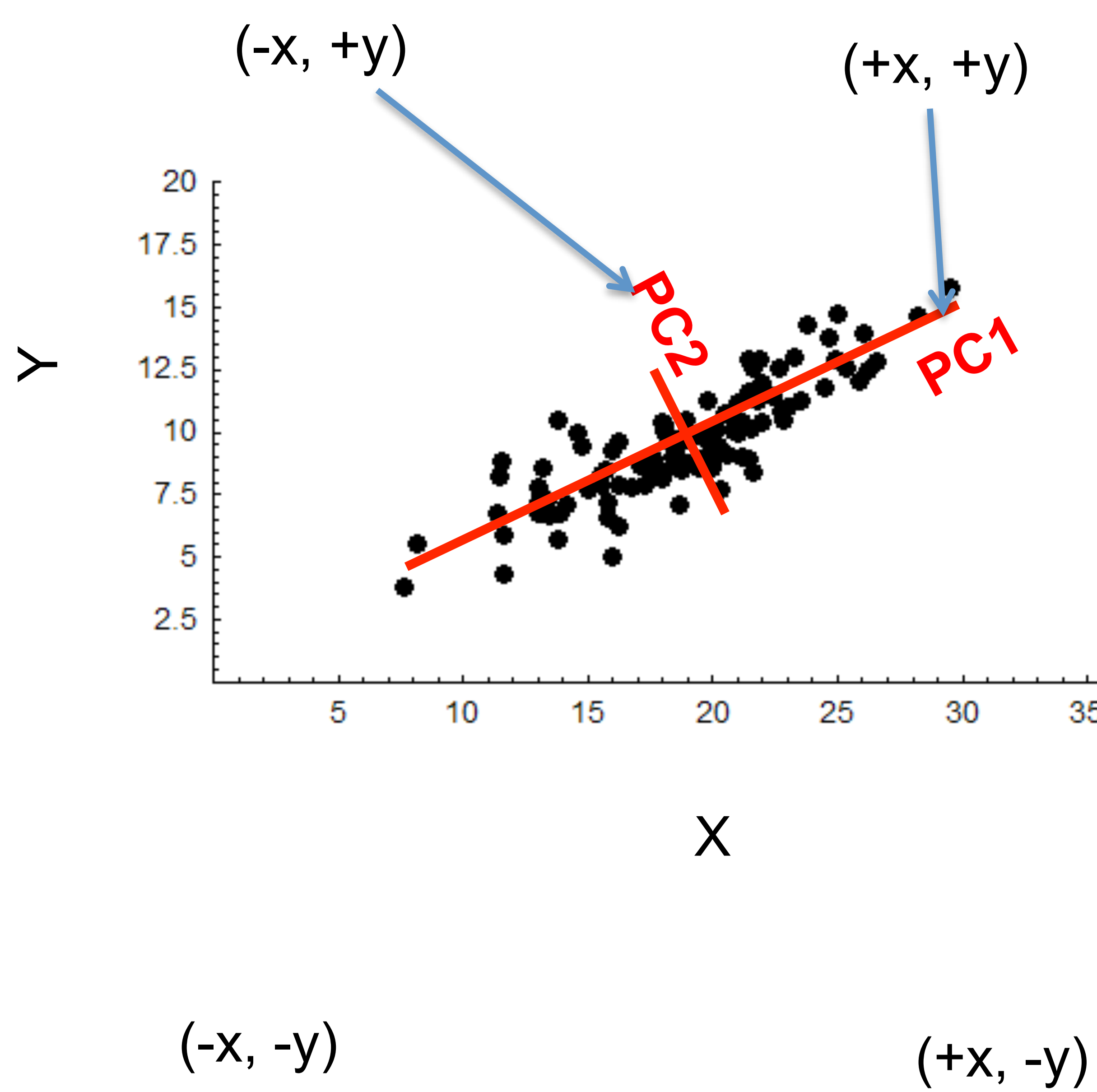
11.08^0.5 = 3.33
1.01^0.5 = 1.00
0.04^0.5 = 0.20

Scree plot



barplot(svd$d)

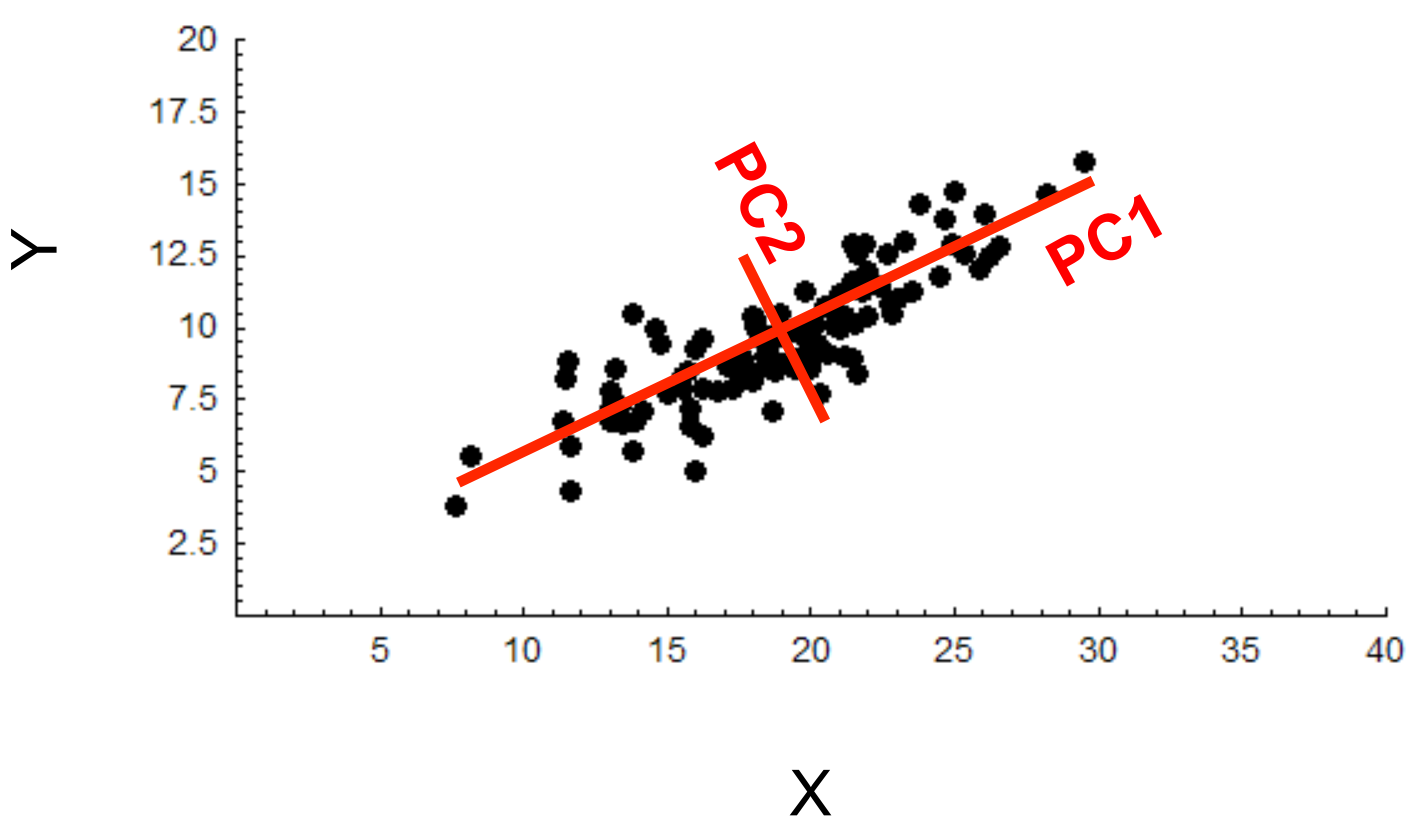
Eigenvector 'loadings' tell how each original variable contributes to the PC



Eigenvector Matrix

	PC1	PC2
X	0.89	-0.44
Y	0.44	0.89

Eigenvectors also describe how to transform data from original coordinate system to PCs and back



Eigenvector Matrix

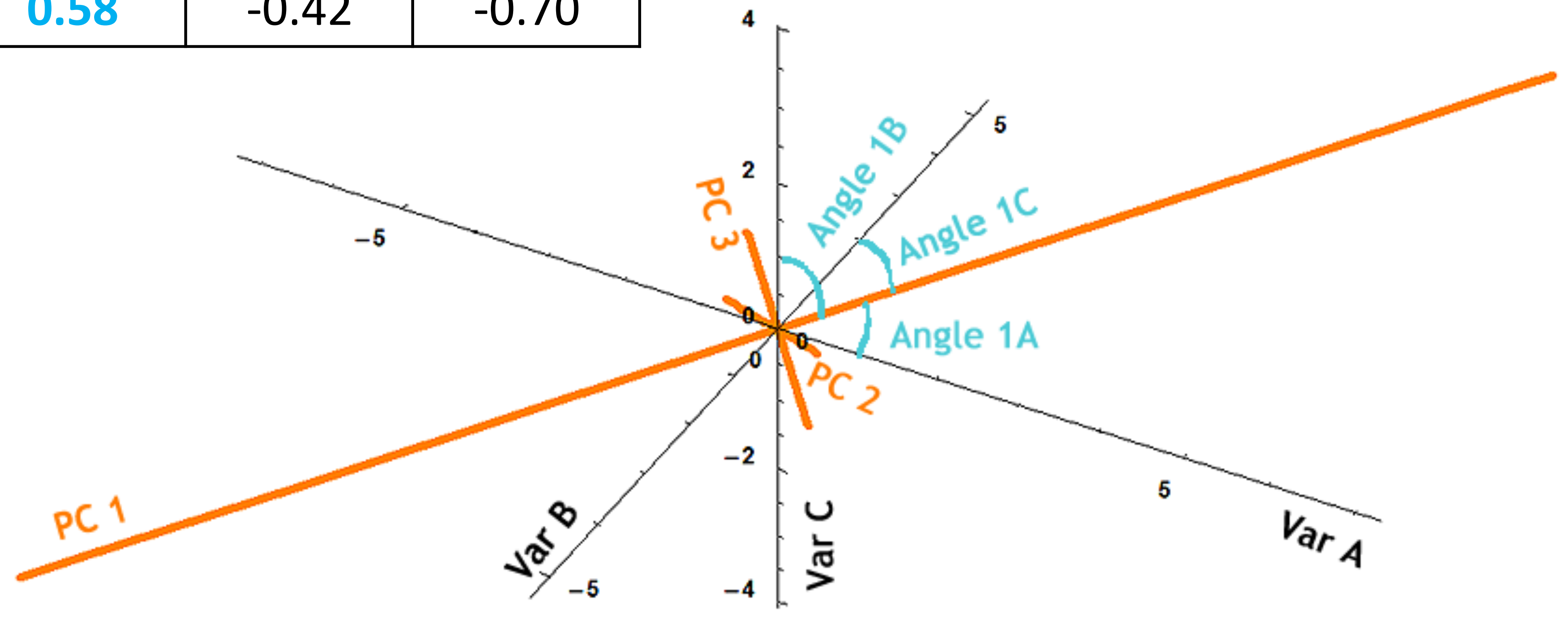
	PC1	PC2
X	0.89	-0.44
Y	0.44	0.89

(multiply PC1 X score by 0.89 and PC1 Y score by -0.44 and add back X,Y meanto get real X,Y)

Eigenvectors: definition 1

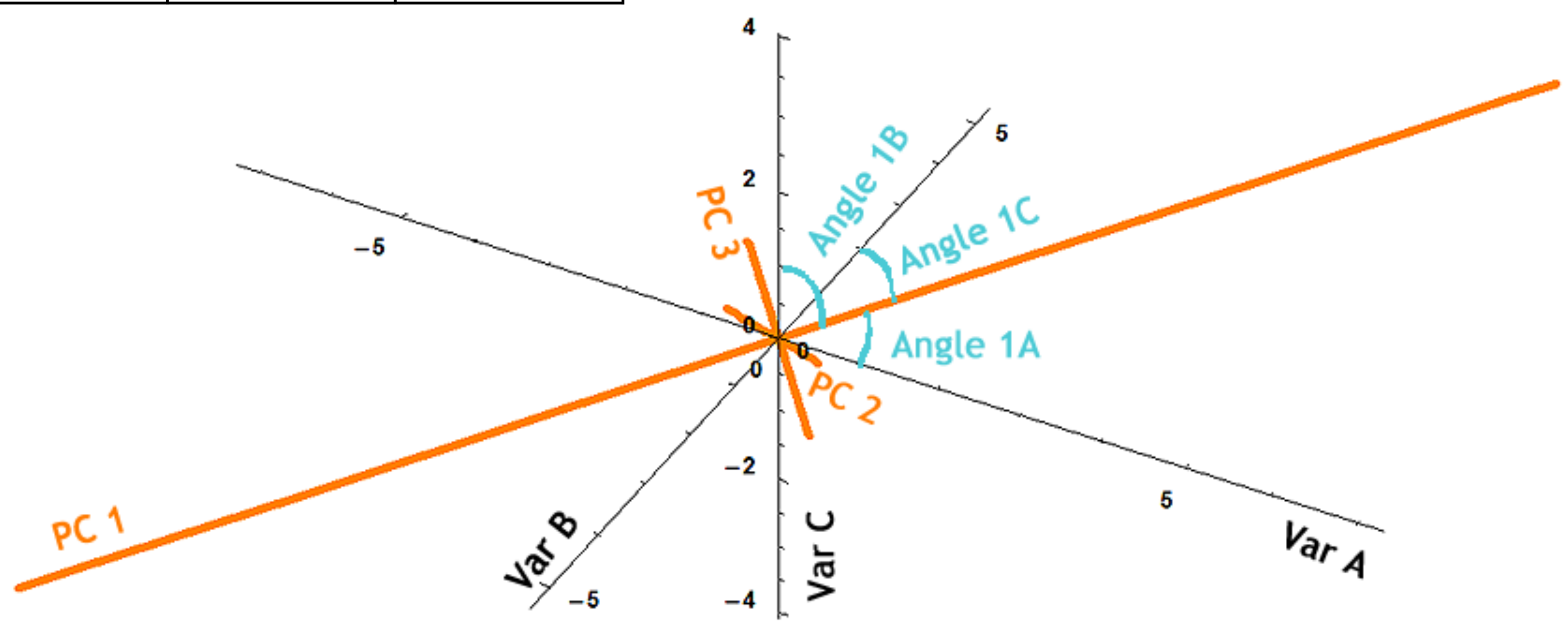
Angles between PC and original variables
(the eigen vector matrix is a rotation matrix in radians)

	PC1	PC2	PC3
Var A	-0.76	-0.58	-0.29
Var B	0.28	-0.69	0.66
Var C	0.58	-0.42	-0.70



Same Eigenvectors converted from radians to degrees

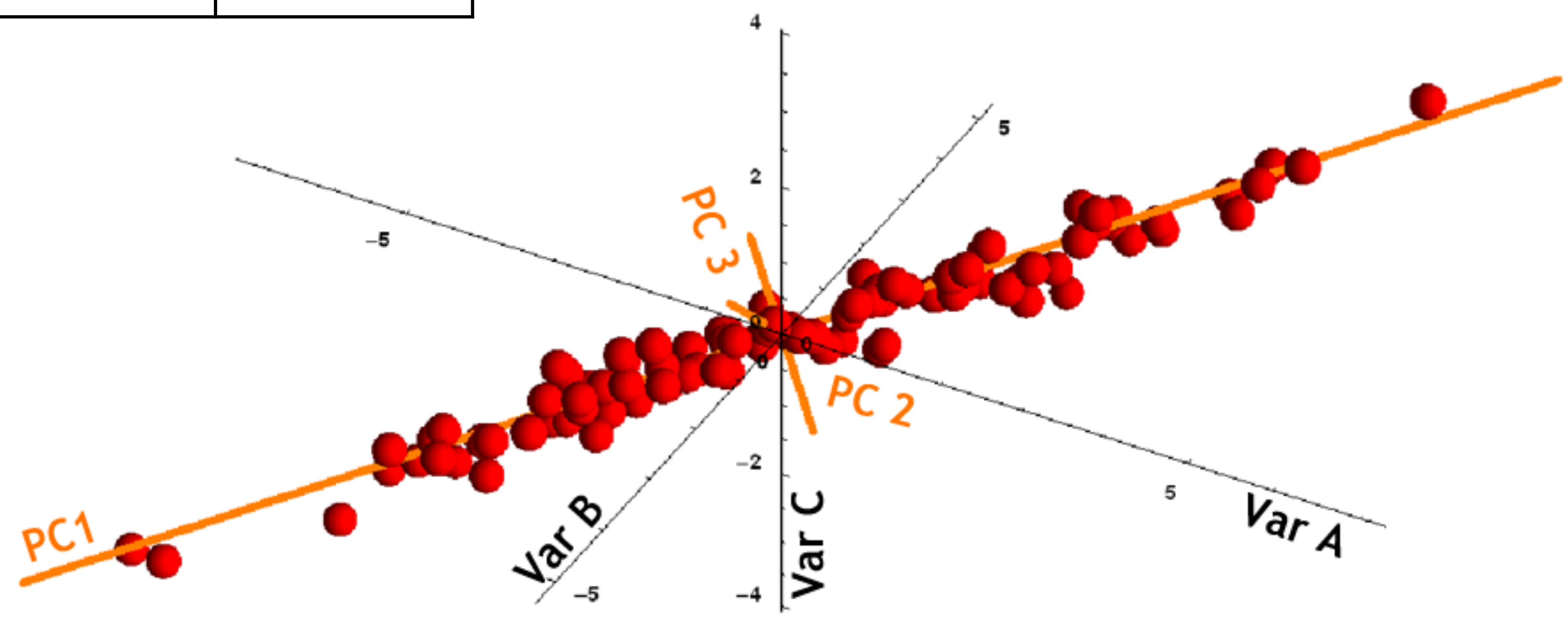
	PC1	PC2	PC3
Var A	**-43.8**	-33.1	-16.4
Var B	**16.2**	-39.9	37.7
Var C	**33.2**	-24.2	-39.9



Eigenvectors: definition 2

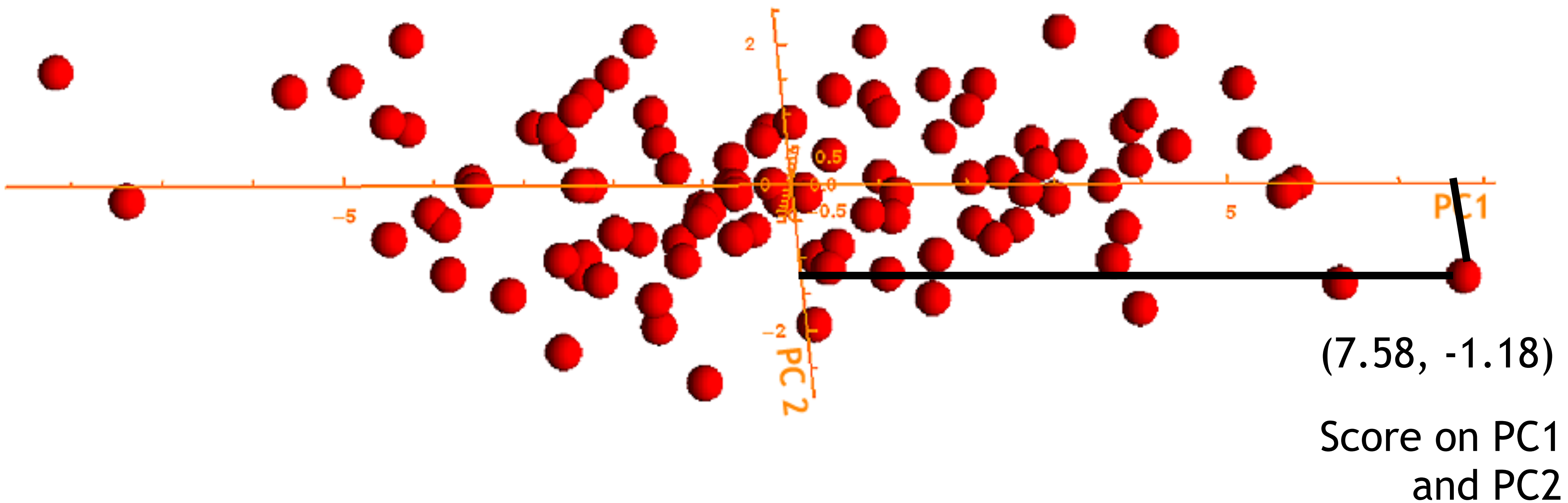
Loading (or importance) of each variable to the PC.
The larger the absolute value, the more important the variable.

	PC1	PC2	PC3
Var A	-0.76	-0.58	-0.29
Var B	0.28	-0.69	0.66
Var C	0.58	-0.42	-0.70

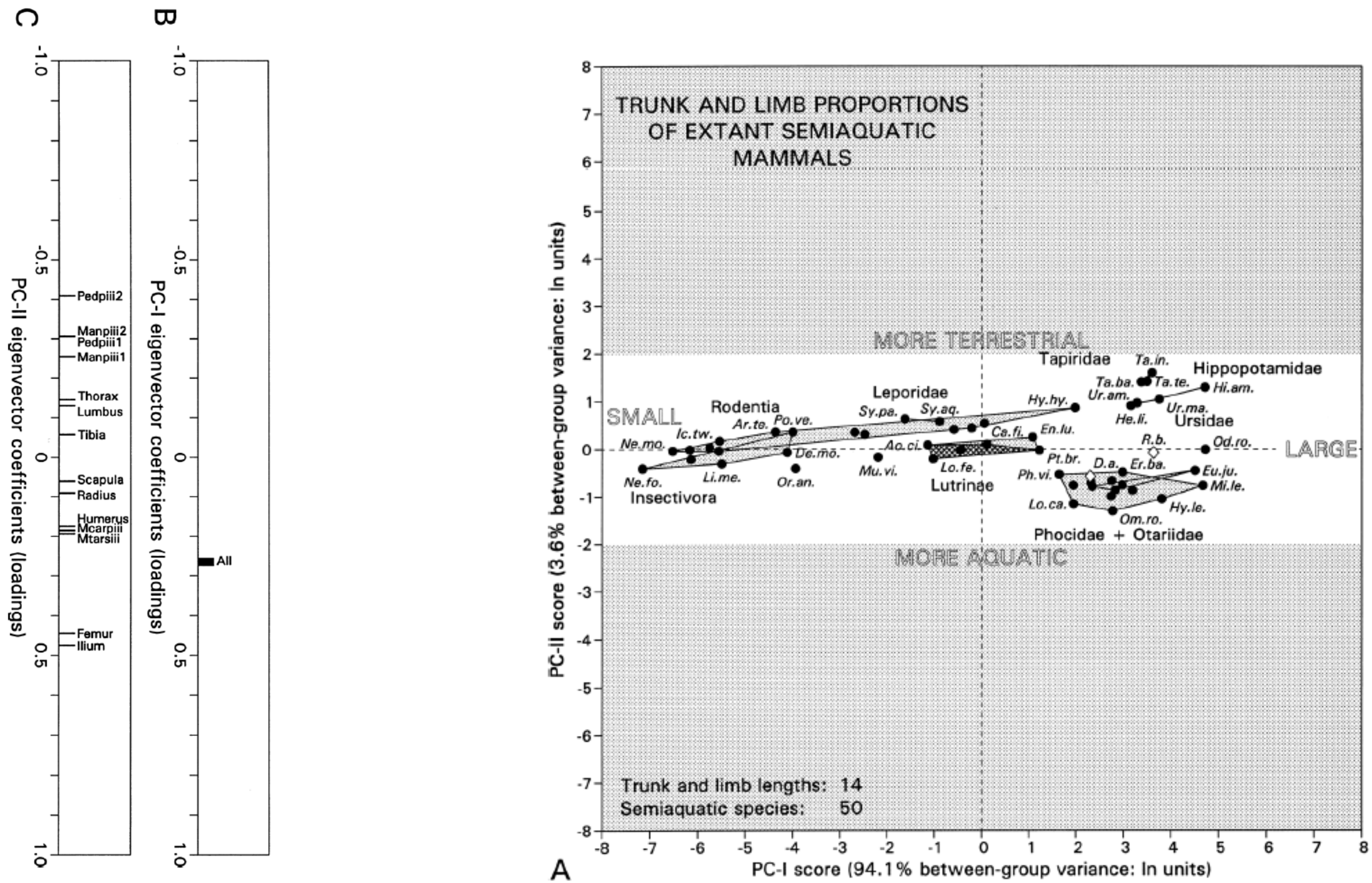


Scores

The coordinates of each data point on the PC axes.
These numbers can be thought of as new variables, or shape variables.



PCA plots have lots of meaning



Gingerich, P.D. 2003. Land-to-sea transition in early whales: evolution of Eocene archaeoceti (Cetacea) in relation to skeletal proportions and locomotion of living semiaquatic mammals. *Paleobiology*, 29: 429-454.

PCA is important in Geometric Morphometrics because....

1. PCA scores are used as shape variables
2. Eigenvectors are convenient axes for shape space
3. Eigenvectors and their scores are uncorrelated as variables
4. Variance (eigenvalues) is partitioned across eigenvectors and scores in descending order
5. Scores can be safely used for all other statistical analyses, including tree building
6. Eigenvectors can be used to build shape models

PCA vs Relative Warps vs Partial Warps

Relative warps = Principal components

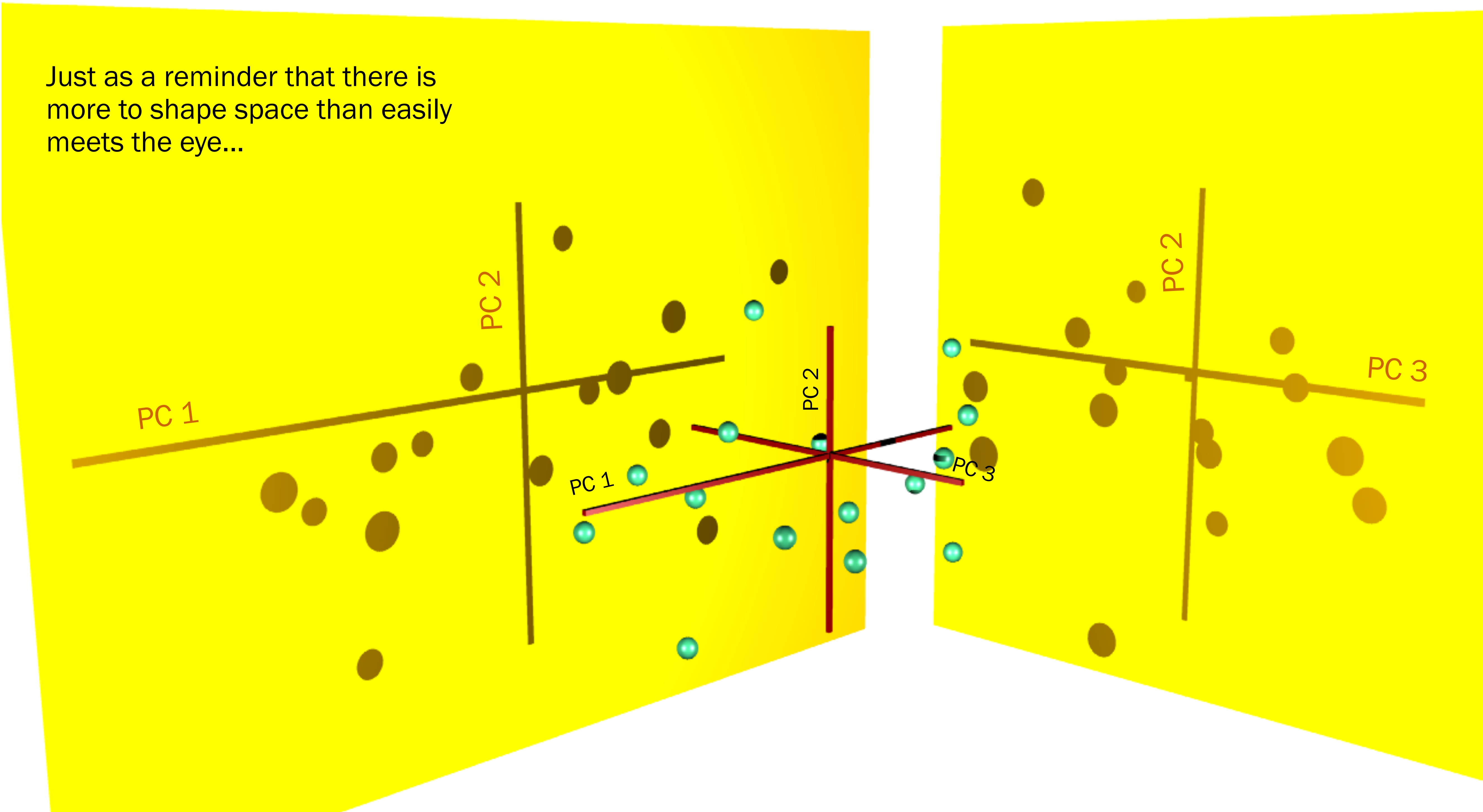
Relative warps/Principal components organize shape variation so that the greatest amount is explained on PC1, second greatest on PC2, etc. Also PC1 is uncorrelated with PC2 is uncorrelated with PC3, etc.

Partial warps (can safely be ignored)

Partial Warps measure the “scale” of shape variation over the entire object down to a small part of the object. NOT principal components (even though the plots look alike). Partial warp 1 explains variation in ALL the landmarks, Partial warp 2 explains variation in part of the landmarks, Partial warp 3 in a smaller number, etc. Partial Warp 1 MAY be correlated with Partial warp 2, etc.

Principal components space is multidimensional

Just as a reminder that there is more to shape space than easily meets the eye...



PC graph plots are two dimensional *projections*, or shadows, of multidimensional space

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$$

Singular Value Decomposition (svd)

SVD is a method for calculating eigenvectors and eigenvalues from a covariance that is “singular” (has fewer degrees of freedom than variables).

To carry out SVD on shape coordinates:

1. Procrustes superimpose the shapes (Procrustes coordinates)

2. Subtract the consensus to center the data on the mean (Procrustes residuals)

3. Calculate covariance matrix

4. Perform singular value decomposition on covariance matrix to give:

    4.1. U matrix = eigenvectors

    4.2. D matrix = eigenvalues

    4.3. V matrix = conjugate transpose of eigenvectors

PCA in *R* using svd

library(svd)

1. Obtain Procrustes coordinates
```
proc <- gpagen(lands)
```

2. Convert coordinates to two-dimensional matrix
```
coords2d <- two.d.array(proc$coords)
```

3. Calculate consensus and flatten to single vectors
```
consensus <- apply(proc$coords, c(1,2), mean)
consensusvec <- apply(coords2d, 2, mean)
```

4. Calculate Procrustes residuals (Procrustes coordinates - consensus)
```
resids <- t(t(coords2d)-consensusvec)
```

5. Calculate covariance matrix
```
P <- cov(resids)
```

6. Calculate eigenvector and eigenvalues with SVD
```
pca.stuff <- svd(P)
```

7.
```
eigenvalues <- pca.stuff$d
eigenvectors <- pca.stuff$u
```

8. Calculate PCA scores
```
scores <- resids%*%eigenvectors
```

Check that you did the PCA correctly

1. Check that everything worked by comparing variances

```
sum(apply(coords2d,2,var))  #  total variance of Procrustes coordinates
sum(apply(resids,2,var)) # total variance of Procrustes residuals
sum(pca.stuff$d)  #  total variance of singular values
sum(apply(scores, 2, var)) # total variance of scores
```

all of the above should be equal

2. Also check that the scores calculated here equal scores from plotTangentSpace() in geomorph

3. Create PCA plot by plotting columns of scores (first column = PC 1, etc.)

```
plot(scores[,1:2],asp=1, pch=20,cex=2)
```