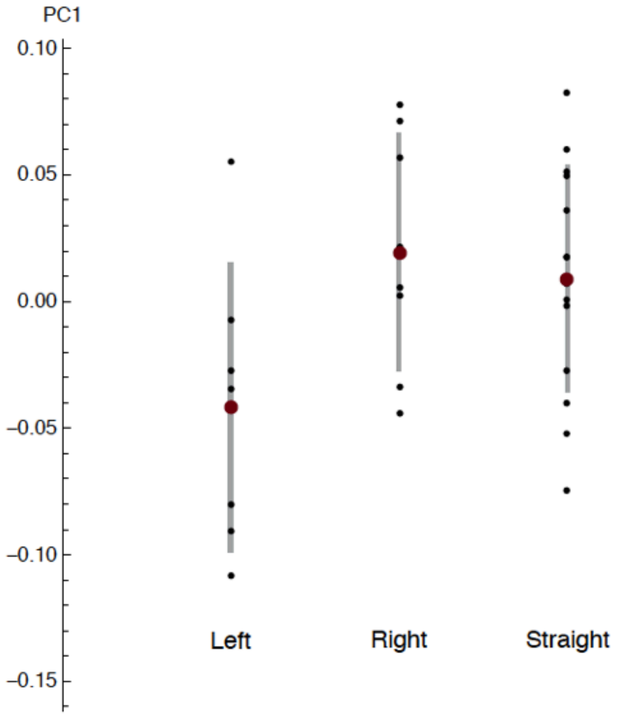
Basic statistical tests for GMM

Day 2



	SS	df	MS
Between	0.0060	2	0.0030
Within	0.1820	27	0.0911
Total	0.1880	29	0.0065

P. David Polly
Department of Earth and Atmospheric Sciences
Adjunct in Biology and Anthropology
Indiana University
Bloomington, Indiana 47405 USA
pdpolly@indiana.edu

Outline

Ordination methods

- Principal Components Analysis (PCA)
- Canonical Variates Analysis (CVA) / Discriminant Function Analysis (DFA)
- Multidimensional scaling (MDS)
- Between-groups PCA (BG-PCA)

Three major classes of statistical test

- Multivariate regression (shape and a continuous variable)
- MANOVA (shape and a grouping variable)
- Two block partial-least squares (shape and many variables, shape and shape)

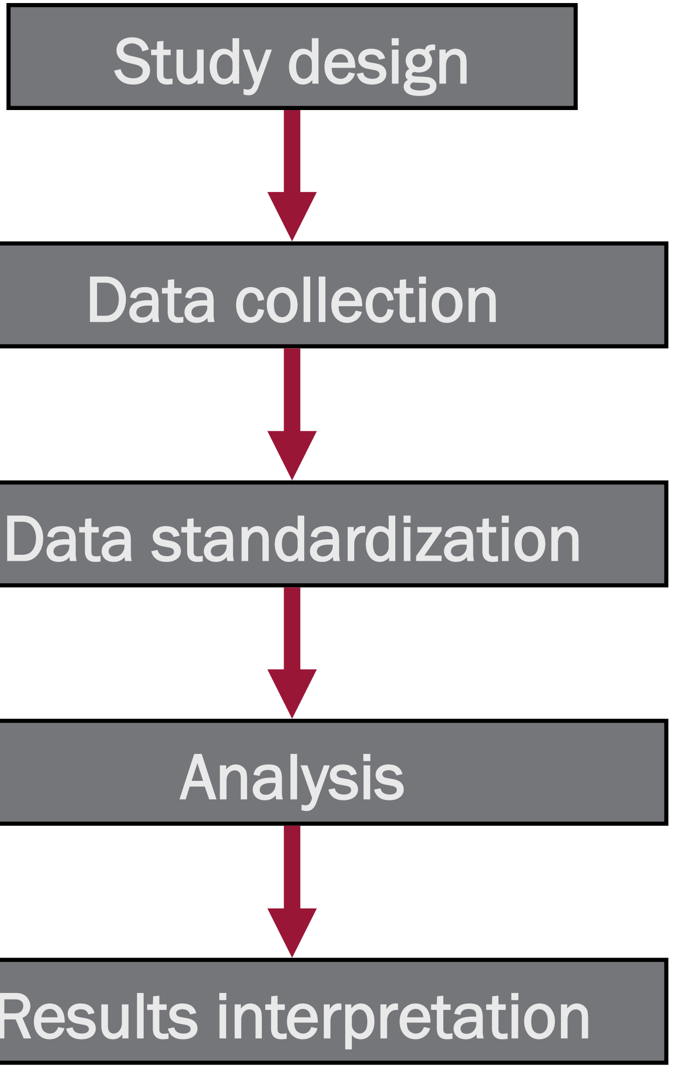
Randomization versions of tests are preferred for GMM data

Things to keep in mind

- PC axes are sample dependent and do not align with real processes: beware analyzing only a subset of axes
- Objects in two-dimensional morphospace may be farther than they appear

Phylogeny may need to be taken into account

Steps in a geometric morphometric study



Study design

Data collection

Data standardization

Analysis

Results interpretation

Careful thought

Landmarking

Procrustes, PCA

Other statistical tests

Careful thought

Basic components of GMM

Whatever is analyzed together must be superimposed together

Procrustes

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

PCA (primary use)

This creates a shape space in which shape similarities and differences are easily seen. The first principal component (PC) accounts for most of the variation in shape, the second PC for the second most variation, etc. Variation on one PC is statistically uncorrelated with variation on another (they are independent components of shape variation).

PCA (secondary use)

PCA is also useful because it gives:

1. **scores** (the coordinates of each object in the shape space) that can be used as shape variables in other statistical analysis

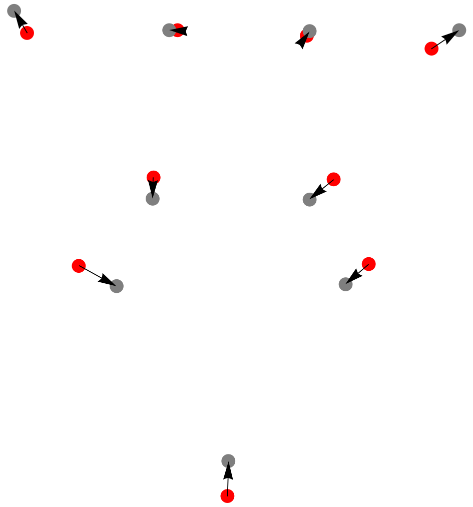
2. eigenvalues that are the variances of the objects on each PC. The eigenvalues sum to the total variance in the data set

3. eigenvectors that are the rotation vectors from the PC space back to the landmark shape. Eigenvectors * scores + consensus give you a shape model for a particular point in the PC shape space.

Procrustes distances same in object space and shape space

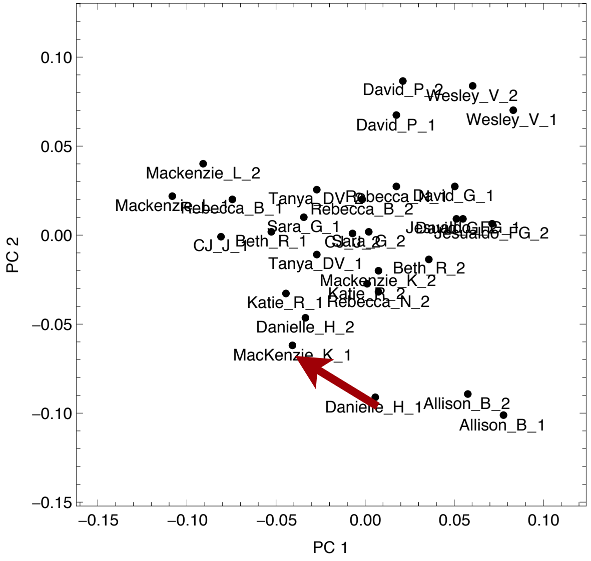
Distances between objects are the same regardless of whether you measure between their landmarks or their PC scores....

....provided you use the scores of all the PCs to do the calculation. You should always include all PCs in statistical tests (or have a good justification why not).



Procrustes distance between two objects

=

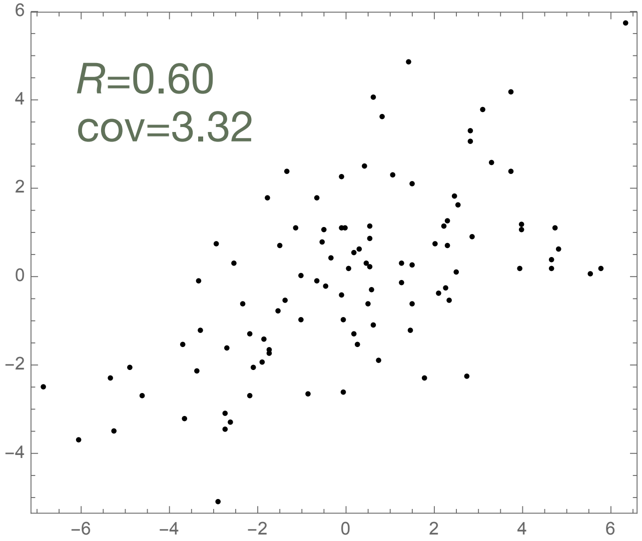


Euclidean distance between two locations in shape space

GMM is (almost) always based on covariance matrices

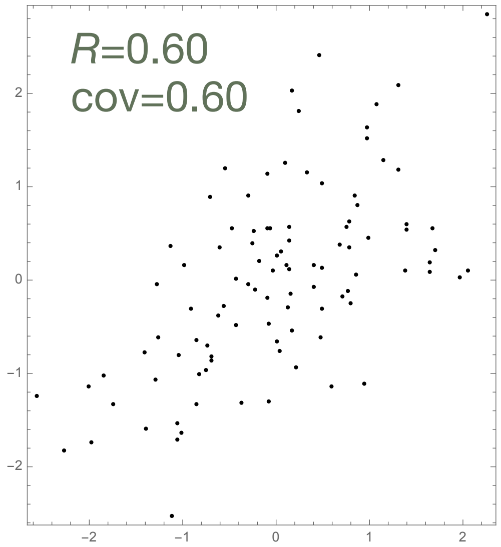
Covariance

$$\mathrm{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right],$$



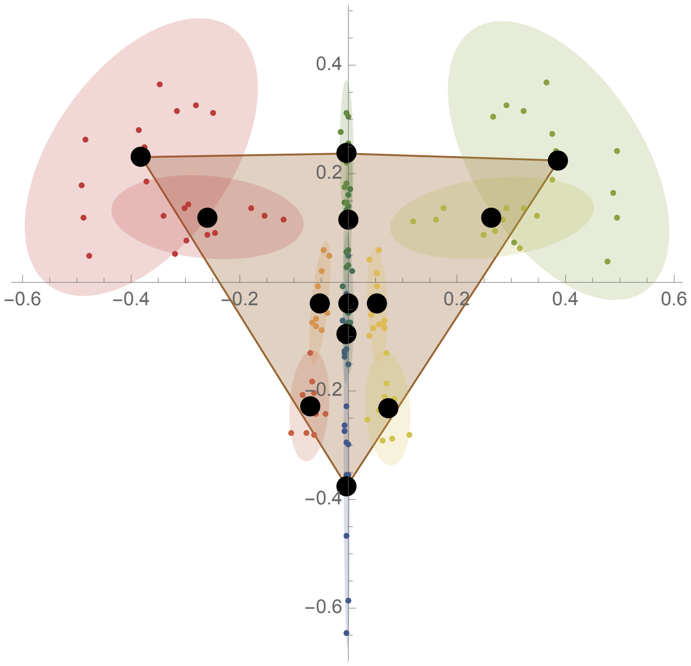
Correlation

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

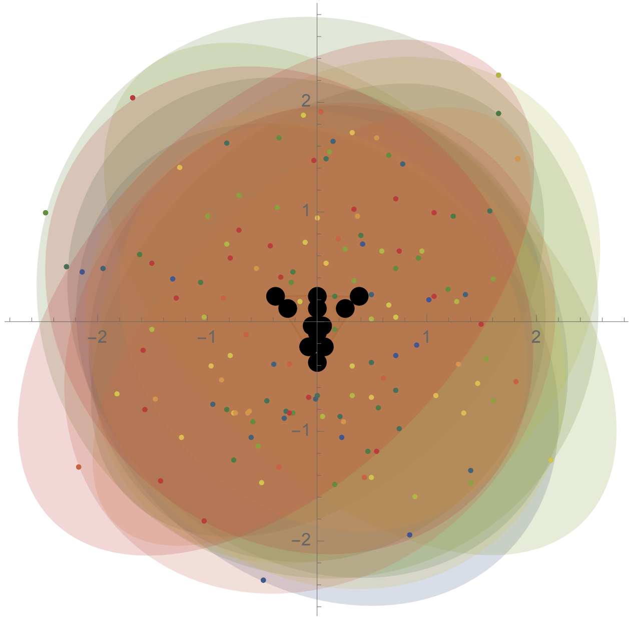


Applied to Landmarks

Natural Units



Standardized Units



Definitions

Statistic – a measure that summarizes some feature of a set of data (e.g., mean, standard deviation, skew, coefficient of variation, regression slope, correlation, covariance, principal component, eigenvalue, f-value).

Statistical parameter – the value of a particular statistic for the entire population (= statistical estimate).

Statistical estimate – the value of a particular statistic for a sample of a population (= statistical parameter).

Statistical test (or analysis) – an assessment of how likely a null hypothesis is to be true given the data at hand, or how probable it is that the data fit the null hypothesis given a random sample of the population.

Null hypothesis – usually the hypothesis that the estimated statistics from two or more samples result from two or more random samplings of the same population. In other words, the hypotheses that the sample do not come from different populations.

Confidence interval or standard error – a metastatistic that expresses how closely a statistical estimate is likely to match the population parameter.

Ordination (common GMM methods)

Ordering specimens along new variables

Principal Components Analysis (PCA)

- Used to show variation in its natural scale
- Used to show multivariate differences on small number of axes (dimension reduction)
- Arranges data by major axes based on measured variables
- Used in GMM to produce shape variables (=PCA scores)
- Axes are aligned along the major axes of variance
- Distances in PCA space = Procrustes distances

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

- Used to identify differences between known groups
- Discriminant functions are used to assign unknown objects to one of the groups
- Can be combined with cross-validation (leave-one-out) experiments to produce stats on how well the variables are able to classify objects
- Axes are aligned to the group means
- Distances in CVA space = Mahalonobis distances

Ordination (usually inappropriate for GMM)

Principal Coordinates Analysis (PCO)

- Arranges data by major axes based on distance measures
- Not normally used with GMM because quantitative variables are always available
- PCO on Euclidean distances = PCA on variables

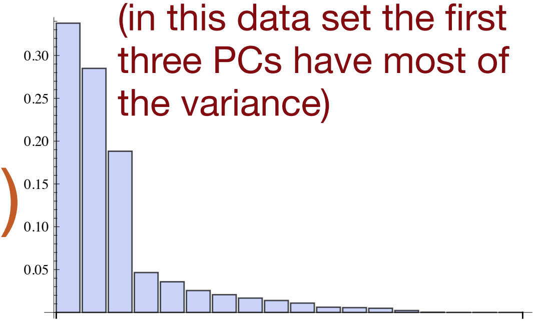
Non Metric Multidimensional Scaling (NMDS)

- Arranges data so the distances on 2D plot are as similar as possible to original multivariate distances
- Does not preserve between-object distances
- For visualization only, do not analyze NMDS scores

[note: PCO is sometimes referred to as "multidimensional scaling" even though it is a quite different kind of ordination. Know what you're using. PCO scores can be analyzed]

Between-groups PCA (BG-PCA)

- Projection of all data into a PCA based on group means
- Increasingly popular, but dubious ordination
- In GMM can produce highly misleading plots in which overlapping groups appear to be distinct
- BG-PCA scores do not have the properties of ordinary PCA (most of the variance may be on the higher axes; axes will be correlated)
- CVA is preferred if the goal is to identify group differences



Variance explained

(in this data set the first three PCs have most of the variance)

Principal components space (morphospace)

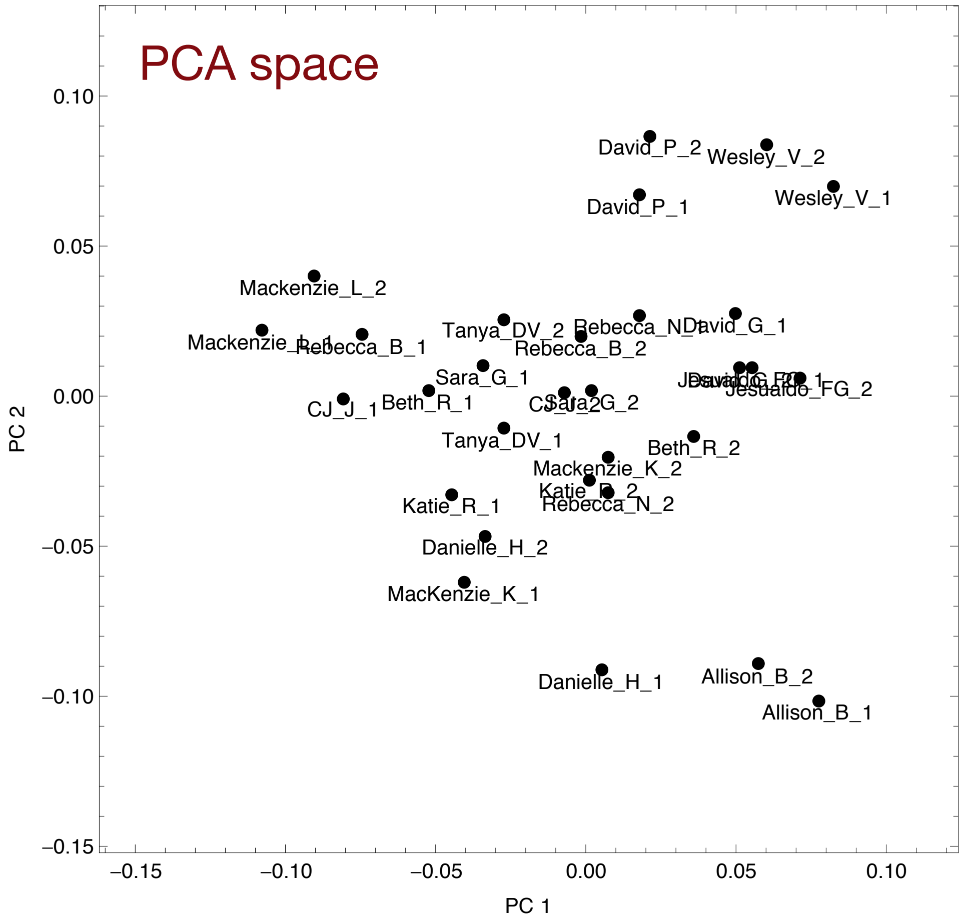
Arranges shapes to show the most variance possible on the first two axes (PCs)

Each point in shape space corresponds to a shape model

The origin ({0,0}) point corresponds to a shape model of the consensus, or mean shape

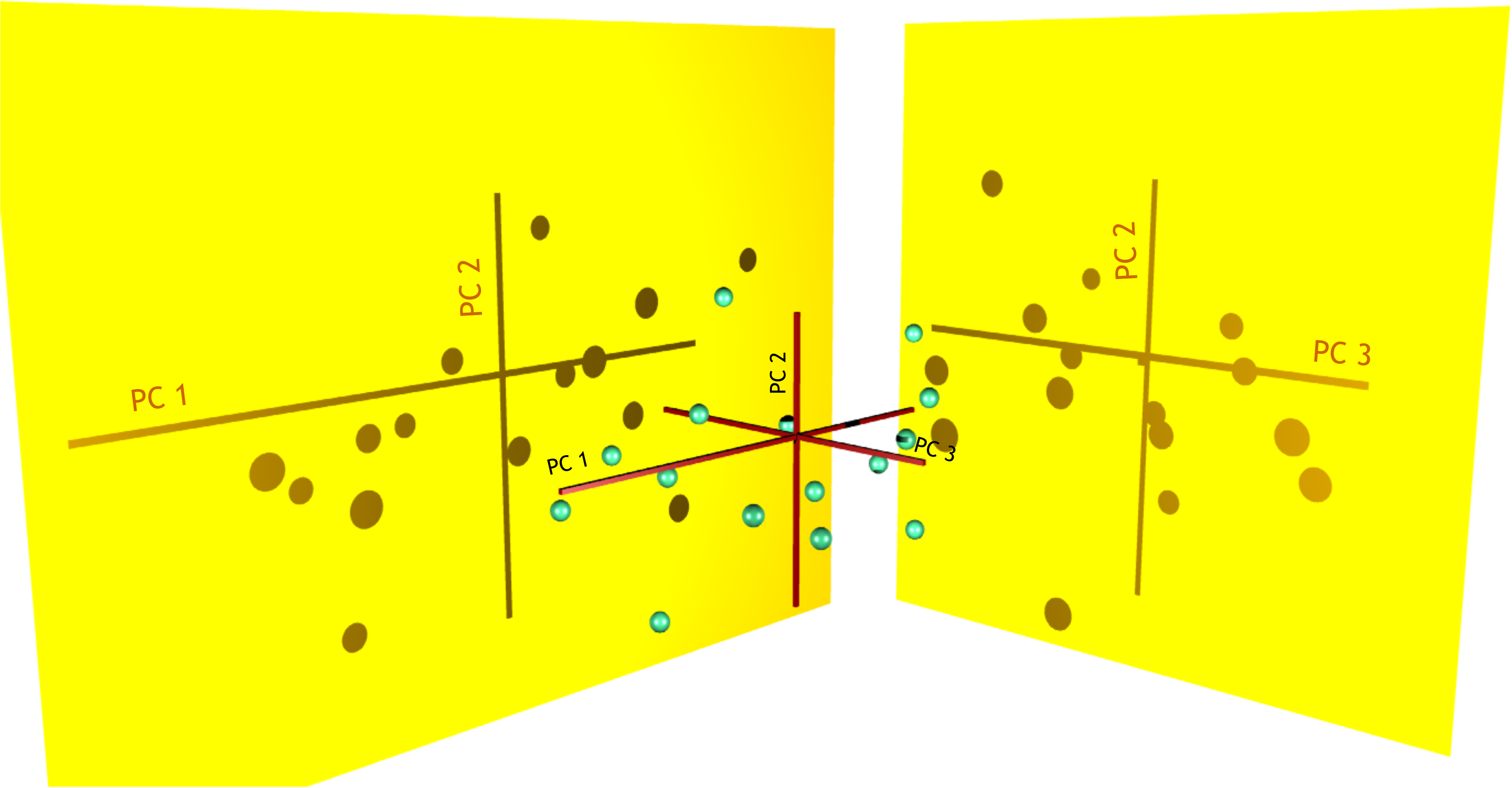
Coordinates of the objects in the morphospace are their scores

Highly multivariate. Check proportion of variance explained to discover how many are likely to be relevant



Morphospace is highly multivariate

The two-dimensional graphs we see are “shadows”, or **projections**.
Projections show a lot of the similarity in shape, but not all of it.



Statistical analysis: partitioning variance

Purpose: to ascertain to what extent part of shape variance is associated with a factor of interest, aka partitioning variance.

Typical parameters of a statistical analysis:

1. **P-value:** indicates whether the association is greater than expected by random chance

2. **Regression parameters (slopes, intercepts):** indicate the axis in shape space associated with the factor, useful for modeling the aspect of shape associated with the factor

3. **Correlation coefficient (*R*):** indicates the strength of the association between the variance and the factor

4. **Coefficients of determination (*R*^{2}):** indicates the proportion of the variance that is associated with the factor

[note: In GMM, variance / covariance = variation in shape]

Common statistical analyses for GMM

Regression
for use with factors that are continuous variables

Analysis of Variance (ANOVA)
for use with factors that are categorical (MANOVA if the test is multivariate)

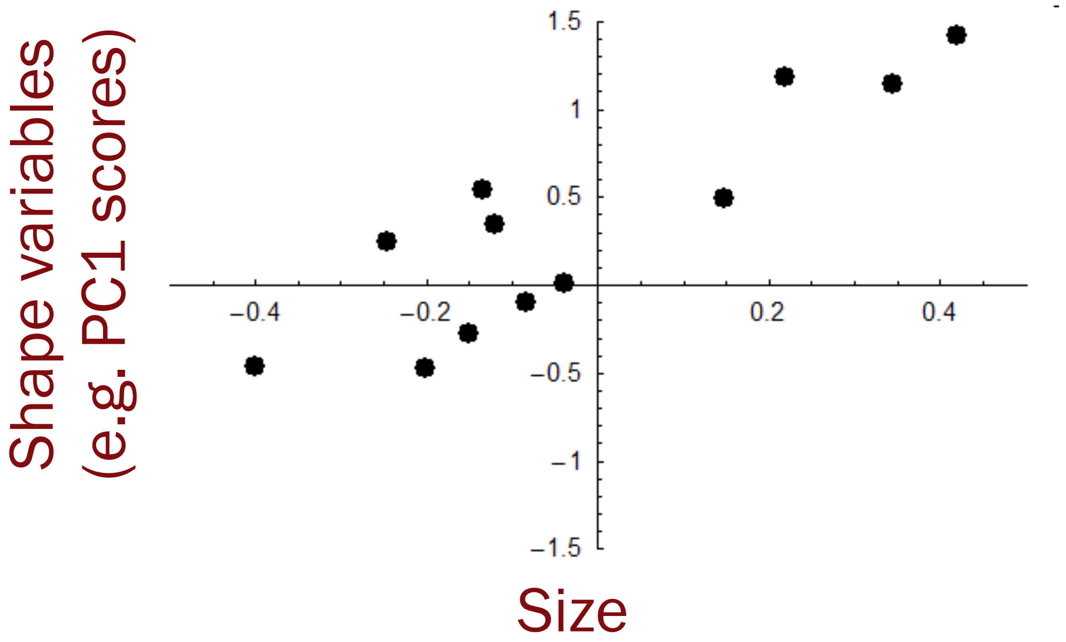
Two-block partial least squares (2B-PGLS)
for use with multiple factor variables (technically 2B-PGLS is an ordination method, not a statistical analysis)

Regression

Regression measures (and assesses) the relationship between geometric shape and another continuous predictor variable.

Continuous variables are ones that can take nearly any value (e.g., temperature, latitude, body mass, age, etc.).

Results of regression can be used to predict geometric shape for new values of the predictor variable.

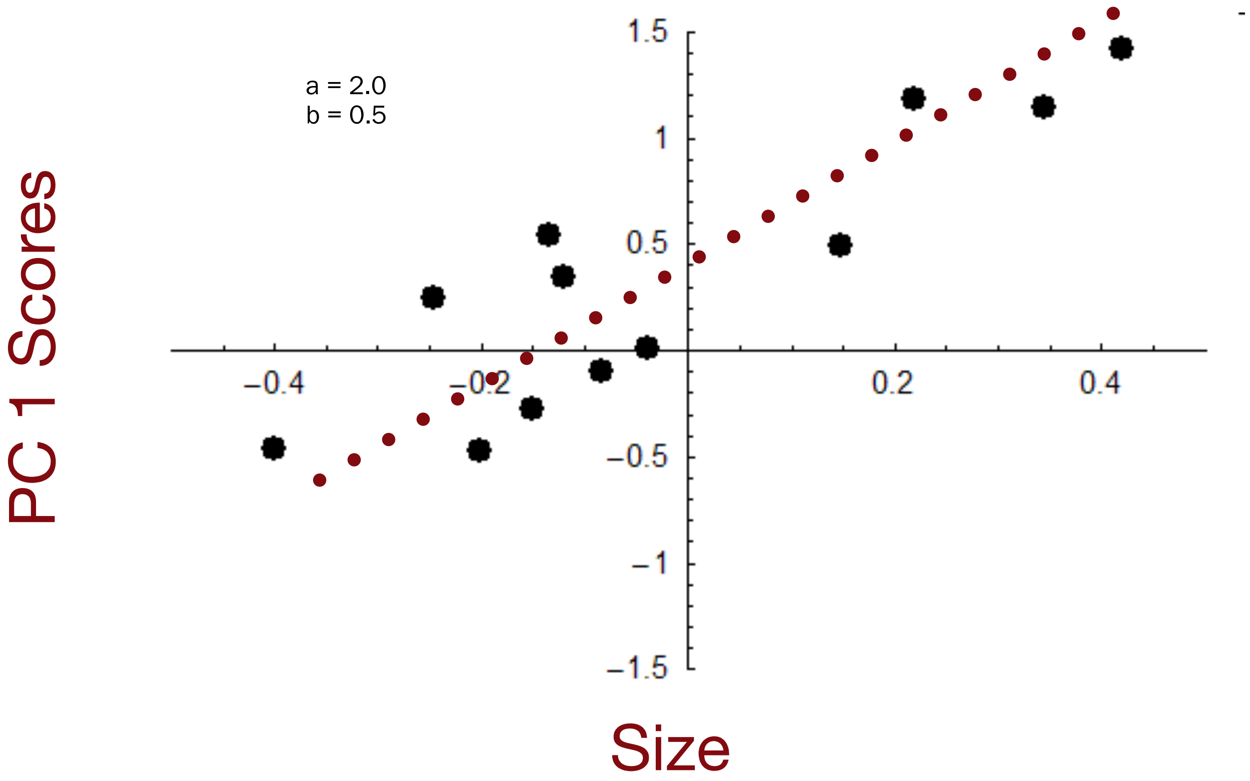


Linear Regression Basics

Linear regression finds the regression line that predicts variable Y from variable X.

$$Y = a X + b + E ,$$

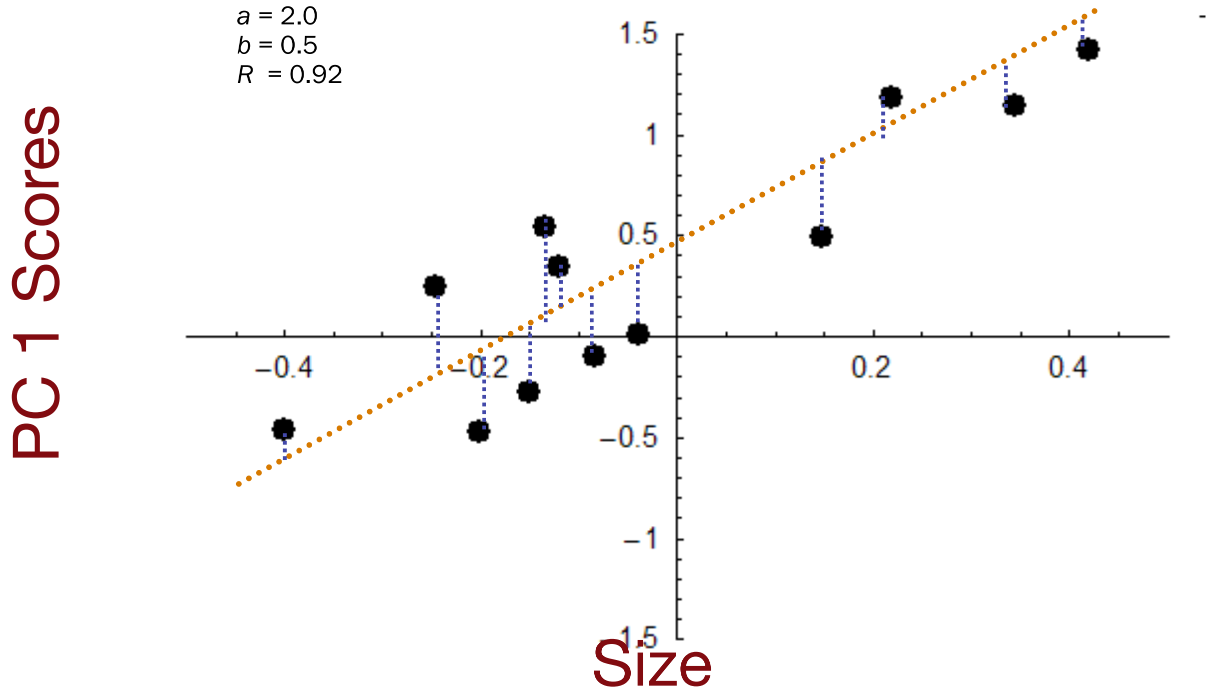
where a is the slope of the line, b is the intercept on the Y axis, and E is the residual error around the regression line.



The Correlation Coefficient (R) is a measure of E

$$Y = a X + b + E$$

R measures the tightness of fit to the regression line, roughly 1-E when E is measured as the standard deviation of the points from the line in the X axis and the data have been standardized so that E is never greater than 1.

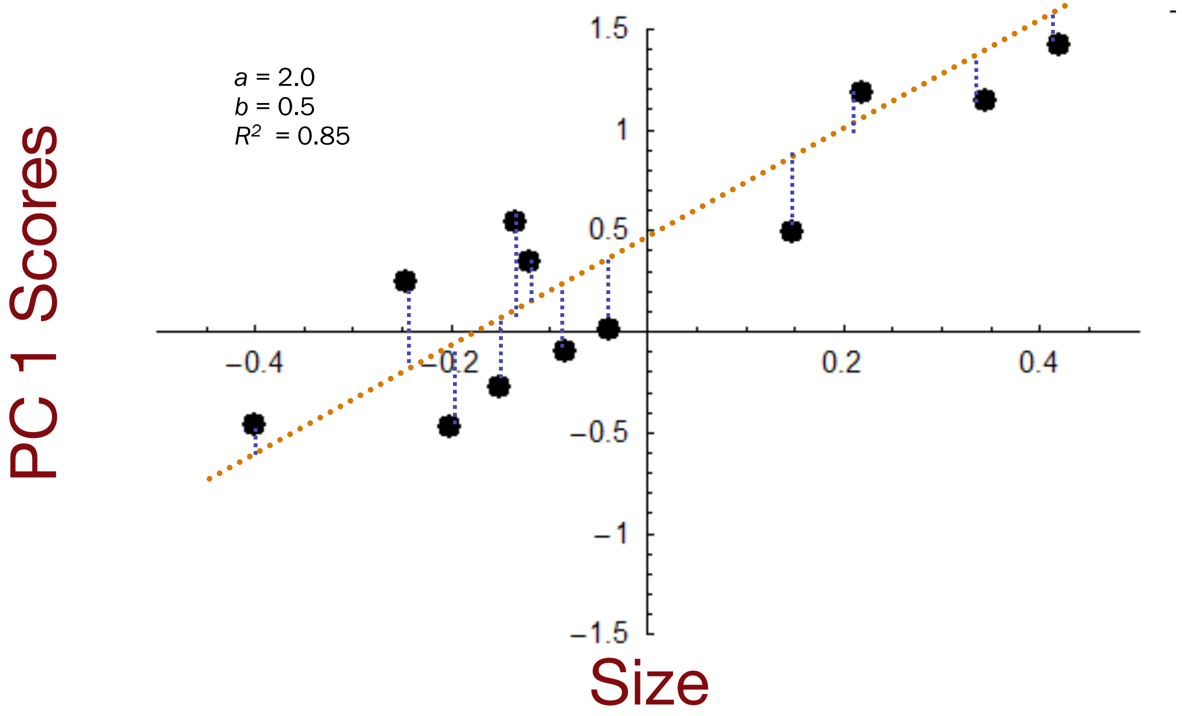


R ranges from 1.0 (perfect correlation) to 0.0 (no correlation).

The Coefficient of Determination (R2) is related to E

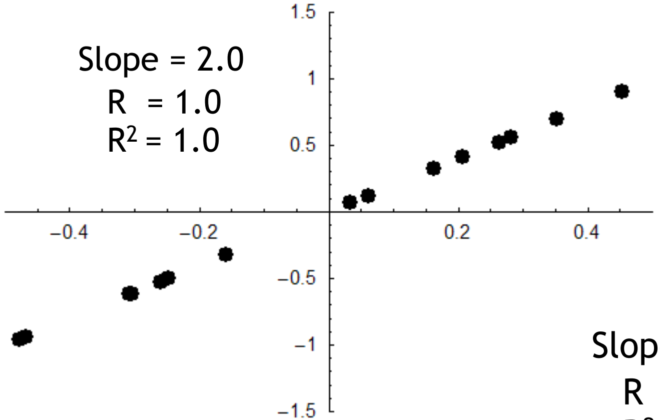
$$Y = a\,X + b + E$$

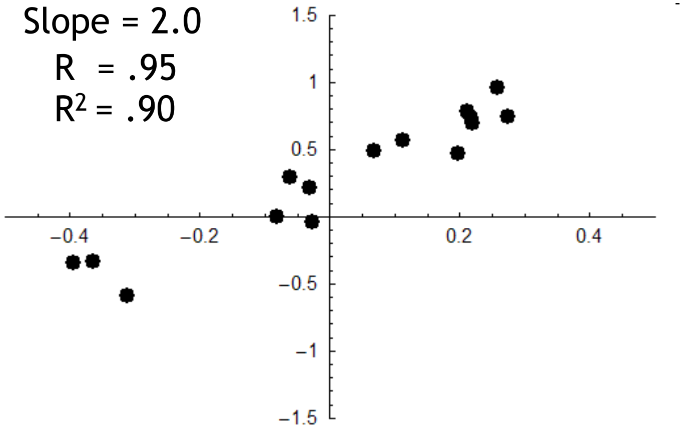
R^2 also measures the tightness of fit to the regression line, but as the variance rather than the standard deviation of the points around the line.
R^2 can be interpreted as the proportion of the variance in Y that is explained by X.

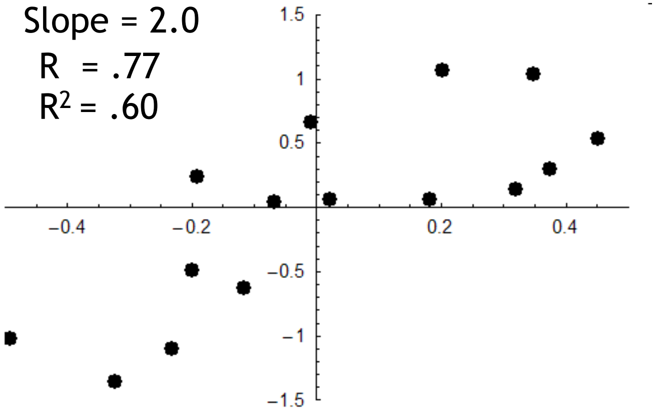


R^2 also ranges from 1.0 (100% explained) to 0.0 (0% explained).

Correlation and Coefficient of Determination measure scatter


Slope = 2.0
R = 1.0
R^2 = 1.0


Slope = 2.0
R = .95
R^2 = .90


Slope = 2.0
R = .77
R^2 = .60

Regression analysis thus reveals the following

$$Y = a\,X + b + E$$

- The slope of the relationship between x and y (coefficient *a*, often symbolized as *β*)
- The statistical significance and standard error of the slope
- The intercept of the regression line (coefficient *b*, often symbolized as *c*)
- The statistical significance and standard error of the intercept
- The amount of variance in Y explained by X (Sum of Squares of the Model)
- The amount of variance in Y not explained by X (SS Error, also known as SS Residual)
- The percentage of variance explained by the regression line (R^2 can be interpreted as the percentage of the variance in Y that is explained by X

Note that regression equations can be varied to suit new problems

$$Y = a\,X + b + E$$

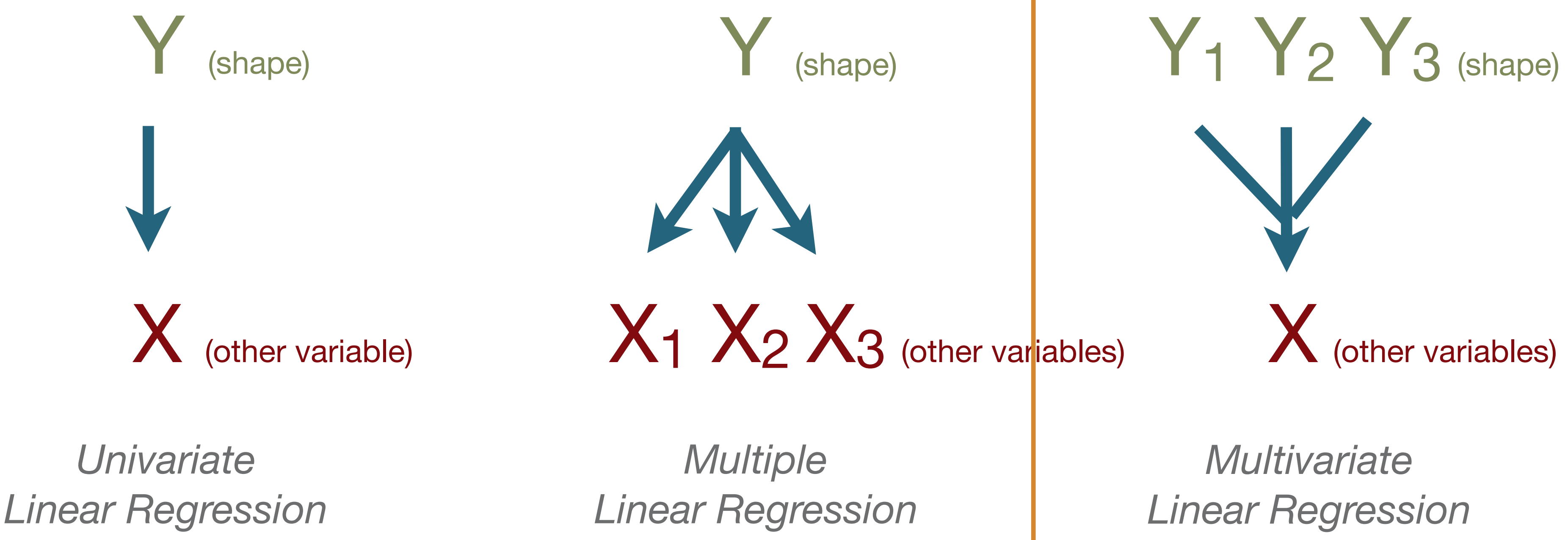
$$Y = a\,X^2 + b + E$$

$$Y = a_1\,X_1 + a_2\,X_2 + b + E$$

$$Y_2 + Y_2 = a\,X + b + E$$

$$Y = a\,X * Z + b + E$$

Types of regression



Y (shape)

X (other variable)

Univariate Linear Regression

Y (shape)

X_1 X_2 X_3 (other variables)

Multiple Linear Regression

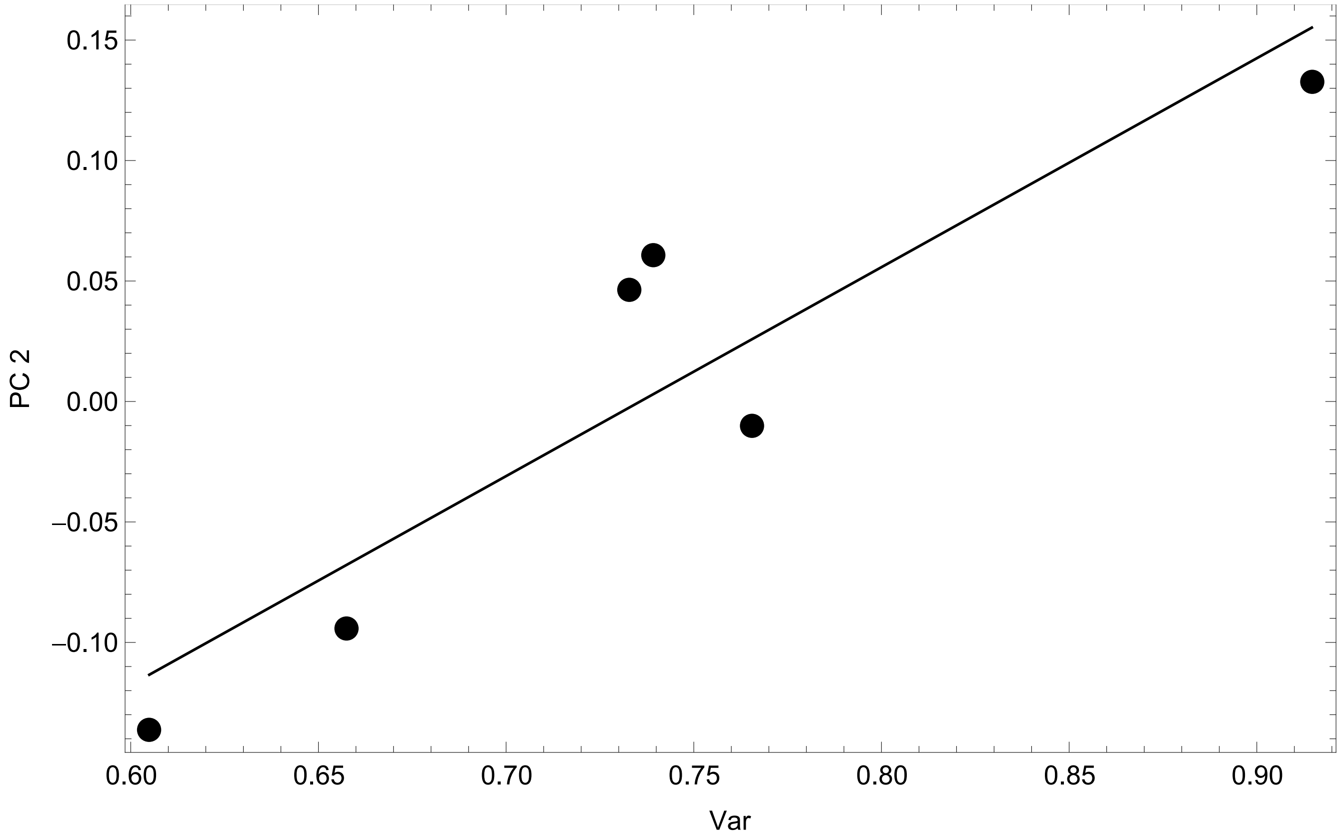
Best match for GMM

Y_1 Y_2 Y_3 (shape)

X (other variables)

Multivariate Linear Regression

Multivariate regression of shape

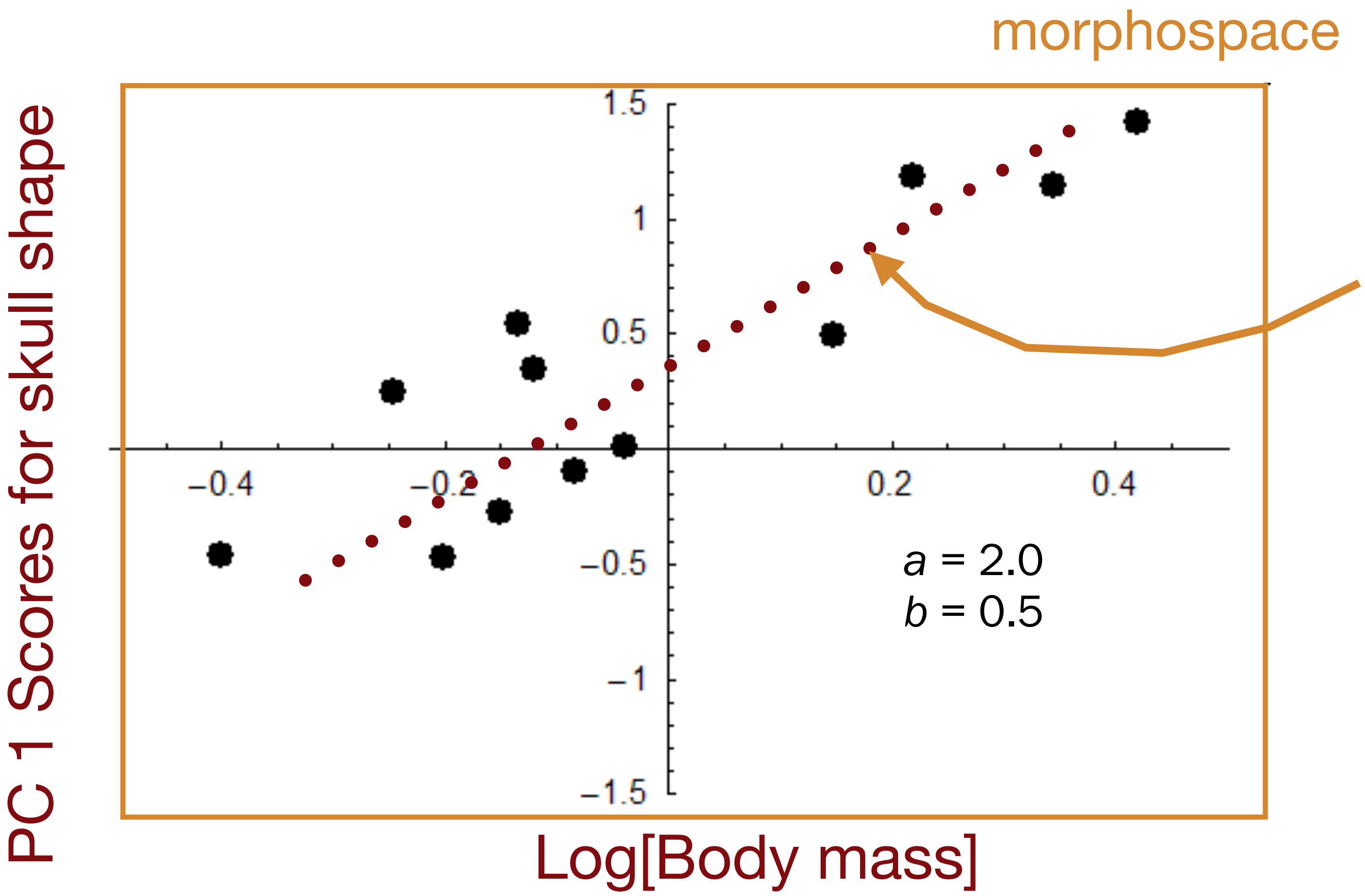


```
R-square (all PCs) = 0.21
P[R-square is random] = 0.34
                     │ PC1         PC2         PC3          PC4         PC5
─────────────────────┼──────────────────────────────────────────────────────────────
Intercept            │ -0.405353   -0.63799    -0.0227621   -0.07787    0.0231763
Slope                │ 0.550962    0.867165    0.0309385    0.105842    -0.0315015
Univariate R-square  │ 0.09        0.83        0.00         0.07        0.01
```

Regression example for GMM

$y = 2\,x + 0.5$

PC 1 score = 2 body mass + 0.5



describes a sequence of shapes that varies with body mass

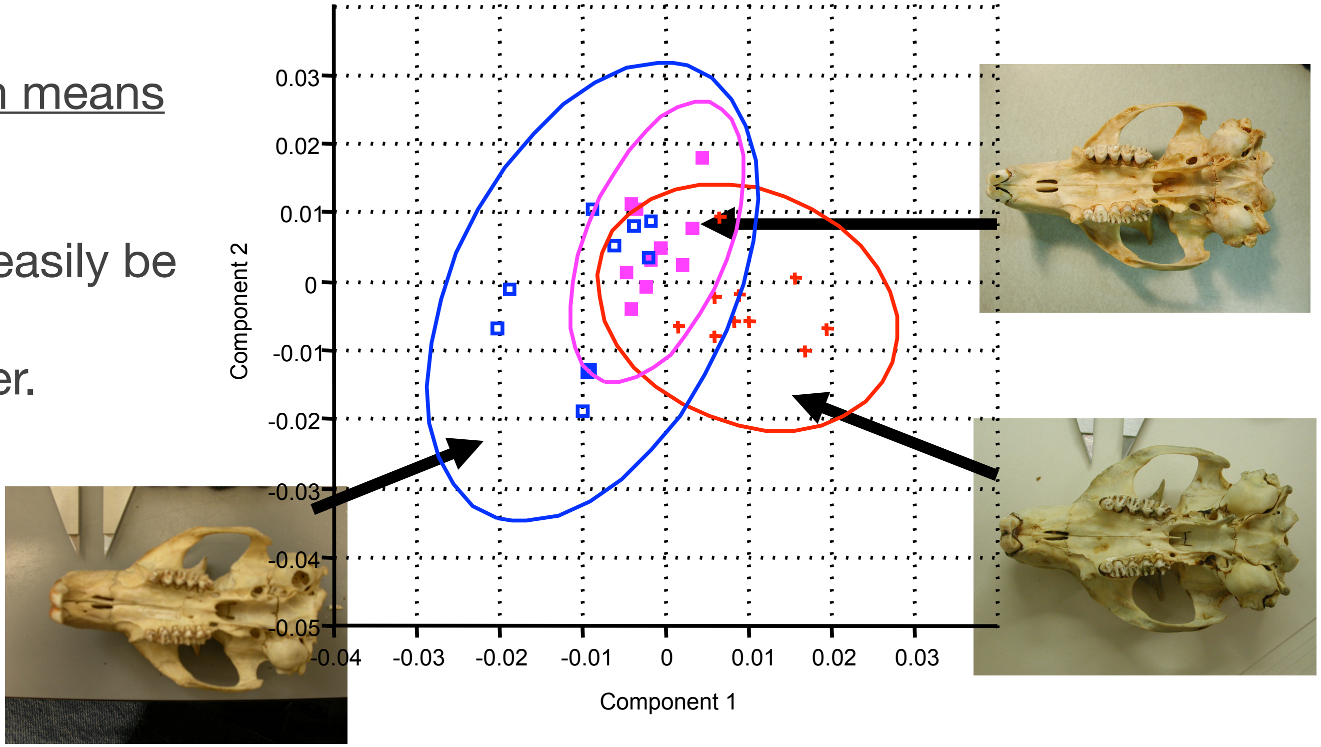
Multivariate Analysis of Variance (MANOVA)

Assesses the relationship between geometric shape and a categorical predictor variable.

Categorical variables are ones that define groups and are not ordered (e.g., male/female, herbivore/carnivore, island/continent)

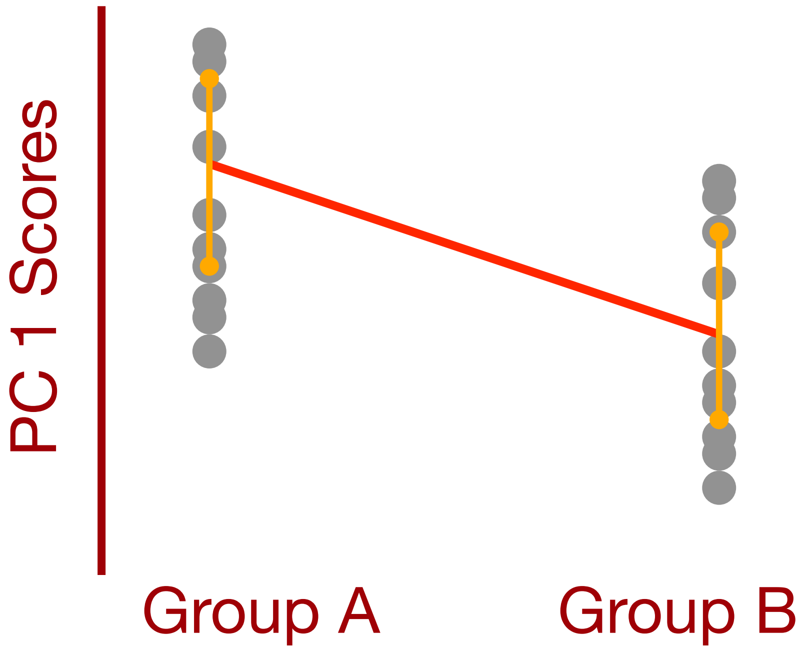
ANOVA tests for differences in means between the groups.

The mean of each group can easily be modelled and illustrated as a deformation of one to the other.



MANOVA

MANOVA uses F-test for overall difference among groups, where F is based on proportion of variance within groups to that between groups



F=10.29
p < 0.01

Overall results relate to differences between <u>any</u> group

<u>Pairwise comparisons</u> give *p* values for differences between specific pairs of groups if there are more than two groups.

MANOVA provides the following useful results

- means of the groups
- among-group variance (SS Model)
- within-group variance (SS Residual)
- total variance (SS Total)
- statistical significance of the difference between the two
- proportion of variance explained by the group difference (R^2)

On statistical tests

A difference to be a difference must make a difference. - Gertrude Stein

Always consider two aspects of a statistical test:

1. Does the P-value show the association to be significantly stronger than random?

2. Does the R^2 value show that a substantial part of the variance is associated with the factor?

What test to use?

How do we determine whether mandible shape is related to skull length?

How do we determine how much of mandible shape is related to skull length?

How do we determine whether mandible shape is related to sex?

How do we determine how much of mandible shape is related to sex?

How do we determine if mandible shape is related to depositional environment?

Bootstrapping and randomization tests are preferred for GMM data

- randomization tests take biases, non-normality, etc. into account automatically
- useful when assumptions of ordinary (parametric) statistical tests are not met, or when they are not known (such as with shape data)
- the tests randomize the data with respect to the statistic being measured
- the randomization is repeated a large number of times (e.g., 10000) and a distribution of the randomized statistic is generated.
- the observed value from the real data is compared to see whether it falls within the range of randomized values

Types of randomization tests

Bootstrap. Random resampling of original data, recalculation of test statistics to determine standard errors.

Jackknife. Same as bootstrap, but where each individual data point is left out in turn and the test statistic recalculated each time to determine standard error.

Randomization. Randomizing original data, test observed compared to randomized samples.

Monte Carlo. Data are simulated based on a particular hypothesis or model, real data are tested against the simulated data to see if the model holds.

Example: Test for difference in mean

1. Choose a statistic that describes difference in mean:

   D = Sqrt[Mean[sample1]-Mean[sample2]^2]

2. Pool samples and randomly draw new sample 1 and 2 with replacement.

3. Calculate D for randomly drawn samples

4. Repeat 10,000 times

5. Compare real D with randomized D distribution

6. P-value is the proportion of randomized D smaller than real D.

Bootstrap replicates vs. theoretical normal density

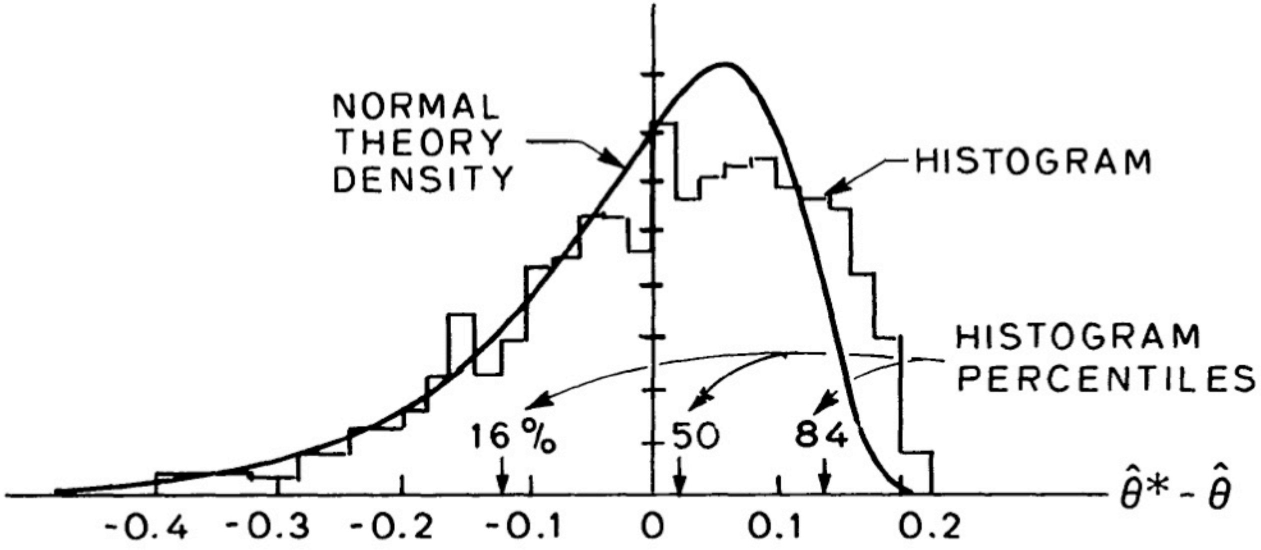


FIG. 2. *Histogram of $B = 1000$ bootstrap replications of $\hat{\theta}^*$ for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.*

Efron and Tibrishani, 1986

Important considerations

- **Always include all PC axes in statistical tests.** PC axes are sample dependent and do not align with real processes: beware analyzing only a subset of axes. If you use fewer axes, have a well-informed justification.

- **Visual inspection of PC plot may be misleading.** Variation is present on the higher axes... objects in two-dimensional morphospace may be farther than they appear.

- **Phylogeny may need to be taken into account.** If your data consist of more than one taxon, statistical tests will probably need adjustment for phylogenetic relatedness.