Statistical Analysis
Basic components of GMM

Whenever shapes are analyzed together, they 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.
Review

Eigenvalues
variance on each PC axis
(In Mathematica: \texttt{Eigenvalues[CM]})

Eigenvectors
loading of each original variable on each PC axis
(In Mathematica: \texttt{Eigenvectors[CM]})

Scores (=shape variables)
location of each data point on each PC axis
(In Mathematica: \texttt{PrincipalComponents[resids]})

resids are the residuals of the Procrustes coordinates
CM is the covariance matrix of the residuals
Procrustes distances same in object space and shape space

Distance between objects is 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)

Procrustes distance between two objects

Euclidean distance between two locations in shape space

\[
\text{In[25]} := \sqrt{\left(\langle \text{proc}[1] \rangle - \langle \text{proc}[2] \rangle \right)^2}
\]

\[
\text{Out[25]} = 0.14564
\]

\[
\text{In[26]} := \sqrt{\left(\langle \text{Scores}[1] \rangle - \langle \text{Scores}[2] \rangle \right)^2}
\]

\[
\text{Out[26]} = 0.14564
\]
Positive and negative ends of GMM PCs are arbitrary
PCA (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.

*Always* use all of the PC scores as shape variables for statistical analyses if possible. Avoid carrying out analyses on just one PC like the plague.
Statistical analysis: partitioning variance

In GMM, variance / covariance = variation in shape

Purpose of statistical analysis is to ascertain to what extent part of that variance is associated with a factor of interest, aka partitioning variance.

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
Statistical analysis

Regression analysis: 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)

Regression and MANOVA are the most common methods because they test for the association between shape and some other variable(s).
Regression analysis

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.
Regression example

**Response variable** (y)  
(aka, the dependent variable)  
[geometric shape as measured by scores]

**Predictor variable** (x)  
(also known as the independent variable)

**PC 1 Scores**

**Size**
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.

![Graph showing linear regression with PC 1 Scores and Size axes, with $a = 2.0$ and $b = 0.5$.]
The Correlation Coefficient (R) is a measure of E

\[ Y = a \times X + b + E \]

*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.

\[ a = 2.0 \]
\[ b = 0.5 \]
\[ R = 0.92 \]

*R* ranges from 1.0 (perfect correlation) to 0.0 (no correlation).
The Coefficient of Determination ($R^2$) 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 $X$.

$a = 2.0$
$b = 0.5$
$R^2 = 0.85$

$R^2$ also ranges from 1.0 (100% explained) to 0.0 (0% explained).
Correlation and Coefficient of Determination measure scatter of data around a linear relationship.

- **Slope = 2.0**  
  - $R = 1.0$  
  - $R^2 = 1.0$

- **Slope = 2.0**  
  - $R = 0.95$  
  - $R^2 = 0.90$

- **Slope = 2.0**  
  - $R = 0.77$  
  - $R^2 = 0.60$
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 \( \beta \))
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

\[ Y = a \ X + b + E \]
\[ Y = a \ X^2 + b + E \]
\[ Y = a_1 \ X_1 + a_2 \ X_2 + b + E \]
\[ Y_1 + Y_2 = a \ X + b + E \]
\[ Y = a \ X \times Z + b + E \]
Regression example for GMM

\[ y = 2x + 0.5 \]

PC 1 score = 2 body mass + 0.5

\[ a = 2.0 \]
\[ b = 0.5 \]
Univariate Regression in Mathematica

In[70]:= bodysize = {0.6096434840068126, 0.22768009261349156, 0.14267428923459877, 0.052650361541709906, 0.2485115495496144, 0.7843869946729531, 0.11408601516768813, 0.3612797615941762, 0.675457185239736, 0.5041337110519019, 0.9652503488936305, 0.643272655385519, 0.335279432158183645, 0.3230842158183645, 0.3243405473692482, 0.8043469600204844, 0.1451815203501622, 0.246509913279418, 0.2896148695669335, 0.8716828440086689, 0.1433024031749364, 0.7513222890447498, 0.7764383283356451, 0.7566633994122778, 0.533658919167681, 0.5236421964312582, 0.6337640391010464, 0.7040130378705679, 0.28514736961806664, 0.7392552017520052};

In[71]:= regressiondata = Transpose[{bodysize, Scores[[1 ;;, 1]]}]

Out[71]= {{0.609643, 0.0745046}, {0.22768, 0.0762214}, {0.142674, -0.0826827}, {0.0526504, 0.0274822}, {0.248512, 0.107981}, {0.784387, -0.0499361}, {0.114086, 0.0855598}, {0.36128, -0.0552454}, {0.675457, -0.00524725}, {0.504134, 0.0444181}, {0.96525, -0.0773846}, {0.643273, 0.0523366}, {0.335279, -0.0176351}, {0.323084, 0.0341239}, {0.324341, -0.179484}, {0.804347, -0.0572031}, {0.145181, -0.00102485}, {0.24651, 0.0334313}, {0.289615, -0.0711676}, {0.871683, 0.00687555}, {0.143302, -0.0514629}, {0.751322, 0.0906694}, {0.776438, 0.0273271}, {0.756663, -0.0601637}, {0.533659, -0.00741869}, {0.523642, 0.00162767}, {0.633764, -0.00755168}, {0.704013, -0.00206838}, {0.285147, -0.0361198}, {0.739255, -0.0214568}}

In[77]:= lm = LinearModelFit[regressiondata, {x, 1}, x]

Out[77]= FittedModel[0.0126624 - 0.0261688 x]

In[73]:= lm["ANOVATable"]

Out[73]= DF SS MS F–Statistic P–Value
1 1 0.00139567 0.00139567 0.512699 0.479902
Error 28 0.0762214 0.00272219
Total 29 0.0776171

In[82]:= R2 = 0.0013956656150502095 / 0.07761705710408748

Out[82]= 0.0179814

R² is Model SS / Total SS. Tells us 1.8% of variance on PC1 is explained by body size.
Plot the regression

List plot of regression data

Plot of the regression equation over interval $x=0.0$ to $x=0.9$

```
In[74]:= Show[{ListPlot[regressiondata], Plot[lm[x], {x, 0, .9}]}]]
```

```
Out[74]=
```

PC 1 Score

Body size
Regression - a multivariate look

Univariate Linear Regression

Multiple Linear Regression

Multivariate Linear Regression
ShapeRegress[proc, variable (, PCs)]

Function for multivariate regression of entire shape onto a single independent variable

- **proc** is a matrix of Procrustes superimposed landmark coordinates with objects in rows and coordinates in columns.
- **variable** is the variable onto which proc is to be regressed. It is a vector containing observations for each object from a continuous variable.
- **PCs** is an optional parameter specifying which PC should be shown in the graph. By default the regression on PC1 is shown.
Example of ShapeRegress[]

Graph shows one dimension of the regression

R-square shows total amount of shape explained by x variable

Table gives regression coefficients and univariate r-squared

*R-square (all PCs) = 0.21
P[R-square is random] = 0.34*

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.405353</td>
<td>-0.63799</td>
<td>-0.0227621</td>
<td>-0.07787</td>
<td>0.0231763</td>
</tr>
<tr>
<td>Slope</td>
<td>0.550962</td>
<td>0.867165</td>
<td>0.0309385</td>
<td>0.105842</td>
<td>-0.0315015</td>
</tr>
<tr>
<td>Univariate R-square</td>
<td>0.09</td>
<td>0.83</td>
<td>0.00</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Multivariate Regression in Mathematica

Requires the Morphometrics package
Analysis of Variance (ANOVA)

Also, Multivariate Analysis of Variance (MANOVA)

ANOVA 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.
ANOVA

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

Pairwise comparisons give p values for differences between specific pairs of groups if there are more than two groups.

Overall results relate to differences between any group.
ANOVA gives the following useful results

The means of the groups

The among-group variance (SS Model)

The within-group variance (SS Residual)

The total variance (SS Total)

The statistical significance of the difference between the two

The percentage of variance explained by the group difference ($R^2$).
**Example testing for camera angle differences in face shape**

Univariate ANOVA in Mathematica

```
In[35]:= << ANOVA`  

In[39]:= Groups = associateddata[[2 ;;, 4]]

Out[39]= {Straight, Straight, Straight, Left, Left, Straight, Left, Left, Right, Right, Right, 
  Straight, Straight, Left, Straight, Right, Straight, Right, Right, Left, Straight, 
  Left, Straight, Straight, Straight, Straight, Right, Right, Straight, Right}  

In[40]:= anovainput = Transpose[{Groups, Scores[[1 ;;, 1]]}]

Out[40]= {{Straight, 0.0745046}, {Straight, 0.0403798}, {Straight, -0.0826827}, {Left, 0.0274822}, 
  {Left, 0.107981}, {Straight, -0.0499361}, {Left, 0.0805598}, {Left, -0.0552454}, {Right, -0.00524725}, 
  {Right, 0.0444181}, {Right, -0.0773846}, {Straight, 0.0523366}, {Right, -0.0176351}, 
  {Left, 0.0341239}, {Straight, -0.0179484}, {Right, -0.0572031}, {Straight, -0.00102485}, 
  {Right, 0.0334313}, {Right, -0.0711676}, {Left, 0.00687555}, {Straight, -0.0514629}, {Left, 0.0906694}, 
  {Straight, 0.0273271}, {Straight, -0.0601637}, {Straight, -0.00741869}, {Straight, 0.00162767}, 
  {Straight, -0.00755168}, {Right, -0.00268388}, {Straight, -0.0361198}, {Right, -0.0214568}}

In[41]:= ANOVA[anovainput, PostTests -> Bonferroni]

Out[41]= ANOVA` ANOVA``

```
Model 2 0.0165152 0.00825762 3.64892 0.0395649
Error 27 0.0611018 0.00226303
Total 29 0.0776171

{All, 2.24358 \times 10^{-17}}``

```
CellMeans`  

Model[Left] 0.0417781
Model[Right] -0.0195848
Model[Straight] -0.00905121

\{ANOVATable\}
```

```
R2 = 0.01651523093155123 \text{\textasciitilde} 0.07761705710408748

Out[42]= 0.212778
```

\(R^2\) is Model SS / Total SS. Tells us 21.3% of variance on PC1 is explained by face direction. PC1 accounts for 33.7% of total face variance, so we have accounted for 7.2% of total variance.
Plot ANOVA data

You can plot the PC scores of your groups to see how different they are. Substitute the group labels with numbers to do this.

In[53]:= ListPlot[anovainput /. {"Left" -> 1, "Straight" -> 2, "Right" -> 3}, PlotRange -> {{0, 4}, Automatic}]

Out[53]=

You can plot the PC scores of your groups to see how different they are. Substitute the group labels with numbers to do this.
Multivariate MANOVA in Mathematica

Needs the Morphometrics Package

- Categorical predictor variable
- Procrustes superimposed coordinates
- P-value for overall shape differences between groups (no assumptions about normality or balanced sample sizes)
- P-value for overall shape differences between groups (based on f-test, which assumes normality and balanced samples sizes, which are probably not valid assumptions for most GMM data)
- R-squared value is the proportion of variance in shape that is explained by the groups
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?