Multivariate Data & GLM

Advanced Biostatistics

Dean Adams

Lecture 6

EEOB 590C

GLM is the fitting of models

-Fit null model (e.g., H_0 = Y~1) and then more complicated models
-Assess fit of different models via SS (i.e,. LRT)

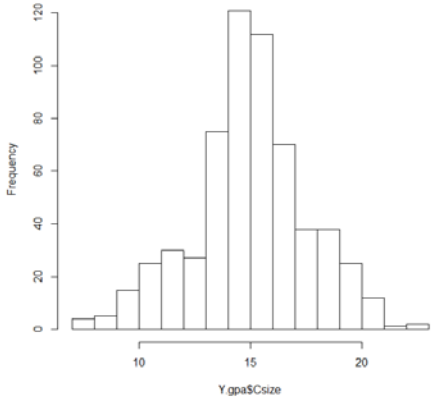
ANOVA and regression all the same model (GLM)
(key difference is what is used in X)

GLM parameters found using matrix algebra

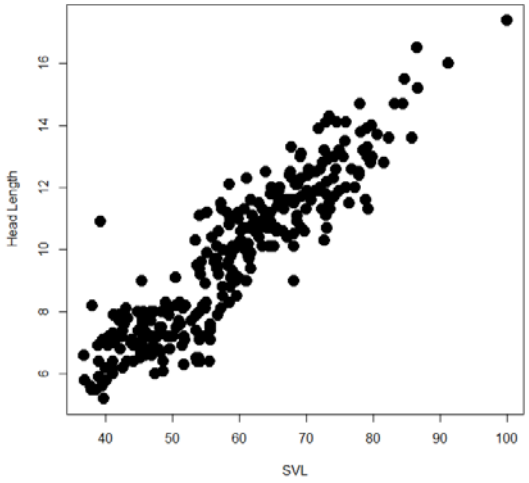
$$\mathbf{B} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

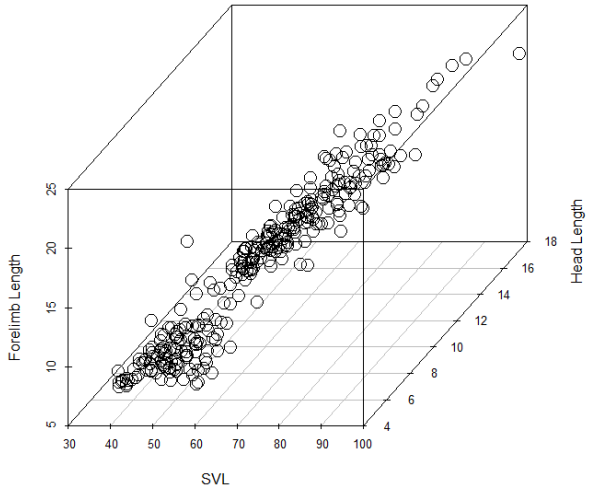
Question: What about multivariate-Y?

Univariate statistics: Assess variation in single Y (obtain scalar result)



Multivariate statistics: Assess variation in multiple Y simultaneously



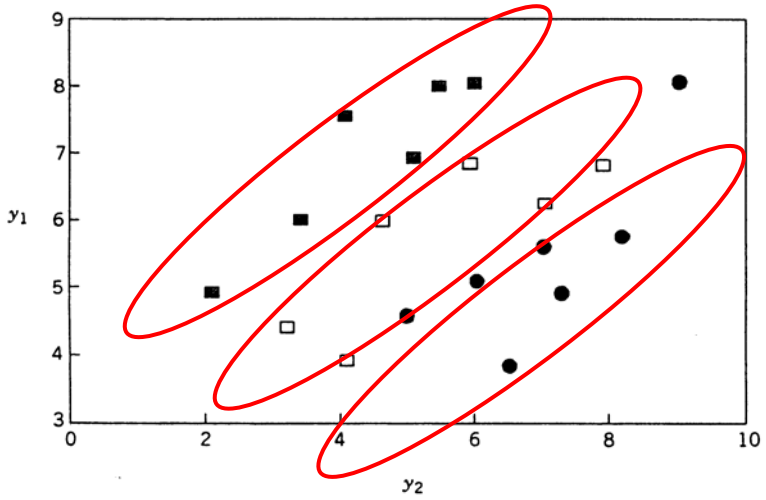


Multivariate methods are mathematical generalizations of univariate (ACTUALLY, univariate methods are special cases of multivariate!)

More complete description of pattern

Biological data are often multivariate, so treat as such

Separate univariate analyses misses covariation signal*



ANOVAs on y_1 and y_2 separately would fail to identify group differences from covariation

*Covariation ***IS*** biology; so one must think multivariately!

Increasing # dimensions of data (Y) means more information

However, for a given *n* the statistical power decreases

Eventually, too few *n* for # variables in Y

How large should sample size be?

Many suggestions:

- $n = 2*\#vars$
- $n = 4*\#vars$
- $n = \#vars^2$
- $n_{gp} = 2*\#vars$
- $n_{gp} = 4*\#vars$

*As we'll see, Q-mode inferential statistics depend less on this

Several ways to identify patterns in $Y_{n \times p}$ (Y-matrix of *n* objects × *p* variables)

1: Linear models: MANOVA/regression to assess patterns

2: R-mode analyses: Summarize by columns (VCV matrix of variables)

3: Q-mode analyses: Summarize by rows (distance matrix for objects)*

First one needs to *DESCRIBE* the multivariate data!

*Many Q-mode & R-mode methods yield identical results: PCA (R-mode) vs. $PCoA_{DEuclid}$ (Q-mode)

Describing multivariate data = understanding it

Data are dots in space, so goal is to describe point cloud

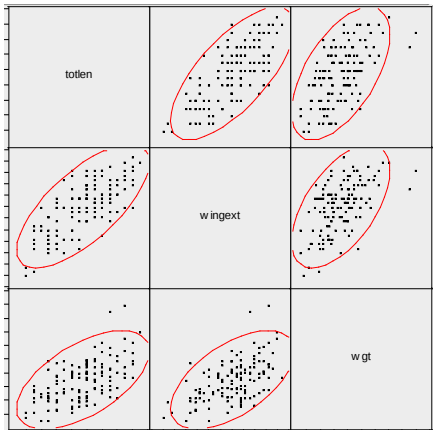
VCV (S): Covariance matrix of variances and covariances ('multivariate variance')

$$s^2 = \hat{\sigma}^2 = \frac{\sum (\mathrm{Y}_i - \overline{\mathrm{Y}})^2}{(\mathrm{n}-1)}$$

Correlation matrix: matrix of pairwise variable correlations (standardized covariance matrix)

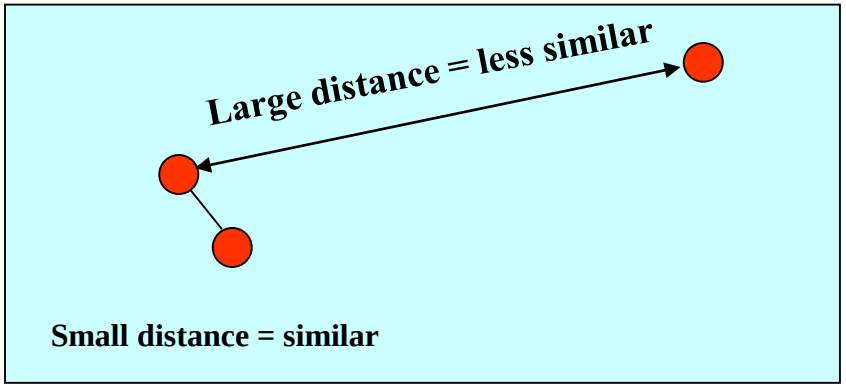
$$\mathbf{S} = \begin{bmatrix} s_{11} & & \\ s_{21} & s_{22} & \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & & \\ r_{21} & 1 & \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

Bivariate correlation plots are also useful



Data are ‘dots’ in multivariate data space

Distances between objects describe similarity (or difference)



Distance (or similarity) measure used depends on the type of data

NOTE: Distances (D) can be converted to similarities (S) and vice-versa

When scaled to 0→1, relationship is: $D = 1 - S$ or $D = \sqrt{1 - S}$ or $D = \sqrt{1 - S^2}$

Data are 0/1

Generate 2×2 frequency table for each pair of specimens

Specimen 1 \ Specimen 2	1	0
1	a	b
0	c	d

Similarity/distance based on a,b,c,d (# traits in each category)

Simple matching coefficient: $S_1 = \frac{a+d}{a+b+c+d}$

Jaccard's coefficient: $S_2 = \frac{a}{a+b+c}$

Hamming distance: $D_1 = b+c$ (#differences)

Choice depends on data and assumptions (e.g., are shared absences (0,0) meaningful?)

For S→ D conversions see Legendre & Legendre 2012

Multi-state data requires different S/D measures

Spec 1	9	3	4	6	2	1	6	8	7	8
Spec 2	5	3	3	4	3	1	6	4	6	8
$Y_{Agreements}$	0	1	0	0	0	1	1	0		1

Percent matching: $S_4 = \sum Y_{agree} / length(Y)$

Note: can extend all binary descriptors in this fashion

Gower's general similarity: $S_5 = \frac{1}{p} \sum s_{12j}$

Contribution of each trait (s_j) is: 0/1 for binary OR multi-state

For S→ D conversions see Legendre & Legendre 2012

Common in ecology (species abundance)

Bray-Curtis: commonly used

$$D = \frac{\sum \left| y_{1j} - y_{2j} \right|}{\sum y_{1.} + \sum y_{2.}}$$

Sum(absolute difference in abundance species by species relative to total counts at each site

Note: not a metric distance (described below)

Chi-Square distance:

$$D_{X^2} = \sqrt{\sum_k \frac{1}{y_k} \left(\frac{y_{ki}}{y_i} - \frac{y_{kj}}{y_j} \right)^2}$$

Is metric (square-root of counts commonly used)

where $y_k = \sum y_{+k} \Big/ \sum y_{++}$

and $y_i = \sum y_{i+}$

For S→ D conversions see Legendre & Legendre 2012

Continuous data common in E&E

MANY possible distance measures

Euclidean distance: $$D_{Euclid} = \sqrt{\sum \left(y_{1j} - y_{2j} \right)^2} = \sqrt{\left(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2 \right)^t \left(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2 \right)}$$

Manhattan distance: $$D_{Manhat} = \sum \left| y_{1j} - y_{2j} \right|$$

Canberra distance: $$D_{Canberra} = \sum \left(\frac{\left| y_{1j} - y_{2j} \right|}{\left(y_{1j} - y_{2j} \right)} \right)$$

Note double 0 must be removed

Mahalanobis distance: $$D^2_{Mahal} = \left(\mathbf{Y}_1 - \mathbf{Y}_2 \right)^t \mathbf{S}^{-1} \left(\mathbf{Y}_1 - \mathbf{Y}_2 \right)$$

Some distances (e.g., D_{euclid}) generate a METRIC space

All distance measures require data in commensurate units

D_{euclid} requires all Y are continuous

$D_{hamming}$ requires all Y are 0/1

Researchers sometimes combine data types

Y=SVL, #bristles, presence of nose (0/1)

Y=elevation, #individuals/km, presence of competitor (0/1)

THIS IS GIGO!!!

A program may calculate the distance, but it has no meaning

(variables in incommensurate units, not weighted properly, etc.)

Could convert characters to common unit & combine, but still have the weighting problem

Generally not advisable to combine data types for obtaining distances

Not all distance measures are the same: they fall into different classes

Metric: A distance is a metric IFF:

1: minimal: min(d_{11}=0)

2: symmetry: (d_{12}=d_{21})

3: Triangle inequality: (d_{12}+d_{13} ≥ d_{23})

Semimetric (pseudometric): Triangle inequality not satisfied

(e.g., Bray-Curtis distance & Sørenson's similarity)

Nonmetric: min(d_{11}<0): i.e. has negative distances

(e.g.,Kulczynski's coefficient)

Some distance measures are metric (e.g., D_{Euclid}, D_{Manhat}), others not (see above)

For discussion of common ecological distance/similarity coefficients, see Legendre & Legendre, 2012 *Numerical Ecology*

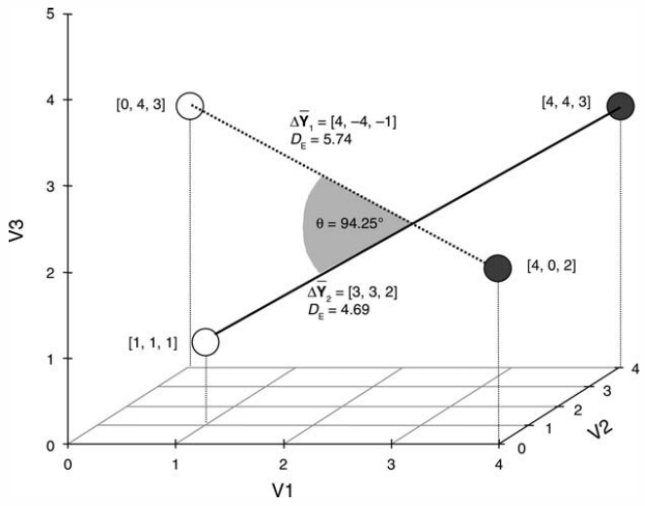
Euclidean spaces are defined by the Euclidean metric (D_{euclid})

Euclidean spaces satisfy:

- 3 metric space conditions: 1: min(d_{11}=0); 2: d_{12}=d_{21}; 3: d_{12}+d_{13} ≥ d_{23}
- **Axis Perpendicularity**: if $\sum x_i y_i = 0$ *x* & *y* are perpendicular (orthogonal)

In Euclidean spaces, distances, directions, and angles can be defined

- Thus they can be examined and compared for biological interpretation



NOTE: most multivariate studies assume a metric (typically Euclidean) geometry

-ALL previous GLM models (ANOVA, factorial, nested, multiple regression, ANCOVA, etc.) can be completed using the matrix form

$$\mathbf{Y} = \mathbf{XB} + \varepsilon \qquad \mathbf{B} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

	1 Categorical X	>1 Categorical X	1 Continuous X	>1 Continuous X	Both
1 Continuous Y	ANOVA	Factorial ANOVA	Regression	Multiple Regression	ANCOVA
>1 Continuous Y	MANOVA	Factorial MANOVA	Multivariate Regression	Multivariate Multiple Regression	MANCOVA

The 'leap' to multivariate GLM is accomplished simply by incorporating additional columns to **Y** (i.e., examine more than 1 Y-variable simultaneously)

The algebra enables us to estimate the regression coefficients for a single (univariate) response variable (Y)

Now, let's generalize

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times k\cdot k\times 1}{\mathbf{X}\mathbf{b}} + \underset{n\times 1}{\mathbf{e}}$$

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times k\cdot k\times p}{\mathbf{X}\mathbf{B}} + \underset{n\times p}{\mathbf{E}}$$

$$\hat{\mathbf{b}} = \left(\mathbf{X}^{\mathbf{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{y}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}^{\mathbf{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{Y}$$

$$\hat{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk-1} \end{bmatrix} \qquad \hat{\mathbf{B}} = \begin{bmatrix} b_{01} & b_{02} & \cdots & b_{0p} \\ b_{11} & b_{12} & \cdots & b_{1p} \\ \vdots & \vdots & & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kp} \end{bmatrix}$$

ALL computations the same as in univariate

$$\mathbf{X_R} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \mathbf{X_F} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

1) Estimate coefficients $\hat{\mathbf{B}}_\mathbf{R} = \left(\mathbf{X}_\mathbf{R}^\mathbf{T}\mathbf{X}_\mathbf{R}\right)^{-1}\left(\mathbf{X}_\mathbf{R}^\mathbf{T}\mathbf{Y}\right)$ $\hat{\mathbf{B}}_\mathbf{F} = \left(\mathbf{X}_\mathbf{F}^\mathbf{T}\mathbf{X}_\mathbf{F}\right)^{-1}\left(\mathbf{X}_\mathbf{F}^\mathbf{T}\mathbf{Y}\right)$

2) Estimate predicted values $\hat{\mathbf{Y}}_\mathbf{R} = \mathbf{X}_\mathbf{R}\hat{\mathbf{B}}_\mathbf{R}$ $\hat{\mathbf{Y}}_\mathbf{F} = \mathbf{X}_\mathbf{F}\hat{\mathbf{B}}_\mathbf{F}$

3) Estimate Error $\hat{\mathbf{E}}_\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{R}$ $\hat{\mathbf{E}}_\mathbf{F} = \mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{F}$

4) SSCP $\mathbf{S}_\mathbf{R} = \hat{\mathbf{E}}_\mathbf{R}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{R}$ $\mathbf{S}_\mathbf{F} = \hat{\mathbf{E}}_\mathbf{F}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{F}$

What about test statistics here?

Problem: For multivariate data, SS_R & SS_F are now SSCP matrices, so no univariate F-ratio

Need to summarize variation explained by S_R & S_F matrices

"Traditional" test statistics

(Require *n* >> *p*, and conversion to parametric *F* values)

Based on $\mathbf{S}_F^{-1}(\mathbf{S}_R - \mathbf{S}_F)$

And finding canonical vectors and eigenvalues

$$(\mathbf{S}_f^{-1}(\mathbf{S}_r - \mathbf{S}_f) - \lambda_i \mathbf{I})\mathbf{u}_i = 0$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{\Delta k} \end{bmatrix}$$

$$\mathbf{C} = \mathbf{U}\left(\mathbf{U}' \frac{1}{\Delta k}(\mathbf{S}_r - \mathbf{S}_f)\mathbf{U}\right)^{-1/2}$$

To determine multivariate statistics

$$\Lambda_{Wilks} = \prod_{i=1\ldots p}\left(\frac{1}{1+\lambda_i}\right) = \frac{|\mathbf{S}_f|}{|\mathbf{S}_r|}$$

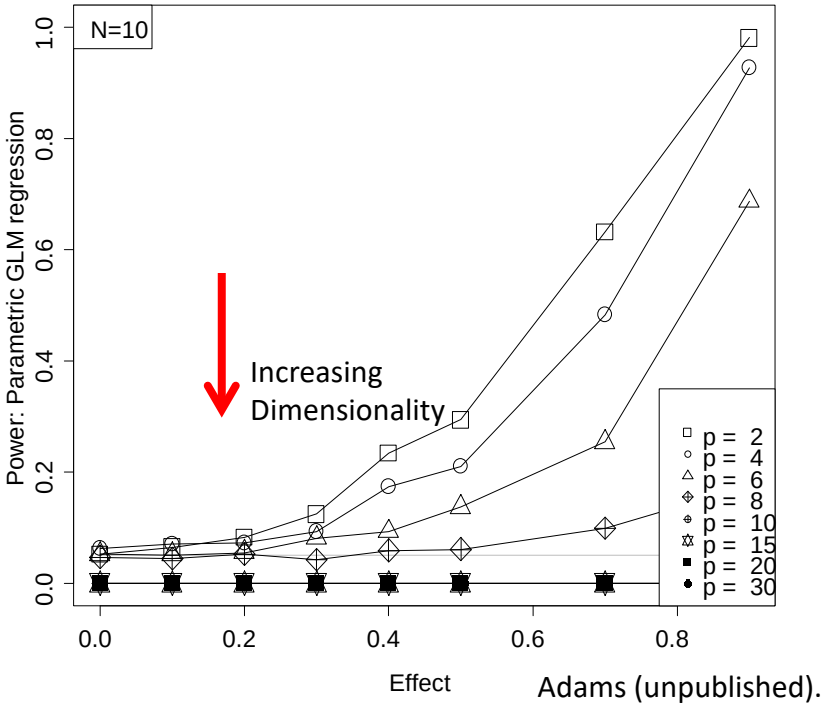
$$\Lambda_{Pilai} = \sum_{i=1}^{p}\left(\frac{\lambda_i}{1+\lambda_i}\right) = tr\left(\mathbf{S}_r^{-1}(\mathbf{S}_r - \mathbf{S}_f)\right)$$

$$\Lambda_{Hotelling-Lawley} = \sum_{i=1}^{p}\lambda_i = tr\left(\mathbf{S}_f^{-1}(\mathbf{S}_r - \mathbf{S}_f)\right)$$

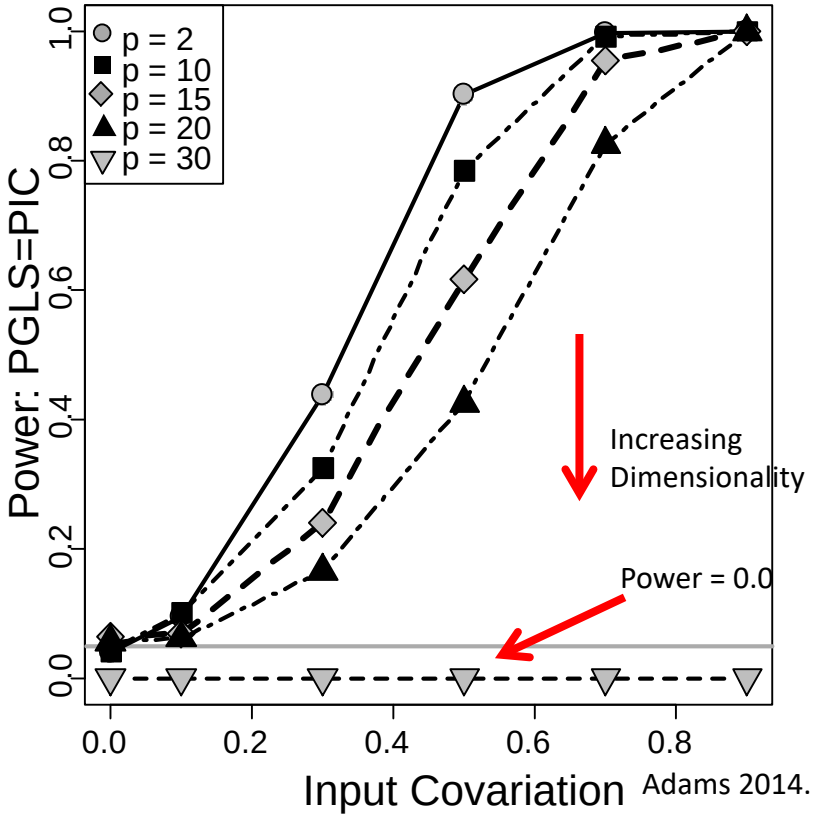
$$\Lambda_{Roy} = \max(\lambda_i)$$

These test stats can all be converted to different approximate *F* values, using different formulae. This is the approach of most "canned" MANOVA programs.

Traditional approaches suffer loss of power as P increases



Adams (unpublished).



Adams 2014. *Evol.*

Need an alternative analytical procedure

First Principles:

- If there was no covariation between **Y** and *X*, where *X* is the variable described by the second vector of **X**, then adding *X* to the **X** matrix should have little to no change in the amount of error produced by the full model.
- Thus, the trace of two SSCP matrices should be about the same

$$\mathbf{S_R} = \hat{\mathbf{E}}_\mathbf{R}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{R} \qquad \mathbf{S_F} = \hat{\mathbf{E}}_\mathbf{F}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{F}$$

- If this is true, we might agree with a null hypothesis that the data in matrix **Y** does not depend on the covariate, *X*.

The null and alternative hypotheses can be written relative to a 'model' covariance matrix

$$H_0 : tr\left(\boldsymbol{\Sigma}_M\right) = 0$$

$$H_A : tr\left(\boldsymbol{\Sigma}_M\right) > 0$$

$$\mathbf{C}_M = \hat{\boldsymbol{\Sigma}}_M = \frac{1}{k-1}\left(\mathbf{S_R} - \mathbf{S_F}\right)$$

Null Hypothesis Tests

$$H_0 : tr\left(\boldsymbol{\Sigma}_M\right) = 0$$

$$H_A : tr\left(\boldsymbol{\Sigma}_M\right) > 0$$

$$\mathbf{C}_M = \hat{\boldsymbol{\Sigma}}_M = \frac{1}{n-k}\left(\mathbf{S_R} - \mathbf{S_F}\right)$$

Note "*null*" means just an intercept. This is true, even if the reduced model has more parameters than the intercept, which could happen for hypothesis tests of multiple factor or covariate models.

Valid "Test" statistics

$$tr\left(\mathbf{S_R} - \mathbf{S_F}\right)$$

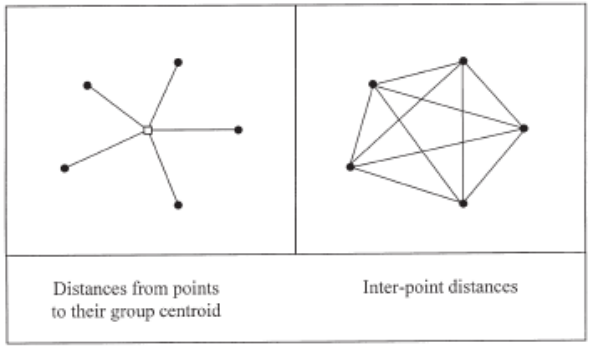
$$tr\left(\mathbf{C}_M\right)$$

$$R^2 = \frac{tr\left(\mathbf{S_R} - \mathbf{S_F}\right)}{tr\left(\mathbf{S}_{null}\right)}$$

$$F = \frac{\frac{1}{k-1} tr\left(\mathbf{S_R} - \mathbf{S_F}\right)}{\frac{1}{n-k} tr\left(\mathbf{S_F}\right)}$$

Note: none of these test statistics requires matrix inversion. Thus, using these test statistics with some other procedure, such as resampling, avoids the pitfalls shown earlier for high-dimensional data.

Evaluate SS using permutation



1. Estimate coefficients $\hat{\mathbf{B}}_R = \left(\mathbf{X}_R^T\mathbf{X}_R\right)^{-1}\left(\mathbf{X}_R^T\mathbf{Y}\right)$
2. Calculate predicted values $\hat{\mathbf{Y}}_R = \mathbf{X}_R\hat{\mathbf{B}}_R$
3. Obtain SSCP $\hat{\mathbf{E}}_R = \mathbf{Y} - \hat{\mathbf{Y}}_R$

   $\mathbf{S}_R = \hat{\mathbf{E}}_R^T\hat{\mathbf{E}}_R$
4. Observed test statistic 'TS_{obs}'

$$tr\left(\mathbf{S}_R - \mathbf{S}_F\right)$$

$$tr\left(\mathbf{C}_M\right)$$

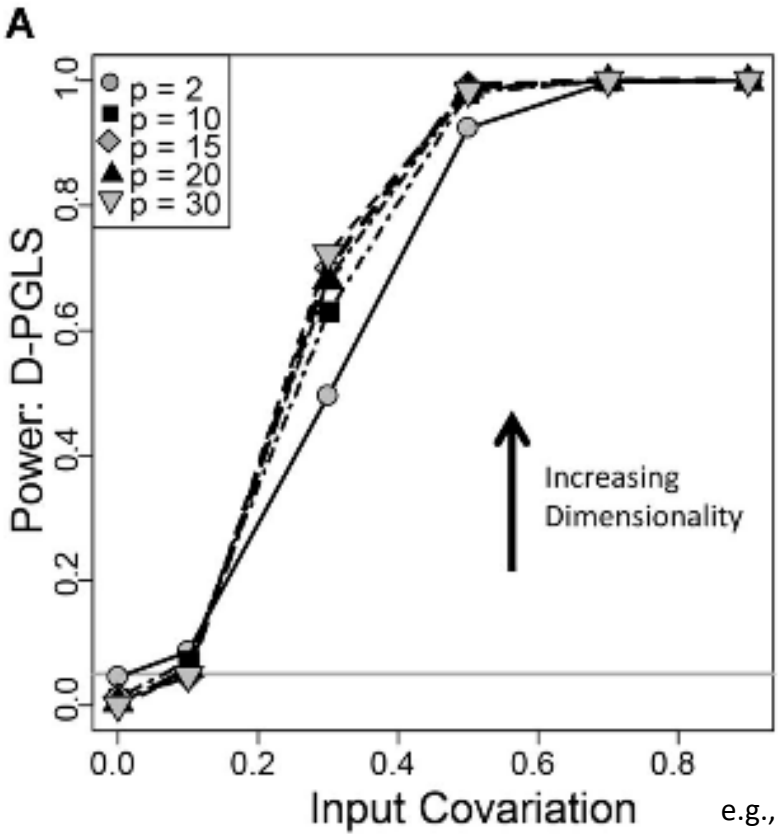
$$R^2 = \frac{tr\left(\mathbf{S}_R - \mathbf{S}_F\right)}{tr\left(\mathbf{S}_{mdl}\right)}$$

$$F = \frac{\frac{1}{k-1}tr\left(\mathbf{S}_R - \mathbf{S}_F\right)}{\frac{1}{n-k}tr\left(\mathbf{S}_F\right)}$$

5. Shuffle data; estimate TS_{rand} and compare to TS_{obs}
6. Repeat

Doesn't require inverting covariance matrix, so general solution

Permutation-GLM has increased power as P increases



e.g., Adams 2014. *Evol.*

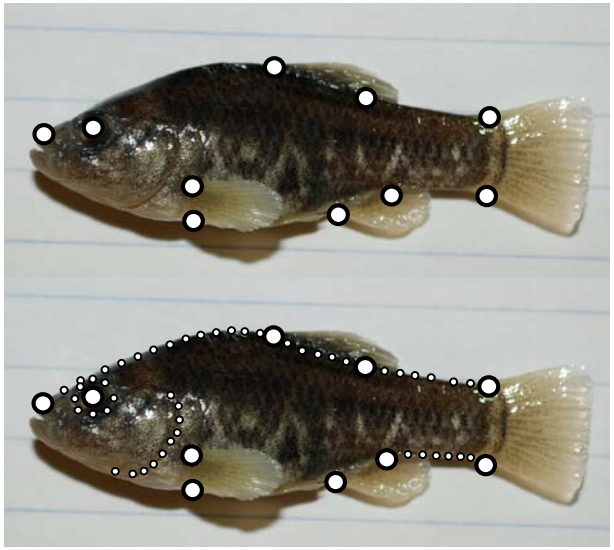
Reason: no additional parameters estimated, so as variables are added, if any variation in those variables associates with the Y~X pattern, it can more easily detect it.

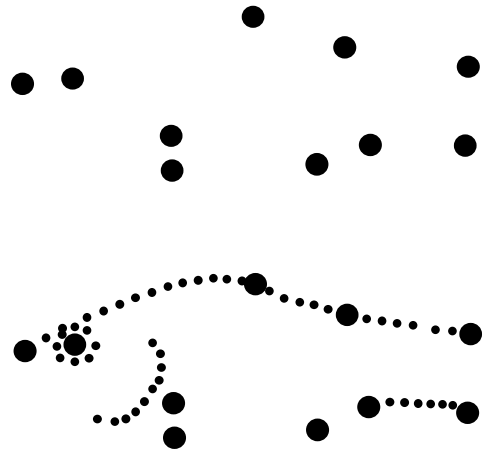
Example

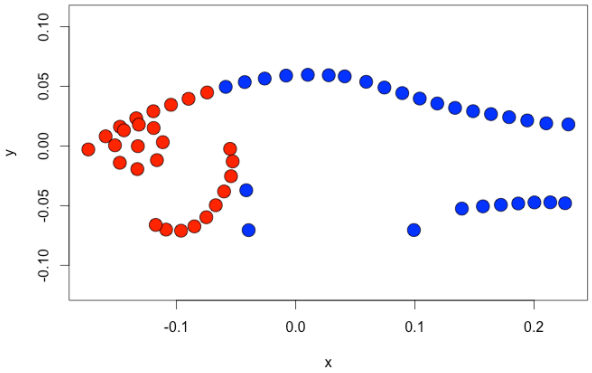
Pecos pupfish
(*Cyprinodon pecosensis*)

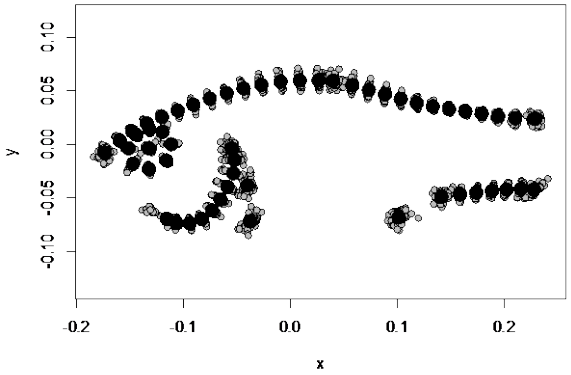
- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from fit of 56 landmarks (using minimized squared Procrustes distance method)
- 27 head landmarks
- 29 body landmarks

- NOTE: N = 54, *p* = 112!



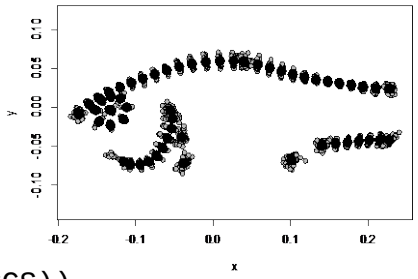






Collyer, Sekora and Adams. 2015. *Heredity*

Hypothesis test on a linear regression slope



Pecos pupfish
(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from fit of 56 landmarks (using minimized squared Procrustes distance method)
- 112 Procrustes residuals (56 landmarks)
- Log(CS)
- I.e., allometry

```
> procD.lm(pupfish$coords ~ log(pupfish$CS))

Type I (Sequential) Sums of Squares and Cross-products

Randomization of Raw Values used

                df       SS         MS     Rsq      F     Z P.value
log(pupfish$CS)  1 0.014019 0.0140193 0.24887 17.229 9.972   0.001
Residuals       52 0.042314 0.0008137
Total           53 0.056333
```

$$tr(\mathbf{S}_R - \mathbf{S}_F)$$

This is the actual test statistic

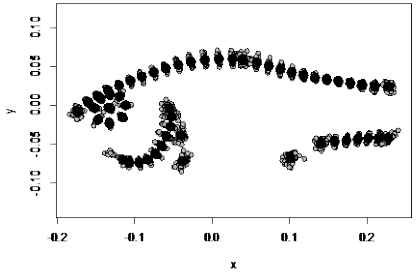
$$F = \frac{\frac{1}{k-1} tr(\mathbf{S}_R - \mathbf{S}_F)}{\frac{1}{n-k} tr(\mathbf{S}_F)}$$

This is a transformed test statistic

The standard deviation and P-value in the sampling distribution of SS under the null hypothesis. This is the part that speaks of the relevance of the observed test stat

Collyer, Sekora and Adams. 2015. *Heredity*

Hypothesis test on a linear regression slope

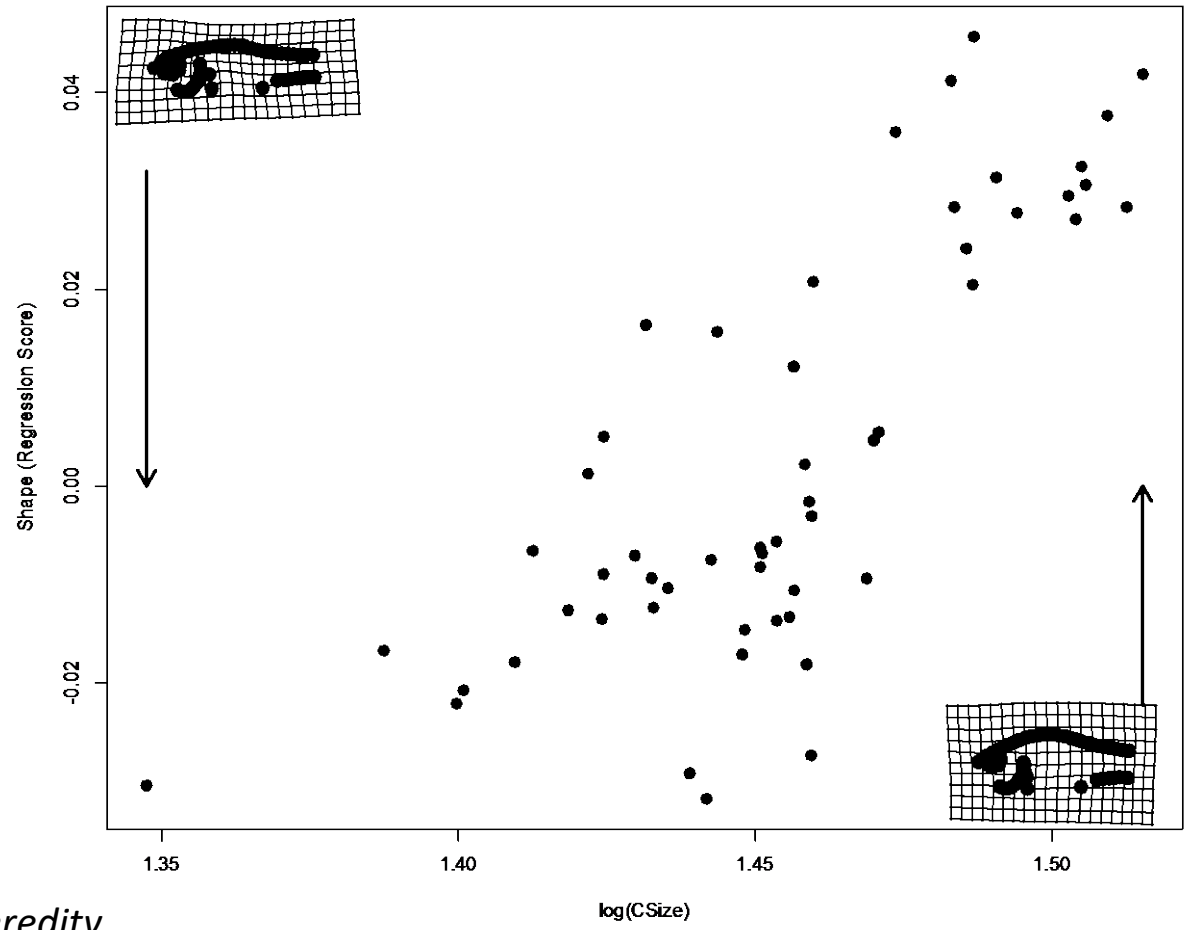


Explanation: Fish shape changes as they get larger

Pecos pupfish

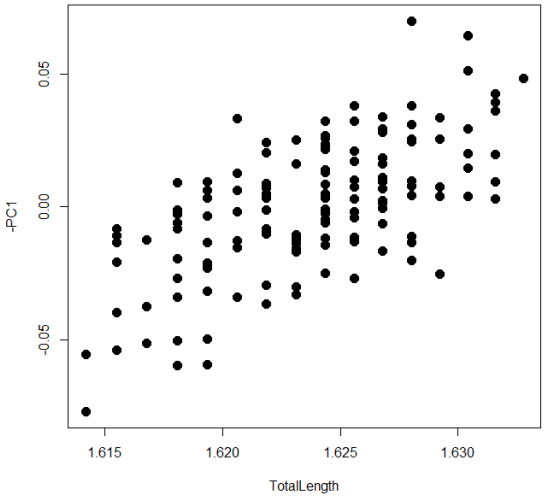
(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from fit of 56 landmarks (using minimized squared Procrustes distance method)
- 112 Procrustes residuals (56 landmarks)
- Log(CS)
- I.e., allometry



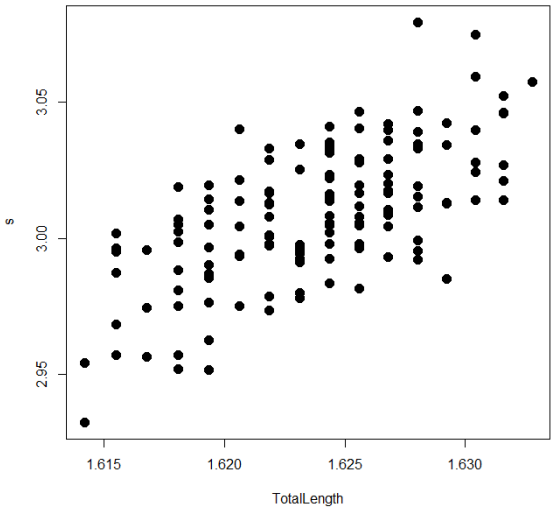
Collyer, Sekora and Adams. 2015. *Heredity*

Visualizing is challenging because **Y** is multivariate
Represent **Y** by some summary axis



PC1 vs. X

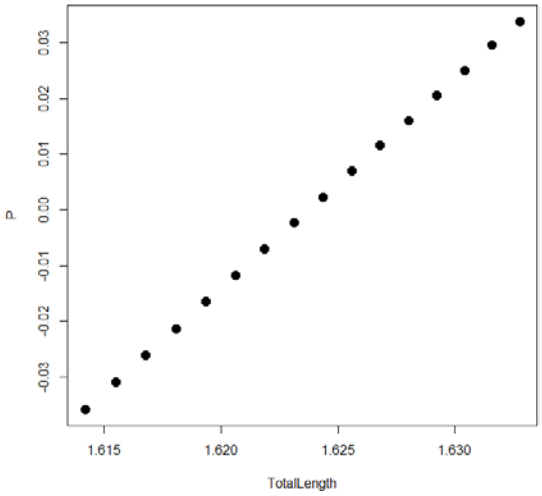
PC1 may not align with direction of covariation



Regression Score vs. X

$$s = \mathbf{Y}\boldsymbol{\beta}^t\left(\boldsymbol{\beta}\boldsymbol{\beta}^t\right)^{-.5}$$

Drake and Klingenberg (2008)
Evolution



Predicted Values vs. X

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

$$P_1 = SVD\left(\hat{\mathbf{Y}}\right)$$

Useful for 2+ groups

Adams and Nistri (2010)
BMC Evol Biol

What about group differences?

-Same procedure algebraically,

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times k\cdot k\times p}{\mathbf{XB}} + \underset{n\times p}{\mathbf{E}}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}^{\mathbf{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{Y}$$

only **X** contains dummy variables designating groups

$$\mathbf{X}_{\mathbf{F}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}$$

Group 1

Group 2

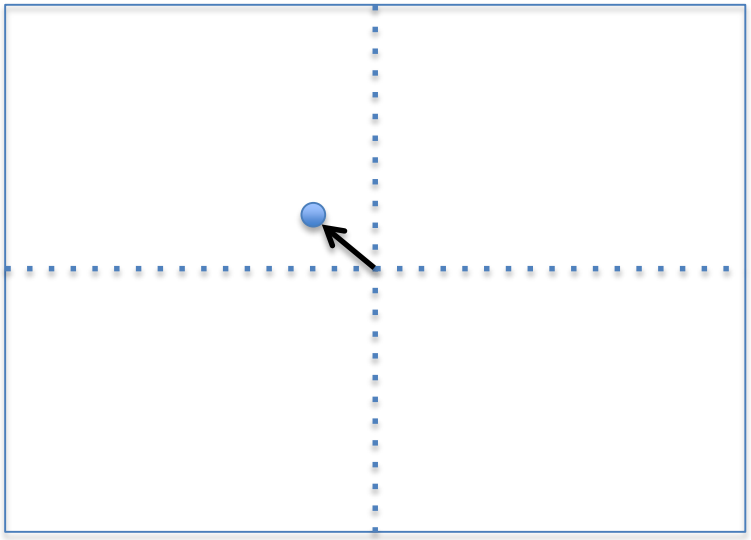
Group 3

Grand mean

Groups

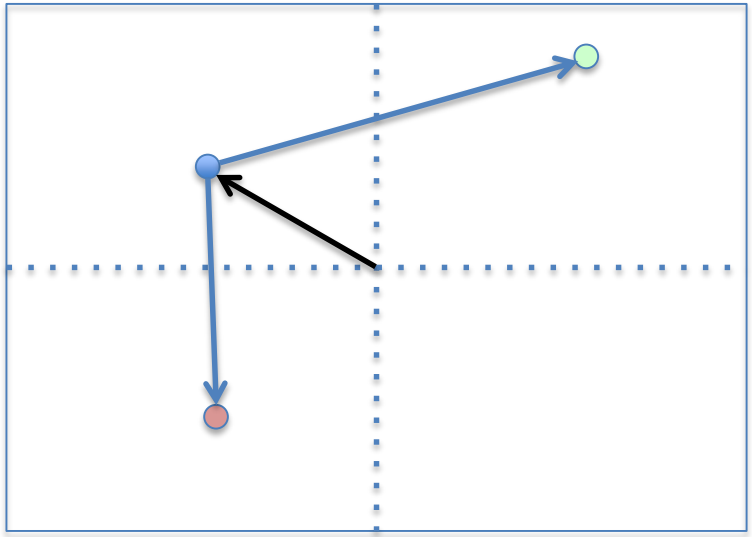
When there is only a column of ones, the vector of coefficients is the mean vector

$$\mathbf{X}_R = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \hat{\mathbf{B}} = \left(\mathbf{X}_R^T \mathbf{X}_R\right)^{-1}\left(\mathbf{X}_R^T \mathbf{Y}\right) = \bar{\mathbf{y}}^T$$



When there are additional columns of dummy variables, the first row vector of coefficients is the group 1 mean; the rest are change vectors to describe other group means.

$$\mathbf{X}_F = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \qquad \hat{\mathbf{B}} = \left(\mathbf{X}_F^T \mathbf{X}_F\right)^{-1}\left(\mathbf{X}_F^T \mathbf{Y}\right) = \begin{bmatrix} \bar{\mathbf{y}}_1^T \\ \Delta\bar{\mathbf{y}}_{12}^T \\ \Delta\bar{\mathbf{y}}_{13}^T \end{bmatrix}$$

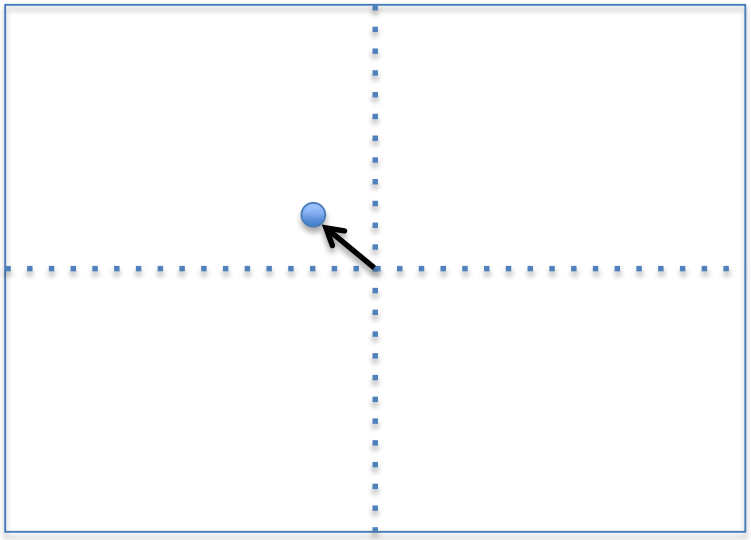


Predicted values are group means, whether there is one group or more.

$$\mathbf{X}_R = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}_R^T\mathbf{X}_R\right)^{-1}\left(\mathbf{X}_R^T\mathbf{Y}\right) = \bar{\mathbf{y}}^T$$

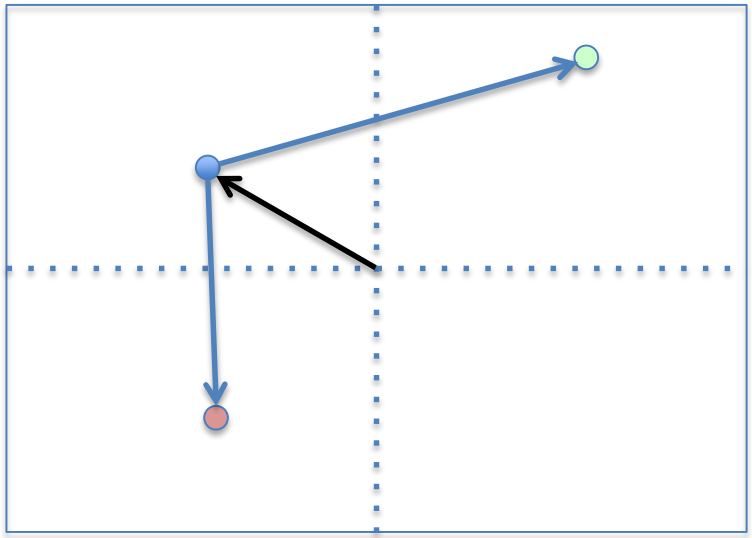
$$\mathbf{X}_R\hat{\mathbf{B}}_R = \begin{bmatrix} \bar{\mathbf{y}}^T \\ \bar{\mathbf{y}}^T \\ \vdots \\ \bar{\mathbf{y}}^T \end{bmatrix}$$



$$\mathbf{X}_F = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}_F^T\mathbf{X}_F\right)^{-1}\left(\mathbf{X}_F^T\mathbf{Y}\right) = \begin{bmatrix} \bar{\mathbf{y}}_1^T \\ \Delta\bar{\mathbf{y}}_{12}^T \\ \Delta\bar{\mathbf{y}}_{13}^T \end{bmatrix}$$

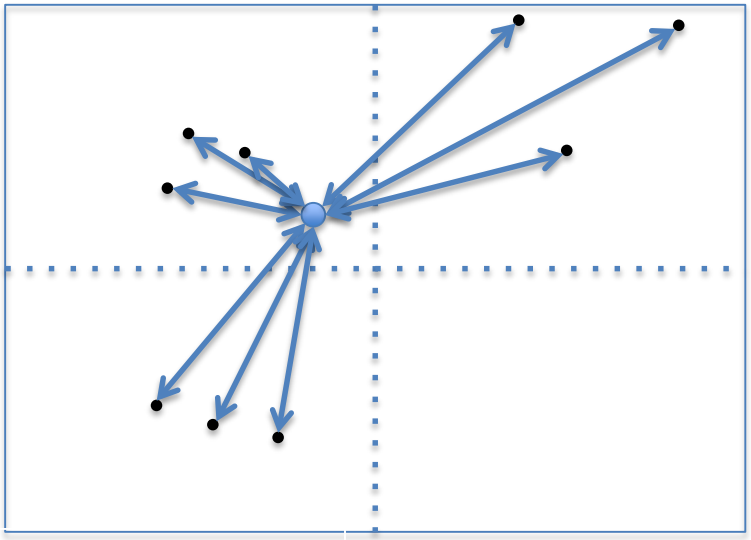
$$\mathbf{X}_F\hat{\mathbf{B}}_F = \begin{bmatrix} \bar{\mathbf{y}}_1^T \\ \vdots \\ \bar{\mathbf{y}}_2^T \\ \vdots \\ \bar{\mathbf{y}}_3^T \\ \vdots \end{bmatrix}$$



Residuals describe within-group dispersion.
Diagonal of outer-product of error represents distances from mean

$$\hat{\mathbf{E}}_\mathbf{R} = \mathbf{Y} - \mathbf{X}_\mathbf{R}\hat{\mathbf{B}}_\mathbf{R} \qquad d_i^2 = diag\left(\hat{\mathbf{E}}_\mathbf{R}\hat{\mathbf{E}}_\mathbf{R}^\mathbf{T}\right)$$

$$SS_{W_R} = tr\left(\hat{\mathbf{S}}_\mathbf{R}\right) = tr\left(\hat{\mathbf{E}}_\mathbf{R}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{R}\right) = tr\left(\hat{\mathbf{E}}_\mathbf{R}\hat{\mathbf{E}}_\mathbf{R}^\mathbf{T}\right) = \sum d_i^2$$

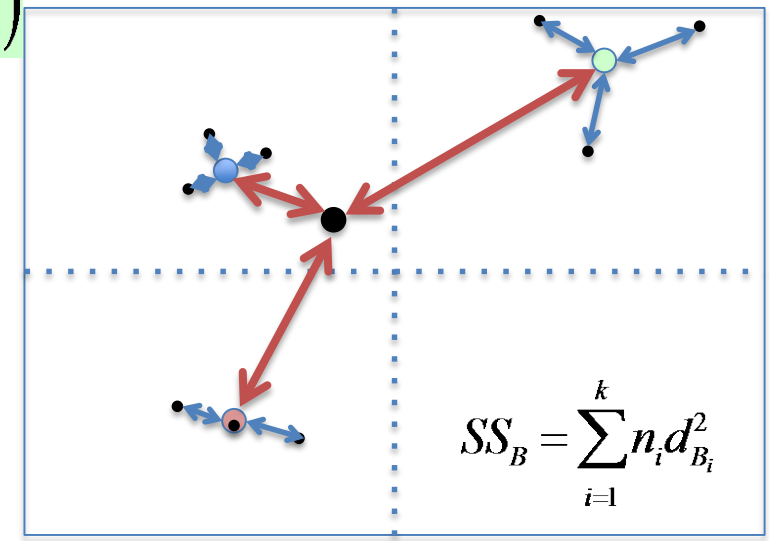


Important! SS_W identical from Σcovariances or Σdistances! (remember Gower, 1966)

$$SS_B = tr\left(\hat{\mathbf{S}}_\mathbf{R}\right) - tr\left(\hat{\mathbf{S}}_\mathbf{F}\right) = tr\left(\hat{\mathbf{S}}_\mathbf{R} - \hat{\mathbf{S}}_\mathbf{F}\right) = \Delta\left(\sum d_i^2\right)$$

$$\hat{\mathbf{E}}_\mathbf{F} = \mathbf{Y} - \mathbf{X}_\mathbf{F}\hat{\mathbf{B}}_\mathbf{F} \qquad d_i^2 = diag\left(\hat{\mathbf{E}}_\mathbf{F}\hat{\mathbf{E}}_\mathbf{F}^\mathbf{T}\right)$$

$$SS_{W_F} = tr\left(\hat{\mathbf{S}}_\mathbf{F}\right) = tr\left(\hat{\mathbf{E}}_\mathbf{F}^\mathbf{T}\hat{\mathbf{E}}_\mathbf{F}\right) = tr\left(\hat{\mathbf{E}}_\mathbf{F}\hat{\mathbf{E}}_\mathbf{F}^\mathbf{T}\right) = \sum d_i^2$$



Anderson MJ (2001). *Austral Ecology* 26: 32-46

As before:

Null Hypothesis Tests

$$H_0 : tr\left(\boldsymbol{\Sigma}_M\right) = 0$$

$$H_A : tr\left(\boldsymbol{\Sigma}_M\right) > 0$$

$$\mathbf{C}_M = \hat{\boldsymbol{\Sigma}}_M = \frac{1}{n-k}\left(\mathbf{S_R} - \mathbf{S_F}\right)$$

'Traditional' Test Statistics (Λ_{Wilks}, Λ_{Pilai}, etc.) (sensitive to small N: large *p*)

'Robust' Test Statistics

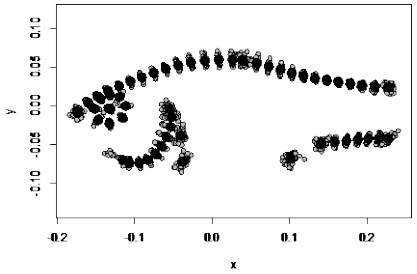
$$tr\left(\mathbf{S_R} - \mathbf{S_F}\right)$$

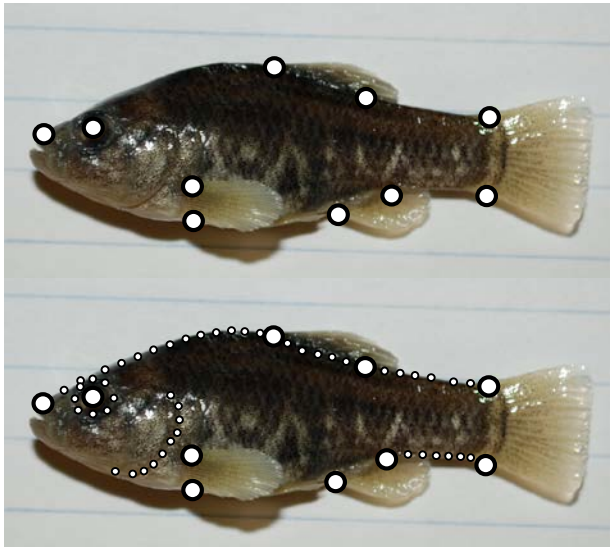
$$tr\left(\mathbf{C}_M\right)$$

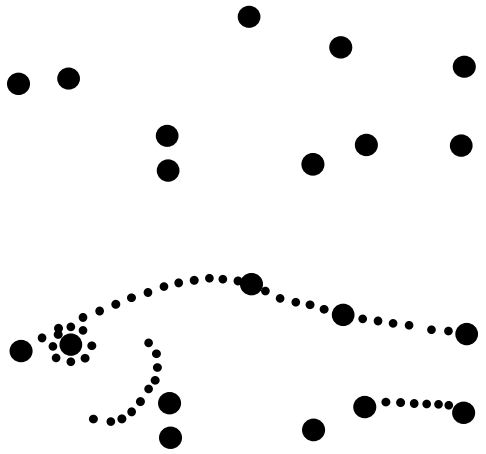
$$R^2 = \frac{tr\left(\mathbf{S_R} - \mathbf{S_F}\right)}{tr\left(\mathbf{S}_{null}\right)}$$

$$F = \frac{\frac{1}{k-1} tr\left(\mathbf{S_R} - \mathbf{S_F}\right)}{\frac{1}{n-k} tr\left(\mathbf{S_F}\right)}$$

Note: none of these test statistics requires matrix inversion. Thus, using these test statistics with some other procedure, such as resampling, avoids the pitfalls shown earlier for high-dimensional data.





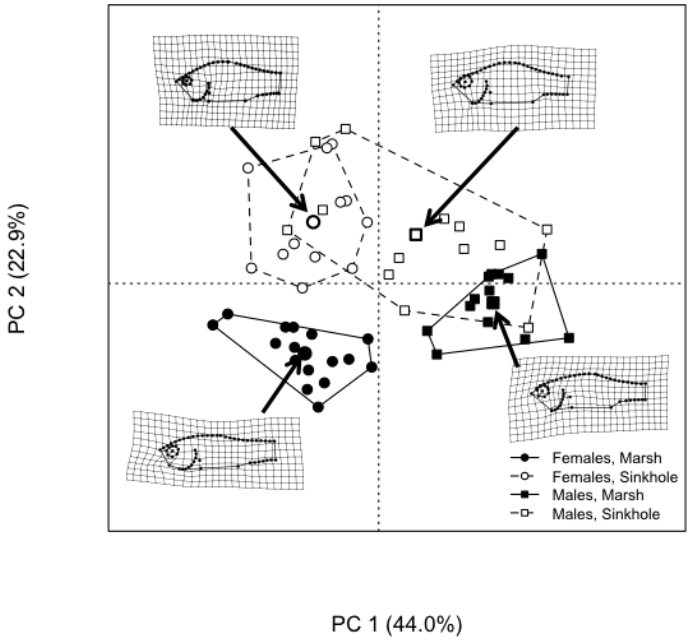


Example

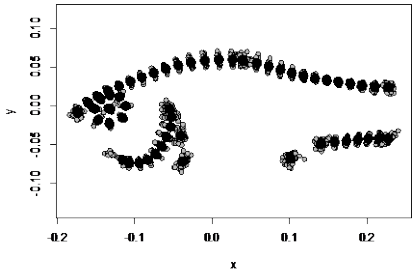
Pecos pupfish

(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from a geometric morphometric analysis
- 112 Procrustes residuals
- Groups defined by population and sex (Marsh-females [16], Marsh-males [13], Sinkhole-females [12], Sinkhole-males [13]



Collyer, Sekora and Adams. 2015. *Heredity*



Example

Pecos pupfish

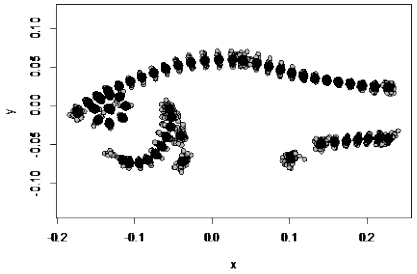
(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from a geometric morphometric analysis
- 112 Procrustes residuals
- Groups defined by population and sex (Marsh-females [16], Marsh-males [13], Sinkhole-females [12], Sinkhole-males [13]

```
> fit.group = lm(Y~group, x=T)
> fit.group$x
     (Intercept) groupF.Sinkhole groupM.Marsh groupM.Sinkhole
229            1                0            0                0
230            1                0            0                0
231            1                0            0                0
232            1                0            0                0
278            1                0            1                0
279            1                0            1                0
280            1                0            1                0
281            1                0            1                0
1013           1                1            0                0
1014           1                1            0                0
1015           1                1            0                0
1016           1                1            0                0
1017           1                1            0                0
1018           1                1            0                0
1019           1                1            0                0
1020           1                1            0                0
1021           1                1            0                0
1022           1                1            0                0
1023           1                1            0                0
1024           1                1            0                0
1071           1                0            0                1
1073           1                0            0                1
1074           1                0            0                1
1075           1                0            0                1
1076           1                0            0                1
1077           1                0            0                1
1078           1                0            0                1
1079           1                0            0                1
1080           1                0            0                1
1081           1                0            0                1
1082           1                0            0                1
1083           1                0            0                1
1084           1                0            0                1
```

Not all rows are shown

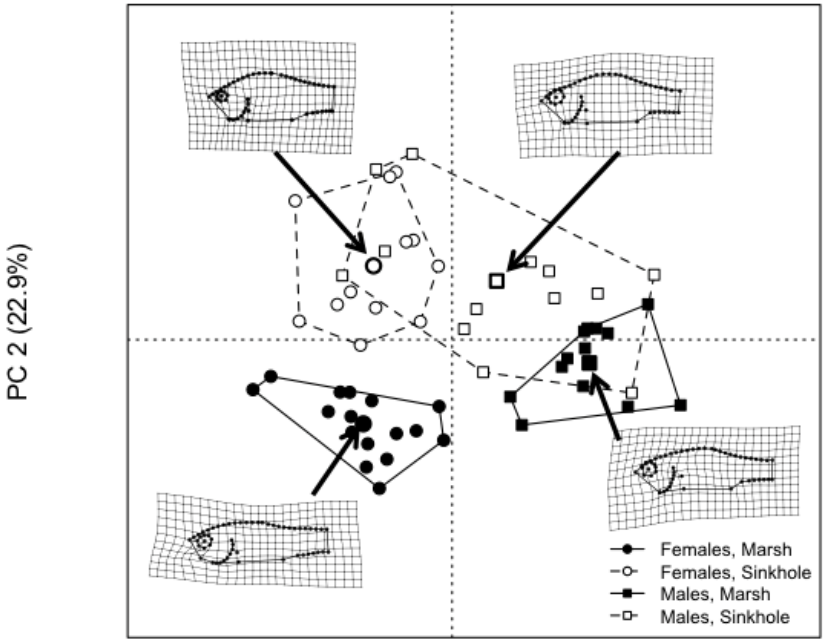
Collyer, Sekora and Adams. 2015. *Heredity*



Example
Pecos pupfish
(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from a geometric morphometric analysis
- 112 Procrustes residuals
- Groups defined by population and sex (Marsh-females [16], Marsh-males [13], Sinkhole-females [12], Sinkhole-males [13]

```
> group = as.factor(paste(pupfish$Pop, pupfish$Sex, sep="."))
> procD.lm(pupfish$coords ~ group)

Type I (Sequential) Sums of Squares and Cross-products

Randomization of Raw Values used

           df       SS         MS     Rsq      F      Z P.value
group       3 0.028363 0.0094543 0.50349 16.901 7.9771   0.001
Residuals  50 0.027970 0.0005594
Total      53 0.056333
```

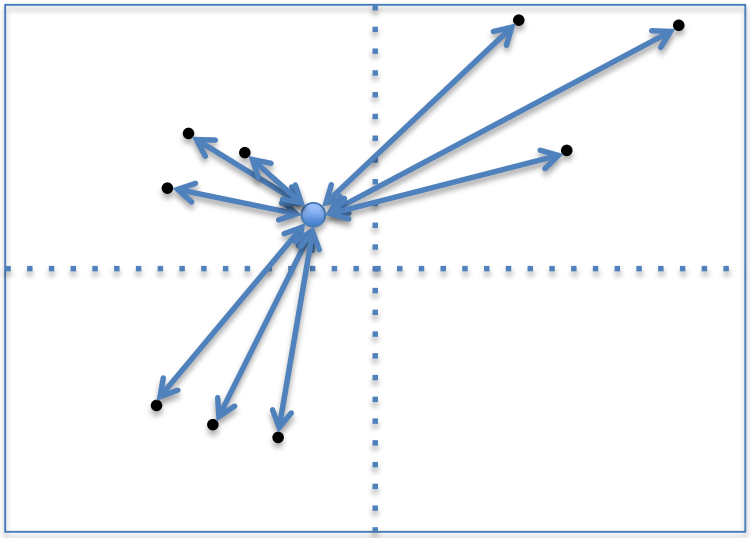


Collyer, Sekora and Adams. 2015. *Heredity*

The null hypothesis for a pairwise comparison of means can be stated as

$$H_0 : (\mu_1 - \mu_2)^{\mathrm{T}} (\mu_1 - \mu_2) = 0$$

$$H_A : (\mu_1 - \mu_2)^{\mathrm{T}} (\mu_1 - \mu_2) > 0$$



Viable test statistics

$$d_{12}^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^{\mathrm{T}} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

$$d_{12} = \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^{\mathrm{T}} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}$$

Traditional test statistics (involves a couple of transformations)

$$d_M^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^{\mathrm{T}} \mathbf{C}_{\mathbf{M}}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

$$T^2 = \frac{n_1 n_2}{\sum n_i - k} d_M^2$$

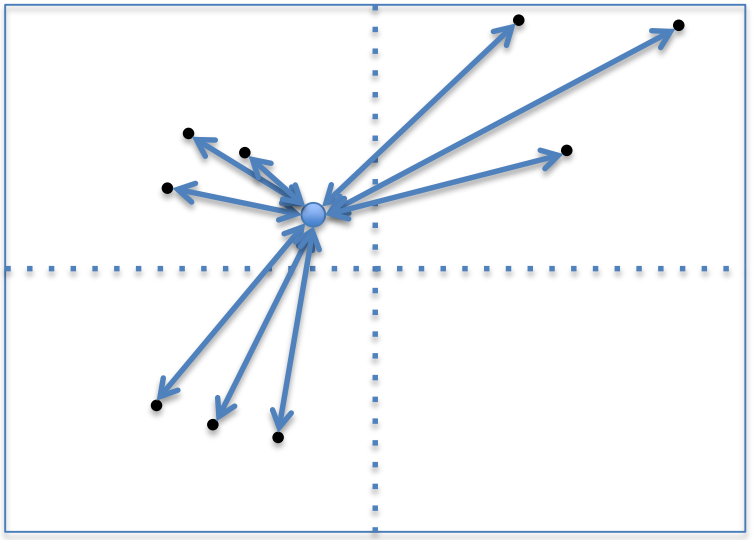
$$F_{12} = \frac{n_1 + n_2 - p - 1}{\left(\sum n_i - k\right) p} T^2 \sim F(p, n_1 + n_2 - p - 1)$$

These steps are calculation of squared Mahalanobis distances, conversion to Hotelling T^2 statistics, and subsequent conversion to *F*-values. It should be obvious that if the number of variables exceeds the number of observations, this approach will not work. This is a good illustration of the limitation of parametric statistical approaches.

The null hypothesis for a pairwise comparison of means can be stated as

$$H_0 : (\mu_1 - \mu_2)^{\mathrm{T}} (\mu_1 - \mu_2) = 0$$

$$H_A : (\mu_1 - \mu_2)^{\mathrm{T}} (\mu_1 - \mu_2) > 0$$

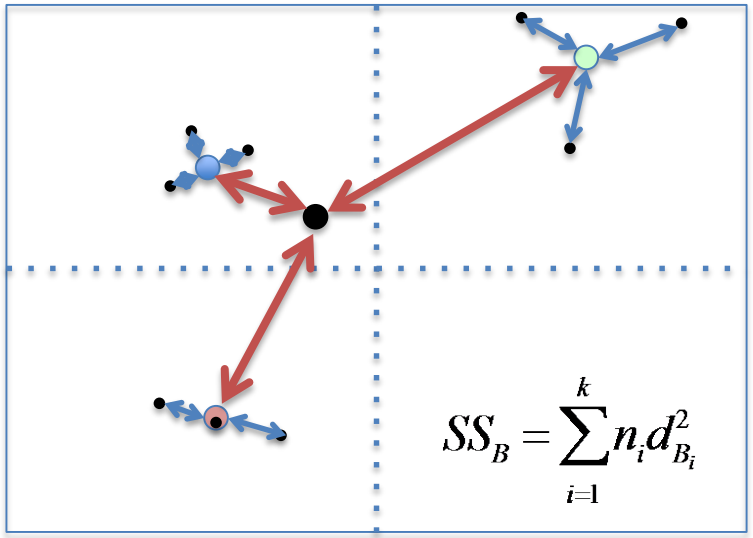


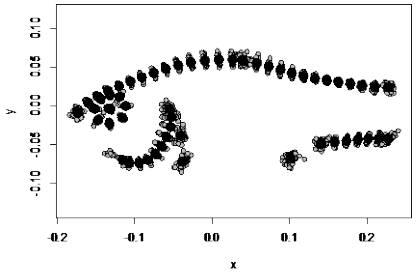
Viable test statistics

$$d_{12}^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^{\mathrm{T}} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

$$d_{12} = \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^{\mathrm{T}} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}$$

Use matrix of d_{ab} as test measures, and the resampling scheme above for testing pairwise comparisons





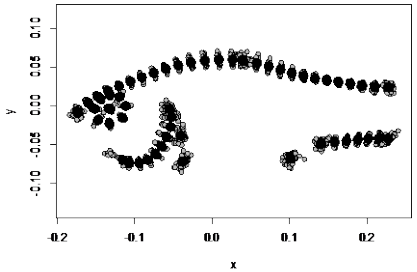
Example
Pecos pupfish
(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from a geometric morphometric analysis
- 112 Procrustes residuals
- Groups defined by population and sex (Marsh-females [16], Marsh-males [13], Sinkhole-females [12], Sinkhole-males [13]

```
> group = as.factor(paste(pupfish$Pop, pupfish$Sex, sep="."))
> procD.lm(pupfish$coords ~ group)

Type I (Sequential) Sums of Squares and Cross-products

Randomization of Raw Values used

          df       SS        MS     Rsq      F      Z P.value
group      3 0.028363 0.0094543 0.50349 16.901 7.9173   0.001
Residuals 50 0.027970 0.0005594
Total     53 0.056333
```

```
> advanced.procD.lm(pupfish$coords~pupfish$Pop*pupfish$Sex,
+                   pupfish$coords~1,
+                   groups = ~pupfish$Pop*pupfish$Sex,
+                   iter = 999)
$anova.table

ANOVA with RRPP

                              df      SSE       SS      F      Z     P
~1                            53 0.056333
~pupfish$Pop * pupfish$Sex    50 0.027970 0.028363 16.901 7.9212 0.001

$Means.dist
              Marsh:F    Marsh:M Sinkhole:F Sinkhole:M
Marsh:F    0.00000000 0.04611590 0.03302552 0.03881514
Marsh:M    0.04611590 0.00000000 0.04605211 0.02802087
Sinkhole:F 0.03302552 0.04605211 0.00000000 0.02568508
Sinkhole:M 0.03881514 0.02802087 0.02568508 0.00000000

$Prob.Means.dist
           Marsh:F Marsh:M Sinkhole:F Sinkhole:M
Marsh:F      1.000   0.001      0.001      0.001
Marsh:M      0.001   1.000      0.001      0.001
Sinkhole:F   0.001   0.001      1.000      0.003
Sinkhole:M   0.001   0.001      0.003      1.000
```

Collyer, Sekora and Adams. 2015. *Heredity*

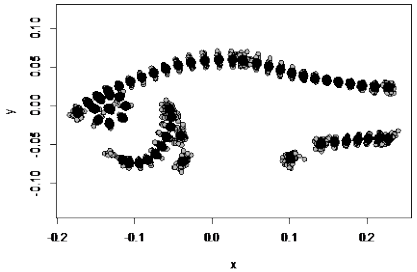


-The pupfish are from two populations and contain 2 sexes.

-Design is: Y~ Pop + Sex + Pop:Sex

-MANOVA works identically here with factors

Collyer, Sekora and Adams. 2015. *Heredity*



Example
Pecos pupfish
(*Cyprinodon pecosensis*)

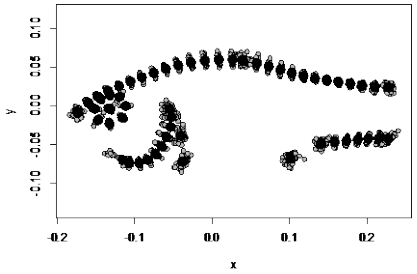
```
> procD.lm(pupfish$coords~pupfish$Pop*pupfish$Sex, RRPP = TRUE)

Type I (Sequential) Sums of Squares and Cross-products

Randomized Residual Permutation Procedure used

                          df       SS        MS      Rsq       F        Z P.value
pupfish$Pop                1 0.008993 0.0089927 0.159635 16.076   6.7169   0.001
pupfish$Sex                1 0.015917 0.0159169 0.282551 28.453  13.0513   0.001
pupfish$Pop:pupfish$Sex    1 0.003453 0.0034532 0.061299  6.173   4.9849   0.001
Residuals                 50 0.027970 0.0005594
Total                     53 0.056333
```

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from fit of 56 landmarks (using minimized squared Procrustes distance method)
- 112 Procrustes residuals
- Factorial approach: Population, Sex, Population × Sex

```
> group = as.factor(paste(pupfish$Pop, pupfish$Sex, sep="."))
> procD.lm(pupfish$coords ~ group)

Type I (Sequential) Sums of Squares and Cross-products

Randomization of Raw Values used

          df       SS        MS     Rsq      F      Z P.value
group      3 0.028363 0.0094543 0.50349 16.901 7.9173   0.001
Residuals 50 0.027970 0.0005594
Total     53 0.056333
```

Note that the factorial model produces the same error as the single-factor model, but allows the error to be partitioned into smaller units.

Collyer, Sekora and Adams. 2015. *Heredity*



Example

Pecos pupfish

(*Cyprinodon pecosensis*)

- 54 specimens
- 10 fixed landmarks
- 46 sliding semilandmarks
- Procrustes residuals from fit of 56 landmarks (using minimized squared Procrustes distance method)
- 112 Procrustes residuals
- Factorial approach: Population, Sex, Population × Sex

```
> advanced.procD.lm(pupfish$coords~pupfish$Pop*pupfish$Sex,
+                   pupfish$coords~pupfish$Pop+pupfish$Sex,
+                   groups = ~pupfish$Pop*pupfish$Sex,
+                   iter = 999)
$anova.table

ANOVA with RRPP

                              df      SSE        SS     F      Z     P
~pupfish$Pop + pupfish$Sex    51 0.031423
~pupfish$Pop * pupfish$Sex    50 0.027970 0.0034532 6.173 4.9874 0.001

$Means.dist
              Marsh:F    Marsh:M Sinkhole:F Sinkhole:M
Marsh:F    0.00000000 0.04611590 0.03302552 0.03881514
Marsh:M    0.04611590 0.00000000 0.04605211 0.02802087
Sinkhole:F 0.03302552 0.04605211 0.00000000 0.02568508
Sinkhole:M 0.03881514 0.02802087 0.02568508 0.00000000

$Prob.Means.dist
           Marsh:F Marsh:M Sinkhole:F Sinkhole:M
Marsh:F      1.000   0.007      0.051      0.718
Marsh:M      0.007   1.000      0.551      0.424
Sinkhole:F   0.051   0.551      1.000      0.987
Sinkhole:M   0.718   0.424      0.987      1.000
```

RRPP: More appropriate permutation that considers significance of interactive effects, given main effects!

```
           Marsh.F Marsh.M Sinkhole.F Sinkhole.M
Marsh:F      1.000   0.001      0.001      0.001
Marsh:M      0.001   1.000      0.001      0.001
Sinkhole:F   0.001   0.001      1.000      0.003
Sinkhole:M   0.001   0.001      0.003      1.000
```

Collyer, Sekora and Adams. 2015. *Heredity*

- Common MANOVA functions include `procD.lm` {geomorph} and `adonis` {vegan}. One can also use `manylm` {mvabund}, followed by `anova` {base}.

- Currently, geomorph is the only published package that uses RRPP. All of the functions above allow randomization of raw values.

- Using `manova` requires having $p > n$ and implicit assumptions met.

- Relates variation in shape to variation in covariate

- X = independent variables (groups & continuous)
- Y = dependent variables
- Solve for b (components of means)

$$\hat{\mathbf{B}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

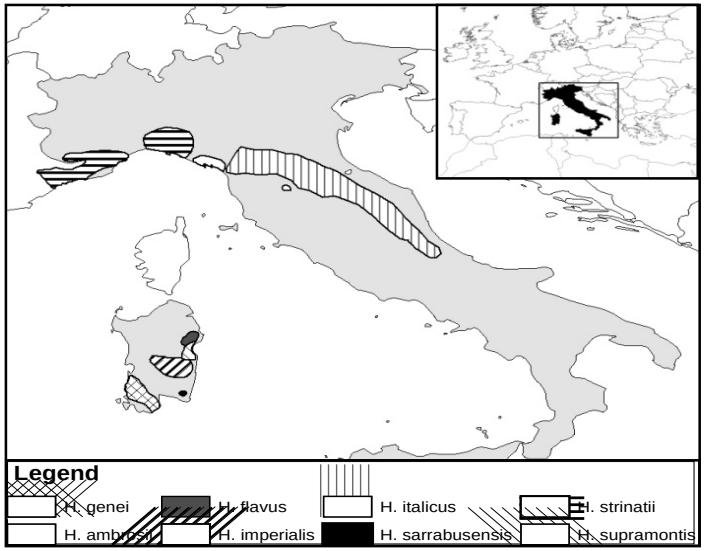
$$\mathbf{SSCP}_{err} = \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)^t\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)$$

- NOTE: MANCOVA is sequential procedure
- Test interactions first (group-specific slopes)
- If NS, remove and compare groups (while accounting for covariate)

**** Implementation point: covariate must be first variable in X-matrix, as R uses Type I SS for H**

Italian *Hydromantes* inhabit caves

Climb walls & ceilings (strong ecological selection)



H_o: Adult foot morphology adapted for climbing (e.g,. Lanza, 1991)

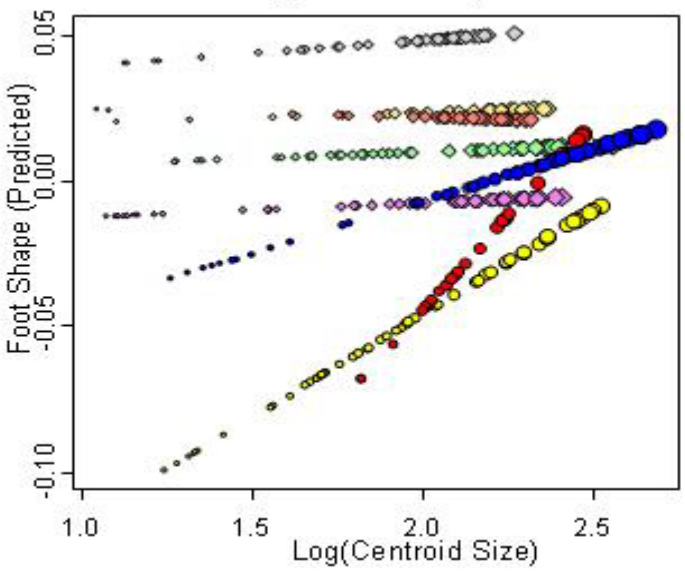
a) never tested empirically, b) ignores developmental influences

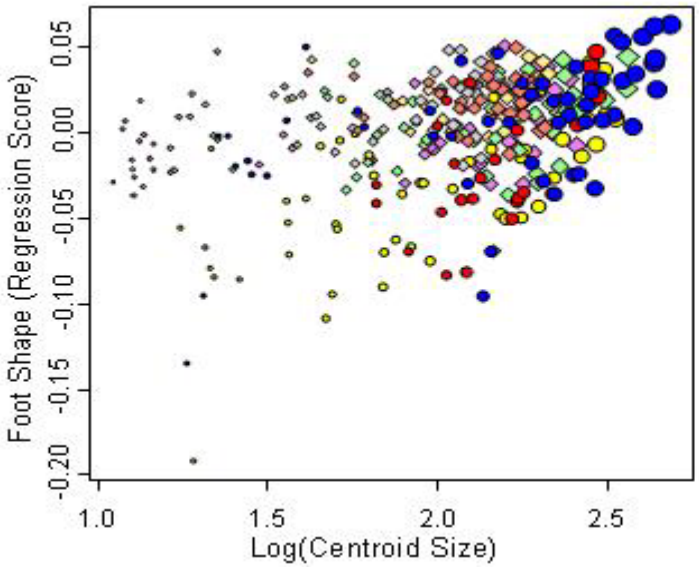
Is there evidence for this hypothesis?

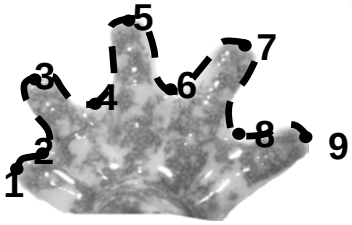
Adams & Nistri (2010) *BMC Evol. Biol.*

Significant foot shape allometry (and convergence)

```
> procD.lm(Y~Csize*Species,RRPP=TRUE)

Type I (Sequential) Sums of Squares and Cross-products
 Randomized Residual Permutation Procedure used

                 df      SS       MS      Rsq      F      Z P.value
Csize             1 0.05542 0.055417 0.027409 11.484 7.2652   0.001
Species           7 0.38175 0.054536 0.188814 11.302 8.6439   0.001
Csize:Species     7 0.06948 0.009926 0.034366  2.057 1.9773   0.002
Residuals       314 1.51519 0.004825
Total           329 2.02184
```







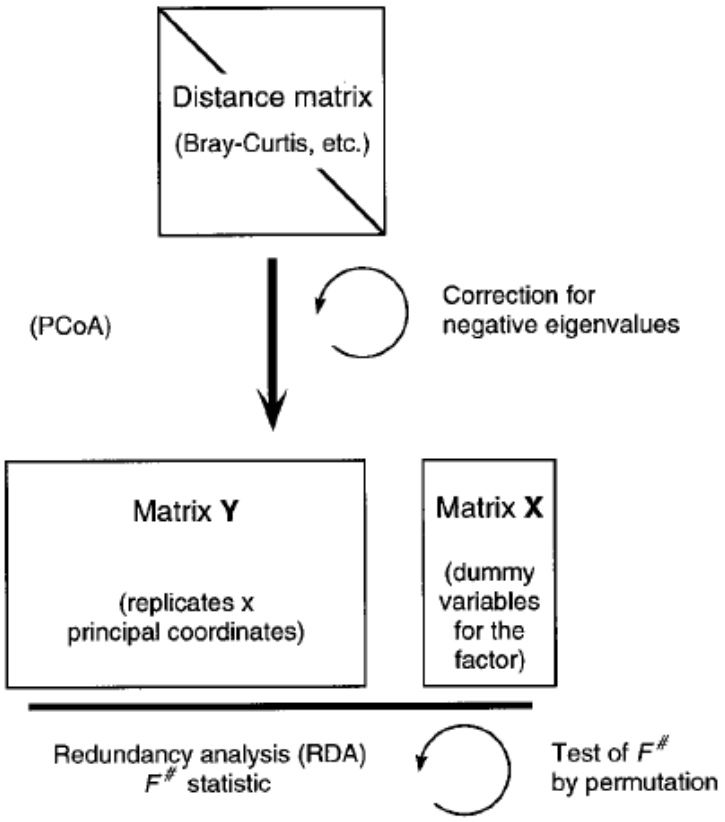
Adams & Nistri (2010) *BMC Evol. Biol.*

General linear models require normally-distributed data: multivariate, continuous
Such data follow the Euclidean metric

For non-Euclidean data:

1: Use permutational-MANOVA with appropriate distance measure (e.g., Bray-Curtis for species abundances)

2: Peform PCoA on data (based on appropriate distance), obtain PCo Scores, and treat these as input for GLM



NOTE: also a way to combine different data types:
(concatenate PCoA columns from separate distance matrices)

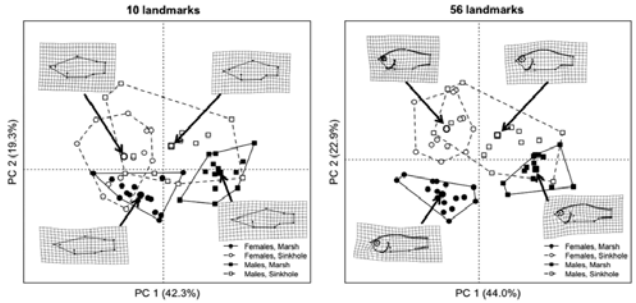
See Legendre and Anderson, 1999.
Ecol. Monogr. 64:1-14

-GLM provides flexible tools for evaluating:

- -Regression/covariation
- -Group differences
- -Both

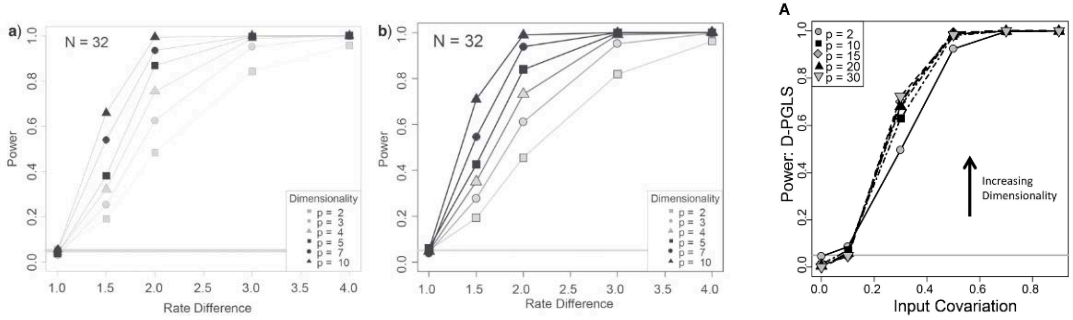
-RRPP has better statistical properties than permuting observed data (*sensu* Anderson and Legendre 1999; Anderson 2001)

-More variables can equal greater statistical power, when using general testing procedures (e.g., Collyer et al. 2015; Adams 2014)



Effect size (Collyer, Sekora, & Adams 2015)

Statistical power (Adams 2014 a,b)



Also:
Anderson MJ (2001). *Can J Fish Aquat Sci* 58: 626–639.
Anderson MJ, Legendre P (1999). *J Stat Comput Simul* 62: 271–303.