Morphometric methods and threshold models

Friday Harbor Laboratories, 9 June 2017

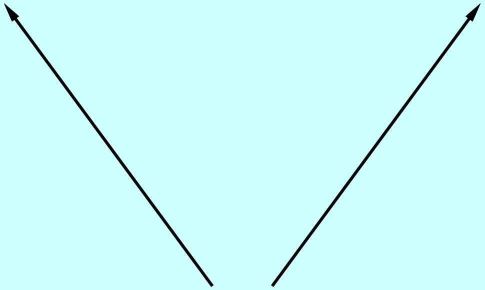
Joe Felsenstein

Workshop on Evolutionary Quantitative Genetics

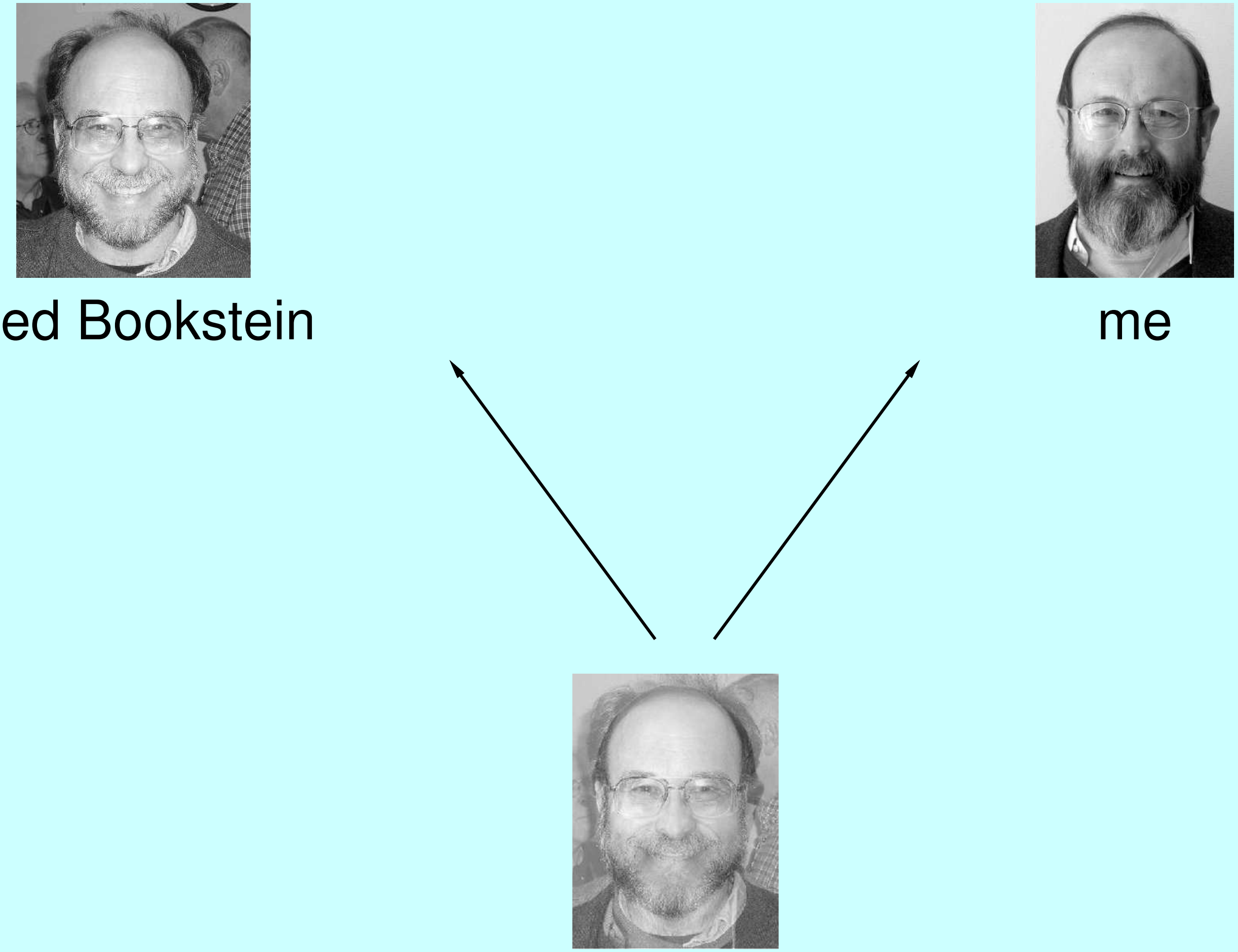
My co-author

Fred Bookstein

me



My co-author



Fred Bookstein

me

“J. F. L. Bookenstein”

(we figure we have a common ancestor ...)

How to use morphometric coordinates on phylogenies?

Is it possible to simply use the coordinates of landmarks $(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$ as continuous phenotypes $x_1, y_1, \ldots, x_p, y_p$ using Brownian motion along a phylogeny?

Yes, but ...

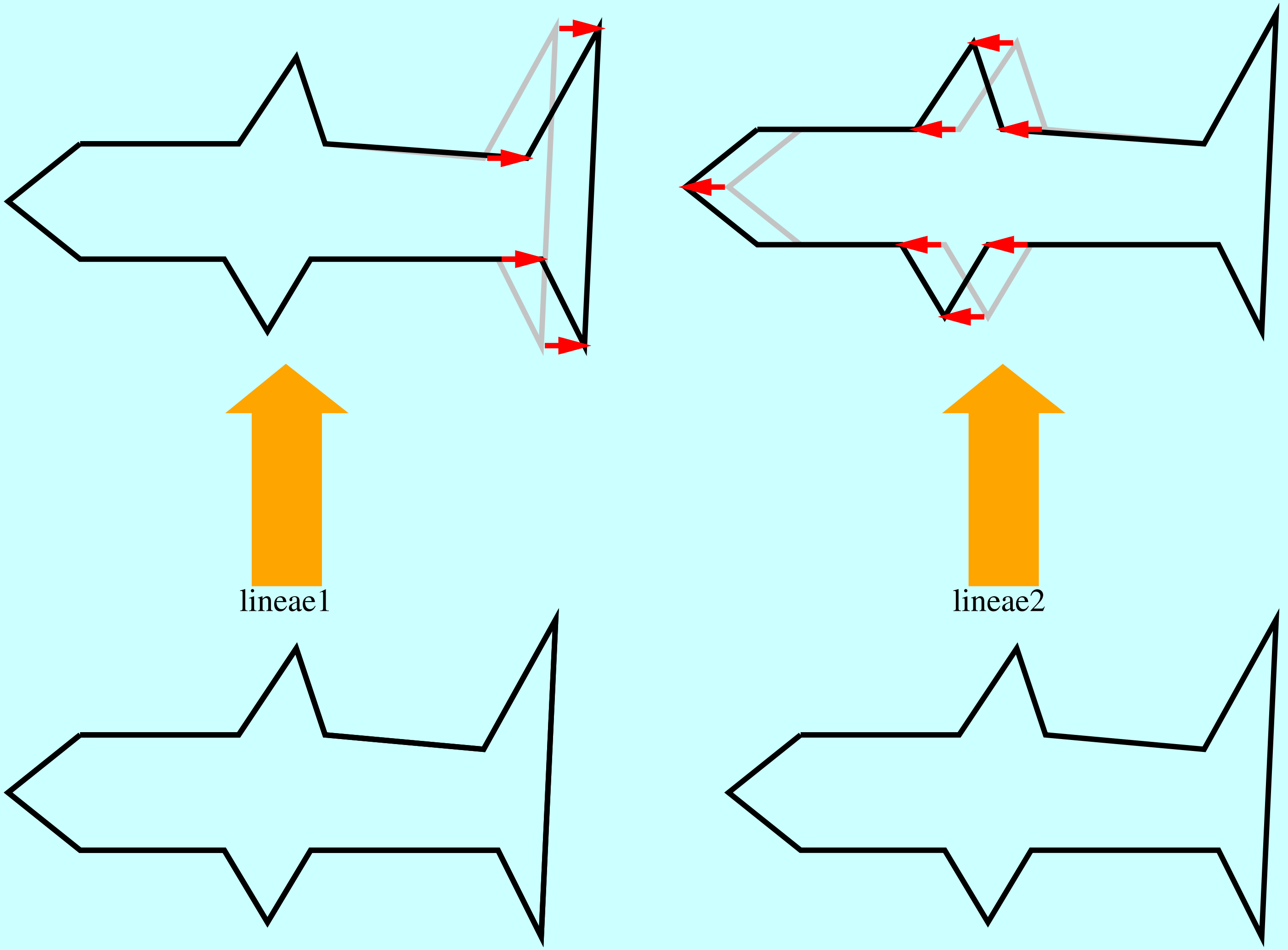
Make sure they are represented in a morphometrically valid way.

Otherwise meaningless translations (shifts) or rotations of the specimens will affect the coordinates.

In effect we are superimposing the specimens properly, although, interestingly, a complete superposition isn't necessary.

Can we superpose specimens?

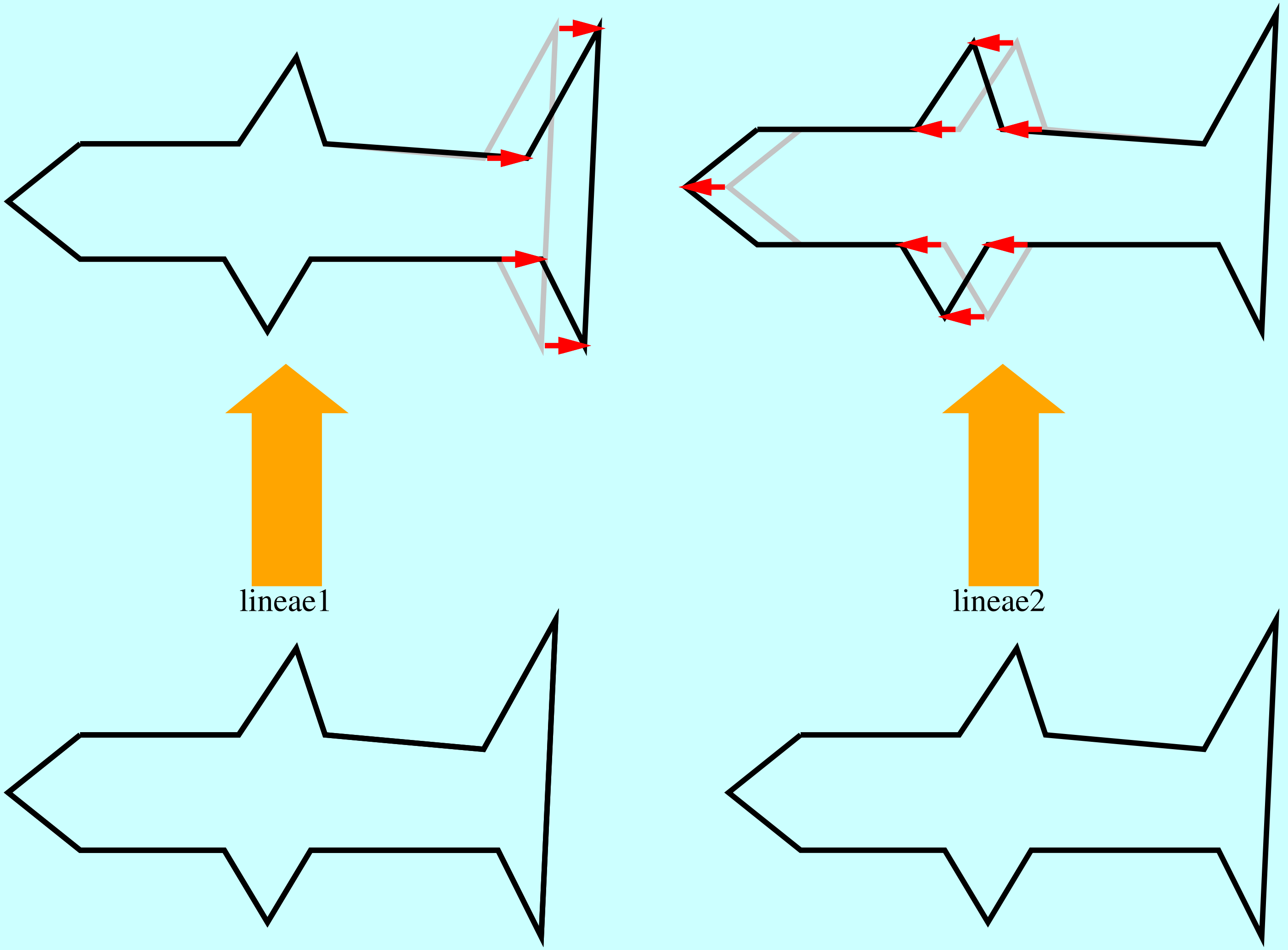
Consider two cases:



Are these different?

Why superposition is in principle impossible

Consider two cases:



Are these different? No!

Dealing with translation

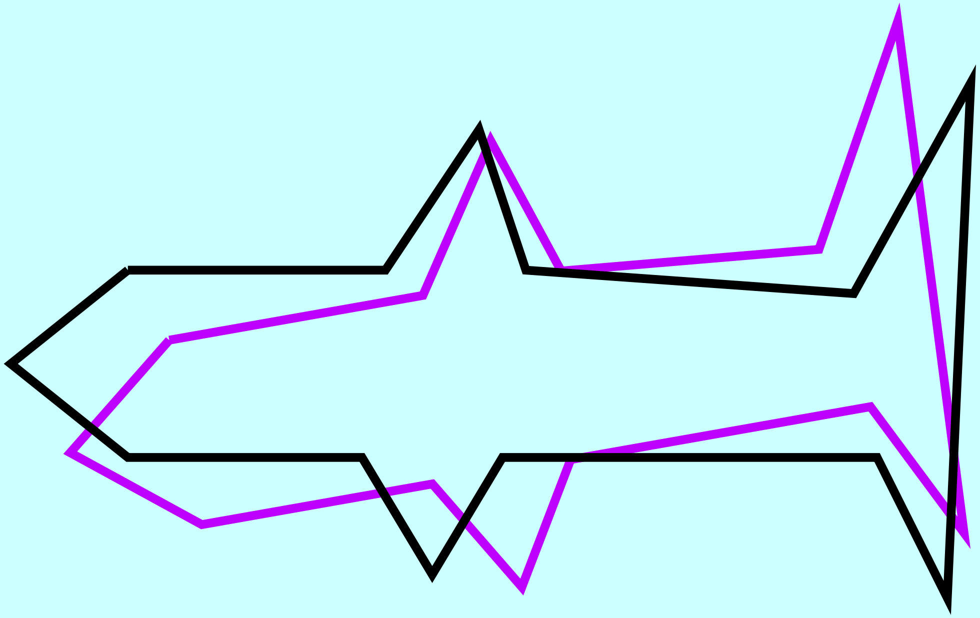
The specimens can be reduced to differences among the `x` coordinates of different points, and differences among the `y` coordinates too, thus losing the grand mean of each specimen.

This amounts to taking contrasts between the different points of one specimen *(a different matter from phylogenetic contrasts, which are for the same coordinate, but between different specimens)*.

In effect one is centering each specimen so that the mean of its points is at $(0, 0)$. (The assumption is that the horizontal and vertical placement of the specimen on the digitizer is not useful information).

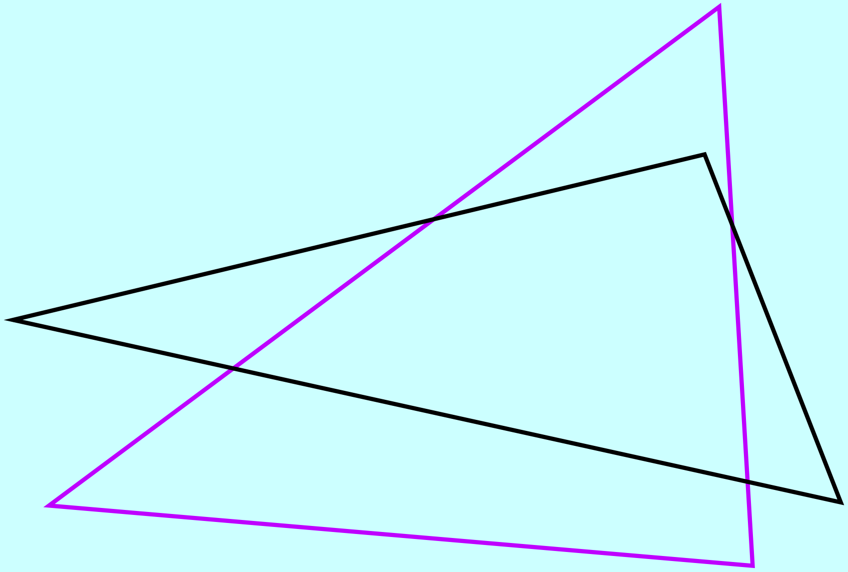
This has the effect of dropping two degrees of freedom so that each specimen now has $2p - 2$ coordinates. It now “lives” in a $(2p - 2)$-dimensional space.

The annoying issue of rotation



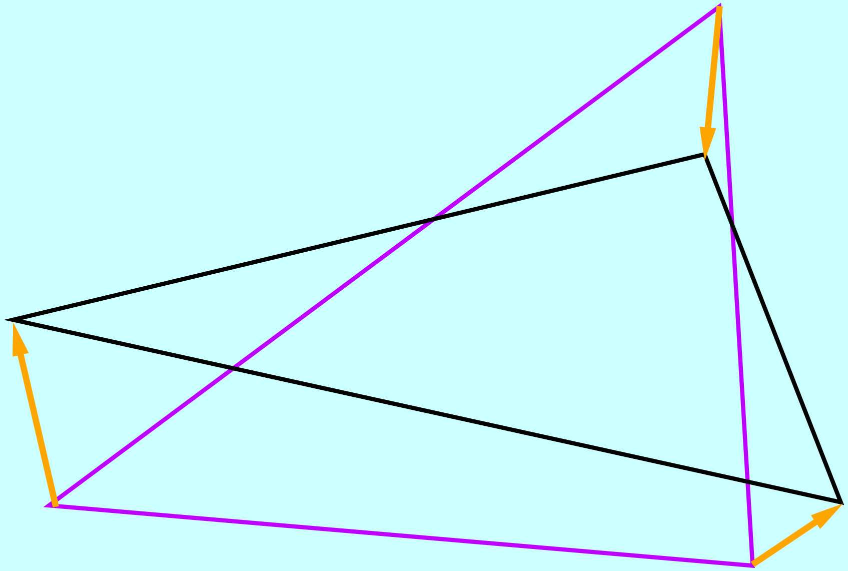
Sadly, there is no corresponding transform that tosses out rotation, as there is for translation.

The Procrustes Transform



Trying to optimally superimpose these forms by translations and rotations so as to minimize some relevant criterion ...

The Procrustes Transform



Achieves a least squares fit by centering and rotating so that the sum of squares of the golden arrows is at a minimum.

Procrustes superposition of multiple forms

One algorithm that works:

- Superpose forms $2, 3, \ldots, n$ each to the first form.

Procrustes superposition of multiple forms

One algorithm that works:

- Superpose forms $2, 3, \ldots, n$ each to the first form.
- Take the averages of the coordinates to get an average form

Procrustes superposition of multiple forms

One algorithm that works:

- Superpose forms $2, 3, \ldots, n$ each to the first form.
- Take the averages of the coordinates to get an average form
- Superpose all n forms to this average.

Procrustes superposition of multiple forms

One algorithm that works:

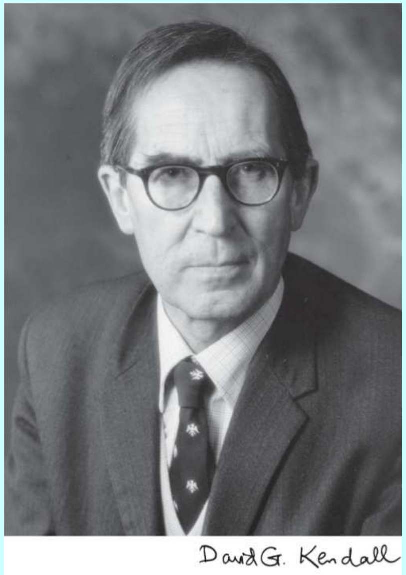
- Superpose forms $2, 3, \ldots, n$ each to the first form.
- Take the averages of the coordinates to get an average form
- Superpose all n forms to this average.
- Now recompute average form and superpose to that.

Procrustes superposition of multiple forms

One algorithm that works:

- Superpose forms $2, 3, \ldots, n$ each to the first form.
- Take the averages of the coordinates to get an average form
- Superpose all n forms to this average.
- Now recompute average form and superpose to that.
- Continue until it converges, which it will, quickly.

The “morphometric consensus”



David Kendall

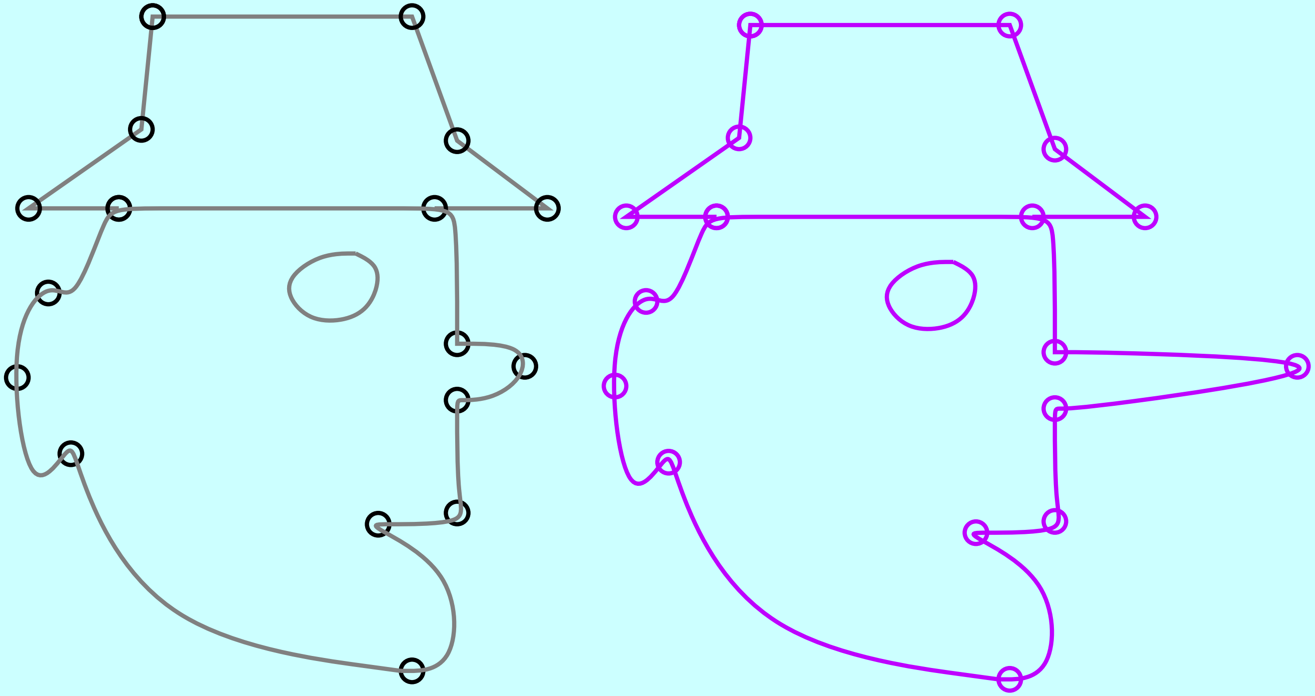
(1918-2006)

- Superpose by a Procrustes fit
- Compute all pairwise distances between forms
- Do Principal Coordinates (not Components) on these
- These PCA coordinates are the ones you then analyze

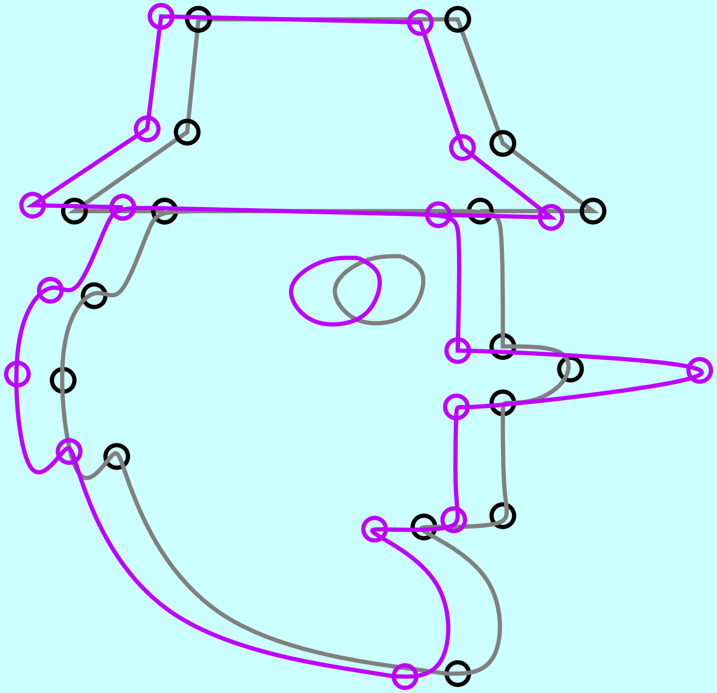
John Gower (in 2008)

Problems: ignores phylogeny, implicitly assumes that the model of statistical error is independent, isotropic noise at all landmarks

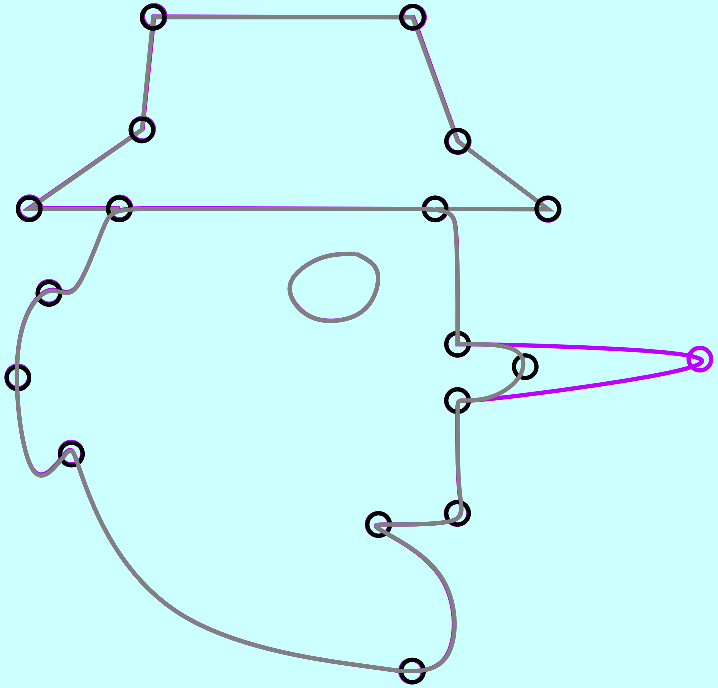
The Pinocchio effect



What a Procrustes fit might do



One would like to get more like this



Doug Theobald's "Theseus Transform"

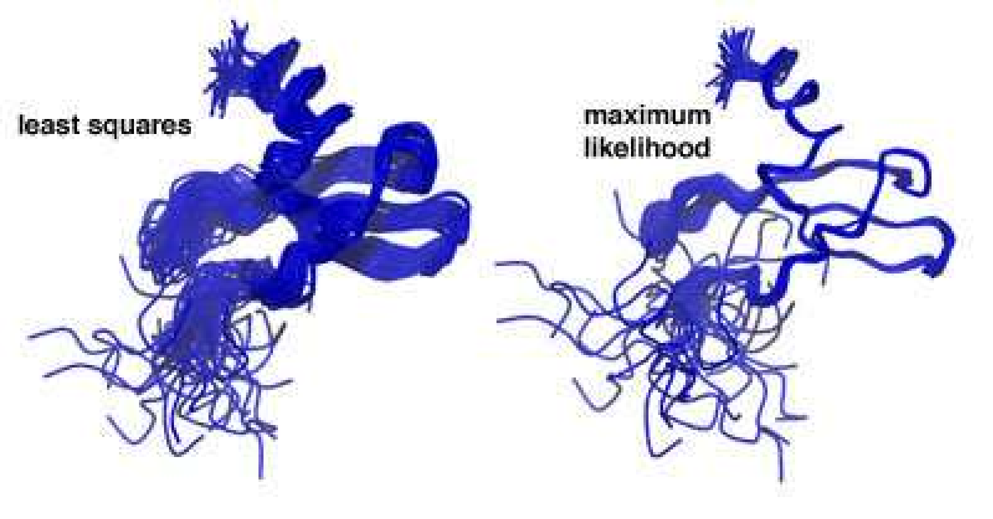


Figure. 2. A maximum likelihood superposition vs least squares, for 30 NMR models of cytokine stromal factor SDF-1.

Doug Theobald   his molecular superposition

Theobald, D. L. and D. S. Wuttke. 2008. Accurate structural correlations from maximum likelihood superposition. *PLoS Computational Biology* **42(2):** e43.

It does not assume isotropic, independent error but estimates a full error covariance matrix, as we will do too.

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model

Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix

Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix
- The rotations are chosen, iteratively, to maximize the likelihood

Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix
- The rotations are chosen, iteratively, to maximize the likelihood
- Size (scale) is incorporated into the linear model as a linearized expansion of the mean form.

Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix
- The rotations are chosen, iteratively, to maximize the likelihood
- Size (scale) is incorporated into the linear model as a linearized expansion of the mean form.
- We are in effect using the phylogenetically independent contrasts on the tree instead of treating the specimens as independent data

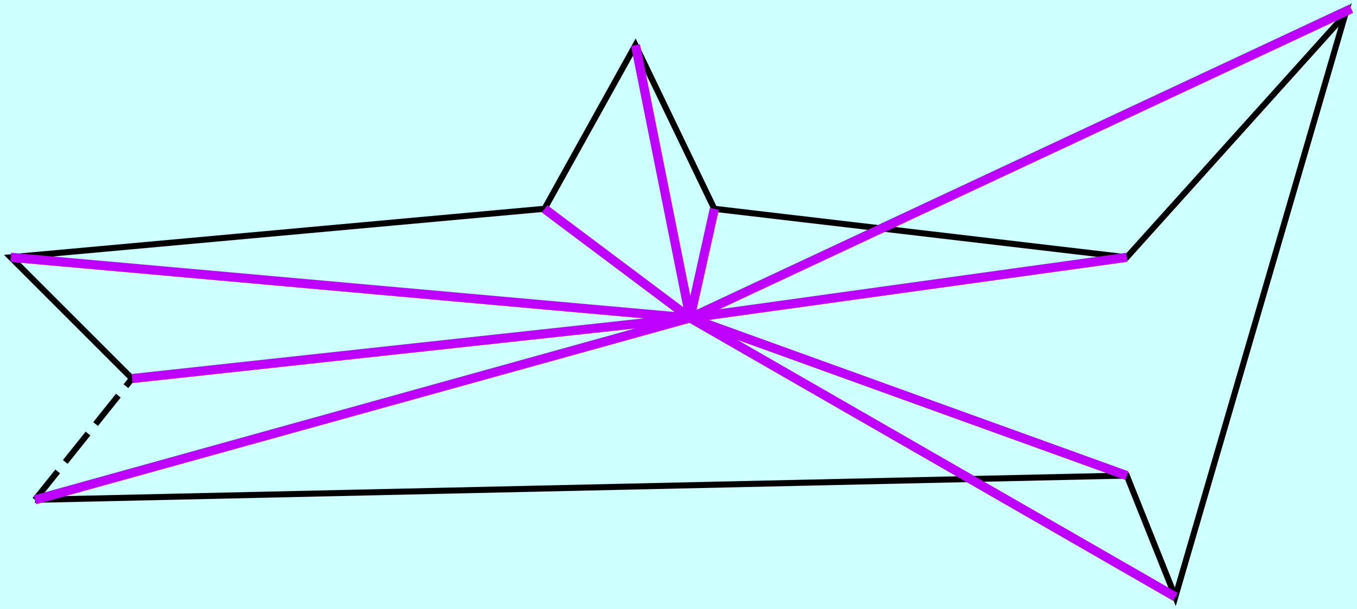
Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Our method uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix
- The rotations are chosen, iteratively, to maximize the likelihood
- Size (scale) is incorporated into the linear model as a linearized expansion of the mean form.
- We are in effect using the phylogenetically independent contrasts on the tree instead of treating the specimens as independent data
- (But wait, weren’t superpositions meaningless?)

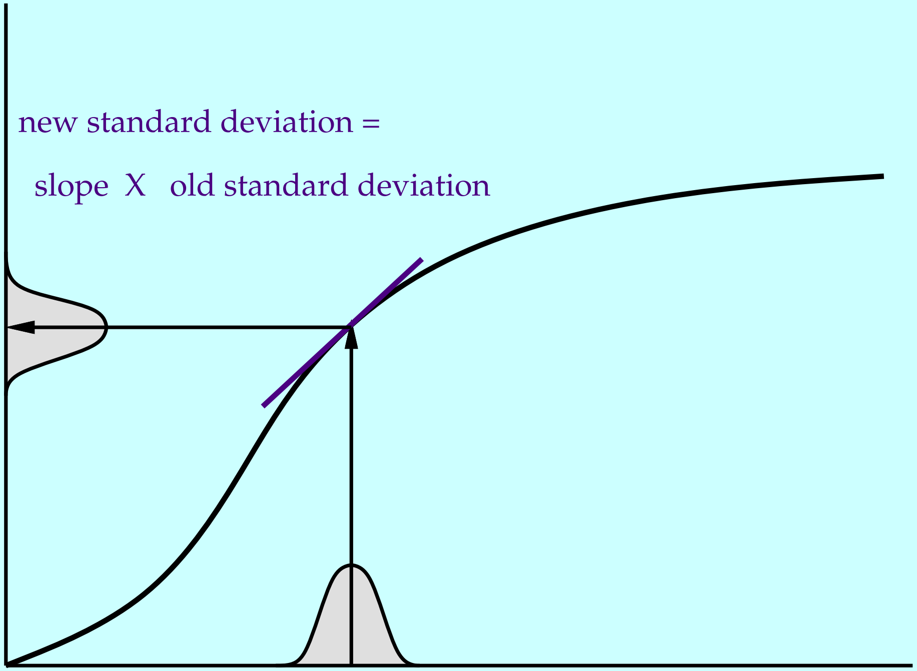
Most steps of the Morphometric Consensus are omitted in this (no Procrustes distances, no Principal Coordinates, maybe even no centroid sizes).

Size as treated in the Morphometric Consensus



In the Consensus, size is computed by (1) taking the centroid of the landmarks, (2) computing the square root of the sum of squared distances from it. For analyses of “shape” rather than “form” one scales the forms so that this is 1.

Why multivariate normality may be OK



The well-known “delta method” uses the fact that a linear transformation of a scale is locally nearly linear (Recall the Taylor Series in your calculus course). Also implies that a normal distribution before is approximately normal afterwards too. This is reasonable if variation is small.

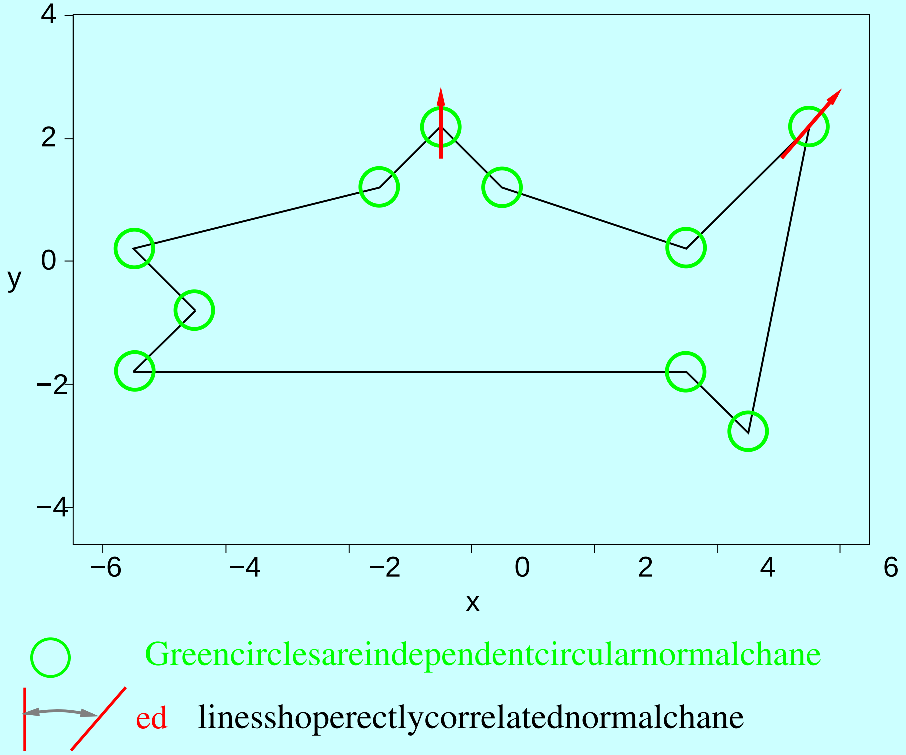
Our method is similar but uses phylogenies

- We compute the likelihood of changes on a phylogeny using a (covarying) Brownian motion model
- We are maximizing the likelihood to infer the covariance matrix
- The translations and rotations are chosen, iteratively, to maximize the likelihood
- We are in effect using the phylogenetically independent contrasts on the tree instead of treating the specimens as independent data
- (We also have a new approach to size which will not be described here).

Basically, none of the steps of the Morphometric Consensus are left except for centering all specimens at their centrois.

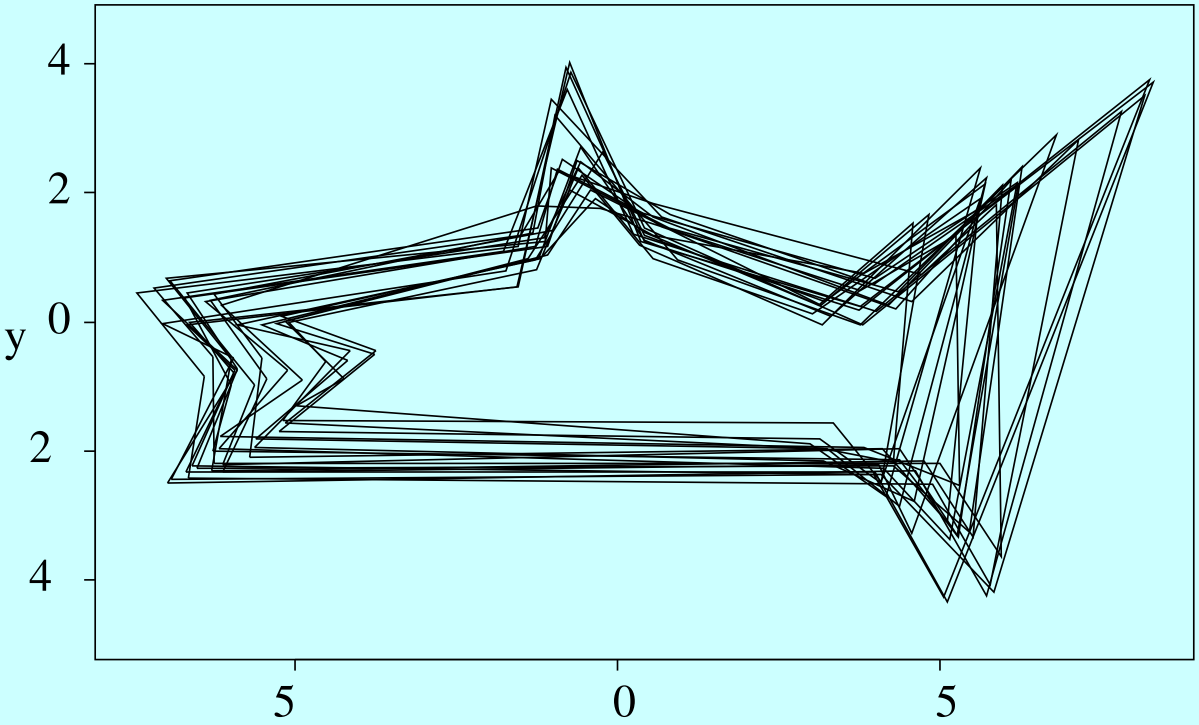
A simulation test

1. Generate 100-species trees by a pure birth process
2. For each evolve forms by (covarying) Brownian Motion up the tree
3. These are the true covariances:

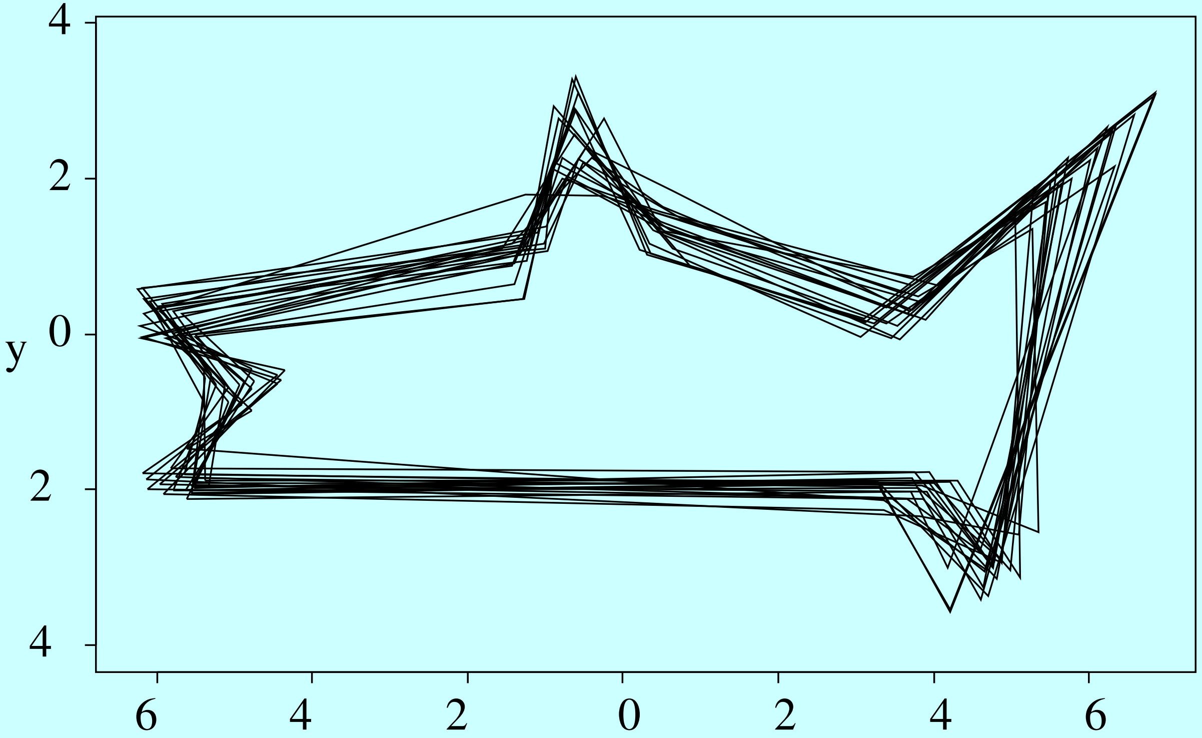


- All 10 landmarks move by independent and equal Brownian Motion of the coordinates with variance (per unit branch length) of 0.001, *plus*
- the vertical coordinate of the pectoral fin and the two coordinates of the top of the tail move in a perfectly correlated change with variance 0.003.

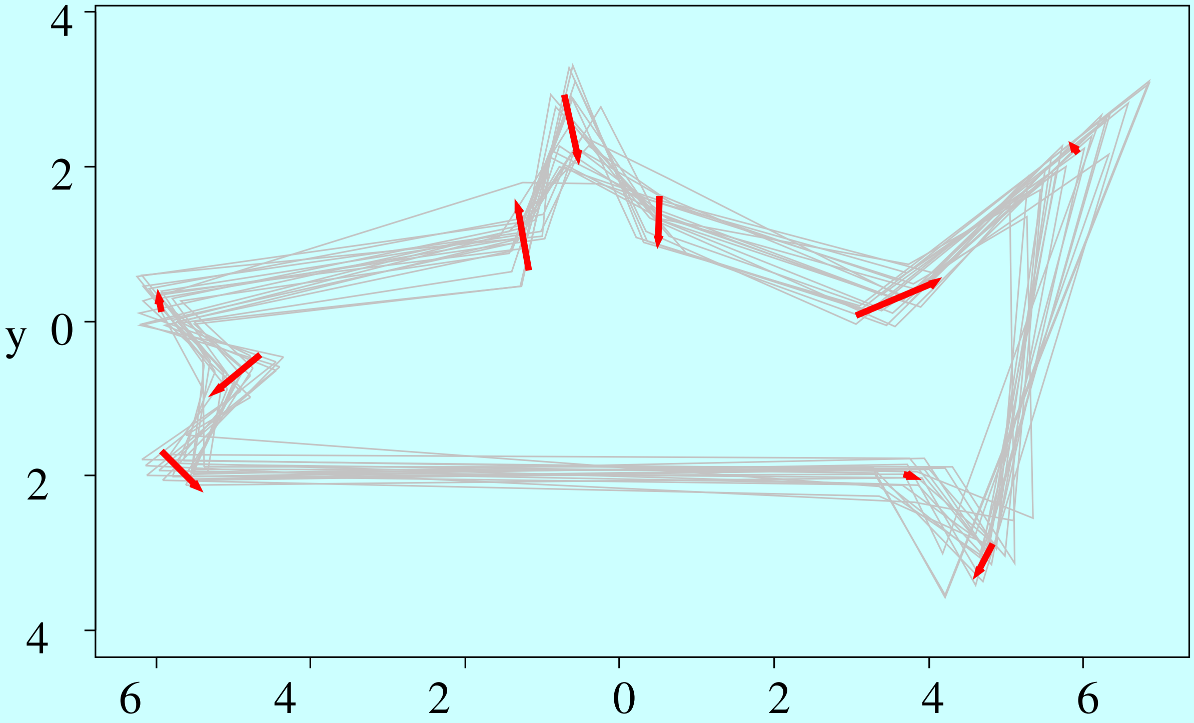
20 of the 100 fishes from data set #2, centered and rotated



20 of the 100 fishes from data set #2, also rescaled

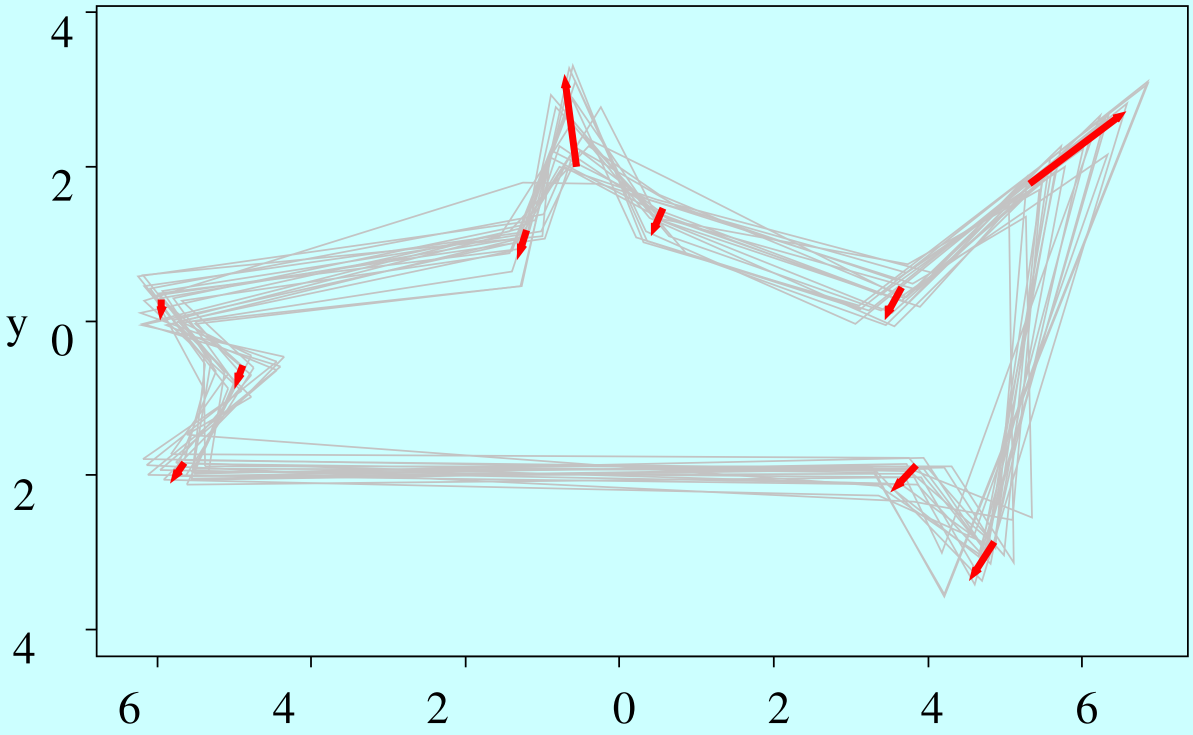


First PC 1 for data set #2



This principal component shows both size changes and the fin extensions, and it is not easy to see which is which.

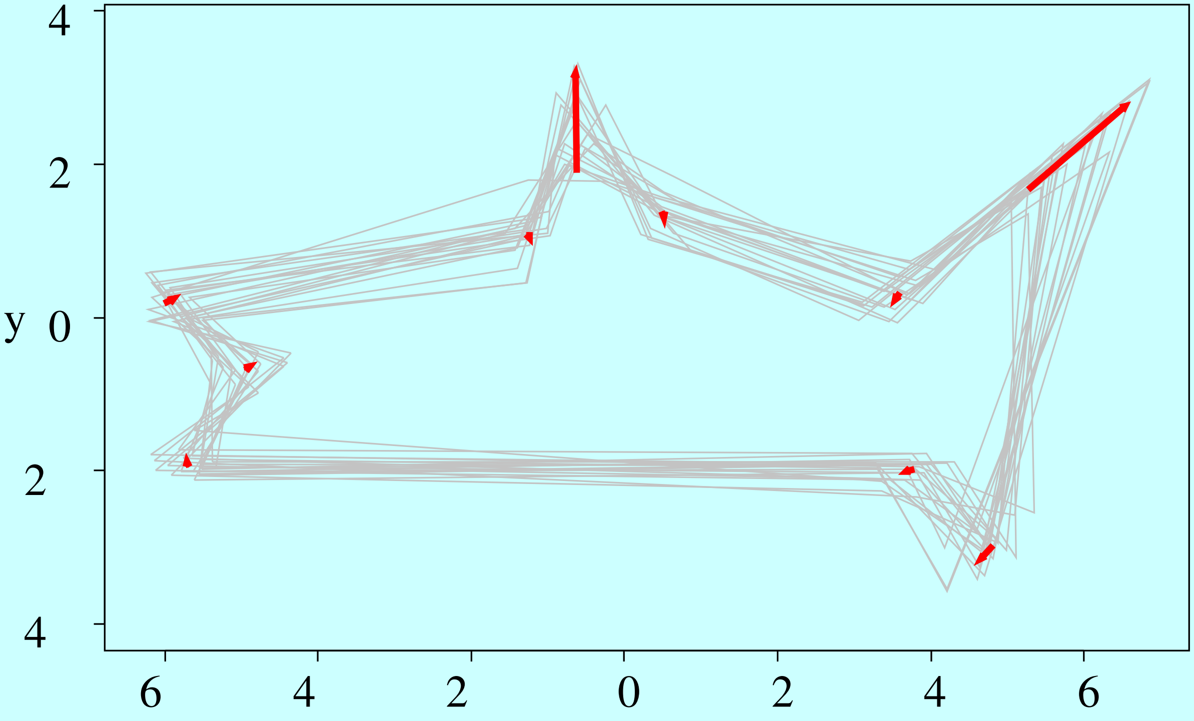
First shape PC 1 for data set #2



Now we’ve inferred a scale (size) component and removed it from the covariances, and then taken the first PC of the residual on size. We can see the fin component more clearly.

Making the first shape PC sparser by “medianizing”

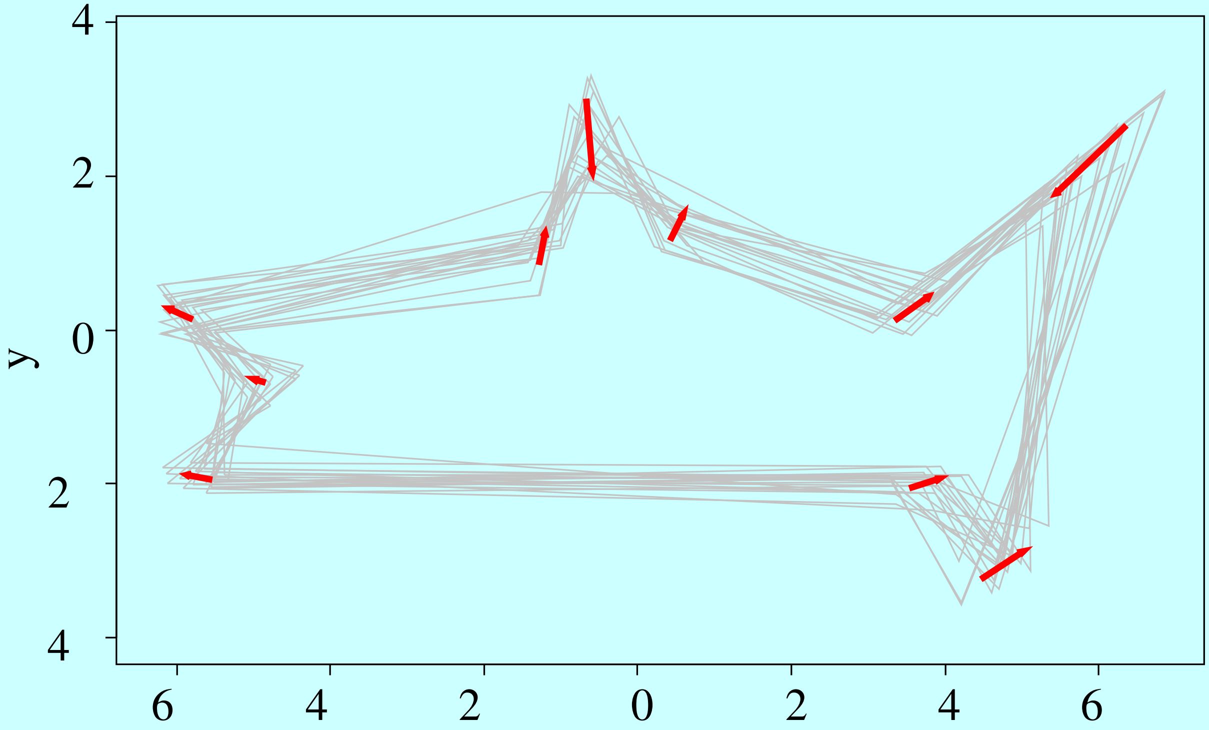
To make PC1 be sparses we can add in a little location (not forcing the changes to maintain the centroid superposition).



This is done by subtracting from the x components, their median, and similarly for the y components. So it minimizes the L^1 norm of the PC coefficients. The result is very clear.

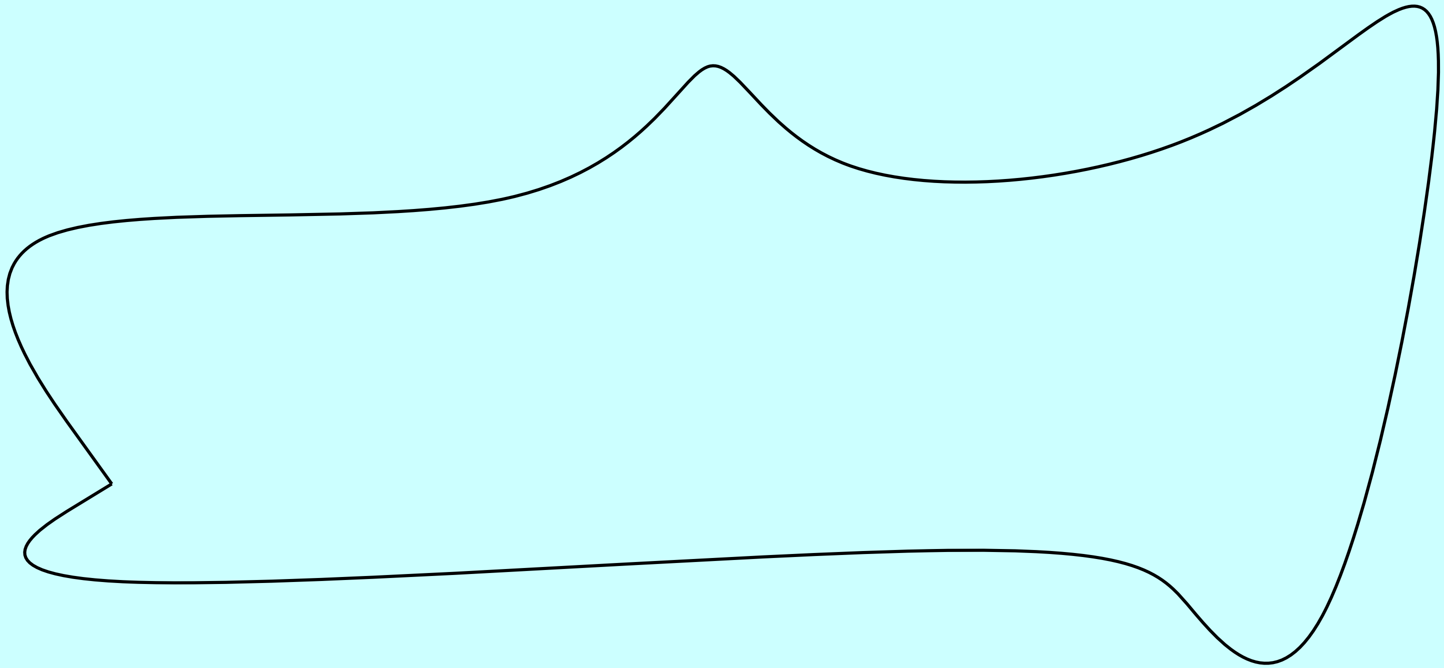
What do we get from the Morphometric Consensus?

Using a Procrustes superposition and assuming the forms are i.i.d. and then computing principal components:



... we get a not-as-clear result with some size still there – we have ignored the tree, taken out size by standardizing centroid size, so fin component gets more mixed up with size.

Outline and semilandmark methods



A number of methods exist to use somewhat-arbitrarily placed pseudolandmarks on a curved shape that does not have any true landmarks. There are also curve-fitting methods such as Elliptical Fourier Analysis (one considers the coefficients of the fit as the phenotype values).

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...
- ... they do not take developmental processes into account

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...
- ... they do not take developmental processes into account
- ... they do not take into account what genetic variation is available or not available.

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...
- ... they do not take developmental processes into account
- ... they do not take into account what genetic variation is available or not available.
- It’s easy to say we should make use of those, but ...

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...
- ... they do not take developmental processes into account
- ... they do not take into account what genetic variation is available or not available.
- It's easy to say we should make use of those, but ...
- But it is important to keep them in mind as goals.

Geometry vs. genetics and development

- All of the above methods have lots of powerful geometry, but ...
- ... they do not take developmental processes into account
- ... they do not take into account what genetic variation is available or not available.
- It’s easy to say we should make use of those, but ...
- But it is important to keep them in mind as goals.
- The hope: compare directions of evolutionary divergence with directions of available genetic variation, a research program we noted before.

References for morphometrics

The older methods:

Bookstein, F. L. 1991. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge, U.K. **[Early statement of the Morphometric Consensus methods]**

Dryden, I. L. and K. V. Mardia. 1998. *Statistical Shape Analysis*. John Wiley and Sons, New York. **[Much mathematical machinery for the Morphometric Consensus (MMC)]**

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53:** 325–338. **[Principal coordinates analysis – principle components for distance matrices, used by MMC]**

Zelditch, M. L., D. L. Swiderski, and H. David Sheets. 2012. *Geometric Morphometrics for Biologists: A Primer*. Academic Press, New York.

The newer methods:

Theobald, D. L. and D. S. Wuttke. 2008. Accurate structural correlations from maximum likelihood superposition. *PLoS Computational Biology* **42(2):** e43. **[Quite similar to ours, though for superposing molecules so does not change sizes]**

(ours to be published soon ...)