Overview: morphometrics and phylogenies

Joe Felsenstein

Bio 550D, Autumn2016

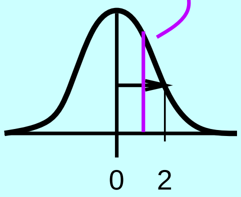
Fred Bookstein is my coteacher in this class

Fred L.Bookstein

A standard quantitative genetics model

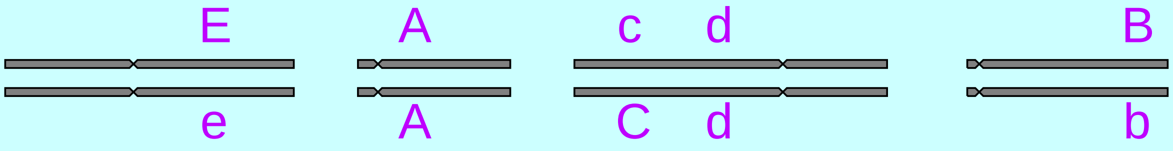
$$P = \mu + \left\{\begin{array}{ll} \mathrm{AA} & 2 \\ \mathrm{Aa} & 4 \\ \mathrm{aa} & 7 \end{array}\right\} + \left\{\begin{array}{ll} \mathrm{BB} & 0.6 \\ \mathrm{Bb} & 0.1 \\ \mathrm{bb} & -0.2 \end{array}\right\} + \left\{\begin{array}{ll} \mathrm{CC} & -1 \\ \mathrm{Cc} & 6 \\ \mathrm{cc} & 6 \end{array}\right\} + \left\{\begin{array}{ll} \mathrm{DD} & 0.3 \\ \mathrm{Dd} & 0.3 \\ \mathrm{dd} & 0.7 \end{array}\right\} + \left\{\begin{array}{ll} \mathrm{EE} & -0.4 \\ \mathrm{Ee} & 0.3 \\ \mathrm{ee} & -0.3 \end{array}\right\} + \text{environmental effect}$$

arbitrary starting point

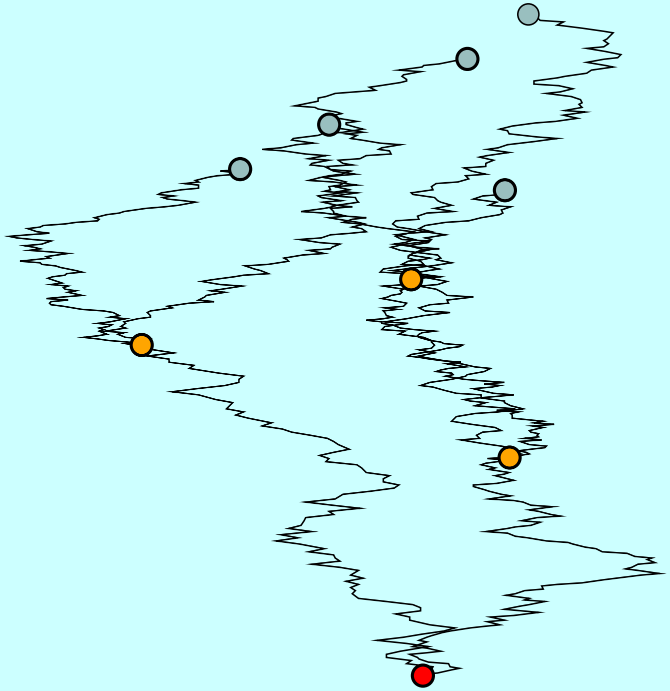


AA	Bb	Cc	dd	Ee	10
Aa	bb	cc	DD	ee	
aa	bb	CC	DD	Ee	
aa	bb	Cc	DD	EE	
Aa	Bb	Cc	DD	Ee	

$$0.3 + 2 + 6 + 0.7 + 0.1 + 0.9$$

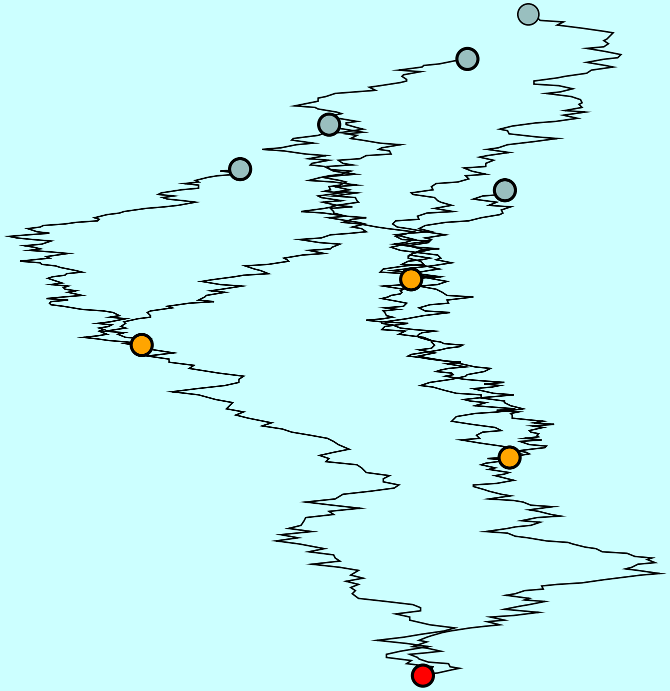


A model of quantitative characters on a phylogeny



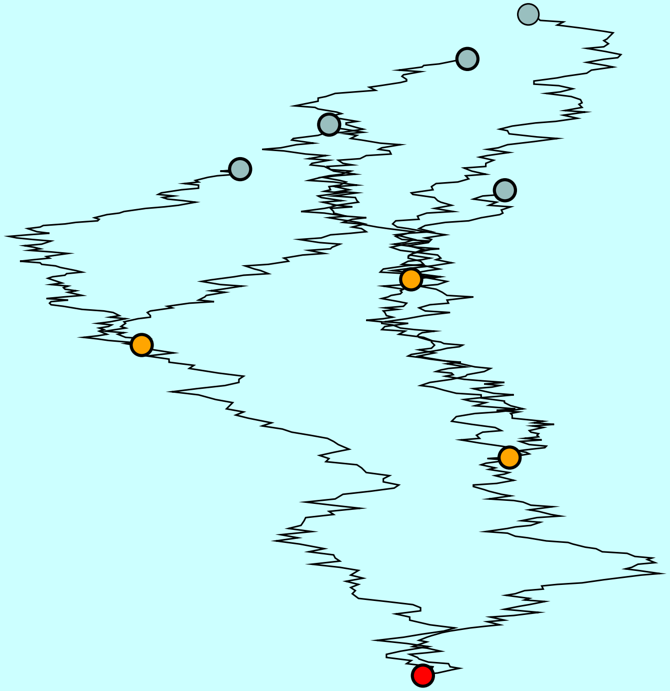
- Brownian motion with multiple characters with different variances and with covariation as well.

A model of quantitative characters on a phylogeny



- Brownian motion with multiple characters with different variances and with covariation as well.
- This started with approximating gene frequencies in the 1960s by Anthony Edwards and Luca Cavalli-Sforza.

A model of quantitative characters on a phylogeny



- Brownian motion with multiple characters with different variances and with covariation as well.
- This started with approximating gene frequencies in the 1960s by Anthony Edwards and Luca Cavalli-Sforza.
- I expanded it to model quantitative characters determined by these geness (1973, 1981, 1988).

Models for long-term evolution

The use of quantative genetics approximations to model long-term evolution in lineages was largely introduced by Russ Lande in the 1980s.

Russell Lande, from his website at Imperial College, U.K., where he has been in recent years.

Where do the covariances come from?

- **Genetic covariances** (the same loci affect two or more traits). Genetic drift or natural selection can change the gene frequencies at these loci, and thus make correlated changes in the two traits.

Where do the covariances come from?

- **Genetic covariances** (the same loci affect two or more traits). Genetic drift or natural selection can change the gene frequencies at these loci, and thus make correlated changes in the two traits.
- **Selective covariances** (Olof Tedin, 1926; G. Ledyard Stebbins 1950) The same environmental conditions can select changes in two or more traits – even though they may have no genetic covariance. This source of evolutionary covariance is widely ignored.

How to use morphometric coordinates on phylogenies?

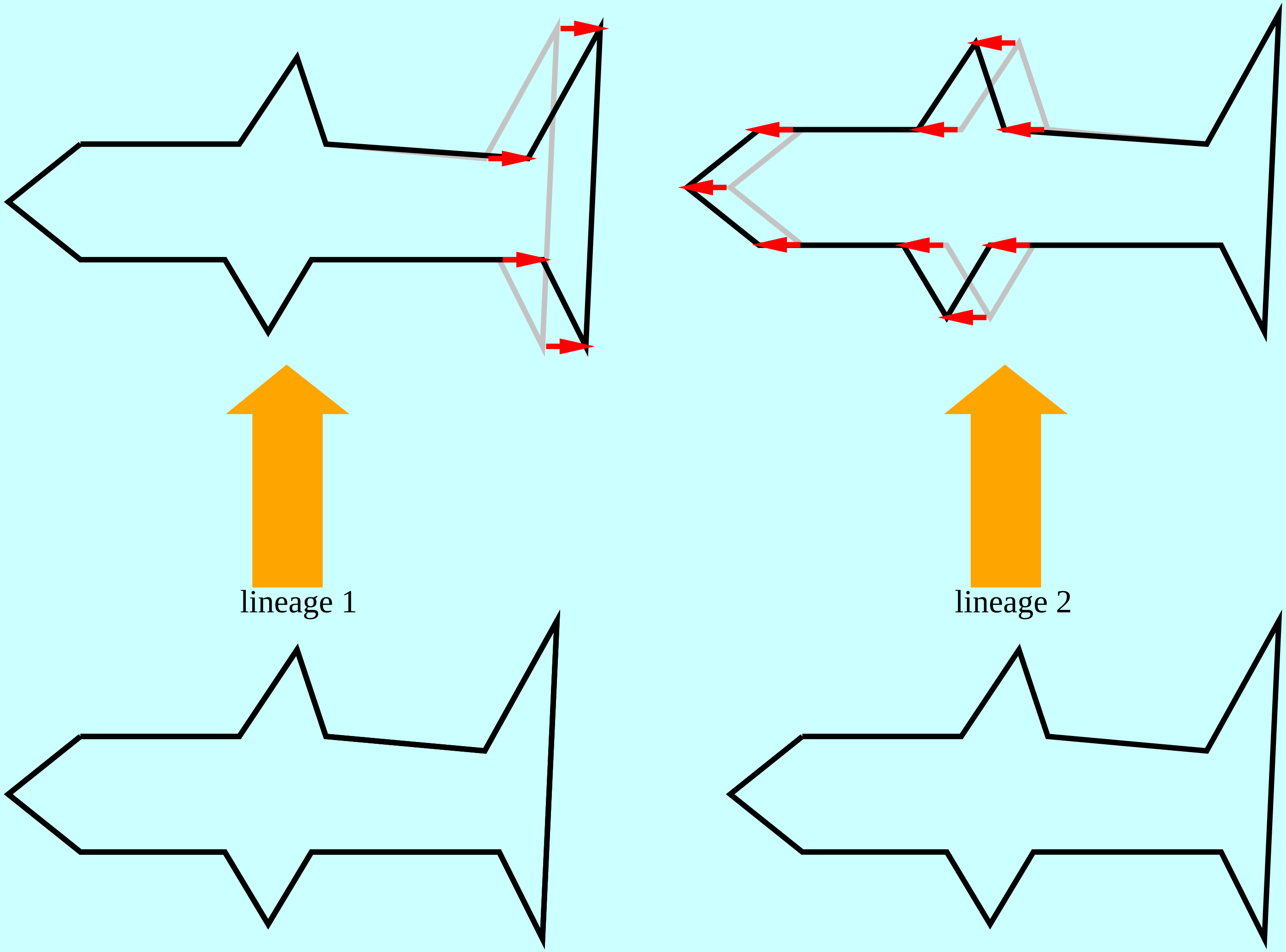
Is it possible to simply use the coordinates of landmarks $(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$ as continuous phenotypes $x_1, y_1, \ldots, x_p, y_p$ using Brownian motion along a phylogeny?

Yes, but ...

We must do proper morphometrics (correct for translation? rotation?)

Can we superpose specimens?

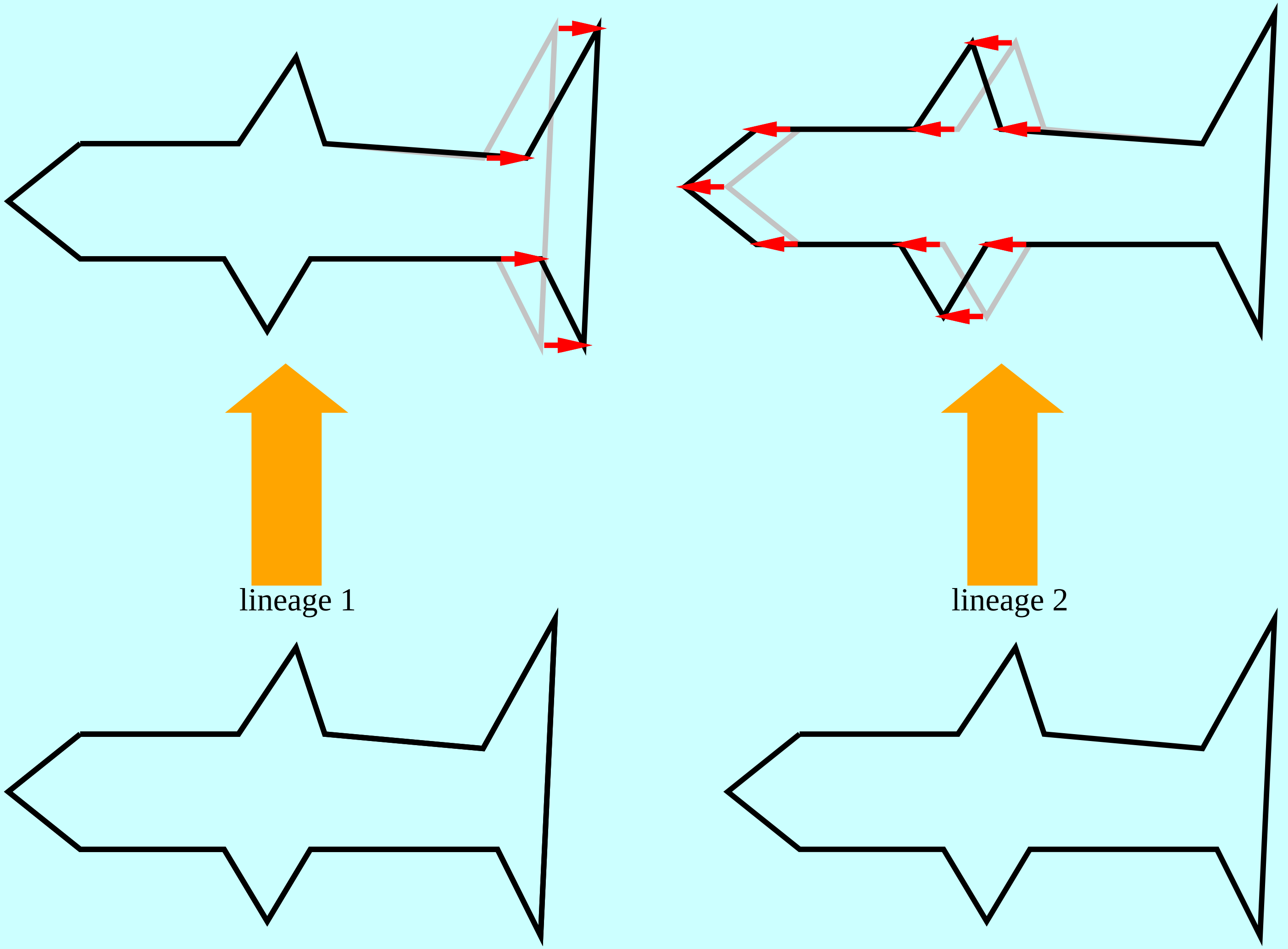
Consider two cases:



Are these different?

Why superposition is in principle impossible

Consider two cases:



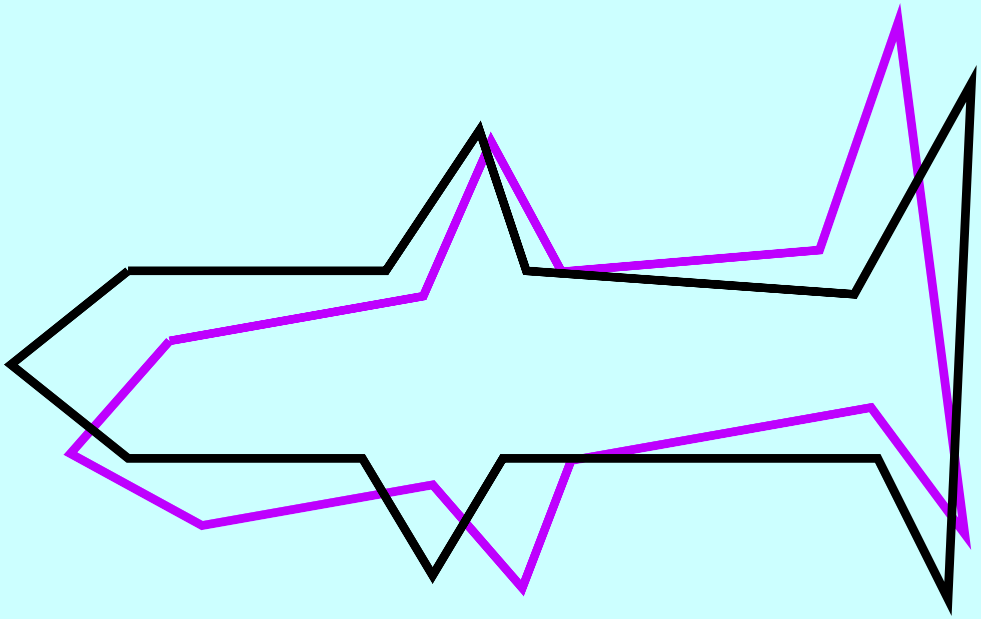
Are these different? No!

Dealing with translation

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).

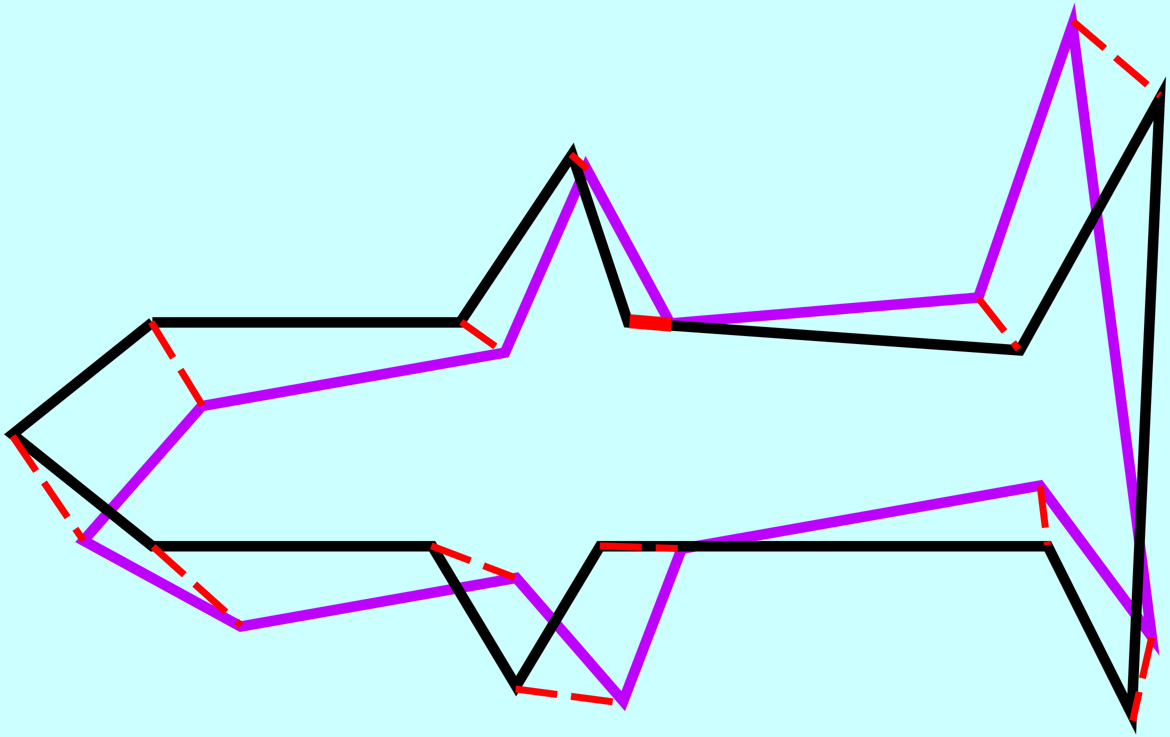
This has the effect of dropping two degrees of freedom so that each specimen now has $2p - 2$ coordinates. It now “lives” in a $(2p - 2)$-dimensional space.

The annoying issue of rotation



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

The Procrustes Transform



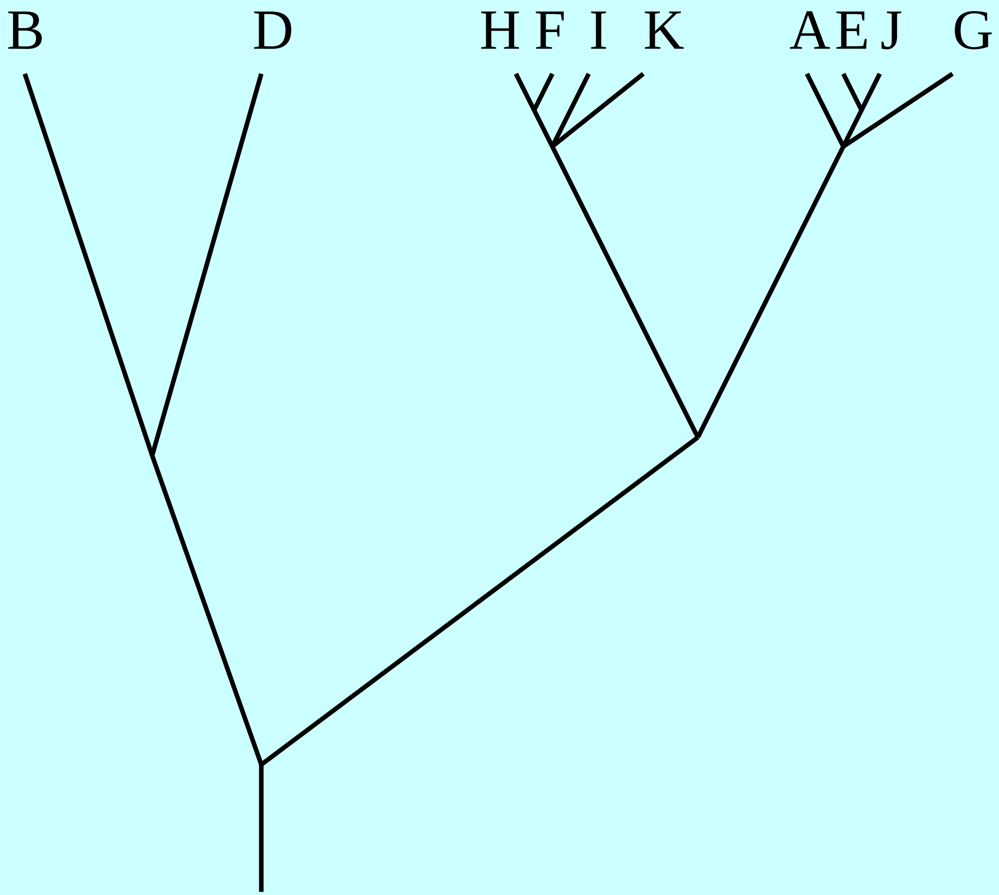
The Procrustes Distance of two forms is the sum of squares of the dashed red lines (connecting corresponding landmarks). The forms are superposed by rotating and translating to minimize this distance. Generalized Procrustes Analysis is simply doing rotations and translations to minimize the sum of squares, summed over all pairs of specimens.

The Morphometric Consensus

- Superpose (and rescale) specimens by a Generalized Procrustes Transform
- Compute (square root of) sum-of-squares distances between all of them
- Get Principal Coordinates from this distance matrix
- Make a space with these PCs as axes, put points in it, one per specimen.

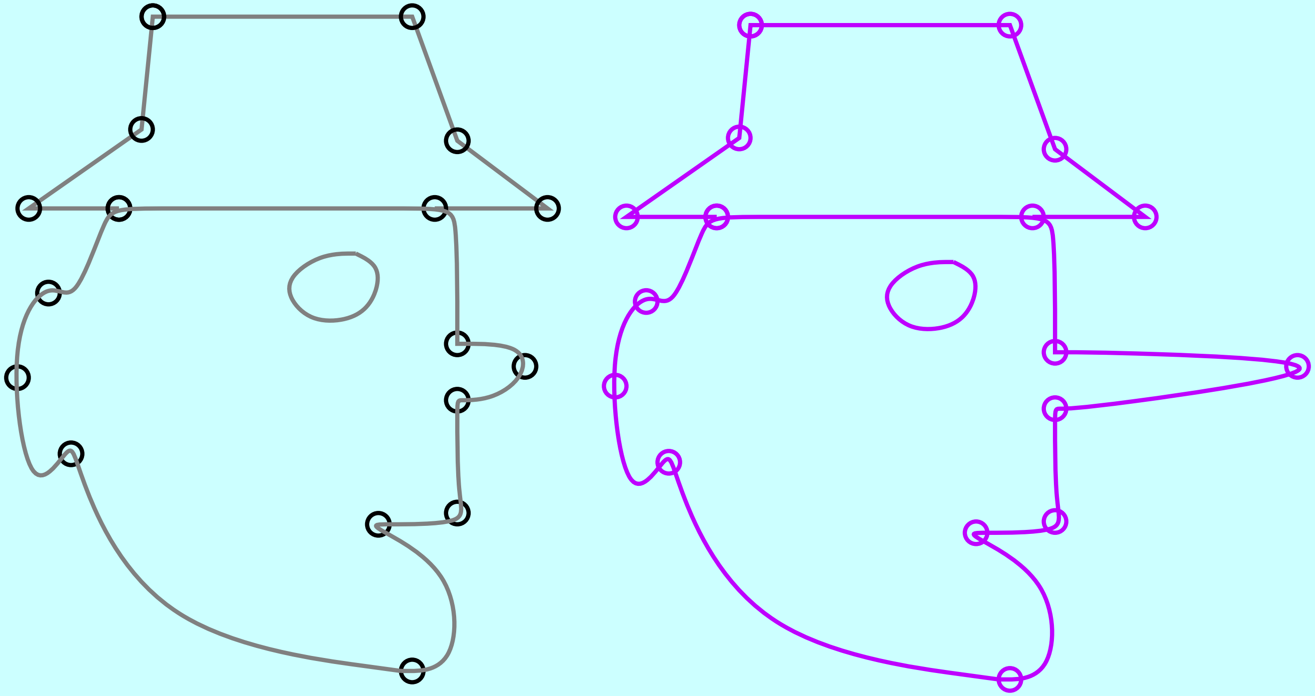
Problems: (1) ignores phylogeny, (2) implicitly assumes that the model of statistical error is independent, isotropic noise at all landmarks

Why you should not ignore phylogenies



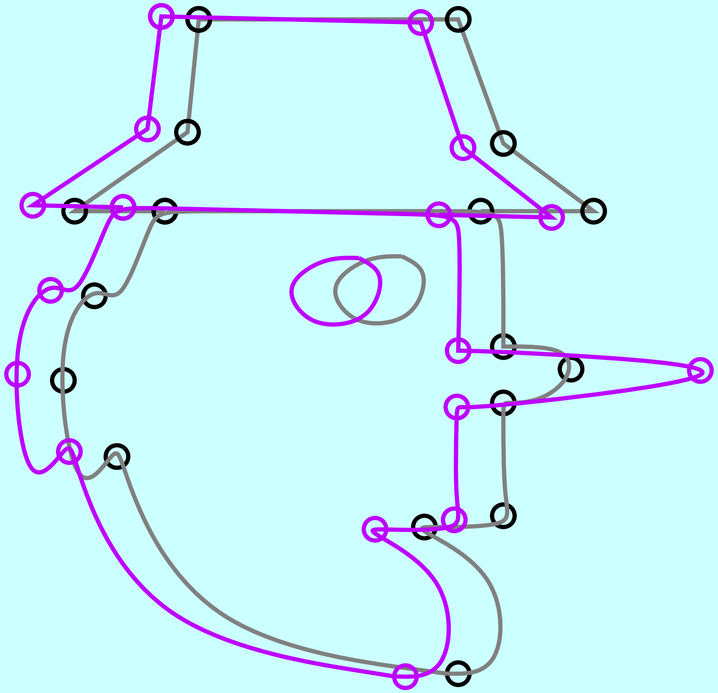
In inferring means (ancestors), variances and covariances, treating the species as i.i.d. independent samples ignores that some are near-duplicates of others. This is the classical phylogenies-and-the-comparative-method problem.

The “Pinocchio effect”



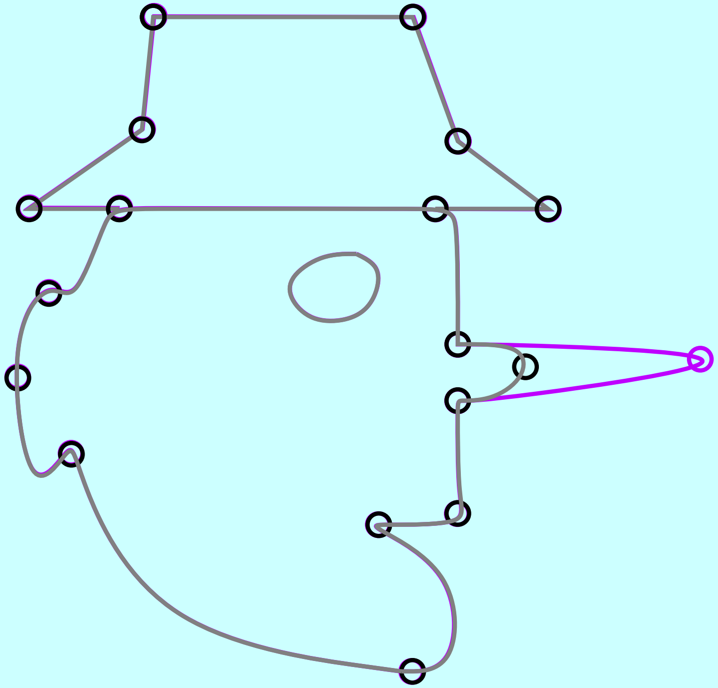
If we have a form for which one part (in this case the nose) changes much more than others, this violates the implicit model of the Procrustes superposition.

The “Pinocchio effect”



... and a Procrustes superposition that does not take this into account will tend to overreact to the changes of that part of the form.

The “Pinocchio effect”



... whereas a more sensible approach would allow that point to deviate farther because it changes more readily.

Removing translation and rotation

- We remove translation by centering each specimen on (0, 0), which drops 2 degrees of freedom in a two-dimensional space.
- We remove rotation by estimating each specimen’s rotation angle by maximizing the log-likelihood for the whole study with respect to it. (Keep rotating them each until the overall log-likelihood is maximum).
- There is something similar that can be done to separate size and shape.

Instead of the “morphometric consensus”

... we do *not* do the usual Procrustes distances and PCs in the tangent space. That implicitly assumes equal and independent noise in each coordinate.

Instead we assume a multivariate Brownian Motion with arbitrary evolutionary covariances, and we estimate them, just as one does in phylogenetic comparative methods.

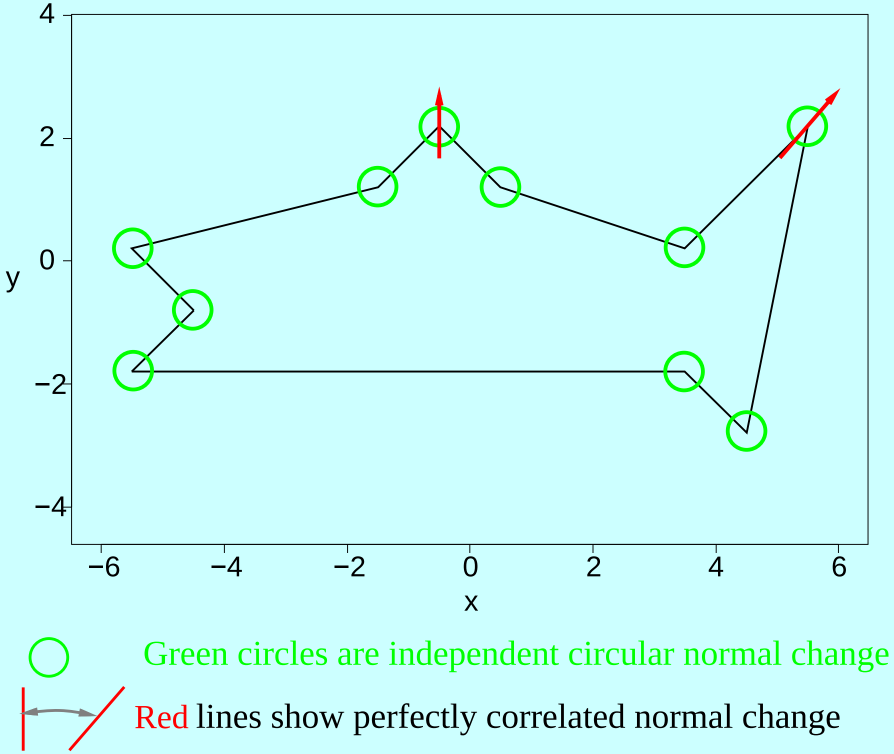
Our method is similar but uses phylogenies

- 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
- We do superpose the centroids of the specimens, just as the MMC does.
- The rotations are chosen, iteratively, specimen by specimen, to maximize the likelihood
- We are in effect using the phylogenetically independent contrasts on the tree instead of treating the specimens as independent data

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

A simulation test

1. Generate 50 100-species trees by a pure birth process
2. For each evolve 100 forms by (covarying) Brownian Motion up the tree
3. These are the true covariances:

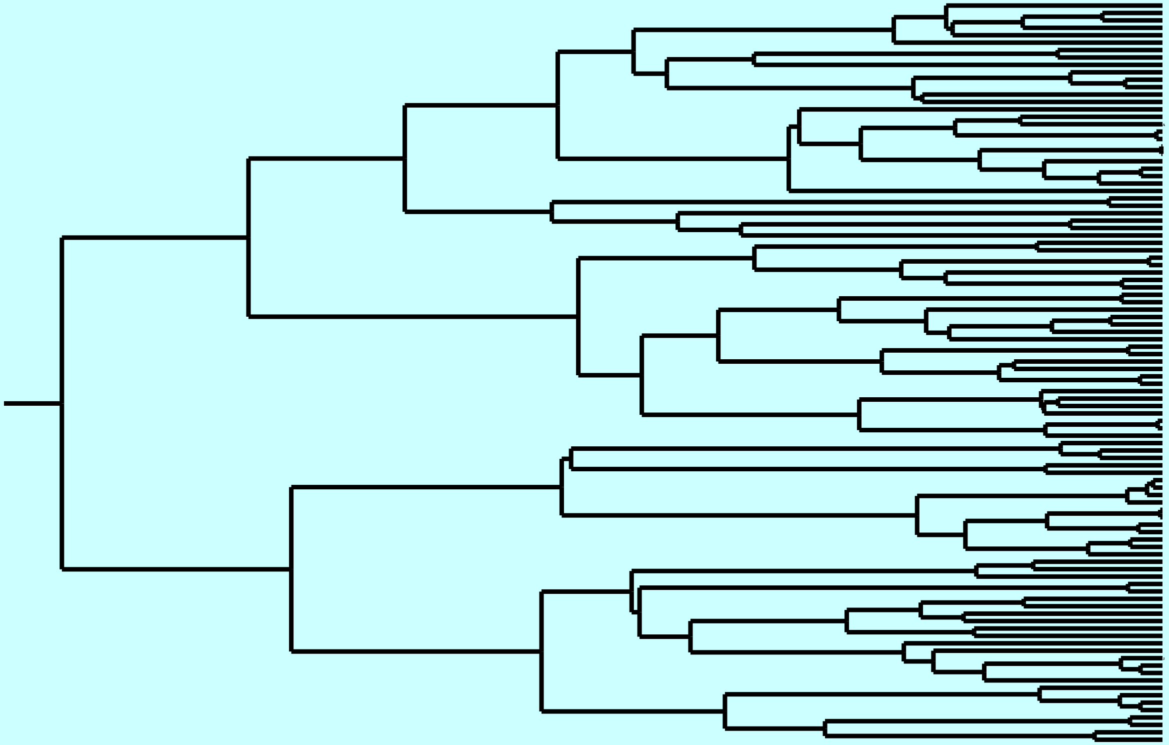


○ Green circles are independent circular normal change

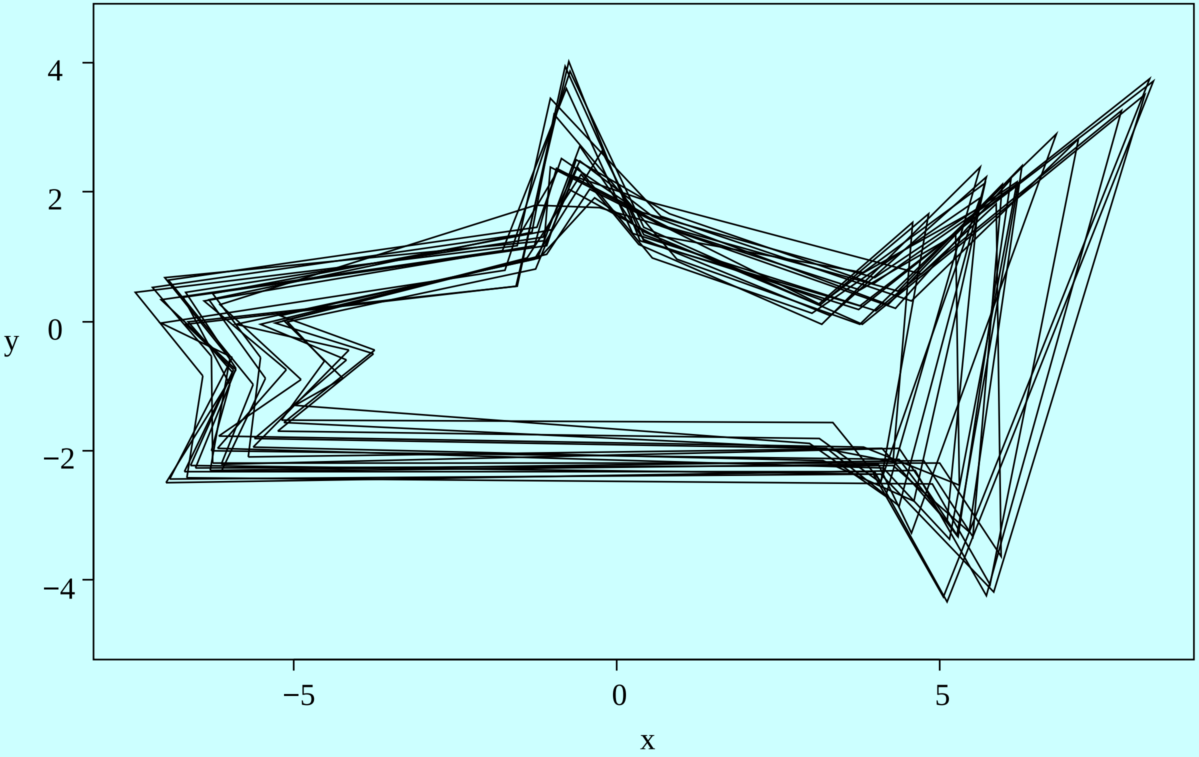
Red lines show perfectly correlated normal change

- 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.

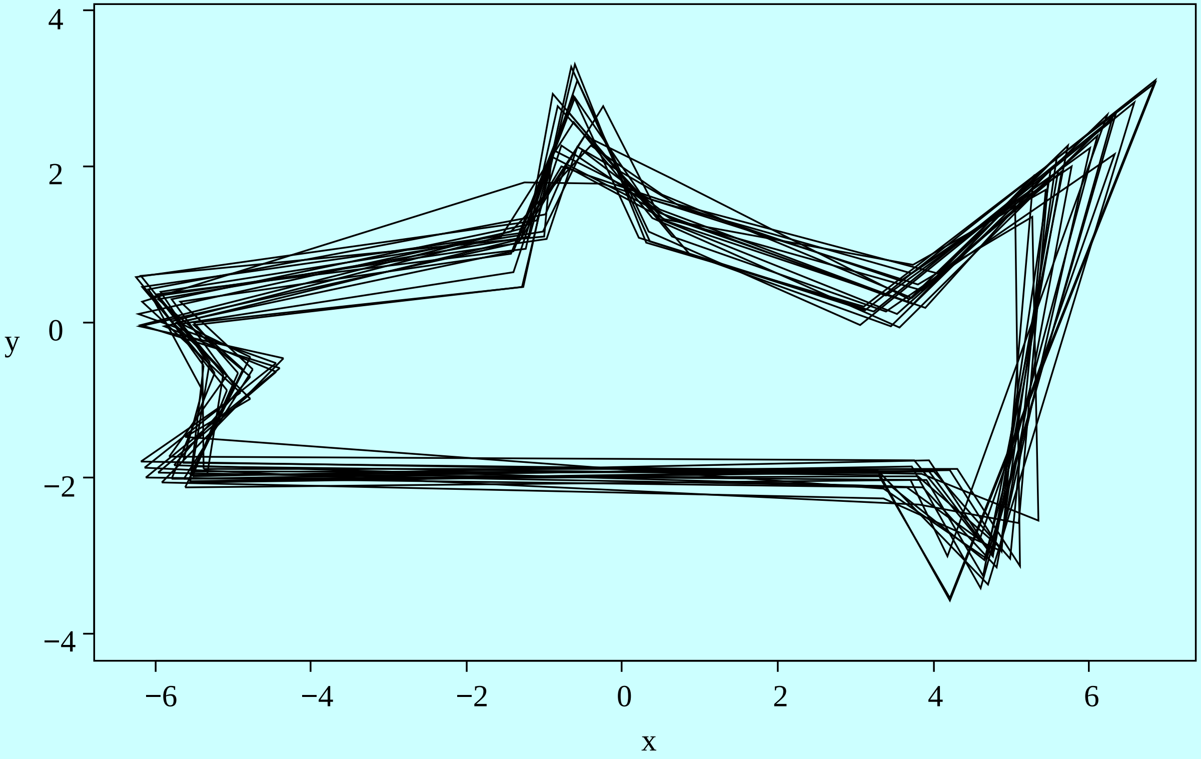
The true tree for one of the 100 data sets



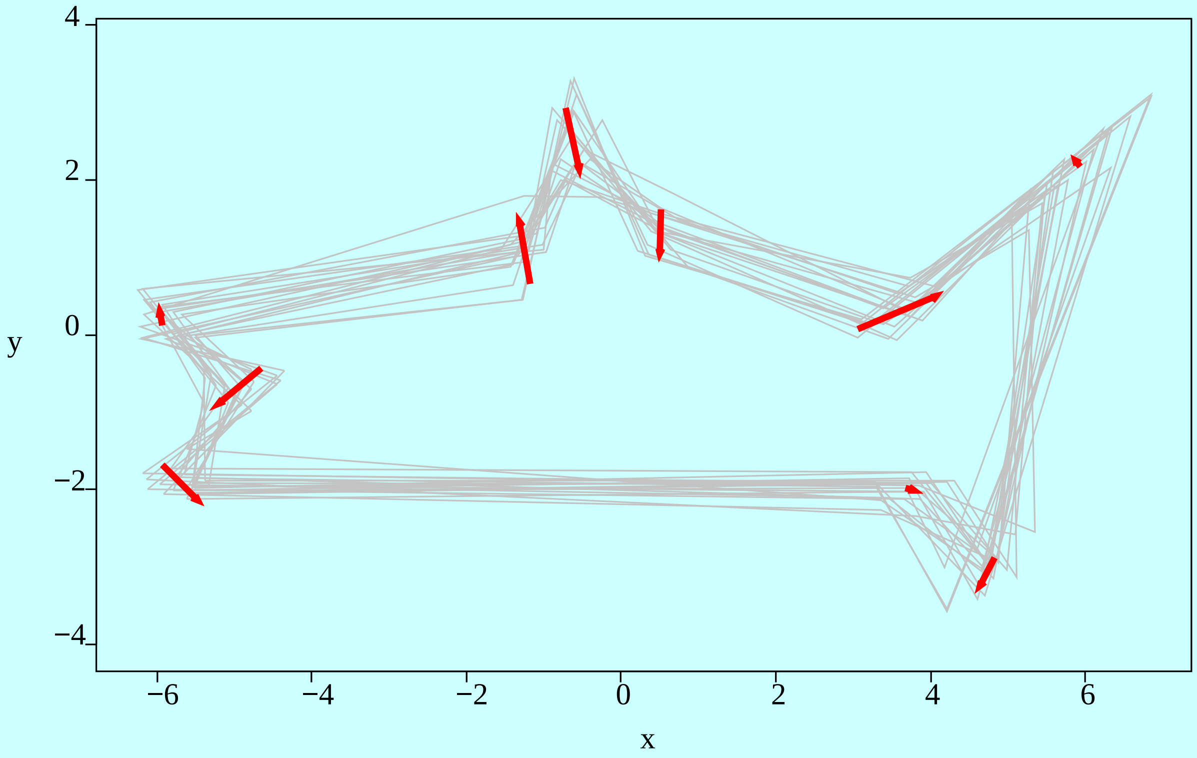
20 of the 100 fishes from data set #2, centered and rotated



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

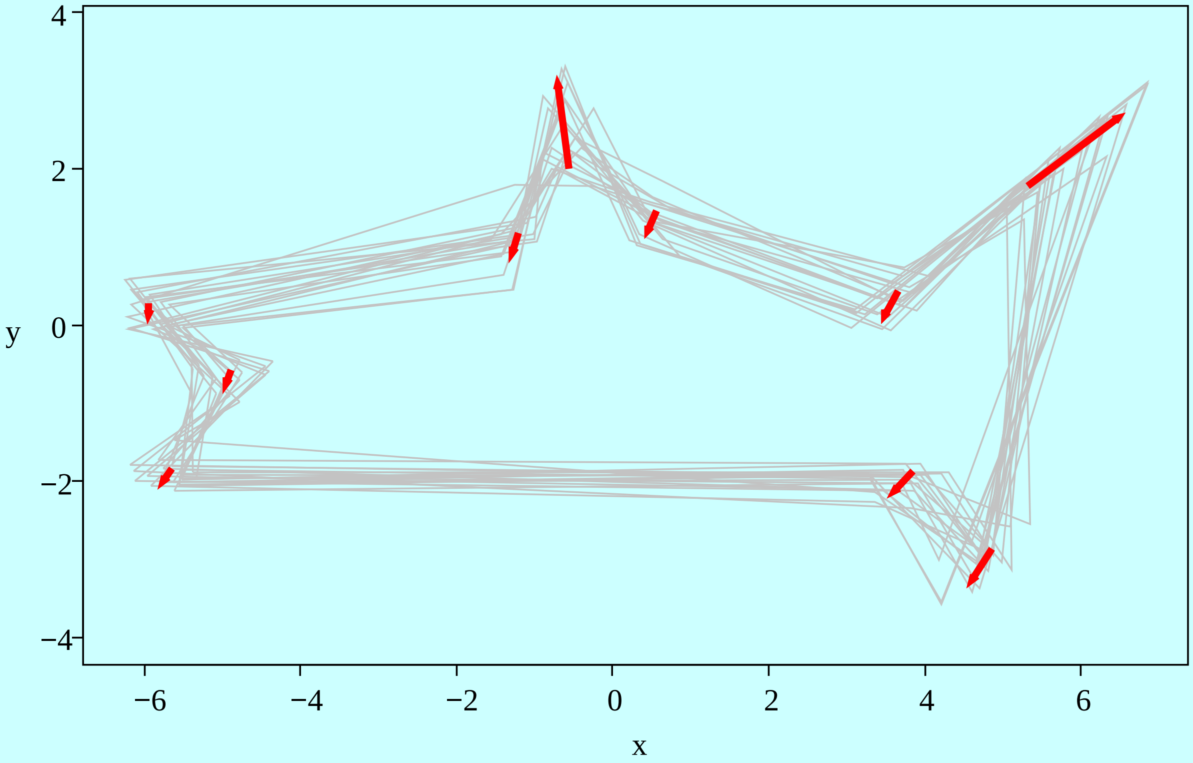


First PC 1 for data set #2



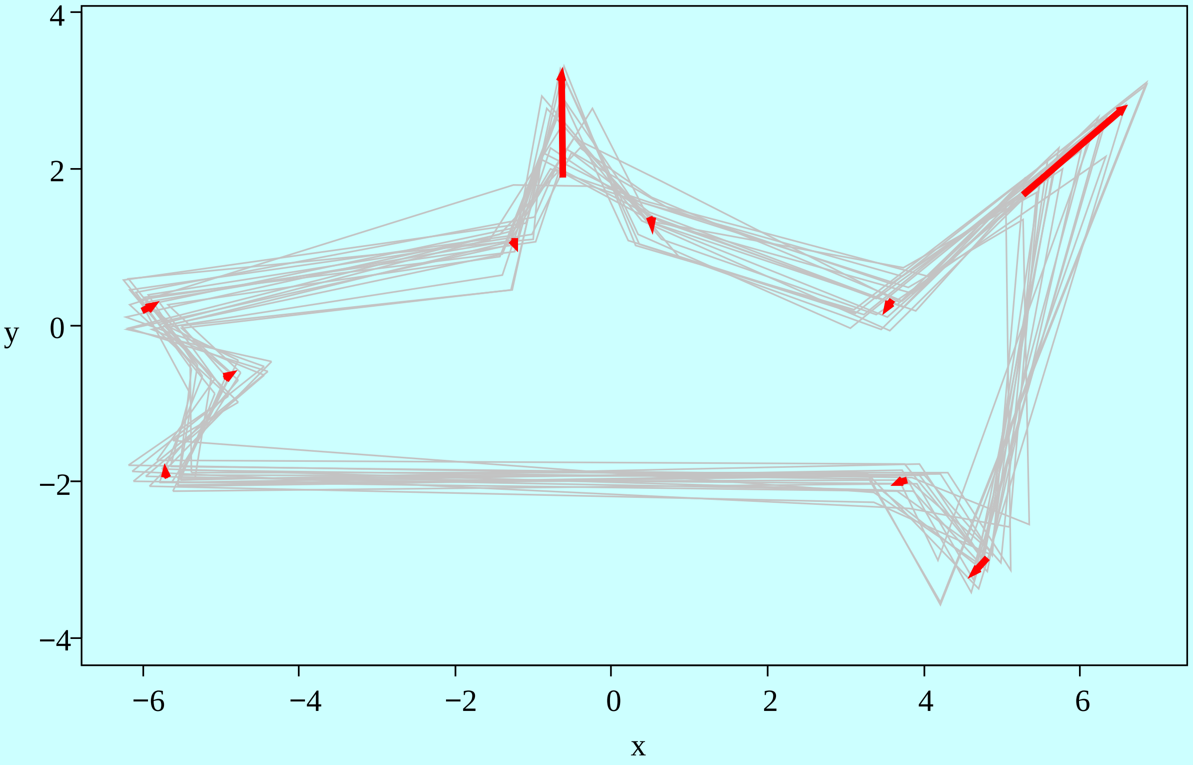
This principal component shows both size changes and the fin extensions, and it is not easy to see which is which.

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.

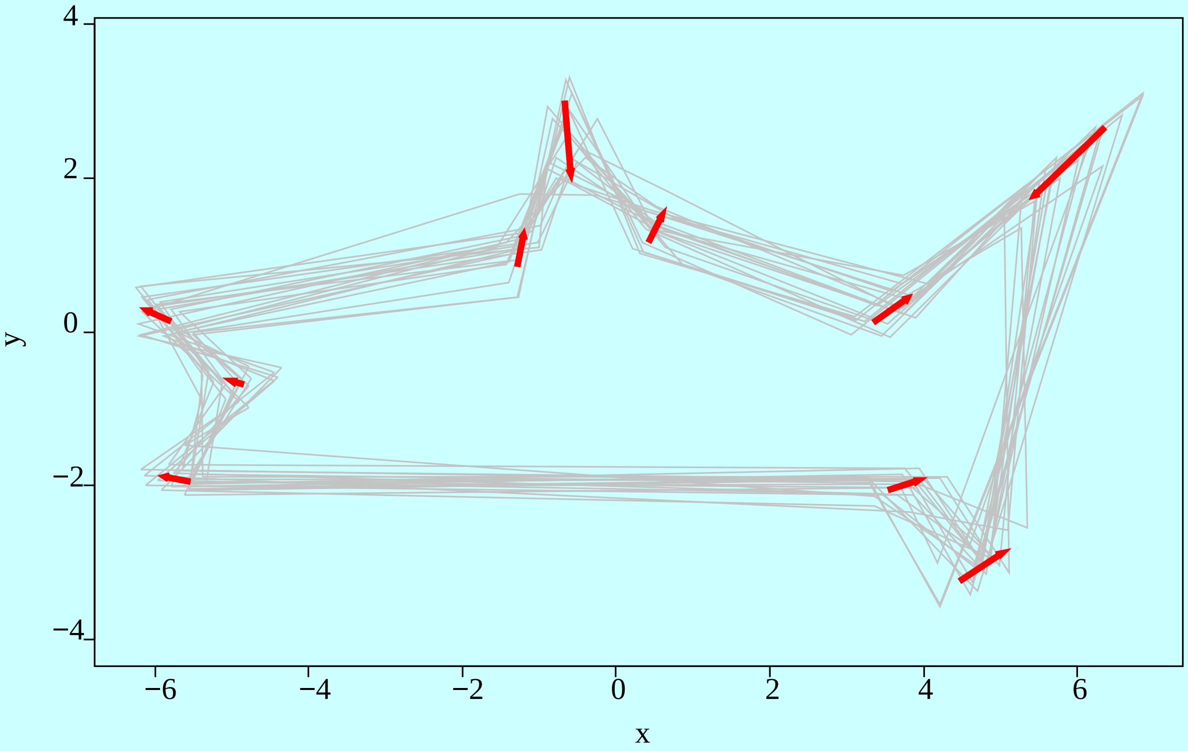
Making the first shape PC sparser by “medianizing”



To make PC1 be sparses we can add in a little location (not forcing the changes to maintain the centroid supeposition). 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 and taken out size by standardizing centroid size, which is affected more by the fin component in the MMC methods.

References

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30 (2):** 314-334. **[One of Russ's major papers using the constant-variances approximation]**

Stebbins, G. L. 1950. Variation and Evolution in Plants. Columbia University Press, New York. **[Describes selective covariance and cites Tedin (1926) for it]**

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19:** 445-471. **[Review]**

Felsenstein, J. 2002. Quantitative characters, phylogenies, and morphometrics.pp. 27-44 in "Morphology, Shape, and Phylogenetics", ed. N. MacLeod. Systematics Association Special Volume Series 64. Taylor and Francis, London. **[Review repeating 1988 material.]**

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.