The Undergraduate Student's Guide to Geometric Morphometrics

Written by Dr. Erika Crispo, Pace University

Version 1 Published November 2020

Acknowledgements

Introduction

Embarking on a new research endeavor can be a daunting task. User guides, books, and published articles are written for an audience that already has some background experience in the field. Undergraduate students like you, who are at the very beginning of their research careers, often struggle to make sense of these documents. Furthermore, students like you often attempt to do so while balancing heavy course loads. Thus, I have written this document to help ease the burden so that you have more time to ponder the interesting scientific questions instead of digging through pages upon pages of documentation.

I assume that you already have some basic familiarity with R before starting this project. I also assume that you have a well devised plan for your experimental design, including the variables you want to collect, the sampling scheme, and of course the questions of interest. Finally, I assume that you have taken a basic course in statistics and have a research mentor that can assist you with more advanced statistical methods. This is not an exhaustive user manual, but rather a guide to help you get started on your journey with geometric morphometrics.

Required materials for photo-taking:

- Any required equipment for capturing and handling the animals
- Camera (cell phone cameras are fine)
- Tray or small transparent cage to hold the animals
- Ruler with mm or cm scale (not inches)
- Two small circular bubble levels that can easily be placed on top of the camera (an example can be found here)

Required (or suggested) materials for analyzing the photos and data:

- Zelditch et al. 2012. Geometric Morphometrics for Biologists: A Primer, 2nd Edition. Academic Press.
  - This book is a thorough guide to the theory behind geometric morphometrics. It also has a companion site including data, scripts, and functions for R.
- Adams & Otárola-Castillo 2013. Methods in Ecology and Evolution 4(4): 393-399.
  - Peer-reviewed paper on the geomorph R package.
- Documentation on the geomorph package (currently version 3.2.1)

- [Vignettes](#) (worked out examples) associated with the geomorph package
- Access to the [Google Group](#) for the geomorph R package
- The latest version of R for [Windows](#) or for [Mac](#)
- The latest version of [RStudio](#)
- Example fish photos and files for practice (available on [GitHub](#)).
  - You can download all the files in a single zipped file by clicking on the green "Code" icon and selecting from a dropdown menu.

Objectives:

The aim of this document is to steer you in the right direction to be able to complete the entire geometric morphometrics component of your project on your own, from start to finish. It points you towards existing documentation and explains steps in an easy-to-understand fashion. It is not intended to be a stand-alone user manual. The ultimate goal is for you to be able to complete the data collection, analysis, and interpretation of your geometric morphometrics project. I will focus only on 2-dimensional (2D) digitization because that is what I have experience with. Once you've mastered 2D geometric morphometrics analysis, you should be able to continue to learn how to perform 3D analysis on your own.

Overview

You are reading this document because you want to study variation in animal body shape. You have interesting questions about differences among populations or species, and you've developed hypotheses you want to test. Thus, the initial step (Step 0) is already complete. But where do you go from there? Zelditch et al. (2012) is an excellent read on the analyses of animal body shape, specifically on geometric morphometric analyses. I strongly recommend that you read it in its entirety if you are interested in a research career that involves geometric morphometrics.

Traditional morphometrics use linear measurements, such as limb length, skull width, body depth, and eye diameter, while controlling for variation in overall body size. You could take many such measurements on each specimen. Each measurement (e.g. eye diameter) constitutes one variable, so you are left with several response (a.k.a. dependent) variables that you are interested in analyzing. You want to conduct only one analysis to obtain a complete picture instead of conducting multiple independent tests for each response variable. When you are analyzing several response variables in a single analysis, this is called a "multivariate analysis" (variables will typically be correlated with each other, which causes a bit of a problem—we'll get to that soon).

The problem is that these variables do not accurately capture an overall representation of the shape of the animal. To better achieve this goal, we can use geometric morphometric analysis, which involves placing dots called "landmarks" on key locations on the animal in photographs, and the relative position of the dots to each other is analyzed for a more complete picture of shape variation. The positioning of the lines connecting each dot to each of the other dots make up the variables (see Figure 1 in Zelditch et al. 2012). The way these dots move in relation to each other among specimens represents shape variation.

We still need to control for variation in overall size among specimens, and so in geometric morphometric analyses we calculate the 'centroid size' for each specimen. Think of the centroid as being the center of a bicycle wheel. The center is still the center regardless of how large the wheel is. But larger wheels will have longer spokes connecting the center to the edge of the wheel. With animal body shapes, we do not have wheels, but we do have an outer frame of the animal. The larger the animal, the longer the 'spokes' that we can draw from the center point to each of the landmarks that we placed.

After placing the landmark dots on the specimens in the photos and using R software to analyze how the dots move relative to each other among specimens, we will notice that, as with traditional morphometrics, some of the variables are correlated with each other. For example, we might notice that as a fish's body gets deeper, its caudal peduncle gets shorter (that's the bony part of the tail before the tail fin). Thus, we can do some fancy statistics that will take all our variables and pull out elements of each of the variables that are uncorrelated with each other, to obtain new variables that are uncorrelated with each other. These new variables are called principle components. It all seems very confusing now, but it will make more sense once you have your data and get to the data analysis stage.

Checklist of steps:

- ☐ Ask interesting question, devise hypotheses, collect specimens
- ☐ Take photos of the specimens
- ☐ Place landmarks on the specimens in the photos at key locations
- ☐ Analyze how the landmarks vary in relation to each other among specimens
- ☐ Perform statistical analyses to test for significant differences in body shape among populations, species, or whatever your hypotheses are about
- ☐ Plot components of variation in shape that are not correlated with each other (i.e. the principle component scores)

STEP 1: Taking photos

I won't spend too much time discussing how to take the photos because the nuances of the photo-taking methods depend on the taxonomic group being studied. I've only worked on fishes before, which are relatively easy to photograph because of their laterally (side-to-side) flattened body plans.

Start with a flat surface such as a tray or a small animal cage. Trays work well for fishes. Make sure that you include two important elements in the photo: (i) a ruler with a millimeter or a centimeter scale, (ii) a label including a unique identifier for the specimen (Figure 1). I use waterproof paper (such as Rite in the Rain) for the label, particularly if photographing aquatic animals. Only use pencil on the waterproof paper because ink will run when wet.

For fishes or any other flat animal, simply place the specimen down on the tray or cage bottom. I find there is no need to anesthetize most fishes. Make sure you do the following:

- Place every specimen on the same side. It does not matter if you choose the left or the right side, just be consistent.
- The most important thing is that you are CONSISTENT in how you place your specimens for the photos. As examples:
  - Fins should be splayed out so that it is easy to see where the fin spines 'insert' (i.e. connect) to the body (Figure 1).
  - Mouths should be closed, and limbs and tails positioned in the same way for each specimen.
- Place the bubble level on the tray or cage bottom to make sure it is completely level with the ground.
- Hold your camera over the tray/cage, with the level placed on top of the camera to make sure that the camera is pointing DIRECTLY down at the specimen.
- Take multiple photos in case one or more do not turn out well.
- Save all the photos as jpeg files in a single folder on your computer.

Figure 1. An example of an adequate, but not excellent, photo. Go through all the above steps to confirm whether they have been successfully achieved for this photo. There is only one problem, and that is that the tail is slightly tilted downwards.

STEP 2: Digitizing the photos

The next step, and the first step in R, is to place the landmarks on the specimens in the photos (i.e. digitization). I have given you four photos from one of my previous studies (Hamel & Crispo 2016) for practice. I have found that the digitization process works better in R than in RStudio. I have had glitches with RStudio crashing during this process, and it is easier to maximize the photo size in R. So, try to do Step 2 in R and not in RStudio.

Install and launch the geomorph package in R.

Go to the GitHub site for the geomorph package (Adams & Otárola-Castillo 2013). Scroll down to the end of the site, where you see instructions to "install the current version of geomorph R-package from Github using devtools." This way you will install the most recent version with any errors fixed. Note that I initially had trouble installing the package, so I installed the most recent version of R (4.0.0 at the time) and then the problem was solved. DO NOT install the Beta version!

```
install.packages("devtools")

devtools::install_github("geomorphR/geomorph", ref = "Stable", build_vignettes = TRUE)

library(geomorph)
```

Organize your photos.

Choose the best photo for each of your specimens and save those files separately in a folder. Make that folder your working directory in R (remember, this is not a manual for how to use R, so I will assume you already know how to do basic things like set a working directory). Chapter 15 of the Quick Guide to Geomorph 3.0 does a nice job at explaining how to digitize landmarks in the geomorph package (although this Quick Guide is no longer updated, so don't use it as your main guide for using the package). Create a file list that contains all your photos. Here is an example using photos from Hamel & Crispo (2016) for you to use for practice:

```
filelist <- c("imageLake5.jpg", "imageLake48.jpg", "imagePond5.jpg", "imagePond6.jpg")
```

Placing landmarks.

You will need to find a resource with information on the best placement for landmarks on your species. Even within a taxonomic group (e.g. fishes), there will be differences with respect to the accepted best practices for landmark placement. My student and I worked with pumpkinseed sunfish (*Lepomis gibbosus*) and used Figure 2 from Jastrebski & Robinson (2004) as a reference. (Also see Figure 2, below.) You will see that this study used 15 'homologous' landmarks, meaning points on the fish that had common evolutionary origins among specimens. The easiest landmarks to place in fishes are those at the fin insertion sites (landmarks 2,3, and 7–11 in Jastrebski & Robinson 2004), or at the end of bones such as those in the jaws (landmarks 1 and 12 in Jastrebski & Robinson 2004) and operculum (landmark 15 in Jastrebski & Robinson 2004). Trickier are the placements of landmarks 4–6 on the tail and 13–14 on the

eye. You will see from Figure 1 (this document, above) that there is no clear insertion of the caudal (tail) fin and so a bit of estimation must come into play for landmarks 4 and 6. Landmark 5 should be placed where the lateral line meets the posterior end of the hypural plate (i.e. the tail bone in the fish). This spot is often impossible to see on photos, so pins can be placed on dead/preserved specimens (but don't put a pin into a live specimen!). To accurately place the landmarks on the eye, you should draw a line on each of the photos using another program such as Paint. You will need to draw a straight line from landmark 1 through the diameter of the eye. Of course, this is just one example using one species of fish and the details of how this should be done might differ for your species.

You will need to determine the best course of action for placing landmarks on your specimens given the available literature. You do not want to use too many landmarks because the more you have, the larger your number of variables, and the greater your chances of 'over-fitting' your data. What that means is that as your number of variables increases, it becomes increasingly probable that you will find significant relationships among variables just by chance alone (and not because there is any real meaning for those relationships).

When placing the landmarks in R, you will get a feel for how it works. First, enter the following code, replacing the number of landmarks (nlandmarks argument) with the number that you'll use. The scale is in mm, currently set at 10mm = 1cm.

```
digitize2d(filelist,
  nlandmarks=15,
  scale = 10,
  tpsfile = "newFile.tps" ,
  MultScale = FALSE,
  verbose = TRUE
)
```

Note that you can change the size of the Graphics window in R without changing the shape of the image. So, go ahead and maximize its size for ease of placing the scale and the landmarks.

Your first step after executing the above command (and after adjusting the image size on your monitor) will be to set the scale. You should see the spinning wheel prompting you to proceed with that step. Locate the ruler on the image and click on one of the scale markers. Then, click again exactly 1cm away from your first click. You will see two red dots where you clicked, connected by a line. If your placement of your clicks does not accurately reflect 1cm, type "n" in the console for no and hit Enter, and repeat until you are satisfied (at which point, type "y" and Enter to proceed to the next step).

R will now ask you to place Landmark 1. Click on the location on the photo where you want Landmark 1 to be placed. If you are happy with this placement, go back to the console and type "y" for yes and hit Enter. Proceed until all the landmarks are placed. You will then be asked if you want to continue with the second photo. Keep going until all your photos are digitized. As you go through the pumpkinseed sunfish photos that are provided to you as an example, think about what is wrong with each of the photos and what actions you will take to ensure that your photos are of better quality.

Open newFile.tps in a text file reader such as NotePad or EditPad Lite. You will see that along with the image ID and a scale for each specimen, there are also two columns of numbers, each

with a length that equals the number of landmarks (15, in this example). Each pair of numbers represents the position of each landmark on the Y and X axis, respectively (if you imagine your photos each being on a grid with a vertical Y and a horizontal X axis). Make sure that you do not have any missing data (NA), which would indicate that you made a mistake in your landmark placements. See below for information on how I correct mistakes.

Correcting mistakes.

You might ask, what happens if I make a mistake with one of the landmarks, for example by typing "y" when I meant to type "n"? There might be an easy way to correct this mistake that I am not aware of, but here is what I do. Let's say I made a mistake digitizing imagePond6:

```
digitize2d("imagePond6.jpg",
  nlandmarks=15,
  scale = 10,
  tpsfile = "newFile2.tps" ,
  MultScale = FALSE,
  verbose = TRUE
)
```

Now I can re-digitize this one photo only, with the data collected in a new file, newFile2.tps. I can then cut and paste ALL the data for this one specimen into my original newFile.tps, replacing the data for which there were one or more mistakes. DO NOT replace data for a single landmark only—it must be all or nothing, due to small differences that might occur in your scale placements.

Do you need to use semilandmarks?

Occasionally, there may not be enough homologous landmarks on a specimen type to accurately capture its shape. When we place landmarks that have no identifiable homologous structure among specimens, we call these 'semilandmarks'. It is possible to place landmarks along the perimeter of the specimen. This can be done by spacing landmarks evenly along the perimeter between two homologous landmarks, or by drawing a line between two homologous landmarks and then rotating that line ninety degrees and placing the semilandmark where the line meets the specimen's perimeter. If this is confusing to you, do not worry! It is best to avoid semilandmarks, if possible. The reason is that the positioning of semilandmarks depends on the positioning of other, homologous landmarks. Each time a landmark is digitized, there is some error that comes into play—if you were to repeat the digitization process, you would be unlikely to get the exact same landmark coordinate measurements each time. So, you are compounding this error by using landmark placements to place semilandmarks. If you do think that semilandmarks are important for your project, see Bardua et al. (2019) for more information.

STEP 3: Processing the data

You now have your data and it needs to be organized and processed before it can be analyzed using statistical methods. After completing Step 2, you have a file called newFile.tps containing two columns (called 'vectors') of data, your two variables. They represent the position of each landmark on each of the two axes (i.e. vertical distance along the Y axis and horizontal distance along the X axis). But, as you will have noticed, each fish was placed slightly differently in each photo. For example, the pond fish were placed closer to the right of their photos than the lake fish. In some cases, specimens might be rotated in the photos such that their ventral surfaces are not aligned to the bottom of the photo. Furthermore, the camera was held at various distances from the specimens, and therefore the scale differed in each photo. Thus, are variables cannot be compared among specimens in their current state. We must transform the data first.

You can now start a project in RStudio. Use the following code for RStudio to read your file with the landmark data. The `readcurves` argument is set to FALSE because we do not have any semilandmarks in our example.

```
myData<-readland.tps("newFile.tps", specID = "ID", readcurves = FALSE, warnmsg = TRUE)
```

Generalized Procrustes Analysis—aligning the landmark coordinates so that data are comparable across specimens.

There is a lot of jargon in science that can feel overwhelming. Generalized Procrustes Analysis (GPA) simply means that we are going to move the landmark positions around as if the specimens were all oriented the same way when we took the photos. The landmark positions are also going to be scaled to account for the fact that our camera was closer to some specimens than to others. This procedure involves 'translating', 'scaling', and 'rotating' the coordinates. See Figure 3.1 in Zelditch et al. (2012) as a visual example of what this means. Perform this procedure using:

```
myGPA<-gpagen(myData, curves = NULL, surfaces = NULL, PrinAxes = TRUE,
        max.iter = NULL, ProcD = TRUE, Proj = TRUE, print.progress = TRUE)
```

To see what each of these arguments refers to, use `?gpagen`. Curves and surfaces refer to semilandmarks along the perimeter and the surface of the specimen, respectively. In our example we do not have any semilandmarks and so these arguments are NULL.

Use the summary function – `summary(myGPA)` – to see what you've just done. You now have new coordinates which are the new standardized positions of each of the landmarks that you placed on your photos. You will get landmark coordinates for the 'consensus configuration' which refers to the average shape—that is, you will get average landmark coordinate positions, along the X and Y axes, for each of the 15 landmarks in our example. Also use `myGPA$coords` to see your new landmark coordinates for each specimen, after translating, scaling, and rotating. Finally, use `myGPA$Csize` to get measurements of size of each of your specimens, after scaling has been accounted for (i.e. considering that the camera was a different distance from each

specimen when photos were taken). To visualize how different specimens are from each other, use:

```
plotAllSpecimens(myGPA$coords,mean=TRUE)
```

You will see in this example that the insertion of the dorsal fin (landmark 2) and the top of the tail varies from specimen to specimen (Figure 2). The variation in the tail (landmark 4) is due to error in placing the fish on a straight line on the tray when taking the photos. As you can see in the photos, the tail for Lake5 is tilted downwards whereas it is tilted upwards for Pond5. DO NOT MAKE THIS MISTAKE! As noted in Step 1, above, you must be consistent in how you place the specimens in the photographs. If you want to compare each specimens' shapes to the average (i.e. consensus) shape, use `mean=TRUE` in the above code (or use `mean=FALSE` if you don't want to see the consensus shape).
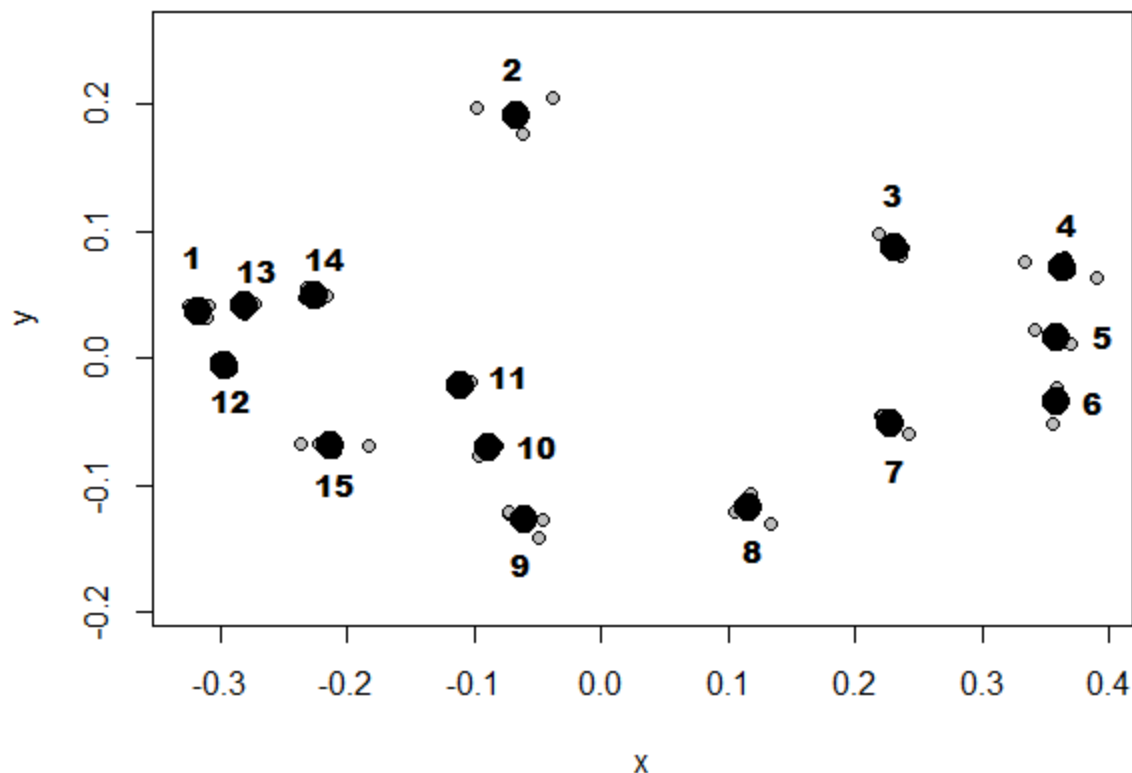

Figure 2. Plot of landmark coordinates for each specimen (grey points) and the consensus (average) landmark coordinates for the data set (black points). The numbers were added in by me in Paint and won't appear in the figure that you make in RStudio.

Questions?

Two questions you might have are: (i) Why are some of the new landmark coordinates negative, and (ii) What does 'size' mean? First, some landmark coordinates are negative because all the landmarks are positioned relative to a baseline located on the specimen during Generalized Procrustes Analysis. So, some landmark coordinates are to the top and right of the baseline and are positive, whereas some landmark coordinates are to the bottom and left of the baseline and are negative. Second, the size represents the distance of landmark coordinates from the center point of the specimen, called the centroid. The centroid is located at the point where the distance of each landmark coordinate to that point is minimized (you can think of it in a similar way to how a regression line is plotted on a scatterplot in a way that minimizes the distances of data points to that line). See the Overview, above, for more information on centroid size.

STEP 4: Analyzing the data

You now have landmark coordinates that can be compared directly among specimens and analyzed using statistical methods. The purpose of this guide is not to teach the fundamentals of statistics. You certainly should take as many courses as you can in statistics, including at the introductory and advanced (multivariable and multivariate) levels. Instead, I will provide you with the practical tools you need to get started with the analysis of geometric morphometric data.

Since everyone's research question is slightly different, there is no 'one size fits all' solution to data analysis. This basic example using pumpkinseed sunfish examines variation in shape between two water bodies—a lake and a pond. You might also want to include other variables such as sex, for example. Or you might want to do something more complex, like construct a phylogeny based on shape variation. You can certainly perform these analyses using geomorph. You can also use other programs and R packages to analyze your data. The goal of this guide is to set you on the path to be an independent learner of geometric morphometric analyses.

Geometric morphometric analyses are multivariate.

In geometric morphometric analyses, multiple homologous data points (i.e. landmark coordinates) are obtained per specimen and that means that these analyses are 'multivariate' (i.e. multiple response variables). Note that this distinction varies from 'multivariable' analyses, which include multiple explanatory (also known as independent or predictor) variables. (Confusing, I know!) In our case we have one categorical explanatory variable and that is the site from which each specimen came (lake or pond). But we also have one additional variable, a numeric explanatory variable, and that is the centroid size of each specimen. Since this variable is numeric, we call it a 'covariate'. And since we have multiple response variables, at least one categorical explanatory variable, and a covariate, we will conduct a multivariate analysis of covariance (MANCOVA).

Creating variables.

We already have a data set (called `myGPA`) that contains the landmark coordinates as well as the centroid size for each specimen. We did not yet indicate which site each specimen was obtained from. The order in which we digitized specimens was: Lake5, Lake48, Pond5, Pond6. So, we need to indicate that the first two specimens are from the lake site and the last two specimens are from the pond site:

```
myGPA$site=c("lake","lake","pond","pond")
```

We also must make sure that RStudio is reading this new variable as a categorical variable called a 'factor':

```
is.factor(myGPA$site)
myGPA$site<-as.factor(myGPA$site)
myGPA$site
```

Specifying the model.

Next, we specify our model to determine effects of site and centroid size on our landmark coordinate positions:

```
model<-procD.lm(myGPA$coords~myGPA$site+log(myGPA$Csize),iter=999)
```

Note here that we log transformed the centroid size. It is usually necessary to do so because as specimens get larger, the possibility for fluctuations in absolute body size increases (as how weight of elephants varies more than does the weight of mice). By log transforming the data, all data values are reduced as a percentage of themselves, and so large data values are reduced more than are small data values. You should construct histograms for centroid size in each site—if data are skewed to the right, log transformation is necessary. Since our example data set is so small, let's lump specimens from both sites together (but DO NOT DO THIS in larger data sets—you should create one histogram per level of the categorical variable). You see that log transformation improves normality:

```
hist(myGPA$Csize,xlab="Centroid Size",ylab="Number of fish",main="")
hist(log(myGPA$Csize),xlab="Log transformed centroid size",
     ylab="Number of fish",main="")
```

We can use `?procD.lm` to determine what arguments are available for this model function. We've used the default settings for each argument except for `iter`. This argument refers to the number of iterations used to randomize the data via a permutation analysis. Permutation simply means that the data get scrambled up (in this case, 999 times), each time calculating the difference between the two sites while controlling for body size, and determining whether the actual differences in body shape between sites is more different than expected based on random chance alone (the randomness reflected in the scrambling up of the data in each of 999 iterations).

The Sums of Squares type (`SS.type`) is something of important consideration but is beyond the scope of this guide. If your sample sizes are equal among groups within your categorical variable (our two sites, in this case), you do not need to worry about this option. If your sample sizes are unequal, the order in which you enter variables into the model (as above) will change your results and you will need to be mindful of which Sums of Squares type you use. Since we have the same number of specimens per site, we can stick to the default for that argument. You can change the other arguments as you see fit, but let's keep it simple for now.

Making sense of the results.

Let's look at our results:

```
anova(model)
```

You will see the degrees of freedom, sums of squares, mean squares, and F values for the effects of site, log transformed centroid size, residuals (i.e. leftover variation), and the total variation. This result output is similar to what you'd get from any ANOVA output. The Rsq is the

R^2 value and represents the proportion of the variation in body size that is related to each of the variables. My output and yours will be a little bit different due to slight differences in where we placed our landmarks on the fish photos. In my analysis, 43.46% of the variation in the data is due to differences between sites, and 42.59% of the variation is due to differences in fish sizes. The Z values are the effect sizes and the P values are a measure of significance (with P<0.05 typically meaning that there is sufficient evidence that the variable in question has a significant effect on the response variable). Our P values are greater than 0.05 for both site and the log of centroid size, and so we cannot conclude that either of these variables has a significant impact on body shape in this study. But remember that significance levels are dependent on sample sizes—with n=2 for each of the sites, it is not likely that we would be able to obtain significant results even if there really were biologically relevant differences among sites in the entire population of fish. We just weren't able to sample enough of that variation.

The effect of body size on body shape.

The issue of body size, or in this case centroid size, deserves more attention. The change in size or shape of a body part as an organism gets larger is referred to as 'allometry' and this phenomenon must be considered in any study of morphometrics. There may not be any evidence of allometry, which would be the case if our P value for the effects of centroid size is >0.05. In this case, we say that the relationship between body shape and size is 'isometric' (but only if we had a reasonably large sample size, which in this example we do not).

Allometry might differ among groups – in this case, it might differ between the two sites. When the effect of size on shape differs among groups, we call this difference in relationship an 'interaction'. With only two specimens per site, it is not possible to obtain significant evidence for an interaction. To test for an interaction in any model, we simply replace the plus symbol with an asterisk, like this:

```
modelInteraction<-procD.lm(myGPA$coords~myGPA$site*log(myGPA$Csize),iter=999)
anova(modelInteraction)
```

We can see that because the P value for the term `myGPA$site:log(myGPA$Csize)` is >0.05, we conclude that we have no evidence for a significant interaction between site and centroid size – that is, no evidence that the effect of centroid size on body shape does not differ significantly between sites. This is expected when our sample sizes are unreasonably small.

STEP 5: Making figures to show shape variation

Last but not least, we want to make figures to visualize shape change among specimens. Given that changes in landmark positions are highly correlated, this process typically involves calculating principle component (PC) scores for each specimen. As an example, elongation of the body leads to elongation of the dorsal fin and a wider spacing between the pectoral and anal fins in fishes. As such, it would not be appropriate to consider landmarks 2, 3, 7, 8, and 9 as being independent of one another (Figure 2 in Jastrebski & Robinson 2004). Thus, we use principle component analysis (PCA) to account for this fact.

Principle components analysis.

A detailed description of how PCA is performed is beyond the scope of this guide. You can perform PCA on any multivariate data, not only geometric morphometric data. In short, we want to take our variables and reorganize the variation associated with these variables so that we have a new set of variables that are uncorrelated with each another. In our case, our variables are our landmark coordinates from the Generalized Procrustes Analysis. Some changes in landmark positions represent a shortening and widening of the body, for example. Others might reflect changes in the positioning of the pectoral fin. The PCA removes variation associated with different components of shape change and puts them into new variables called Principle Components (PCs). Some PCs account for a larger percentage of the overall variation in shape change than others. Often, the first PC will be responsible for allometric variation—that is, shape variation that is due to differences in size among specimens. Perform a PCA:

```
PCA <- gm.prcomp(myGPA$coords)
PCA
```

You should get values for three separate PC variables – PC1, PC2, and PC3. The proportion of variance is the amount of variance (or scatter) in the data that is explained by each of these new variables.

Plotting principle components.

The above information will make the most sense once we start plotting the data. Because our first two specimens are lake fish and the second two are pond fish, we use `bg = c(rep("red", 2), rep("blue", 2)` so that the first two data points will be red and the second two will be blue. After using the code below, you will need to click on your figure where you want the legend to be placed.

```
plot(PCA, pch = 22, bg = c(rep("red", 2), rep("blue", 2)), cex = 1.5)

legend(locator(1),levels(myGPA$site),pch=15, cex=0.8, col=c("red","blue"))
```

You will see that we plotted the first variable, PC1, on the X axis and we plotted the second variable, PC2, on the Y axis (Figure 3).
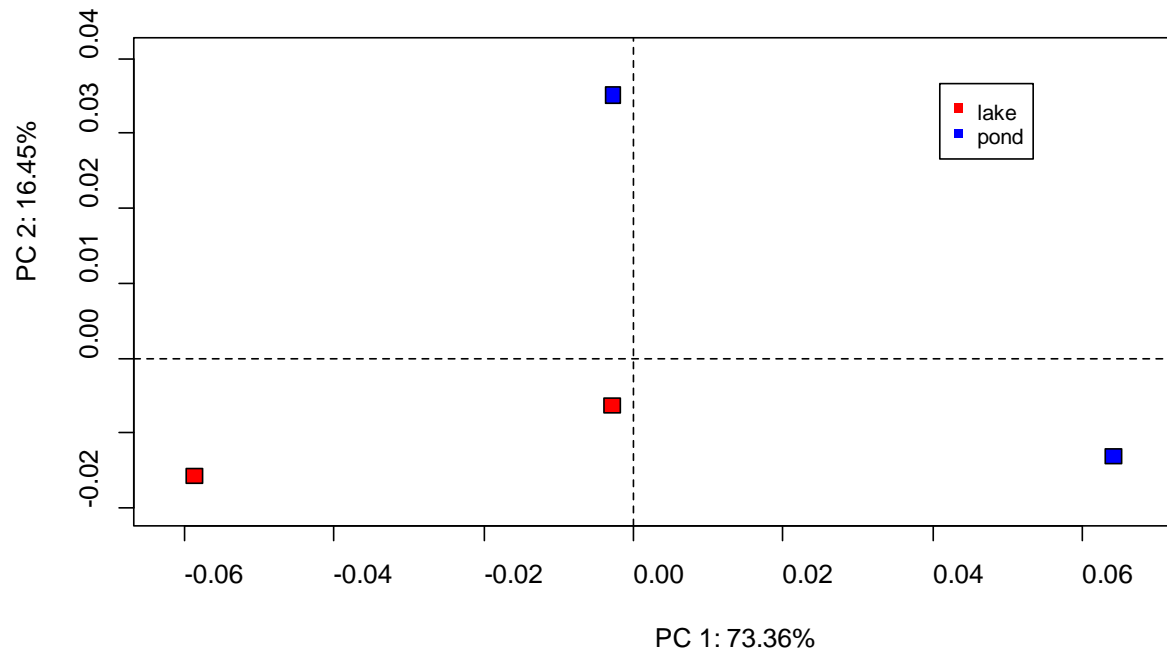
Figure 3. Plot of principle component 1 and 2 scores for each of the four fish specimens, including those from the lake (red points) and from the pond (blue points).

I have not found a way in geomorph to obtain a list of the PC scores presented in Figure 3. I was, however, able to use a different R package called Morpho to do so. You might want to have the PC scores so that you can perform your own analyses separate from what geomorph can perform, or to make graphs outside of the geomorph package. The following code will provide you with the values presented in Figure 3 (`$bescores` for each of the four specimens):

```
library(Morpho)
RW<-relWarps(myData, scale = TRUE, CSinit = TRUE, alpha = 0,
        orp = TRUE,  noalign = FALSE)
RW
```

You will hear about 'relative warps' when you read Zelditch et al. (2012) or other sources about geometric morphometric analyses. When the argument `alpha = 0` is used, as above, the relative warp and PC scores are identical. You can use `?relWarps` to see what the arguments are and how to adjust the options.

Plotting grids showing shape changes.

The graph we just made it not very useful if you don't know what changes in shape PC1 and PC2 represent. You can create grids showing changes in body shape. The code for doing this might differ slightly depending on the version of geomorph and/or R that you're using. I was able

to plot grids showing expected shape for the maximum and the minimum values of each of the principle components using:

```
ref <- mshape(myGPA$coords)
plotRefToTarget(M1=ref, M2=res$pc.shapes$PC1min, method="TPS") #PC1 maximum value
plotRefToTarget(M1=ref, M2=res$pc.shapes$PC1max, method="TPS") #PC1 minimum value
```
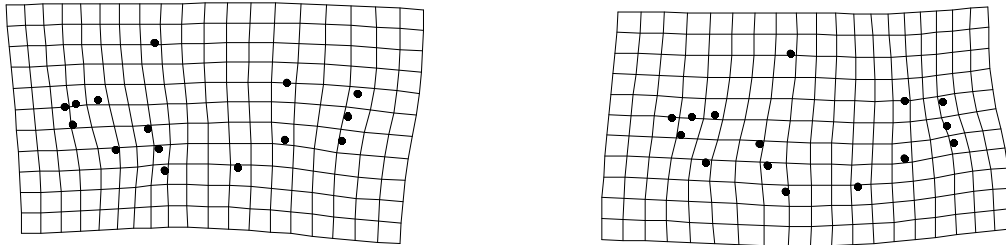

Figure 4. Expected fish shapes for fish with the lowest (left panel) and highest (right panel) principle component 1 scores.

You see that PC1 represents a shift in the tail from a downward tilt (low PC1 value) to an upward tilt (high PC1 value; Figure 4). This shape change is most likely due to error in placing the fish in a straight plane before taking photos. If you look at the photos, you will see that the first specimen (Lake5) has a downward tilting tail, and the third specimen (Pond5) has an upward tilting tail. These specimens correspond to the first and third rows for the PC scores obtained via `RW`, above. You will see that these are the lowest and highest PC1 scores, respectively, confirming what we see in Figure 4.

Now, change the code above to represent PC2 instead of PC1 (by changing `PC1min` and `PC1max` to `PC2min` and `PC2max`, respectively). You will see that PC2 represents a shortening and fattening of the fish. This shape change is more biologically relevant and interesting.

A final analysis we can do is a correlation analysis to determine whether any of the PCs are correlated with body size:

```
variable1<-RW$bescores[,1]
variable1
cor.test(variable1,myGPA$Csize)
variable2<-RW$bescores[,2]
variable2
cor.test(variable2,myGPA$Csize)
```

We see that there is no evidence of a correlation between body size and body shape for each of the first two principle components because the P values for each analysis is >0.05.

OPTIONAL: Traditional morphometrics

This concludes our introduction to geometric morphometrics. Of course, there is much to learn that is not included in this short guide. But I hope that you now feel a little less daunted about the process and how to get started. You might also want to perform traditional morphometrics. I will leave you with a very short description of what would be involved in this type of analysis.

Summary of steps for traditional morphometric analyses:

- ☐ Take measurements from specimen using calipers, or from photos using a program like ImageJ, including total body size
- ☐ Standardize measurements to account for differences in body size
- ☐ Perform a principle components analysis to create a new set of variables that are uncorrelated with one another (optional)
- ☐ Perform statistical analyses to test your hypotheses

Taking measurements.

You could take measurements from photos using a program such as ImageJ. Or, you can take measurements from the actual specimens using calipers. If you plan to perform both geometric and traditional morphometrics, the latter might be the best option to prevent the perpetuation of errors stemming from misplacement of specimens in the photos. The actual type of measurements you take will, again, depend on the type of specimen you are working with.

Size standardization.

After taking the measurements, you will need to standardize them to a common body size. Make sure that you have a measure of body size that is in the same dimension as the measurements of interest. The measurements will typically be in one dimension (i.e. linear measurements such as head width, eye diameter, or limb length), in which case body length is the appropriate measure for size standardization. You then will perform simple linear regression analyses to test for effects of body size (explanatory/independent variable) on each of the measurements of interest (response/dependent variables). The slope of each linear regression will be used for the size standardizations. Standardization is performed using an allometric equation (modified from Reist 1986):

$$Y_{std} = Y_{obs}(M_{avg}/M_{obs})^b.$$

In this equation, Y_{std} is the size standardized measure that you want to obtain independently for each specimen. Y_{obs} is your measurement (e.g. head width) for a given specimen, M_{obs} is the body size of that specimen, M_{avg} is the average body size across all specimens, and b is the slope for the linear regression (as above). You need to perform this standardization for each specimen separately, which can easily be done in R.

Data analysis.

You now have several size standardized variables across a set of individual specimens. Many of these variables will be correlated with each other. For example, individuals who have longer forelimbs might also have longer hindlimbs. This relationship is independent of any characteristics of limb length that is related to body size. Use dog breeds as an example. A mastiff is longer legged than a Yorkshire terrier because the former is larger overall. But both breeds have relatively longer legs than a dachshund or a corgi, given their body sizes. It is not appropriate to consider each of these measurements as independent variables in this instance. You have two options for data analysis: perform a multivariate analysis of variance (MANOVA) or first perform a principle components analysis (PCA) to obtain a suite of new variables that are each uncorrelated with each other. The latter approach might be useful if you want to analyze different aspects of body shape independently of one another. With respect to fishes, one principle component might reflect a deepening of the body (influencing sustained vs. burst swimming ability) whereas another might reflect changes in the positioning of the fins (influencing maneuverability). It is important to read studies that have been published on your study species to determine what is common practice in your subfield, and to be sure you are using the appropriate analyses to address your research question. It is also crucially important that you ensure that all of the assumptions of your statistical test (e.g. normality of the data in each group, homogeneity of variances and covariances) are met so that you can be confident that your results are valid.

References

Adams, D.C., & Otárola-Castillo, E. 2013. geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods in Ecology and Evolution 4(4): 393-399.

Bardua, C., Felice, R.N., Watanabe, A., Fabre. A.-C., & Goswami, A. 2019. A practical guide to sliding and surface semilandmarks in morphometric analyses. Integrative Organismal Biology 1(1): 1-34.

Hamel, C.I., & Crispo, E. 2016. Morphology of pumpkinseed sunfish (*Lepomis gibbosus*) is related to water body size. BIOS 87(2): 56-66.

Jastrebski, C.J., & Robinson, B.W. 2004. Natural selection and the evolution of replicated trophic polymorphisms in pumpkinseed sunfish (*Lepomis gibbosus*). Evolutionary Ecology Research 6: 285-305.

Reist, J. D. 1986. An empirical evaluation of coefficients used in residual and allometric adjustment of size covariation. Canadian Journal of Zoology 64: 1363- 1368.

Zelditch, M.L., Swiderski, D.L., & Sheets, H.D. 2012. Geometric Morphometrics for Biologists: A Primer, 2nd Edition. Academic Press.