Making heads or tails of combined landmark configurations in geometric morphometric data

Michael L. Collyer[1,*], **Mark A. Davis**[2,*], **Dean C. Adams**[3]

05 May, 2020

[1] Department of Science, Chatham University, Pittsburgh, Pennsylvania, USA.

[2] Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL, USA.

[3] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA.

[*] Correspondence: Michael L. Collyer m.collyer@chatham.edu

Keywords: Morphometrics, centroid size, normalized, landmarks

Short Title: Combining landmark configurations

Ethics declarations

Conflicts of interest: The authors declare that they have no known conflicts of interest.

Acknowledgments

The authors wish to thank A Profico, P Piras, C Buzi, A Del Bove, M Melchionna, G Senczuk, V Varano, A Veneziano, P Raia, and G Manzi, for inspiring the update to the `geomorph R` function, `combine.subsets`, to offer centroid size normalization or user-defined weights as alternatives to calculating relative centroid sizes. We received insightful reviews of a previous version by A Kaliontzopoulou, E Baken, B Juarez, E Glynne, and two anonymous reviewers, and we thank them for their efforts to improve this manuscript. This work was sponsored in part by National Science Foundation Grants DEB-1737895 and DBI-1902694 (to MLC) and

DEB-1556379 and DBI-1902511 (to DCA). All analyses in this paper were performed in `R`, using `geomorph` (Adams et al. 2020; Adams and Otárola-Castillo 2013) and `RRPP` (Collyer and Adams 2018, 2020) libraries. The `combine.subsets` function in geomorph has all landmark combining options used in this paper.

Abstract

Researchers using geometric morphometric methods can be confronted with a need to combine separate landmark configurations from the same research subjects as a more holistic description of organismal morphology. Combining configurations might be valid if single configurations represent separate anatomical structures that can change position with respect to each other or have been shown to be phenotypically integrated, and researchers would prefer to recognize these structures as one set, rather than multiple sets. However, generalized Procrustes analysis (GPA) scales separate configurations to unit size, meaning that in combination, some attempt to relativize the size of configurations should be made. A few recent studies have calculated the relative size of separate configurations in different ways but there has been no formal consideration for the implications of *a priori* judgments for how configuration sizes should be weighted, before the synthesis presented here. We offer a general solution for weighting separate configuration centroid sizes when combining them, which captures the intention of different methods thus far proposed. We also demonstrate that under various conditions, weighting via normalized centroid size is fraught with problems, and should be avoided. By contrast, an unweighted approach that seeks to maintain landmark densities in separate configurations provides reliable results. Nevertheless, researchers should realize that combining configurations creates new configurations with landmark covariances that are arbitrary with respect to any real anatomical features. As such, combining landmark configurations should not be a haphazard enterprise under any circumstances.

Introduction

In recent decades, deciphering patterns of shape variation in anatomical features in evolutionary biology research has become the purview of geometric morphometric methods (GMM; Rohlf and Slice 1990; Adams et al. 2013). In GMM, anatomical shape is characterized by a set of landmarks and semilandmarks that represent the relative position of points, curves, and surfaces, from which non-shape variation is removed (Adams et al. 2013). Typically, each object is represented by a single configuration of points, and the shape variables that result from them may be compared statistically, or associated with other variables of interest (e.g., size, phylogenetic position, ecological variables, etc.). However, in some instances, the anatomical objects may comprise multiple structures, and each may be characterized by its own configuration of landmark points. For instance, researchers investigating the ecomorphology of aquatic feeding may quantify the shape of both the head (food capture and manipulation) and tail (swimming strategy) of organisms; in which case, each specimen is represented by two landmark configurations. For some hypotheses, it may be of interest to combine these configurations in some way to arrive at an overall estimate of morphology, especially if the separate configurations are shown to be integrated. However, how to accomplish this properly is rarely straightforward.

The need to combine landmark configurations is generally spurred by recognizing that two or more structures are important components of a subject's morphology, but the spatial relationship of the structures is not fixed. The aforementioned example with heads and tails of an aquatic organism portray two anatomical features that are likely phenotypically integrated (Klingenberg 2010; Olson and Miller 1958) but located at the most anterior and posterior portions of the organism, respectively, with considerable body flexibility in between. One might also consider two structures like the mandible and cranium of an organism, which are articulated but not fixed in one position, yet together are important components of an organism's morphology associated with feeding (see, e.g., Adams 1999). For structures where there are common articulation points between the configurations, one may arrive at a common configuration by standardizing the articulation angle between them in a plane (Adams 1999), or a set of rotational planes in 3-dimensional space (Vidal-Garcı'a et al. 2018). Alternatively, one could consider the configurations to be independent, in which case separate sets of shape variables could be acquired for analysis. Once obtained, sets of shape variables are concatenated to form a single set of variables, with some modification of each structure's relative size in combination (Adams 1999; Davis et al. 2016). This latter approach is the only viable approach if structures are not articulated. Landmark configurations have also been combined in data sets that have multiple 2-dimensional views of a 3-dimensional object (e.g., Davis et al. 2016; Profico et al. 2019).

Recently, how configurations should be scaled when they are combined has been questioned (Profico et al. 2019). Indeed, this is an area of consideration that has not received much conceptual development for evolutionary biology research, other than Adam's (1999) theoretical work for articulated structures, over two decades ago. In the age of high-dimensional data, whereby large numbers of scanned points might be acquired quickly, internally or externally, on different anatomical structures from the same organism, revisiting this topic is certainly warranted. In this synthesis, we explore some of the nuances of scaling landmark configurations for their combination, consider whether combined configurations require alignment, and offer some insights for researchers who might be confronted with such challenges.

Combining landmark configurations, challenges with scaling

We start by outlining a discipline standard; the shape of a structure as characterized by a landmark configuration is not precisely defined but more readily and easily defined by its difference from other configurations of homologous points, *sensu* D'arcy Wentwoth Thompson's Theory of Transformations (Thompson 1917). As such, shape is the property that remains when a configuration's size, position, and orientation have been rendered constant (Rohlf and Slice 1990). Generalized Procrustes analysis (GPA; Rohlf and Slice 1990) is the standard method for generating shape variables, converting landmark configurations to unit size, by dividing landmark configurations by their centroid size (the square root of the summed squared distances of landmarks to their centroid, Bookstein 1991), centering configurations by their configuration centroids, and rotating configurations through a generalized least-squares superimposition, such that variation among configurations is minimized. GPA has become entrenched as a fundamental method within GMM but any researcher who has performed GPA will attest that defining an appropriate landmark configuration is anything but straightforward. Realizing that a full configuration of possible landmarks might comprise multiple sub-configurations that would perhaps be better treated with separate GPAs makes an empirical challenge even less straightforward.

Here we present an example with larval salamanders (Fig. 1), which clearly illustrates how combining separate landmark configurations might be appealing, especially more so than working with single configurations. (In this example, the separate sub-configurations are not fixed in terms of their spatial relationship to each other.) These data were originally summarized in Levis et al. (2016), and are available in the `R` package, `geomorph` (Adams et al. 2020). The data consist of landmarks and semilandmarks for both heads (26 points) and tails

(64 points) of 114 specimens. These data were digitized on whole organisms, meaning we could just perform GPA on single configurations including heads and tails. GPA performed (with sliding of semilandmarks via bending energy, see Bookstein 1997) on such configurations (Fig. 1 A-B) reveals variance in the spatial relationship between heads and tails that obscures shape variation that is better inferred through separate GPAs of these structures (Fig. 1 C-D). Clearly, combining separate configurations from separate GPAs on heads and tails would reduce variation found at any one landmark, owing to the separate fixation of head and tail positions and alignments, which would be rather difficult to manage with full organisms (whose heads and tails can move).

[Insert Fig. 1 here]

When GPA is applied to two or more structures that one wishes to combine, we assume henceforth that configurations have been aligned to their principal axes, such that rotational differences between configurations are also rendered constant (as in Fg. 1 C-D). Resulting Procrustes residuals of the i^{th} configuration of a specimen's morphology are represented in a $p_i \times k$ matrix, for the p landmarks in k dimensions of the configuration, $\mathbf{Z}_i$, and these configurations can be concatenated such that, $\mathbf{Z}$ is a $(p_1 + p_2 + ... + p_g) \times k$ matrix for the g groups of configurations combined, per specimen. It is critical to recognize that landmark configurations combined in this way are new configurations. It might be troubling that the resulting set of landmarks will have some landmark covariances that do not correspond to a real spatial distribution of points on an anatomical structure but instead the arbitrary alignment of each configuration to their principal axes. However, if separate configurations have already been identified to have a problematic association in the single configuration that could comprise them, because the configurations are not fixed in position, such covariances would never be anatomically reliable. It is also certainly troubling that if each $\mathbf{Z}_i$ has been scaled to unit size, the size of $\mathbf{Z}$ is g rather than 1, for each specimen. Combining configurations this way does not consider the relative sizes of configurations that have been combined.

This concern can be alleviated by concatenating for each specimen instead, $CS_i^{'}\mathbf{Z}_i$, where $CS_i^{'}$ is the relativized version of centroid size (CS), found as

$$\frac{w_i CS_i}{\sqrt{\sum_{i=1}^{g} w_i CS_i^2}}, \tag{1}$$

where w_i are *a priori* weights, and the denominator is the pooled (total) centroid size of combined configurations. Relative centroid size, $CS^{'}$, ranges between 0 and 1 and when employed to scale configurations, producing a combined configuration, $\mathbf{Z}$ that is unit size. Davis et al. (2016) introduced a method similar to *equation* 1 but did not include weights and used the sum of CS_i in the denominator. That approach – used only for two configurations – scaled configurations similarly to an unweighted version of *equation* 1 (all w_i are equal) but resulting configurations actually would have a pooled CS of $2^{-1/2}$ rather than 1. (That method is the same as *equation* 1 multiplied by $2^{-1/2}$, and with w_i equal to 1 for both configurations.) Profico et. al (2019) claimed that such an unweighted approach was not reliable, as CS will be related to the number of landmarks and could misrepresent the anatomical size of structures. For example, two structures of similar size, one with dense and one with sparse representation of landmarks, will have vastly different $CS^{'}$, and therefore, misrepresentatively sized configurations once combined. The solution offered by Profico et al. (2019) was that $w_i = (p_i k)^{-1/2}$; i.e., CS should be normalized (Dryden and Mardia 2016) prior to relativization. The appeal of normalized CS is that squared distances of landmarks from their centroid are averaged, rather than summed, in its calculation. (Note that using k in the calculation is a convention for considering the number of variables – rather than the number of landmarks – but when comparing multiple centroid sizes for data in the same dimension, is an unnecessary scalar that could be omitted.) Profico et al. (2019) also did not use the denominator in *equation* 1 in their proposed calculations, meaning all resulting combined configurations would be scaled as $\left(\sum_{i=1}^{g} k^{-1} p_i^{-1} CS_i^2\right)^{1/2}$, and thus, not the same across specimens; i.e., any analysis with such configurations confounds size and shape (see, e.g., Figs. 2 and 6 in Profico et al. 2019).

Alternatively, the w_i in *equation* 1 could be adjusted by trial and error, in an attempt to produce combined configurations that merely seem correct in the eyes of the researcher who has a preference to how large certain portions of combined configurations should be with respect to others. This might not seem ideal; but neither is normalizing CS a general solution, as we show below.

Normalized centroid size is not a universal solution

The proposed solution of normalizing CS from Profico et al. (2019) was offered as a general solution, recognizing that objects of similar anatomical size might be described by different numbers of landmarks. They provided evidence using circles with uniform points on their circumference. For example, a configuration of 10 points (decagon) and a configuration of 100 points (hectogon) should have similar size – despite a 10-fold difference in the number of points – if placed on the same circle, with same circumference and surface

area. We illustrate the concern in Fig. 2 (A-B), with $CS^{'}$ calculated as in *equation* 1. Two configurations with the same number of points lying on circles with the same radius (and, therefore, same surface area) have the same $CS^{'}$, whether calculated via standard (unweighted) CS (SCS) or normalized (weighted) CS (NCS). (Note that two $CS^{'}$, each equal to 0.707 means that the pooled CS is $\sqrt{0.707^2 + 0.707^2} = 1$.) This should be obvious because each point lies exactly one radius length from the configuration centroid, and because the points are uniformly distributed, the configuration centroids and circle centers are identical (Fig.2 A). Whether summing squared distances (SCS) or averaging them (NCS), the equal number of landmarks in both sets makes two $CS^{'}$ calculations unequivocally the same. However, if we change the landmark densities (10 landmarks for circle 1 and 100 for circle 2, both with identical surface area), we see the issue and apparent solution revealed by Profico et al. (2019); circles of the same surface area, characterized by a decagon and hectogon, respectively, have vastly different $CS^{'}$ via SCS but retain matching $CS^{'}$ via NCS (Fig. 2 B).

[Insert Fig. 2 here]

If an evolutionary study required only combining landmark configurations for objects of similar size but different densities, and landmarks were uniformly distributed on the object periphery, normalizing CS might be seen as a universal solution. However, NCS does not scale geometrically with circle size, as can be appreciated with Fig. 2 C. Because using NCS on circles of same size alleviates any concern for landmark number, the change in $CS^{'}$ is predictable with an isometric change in the size of a circle (for uniform points lying on the circumference of the circle). For example, if we measure $CS^{'}$ via NCS of landmark configurations on circles with radius = 1 and radius = 2, the two $CS^{'}$ calculations of the previous illustration change from 0.707 and 0.707 (for two circles with equal radii) to 0.447 and 0.894 (Fig 2. C). In other words, if we double the radius, we double the relative size. If we change the second circle radius to 3, the $CS^{'}$ calculations become 0.316 and 0.948, or thrice the relative size for the second circle. If we change the radius of the second circle to 4, $CS^{'}$ of 0.2425356 and 0.9701425, or a scaling of 4× is observed, consistent with the scaling of the radius. Therefore, the ratio of $CS^{'}$ scales exactly the same as the radius, irrespective of the number of landmarks, which is actually unreasonable. The surface areas of the circles scale isometrically as 4×, 9×, and 16× the original surface area of π for a circle with radius = 1. Volumes of spheres scale 8×, 27×, and 64× for the same 2×, 3×, and 4× increase in radius. Thus, relativization via NCS only seems reasonable if size is a first-order attribute, which is impractical.

Alternatively, what if landmark density (number of landmarks per unit area) is maintained in the scaling

of circles? This would have no effect for NCS but have an interesting effect on SCS. If we use the same example of 10 landmarks for a circle with radius = 1, we would have to use 40, 90, and 160 landmarks to maintain the landmarks/unit area density when increasing the radius to, 2, 3, and 4, respectively. The ratios of $CS^{'}$ via NCS, as mentioned above, are 2, 3, 4, in sequence. The $CS^{'}$ via SCS for beginning and enlarged circles are 0.2425356 and 0.9701425 (ratio of 4, Fig. 2 C), 0.1104315 and 0.9938837 (ratio of 9), and 0.06237829 and 0.9980526 (ratio of 16), for surface areas that increase 4×, 9×, and 16×, respectively. In other words, using SCS preserves a 1:1 relationship between surface area and the ratio of $CS^{'}$, for a consistent landmark density applied to the circles. This relationship is extended for both standard and normalized centroid size ratios in Fig. 3, illustrating the pathology of improper geometric scaling with object size for NCS. Furthermore, when combining disparately sized objects (which might be common when combining landmark configurations), smaller objects will always be more heavily weighted in the final combination when using NCS, and this relationship becomes worse as disparity between circle sizes increases (Fig. 3).

[Insert Fig. 3 here]

Scaling issues are also exacerbated if landmarks are not distributed on the periphery of the circle. The initial example comparing 10- and 100-landmark configurations on circles of the same surface area gave the impression that NCS is not dependent on the density of landmarks. Density-independence is, however, only possible with circles if all landmarks lie on the circle circumference (even for real configurations, only having landmarks on the periphery of a structure might be impractical). This limitation can be appreciated with Fig. 2 D. We might expect with NCS that a landmark configuration with 8 landmarks on the periphery of a circle should have the same relative size as another with 8 landmarks on the periphery of a circle of the same size (radius and surface area). If one of these configurations has more landmarks in the interior of the circle (in this case, 30 uniform points on an interior circle), $CS^{'}$ via SCS will indeed not match the 0.707:0.707 expectation for the two circles of same size (0.825 : 0.566 is observed). This was the concern discussed in Profico et al. (2019), however, $CS^{'}$ via NCS does not offer a solution but, in fact, exacerbates the problem. NCS inappropriately shrinks the landmark-dense configuration because of a $CS^{'}$ that is much smaller by comparison (0.528 : 0849)! This outcome can be appreciated by greater number of landmarks closer to the centroid having large influence on NCS, which averages squared distances of landmarks to the centroid. NCS is therefore quite dependent on the density and distribution of landmarks and in certain conditions can produce $CS^{'}$ that are unnaturally disparate, in a direction that is illogical (like making landmark-rich configurations exceedingly small).

This last issue can be further appreciated with another example that illustrates the effect on data following GPA. In this example, we simulated two configurations of points (Fig. 4). The first has 30 points uniformly distributed on the circumference of a circle with radius = 1 along with 8 points uniformly distributed on the circumference of a concentric circle with radius = 2.7. One might imagine the interior and exterior edges of orbital bones for this configuration. The second configuration has 8 points uniformly distributed on a circle with radius = 1.8. We simulated 100 cases of random multivariate normal, isotropic points ($\mu = [0, 0]$; $\mathbf{\Sigma} = 0.1\mathbf{I}$,where $\mathbf{I}$ is a 2×2 identify matrix) as residuals added to each base landmark, performed GPA on each subset, and combined the subsets using relativization via SCS and NCS (Fig. 4). The results confirm the same important point observed in Fig. 2 D: normalizing CS can arbitrarily and inappropriately mischaracterize the relative sizes of configurations, following GPA. The coordinates of the smaller configuration corresponded to a relatively larger structure after GPA ($CS^{'}$ of 0.54 for the smaller configuration), resulting in an illogical combination of landmarks (Fig. 4 D). Using SCS appears more reasonable, as resulting configurations had the same rank order of size as the initial data (Fig. 4 C). We recognize that neither solution is perfect, if the goal is to maintain a ratio of circle surface areas. The ratio of relative sizes for the standard scaling of centroid size was $0.65 : 0.35 = 1.86$ in this example. If densities of landmarks were maintained, precisely, we might expect a ratio of $2.7^2 : 1.8^2 = 2.25$. While this difference might be troublesome (though likely minimal for statistical analyses on resulting coordinates), $CS^{'}$ via NCS actually changed the small configuration to the large configuration ($0.46 : 0.54 = 0.79$).

[Insert Fig. 4 here]

Thus far our considerations have considered only idealized shapes (circles), with disparate densities of uniformly distributed landmarks. Despite the issues we have already highlighted, NCS would seem less practical if the number of landmarks of similar anatomically-sized objects were comparable and uniformly distributed. For example, if we use used 90 and 100 uniformly distributed landmarks on the circumference of same sized circles, $CS^{'}$ via SCS of 0.688 and 0.725 would be perhaps little cause for alarm, not differing much from 0.707 and 0.707. Furthermore, points were equally spaced around circle circumferences in our examples. Uniformity could also be achieved by perfect reflection of clustered points, such that the configuration centroid is the same as the circle center (Fig. 5 A). In all of our examples, if the distributions were substituted with uniform but unequally spaced distributions as in Fig. 5 A, the results would be the same. Although uniform configurations with reflections of clustered

(but also uniform) points do not keenly resemble a circle as a shape, the center of the configuration and, therefore, CS, are unchanged from a configuration with equally spaced points. Uniformity is, thus, a property that maintains circle center, irrespective of the spacing of landmarks. The logical next question is, therefore, what if landmark configurations are not uniformly distributed on (the periphery of) objects? We can entertain this question with the same 10-fold difference in landmark number used on the same sized circles.

[Insert Fig. 5 here]

Goswani et al. (2019) illustrated with their simulations that increasing landmark number on an object increases centroid stability, a result that makes sense with respect to the Law of Large Numbers (Hsu and Robbins 1947). We illustrate this reality with random samples of points on the circumference of a circle (Fig. 5 B). When sampling 100 points, the centroid of the configuration will tend to change little from the circle center, but when sampling 10 points, it is more easily possible to obtain a centroid that is displaced from the circle center (Fig. 5 B). We have observed already that two circles, each with perfectly uniform landmarks (irrespective of the number of landmarks) will have $CS^{'}$ each equal to 0.707, if estimated via NCS (or SCS if the number of points on both circles are the same). If we randomly draw 10 points on the circumference of a circle (say from 100,000 uniform points equally spaced around the circle circumference) as landmarks, we can measure its $CS^{'}$' in combination with the "perfect" circle of 100 uniform points as landmarks. The expectation from NCS might be that $CS^{'}$ is 0.707 for both circles, as NCS should mitigate the $CS^{'}$ disparity due to the difference in landmark number. Additionally, we can measure the distance of the centroid from circle center as a measure of disuniformity, as a large distance can only be achieved by points clustered disproportionately on one side of a circle. We simulated 100,000 such configurations from circles (radius = 1, p = 10 landmarks, see Fig. 5 A-B), which revealed that $CS^{'}$ decreases in a predictable manner with increased disuniformity (Fig. 5 C). $CS^{'}$ via NCS decreased at a higher rate than $CS^{'}$ via SCS, indicating that larger departures from uniformity were more profound for NCS. Again, this example elucidates that NCS cannot universally mitigate concerns about disparate landmark densities and $CS^{'}$ via NCS is indeed dependent on the density and distribution of landmarks. The same is true for $CS^{'}$ via SCS, but NCS has been proposed as a solution to these issues (Profico et al. 2019).

Our examples illustrate that $CS^{'}$ for configurations are always going to be dependent on the distribution and density of landmarks, and that an attempt to weight CS (*equation* 1), using NCS, rather than an attempt to maintain landmark densities might incur some undesirable issues. We summarize three important points.

First, normalizing CS can arbitrarily and inappropriately mischaracterize the relative sizes of configurations. This was most apparent in the example in Fig. 2 D and Fig. 4. In these examples, the high density of interior landmarks of one configuration caused it to be exceedingly small in $CS^{'}$ via NCS, while SCS produced $CS^{'}$ that were comparatively more logical with respect to circle size. This result is the complete opposite of the expected solution of NCS to mitigate $CS^{'}$ disparity due to disparity in the number of landmarks. Second, using NCS does not mitigate disparity in $CS^{'}$ for landmarks that are not uniformly distributed around the object's periphery. Based on the results shown in Figs. 2 D and 5, one has to wonder if they must first confirm that they only have essentially uniform points, and only on an object's periphery before using NCS as a weighting method for calculation of $CS^{'}$. Third, the ratios of $CS^{'}$ do not scale logically when combining configurations for structures of different size, as we showed in Fig. 3. This example illustrates that if at all possible, maintaining landmark density among configurations might be a better method for assuring appropriate relative sizes of configurations than attempting to fix the disparity in landmark densities by weighting centroid sizes. We recognize that without having true measures of anatomical surface area or volume, maintaining landmark densities (other distributional considerations, notwithstanding) is not really possible. However, either recognizing 10-fold differences in landmark numbers or mitigating such differences with additional landmarks in depauperate configurations should not be insurmountable challenges in most cases.

One might conjecture, especially with the panoply of evolutionary examples that could be considered, that finding contrived, extreme examples is sure to reveal atypical results. However, the probative example of Profico et al. (2019) using 10 points and 100 points uniformly distributed on a circle, is an extreme example that revealed atypical results, and was the sole impetus to offer NCS as a universal solution. To employ normalization of CS as a general solution to any combination of separate landmark configurations means researchers might inadvertently distort the relative sizes of configurations in an unreasonable way. To illustrate this concern with real data, we return to the salamander example (Fig. 1). Using both SCS and NCS in the calculation of $CS^{'}$ yielded quite different results. It is clear that NCS produces combined configurations that suggest salamander heads are unnaturally and inappropriately enlarged (Fig. 6). By contrast, $CS^{'}$ via SCS yielded results that more closely match the actual relative sizes of salamander heads and tails (Fig. 1). Normalizing CS should not be considered a universal solution, as issues with geometric scaling are inescapable.

[Insert Fig. 6 here]

Configuration size is not anatomical size

One should be reminded that CS is the measure of size used to standardize configurations to unit size in GPA (Rohlf and Slice 1990). Additionally, only CS is uncorrelated with shape variables from GPA in the absence of allometry, a crucial property for correct and valid shape analysis. All other possible size measures, including lengths, extents, areas, and volumes, will generate false allometry when used with shape data obtained from GPA (Bookstein 1991). Thus, while some might find areas or volumes more "natural" anatomical size measures, analytically they are not, and when used with shape data, they are prone to generate false patterns of shape covariation in one's data. As such, any alternative that seeks to approximate one's preference for "anatomical size" by altering CS should be scrutinized. By contrast, in evolutionary studies that examine allometric patterns for the size of anatomical features (with respect to orgamism size), using CS as a proxy for anatomical size might be equally ill-advised.

It is easy to embrace the circle example introduced by Profico et al. (2019); it does not make sense that two identically sized circles could have such disparate CS'. However, the shapes used to illustrate the problem – which are defined by the set of digitized landmark points – were, in fact, not circles; they were decagons and hectogons, and for the purpose of GPA, the hectogon has a much larger size, as it has far more landmarks contributing to its generalized least-squares superimposition. An attempt to dismiss the precise definition of CS and its importance in GPA in favor of a weighted CS should be scrutinized as an unwelcomed analytical vicissitude. When combining landmarks on circles of the same size that have a 10-fold difference in landmarks, the researcher has decided *a priori* that the 100-landmark configuration is, in fact, larger by virtue of the greater number of landmarks needed to characterize its shape. This is not an unfortunate miscalculation; it is an analytical necessity. In other words, configurations with more landmarks have implicit greater weight in combination, and this is not an outcome that should be challenged so readily. If landmark density disparities present a problem, one should view this as a digitizing problem rather than an analytical problem for combining them. A weighting scheme that attempts to fix a digitizing problem after superimposition should be done with extreme caution, and one should realize that an approach like NCS is not a guaranteed solution. One has to consider, for example, if the number of semilandmarks on curves or surfaces could be augmented or culled to achieve more comparable landmark densities without compromising the number of landmarks needed to describe shape accurately. Attempting to solve a digitizing issue this way might be preferable to seeking a weighting scheme (w_i in *equation* 1) after GPA.

Should combined configurations be subsequently aligned?

We have until this point avoided an inconvenience. Because GPA applied to separate configurations renders the same centroid of Procrustes residuals for each (e.g., $[0, 0]$ for landmarks in 2 dimensions), which is also the centroid of combined configurations, and because scaling each separate configuration by its $CS^{'}$ causes the combined configuration to have unit size, the combined configuration resembles Procrustes residuals, and therefore, a configuration that confers shape differences among specimens. Furthermore, the denominator of *equation* 1 is the pooled CS (which if divided by $\sqrt{k \sum_i^g p_i}$ returns the pooled normalized CS.) Does this imply that the combined configurations, $\mathbf{Z}$, of all specimens should undergo a generalized least-squares superimposition to account for rotational variation?

This is an interesting question to answer in both theoretical and applied contexts. If we can assume that individual configurations produce small, isotropic scatter around configuration means, and each configuration is aligned to its principal axis, meaning principal axes are parallel among configurations, there should be practically no difference between a combined set of scaled configurations and the Procrustes residuals from GPA performed on the combined configurations. We could actually measure the outcome of this process. If we let $\mathbf{M}$ be the $p \times k$ matrix of coordinate means for the combined configurations, $\mathbf{Z}_j$, from$j = 1$ to n (specimens), and then subject $\mathbf{Z}_j$ to GPA, producing newly aligned coordinates, $\mathbf{Z}_j^a$, with mean, $\mathbf{M}^a$, then we can express a vector of residuals of configuration points from configuration means as $\mathbf{z}_j = vec\,(\mathbf{Z}_j - \mathbf{M})$ and $\mathbf{z}_j^a = vec\left(\mathbf{Z}_j^a - \mathbf{M}^a\right)$, pre- and post-alignment, respectively. For an entire set of combined configurations, we can measure the rotational variance associated with alignment as a fraction of the original variance among landmarks; i.e., the proportion of variance in rotation (pVR), as,

$$pVR = \frac{\sum_{j=1}^{n} \left(\mathbf{z}_j - \mathbf{z}_j^a\right)^T \left(\mathbf{z}_j - \mathbf{z}_j^a\right)}{\sum_{j=1}^{n} \mathbf{z}_j^T \mathbf{z}_j}, \tag{2}$$

where ^{T} means vector transpose. We would expect $pVR = 0$ if GPA does not alter the combined configurations. This is of course a limit that is likely not realized, but if pVR is quite small, it suggests that combined configurations have all the properties we expect for shape variables.

We can demonstrate this as a concept with the salamander example. Using the means of the head and tail configurations, we simulated isotropic residuals from a multivariate normal distribution ($\mu = [0, 0]$;

$\mathbf{\Sigma} = 0.01\mathbf{I}$), purposely to have fairly small scatter around each landmark, for $n = 114$, the sample size of the empirical data. (We visually verified that the result looked similar to the empirical data). This produced $pVR = 0.0085$; i.e., re-aligning the data with GPA incurred less than a 1% change with respect to the variance of points around landmarks. Thus, as expected, the pre- and post-aligned configurations are practically the same. However, in doing this with the real data (via SCS), we observed $pVR = 0.4555$, a profoundly different outcome. (We found no qualitative differences between pVR performed on combined configurations via SCS or NCS, with simulated data or empirical data.) Contrary to expectations, with real data, GPA altered the alignment of combined configurations, substantially. One could perhaps reconcile this with different biological explanations, but analytically, the theoretical example breaks down as soon as there are covariances among landmarks that differ from isotropic scatter, which with any empirical data is a certainty.

More importantly, whether combined data are aligned or not, they are not collectively shape data. We demonstrate this in Fig. 7, with a principal component (PC) plot of combined configurations both before and after alignment (Fig. 7 A-B). (Note that a two-block partial least squares correlation between pre- and post-aligned data was 0.982, which is not surprising, as the relative locations in PC plots of specimens are fairly consistent.) Thin-plate spline (TPS; Bookstein 1991) transformation grids reveal the difficulty with mapping the mean combined configuration on other combined configurations, as lines in the grid can cross (Fig. 7 C-D), which would be a strange occurence for the typically subtler shape transformations for real landmark configurations. Because points are combined and unconstrained in their spatial arrangement, unlike in original configurations, it is possible, for example, for a point in the second configuration to be located "to the right" of a reference point from the first configuration, for one specimen after combination, but located "to the left" for another. For most anatomical confirgurations, this would not make sense (as if the right eye moved left of the nose). Via combination, these relationships are arbitrarily generated but are not true anatomical realities. Despite the different PC projections, however, the TPS transformations of the mean configuration were indistinguishable (when mapped to the scores of the first two PCs), indicating that shapes map the same in PC projections, whether PCA was performed on pre- or post-aligned data even it was not possible to reconcile how head and tail shape are changing in any coordinated way. If we separate the combined configurations into heads and tails, we also see the consistent TPS mapping, whether data were aligned with GPA or not, but also can visualize the coordinated change in head shape and tail shape (Fig. 7 E-H). Therefore, whether data are aligned had no impact on reconciling head and tail shapes, even if there are statistical differences between the approaches. It might come down to a matter of preference whether to re-align the combined configurations.

[Insert Fig. 7 here]

This example emphasizes the crucial point that combined configurations characterize an organism's morphology, but not an organism's shape. By combining configurations, one must realize that the resulting variables are a composite of shapes. There might be an interest in having a composite of shapes for statistical analyses, especially if shapes are integrated, because one might gain better precision in estimating, e.g., group differences in morphology. However, aligning the combined data as if they were shape data from a single configuration seems to be an unneeded step. Additionally, the result depends on the covariances between landmarks that are now a part of single set, as if they are located in relation to each other on an anatomical structure. This means that performing GPA on combined data sets might lead to spurious downstream results. More research is needed to fully understand all of the implications of GPA on combined data sets, but at this time, we do not recommend it as a necessary step in one's analytical pipeline, and encourage others to think not of combined configurations as shapes, but as composites of shapes that can be disjoined for visualizing shape patterns.

Finally, one might wonder then if there is a purpose to combining configurations. We would not dispute that it is perfectly reasonable to restrict analyses to the original configurations, as was done by Levis et al. (2016), in their original treatment of the data. For these salamander data, this would mean two PC plots and two analyses of variance (ANOVA) to test for treatment and allometry effects. Correlations between tail shape and swim speed were also considered by Levis et al. (2016.). As a matter of efficacy, single analyses of combined configurations with multiple TPS transformation grids to correspond to PC points or fitted values from linear models might be preferred, especially if combining multiple configurations would make replicated statistical assesments unwieldy.

Conclusions

In this paper, we demonstrate that a general equation for calculating the relative centroid sizes of configurations can be employed for scaling configurations when they are combined. This equation allows *a priori* determination of weights to use in $CS^{'}$ calculation. We showed that one weighting process, normalizing CS, is fraught with issues, and is only appropriate for the limited (and unrealistic) case where one is combining data

from two structures of similar size and with uniformly distributed landmarks, primarily around the periphery of the object. As such, NCS cannot be viewed as a general solution to the problem. Additionally, NCS does not scale reasonably with object size and has the propensity to make relatively smaller anatomical objects larger in combination, based on disparate landmark distributions in the configurations that characterize them.

Calculating $CS^{'}$ via SCS will sometimes produce undesirable results, but this is more so a digitizing problem than an analytical one, arising from disparate landmark densities (landmarks per unit area or volume). However, only an unweighted version of *equation* 1 (all weights equal: i.e., SCS) can guarantee that the rank order of CS is preserved across the structures being combined. Additionally, CS is the size measure used for standardization of configurations to unit size in GPA and in the absence of allometry, is the only size measure that is uncorrelated with shape (Bookstein 1991). These properties should not be discarded lightly in favor of one's notion of an "anatomical" size measure. Rather, one might instead be more concerned with the implications of disparate landmark densities arising from configurations of similar sized objects for corresponding GPAs, or whether the Procrustes residuals should even be merged for downstream analysis. Alternatively, one might consider a digitizing solution to mitigate $CS^{'}$ disparities, by intensifying or culling landmark or semilandmark densities, rather than seeking a weighting solution. Our perspective is that a digitizing solution that seeks to preserve landmark densities among configurations, utilized in concert with estimation of $CS^{'}$ via SCS, is a better strategy than seeking a weighting scheme that attempts to fix the issues created or ignored in digitizing, after GPA is performed.

Nonetheless, even if after careful consideration, one feels it is imperative to combine landmark configurations from disparate landmark densities, the results presented here demonstrate that NCS is not a universal solution to the problem. In such cases where equal weighting does not provide a satisfactory solution, additional weighting schemes should be envisioned. At present however, we are not aware of any alternative weighting scheme beyond the two presented in this paper. To identify other candidates, trial and error may be considered (for example, making some weights 0.9 and others 1.1, and contrasting results), but this is likely to result in weighting schemes that are not general, and instead are data-specific or restricted to particular scenarios (such as demonstrated above for NCS). For this reason we do not advocate this avenue of pursuit; preferring instead approaches that are more firmly grounded in statistical theory. For instance, a weighting scheme that accounts for landmark variances, much the way weighted least-squares regression is used to account for heteroscedasticity, may provide more reasonable performance for a wider set of cases. For combining landmark configurations with vastly disparate landmark densities, this may be a fruitful research

direction to consider.

Finally, we recommend that empiricists neither perform GPA on combined configurations, nor consider combined configurations to be organismal "shape". Rather, combined configurations are shape composites which can be disjoined for the purpose of shape visualization, but used in conjunction for statistical analyses. Combining configurations often offers statistical efficacy, and this alone is a valid reason to combine configurations for joint-data analysis. But as shown in this paper, such procedures should be considered with care. It is our perspective that whether and how to combine landmark configurations is a topic of increasing concern in morphometrics, as the advent of more advanced data acquisition pipelines (Bardua et al. 2019; Goswami et al. 2019), and automated data collection efficiencies permeate the field and facilitate the generation of such datasets. Faced with this inevitability, we anticipate that procedures such as those investigated here will only increase in utility. Thus, in this era of big data phenomics, how to best utilize morphometric data from distinct substructures will become a pervasive topic in geometric morphometrics for the foreseeable future. Our investigations are but the first step in moving this process forward towards a general approach. For many cases using SCS will suffice, but for others, additional weighting schemes may need to be developed. We hope that our findings regarding SCS, NCS, and configuration weighting, will provide food for thought to both empiricists and theorists alike when considering how to combine landmark configurations.

References

Adams, D. C. (1999). Methods for shape analysis of landmark data from articulated structures. *Evolutionary Ecology Research*, *1*, 959–970.

Adams, D. C., Collyer, M. L., & Kaliontzopoulou, A. (2020). Geomorph: software for geometric morphometric analyses. R package version 3.2.1. https://cran.r-project.org/package=geomorph.

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*, 393–399.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix*, *24*, 7–14.

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.

Bookstein, F. L. (1991). *Morphometric tools for landmark data: geometry and biology*. Cambridge: Cambridge University Press.

Bookstein, F. L. (1997). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, *1*, 225–243.

Collyer, M. L., & Adams, D. C. (2018). RRPP: An R package for fitting linear models to high-dimensional data using residual randomization. *Methods in Ecology and Evolution*, *9*, 1772–1779. Journal Article.

Collyer, M. L., & Adams, D. C. (2020). RRPP: Linear model evaluation with randomized residuals in a permutation procedure, version 0.5.2. R Foundation for Statistical Computing. https://cran.r-project.org/package=RRPP

Davis, M. A., Douglas, M. R., Collyer, M. L., & Douglas, M. E. (2016). Deconstructing a species-complex: Geometric morphometric and molecular analyses define species in the Western Rattlesnake (Crotalus viridis). *PLoS ONE*, *11*(1). http://wwx.inhs.illinois.edu/

Dryden, I. L., & Mardia, K. V. (2016). *Statistical shape analysis: With applications in r* (Vol. 995). John Wiley & Sons.

Goswami, A., Watanabe, A., Felice, R. N., Bardua, C., Fabre, A. C., & Polly, P. D. (2019). High-density morphometric analysis of shape and integration: the good, the bad, and the not-really-a-problem. *Integrative and comparative biology*, *59*(3), 669–683.

Hsu, P. L., & Robbins, H. (1947). Complete Convergence and the Law of Large Numbers. *Proceedings of the National Academy of Sciences*, *33*(2), 25–31.

Klingenberg, C. P. (2010, September). Evolution and development of shape: Integrating quantitative approaches. Nature Publishing Group.

Levis, N. A., Schooler, M. L., Johnson, J. R., & Collyer, M. L. (2016). Non-adaptive phenotypic plasticity: The effects of terrestrial and aquatic herbicides on larval salamander morphology and swim speed. *Biological Journal of the Linnean Society*, *118*(3), 569–581.

Olson, E. C., & Miller, R. L. (1958). *Morphological integration*. Chicago: University of Chicago Press.

Profico, A., Piras, P., Buzi, C., Del Bove, A., Melchionna, M., Senczuk, G., et al. (2019). Seeing the wood through the trees. Combining shape information from different landmark configurations. *Hystrix, the Italian Journal of Mammalogy*, *30*, 157–165.

Rohlf, F. J., & Slice, D. E. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, *39*, 40–59.

Thompson, D. W. (1917). *On growth and form*. Cambridge University Press.

Vidal-Garcı'a, M., Bandara, L., & Keogh, J. S. (2018). ShapeRotator: An r tool for standardized rigid rotations of articulated three-dimensional structures with application for geometric morphometrics. *Ecology and Evolution*, *8*(9), 4669.

Figures

Figure 1. GPA of 114 larval salamanders from Levis et al. (2016) All points and means are showing in (A), but just means in (B), revealing relative head and tail size. GPA was performed separately on tails (C) and heads (D), illustrating the smaller variance around landmarks for separate GPA analyses.

Figure 2. Comparisons of $CS^{'}$ for different distributions of landmarks on circle circumferences. Standard and normalized centroid sizes (SCS and NCS, respectively) return the same $CS^{'}$ for the same distribution of landmarks (A) but only NCS produces equal $CS^{'}$, irrespective of landmark number (10 versus 100), if circles are the same size (B). However, maintaining landmark density (10 versus 40) shows that SCS has more reasonable $CS^{'}$ with respect to circle surface area, when circles are different size (C). Furthermore, NCS yields unreasonable results when the distribution of landmarks is not restricted to a uniform distribution on the circumference of the circle (D). In (D), each circle has 8 exterior landmarks but one has 30 interior landmarks.

Figure 3. Relationship between the ratio of circle surface areas (large:small) and the ratio of relative centroid sizes (large:small) for both standard and normalized centroid sizes.

Figure 4. An example of landmark configurations (A) with 38 points and 8 points, simulated each with random multivariate normal residuals for 100 specimens (B). Following GPA, using standard centroid size scaling (C) and normalized centroid size scaling (D), combined configurations reveal concentric circles. The inability of NCS to guarantee rank order of $CS^{'}$ can be appreciated in D.

Figure 5. Examples of uniform distributions on circles (A) and examples of distributions of landmarks randomly drawn from 100,000 uniform points on a circle circumference (B), showing the issue of small samples. Repeating this sampling scheme for 10 landmarks over 100,000 simulation runs and comparing to a uniform distribution of 100 landmarks affects $CS^{'}$ in predictable ways (C).

Figure 6. Means from combined configurations after SCS scaling (A) and NCS scaling (B). See Fig. 1 for other information.

Figure 7. Principal component (PC) plots for combined configurations (A) and combined configurations subjected to GPA (B). Panels C, E, and G show TPS transformation grids for combined configurations, heads, and tails, respectively, for specimen 20 in the PC plot shown in A. Panels D, F, and H, correspondingly do the same for the plot in B. All TPS transformation grids are deformations of the means with respect to the first two PC scores.

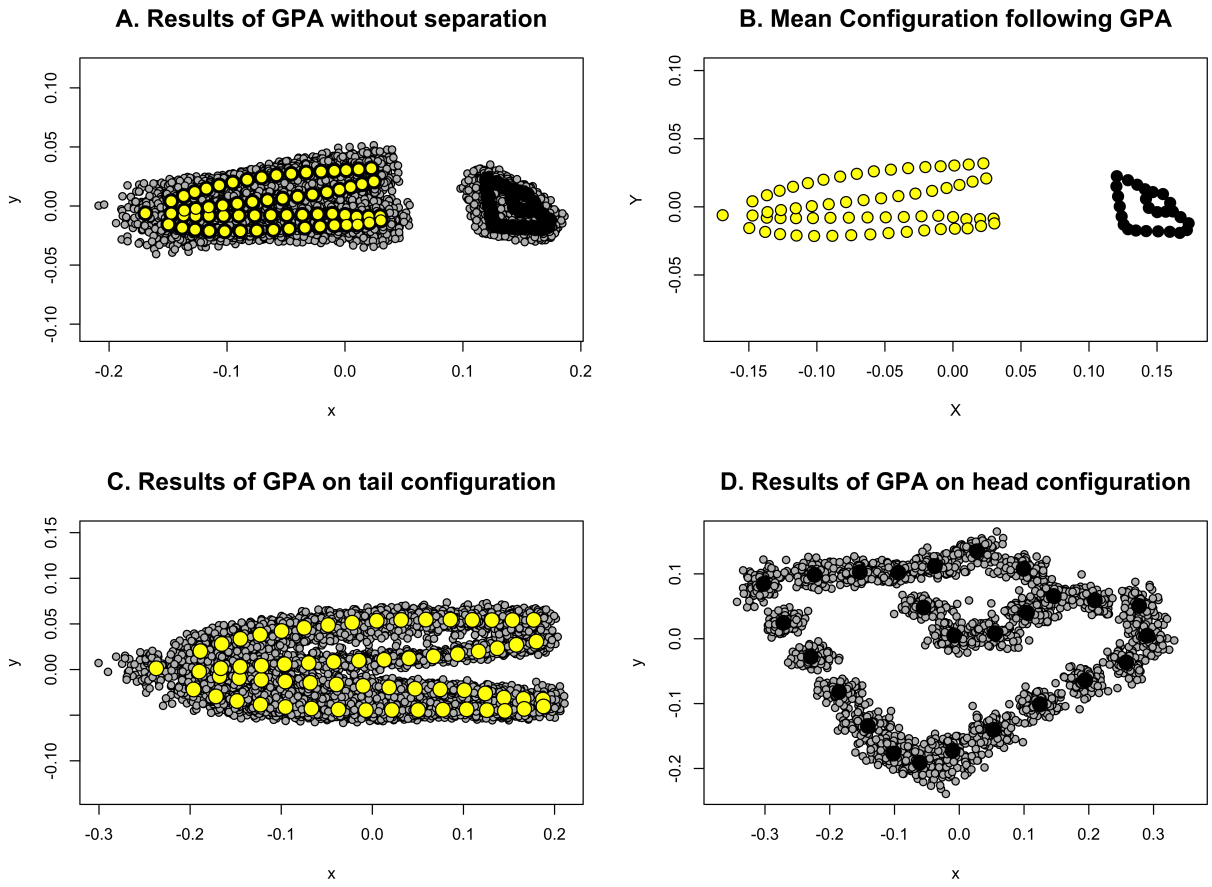


Figure 1: GPA of 114 larval salamanders from Levis et al. (2016) All points and means are showing in (A), but just means in (B), revealing relative head and tail size. GPA was performed separately on tails (C) and heads (D), illustrating the smaller variance around landmarks for separate GPA analyses.

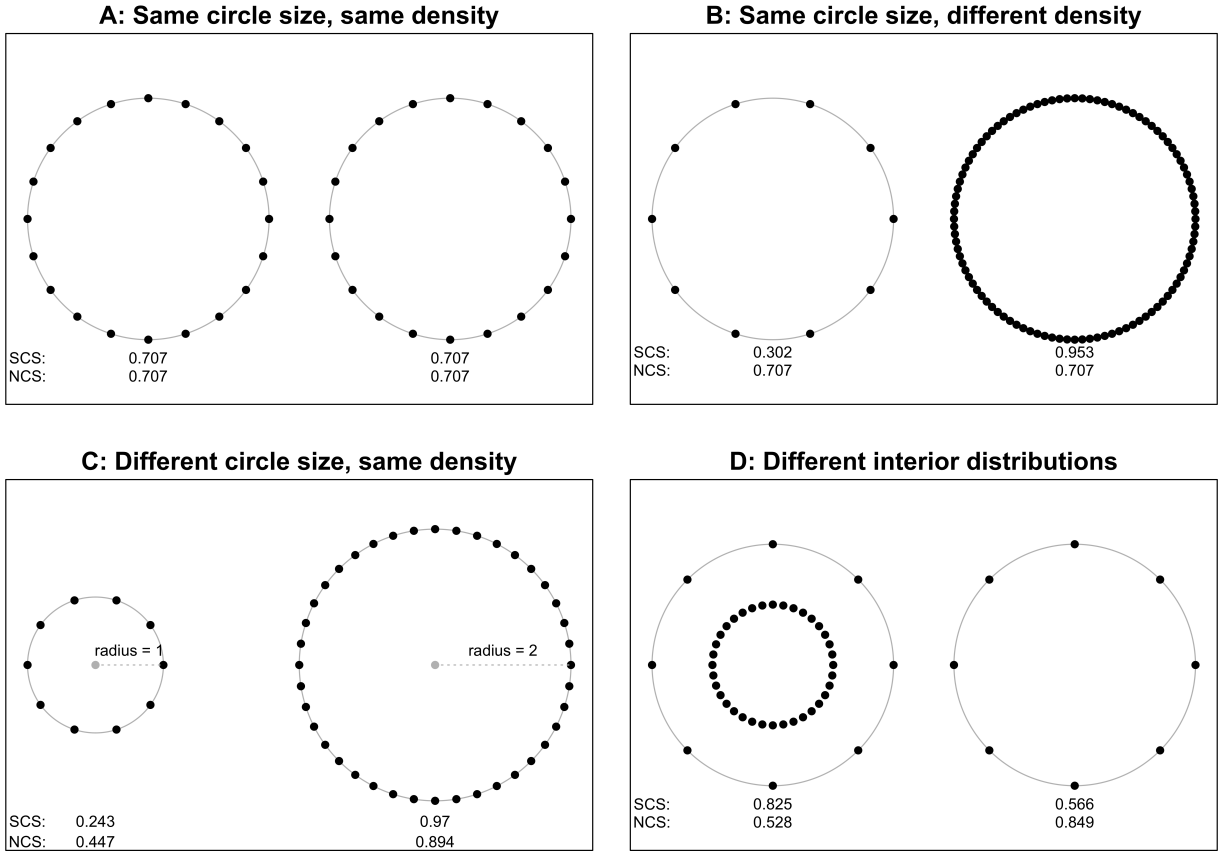


Figure 2: Comparisons of CS' for different distributions of landmarks on circle circumferences. Standard and normalized centroid sizes (SCS and NCS, respectively) return the same CS' for the same distribution of landmarks (A) but only NCS produces equal CS', irrespective of landmark number (10 versus 100), if circles are the same size (B). However, maintaining landmark density (10 versus 40) shows that SCS has more reasonable CS' with respect to circle surface area, when circles are different size (C). Furthermore, NCS yields unreasonable results when the distribution of landmarks is not restricted to a uniform distribution on the circumference of the circle (D). In (D), each circle has 8 exterior landmarks but one has 30 interior landmarks.

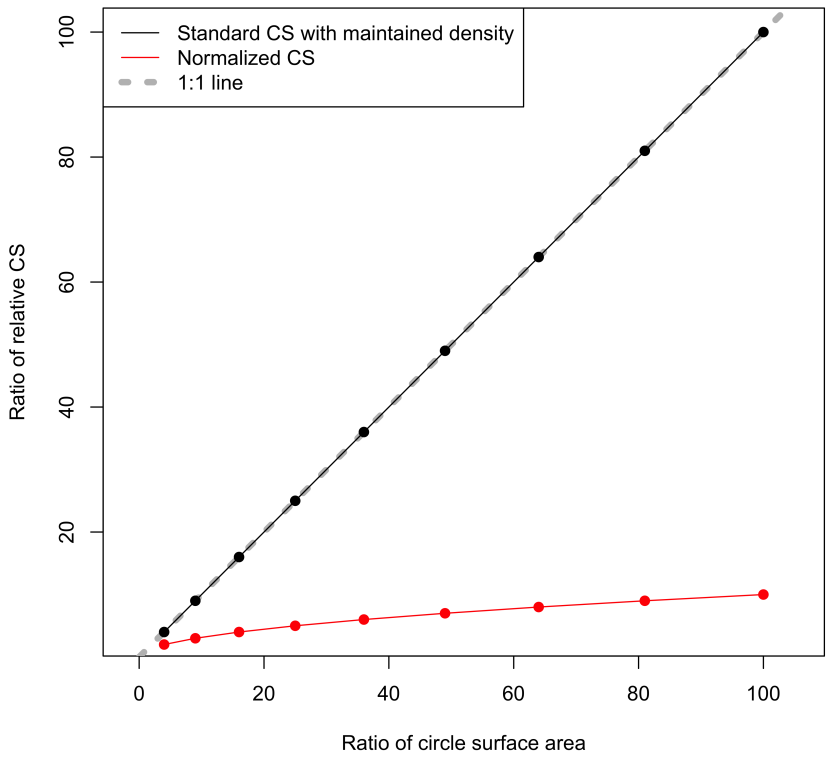


Figure 3: Relationship between the ratio of circle surface areas (large:small) and the ratio of relative centroid sizes (large:small) for both standard and normalized centroid sizes.

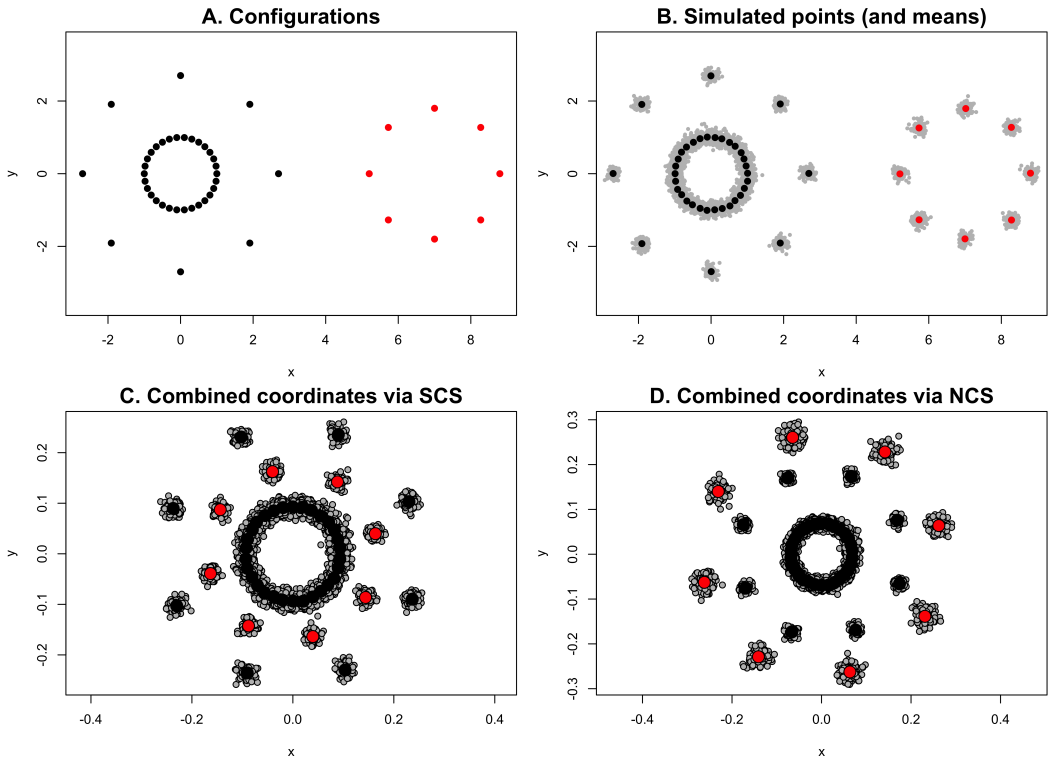


Figure 4: An example of landmark configurations (A) with 38 points and 8 points, simulated each with random multivariate normal residuals for 100 specimens (B). Following GPA, using standard centroid size scaling (C) and normalized centroid size scaling (D), combined configurations reveal concentric circles. The inability of NCS to guarantee rank order of CS' can be appreciated in D.

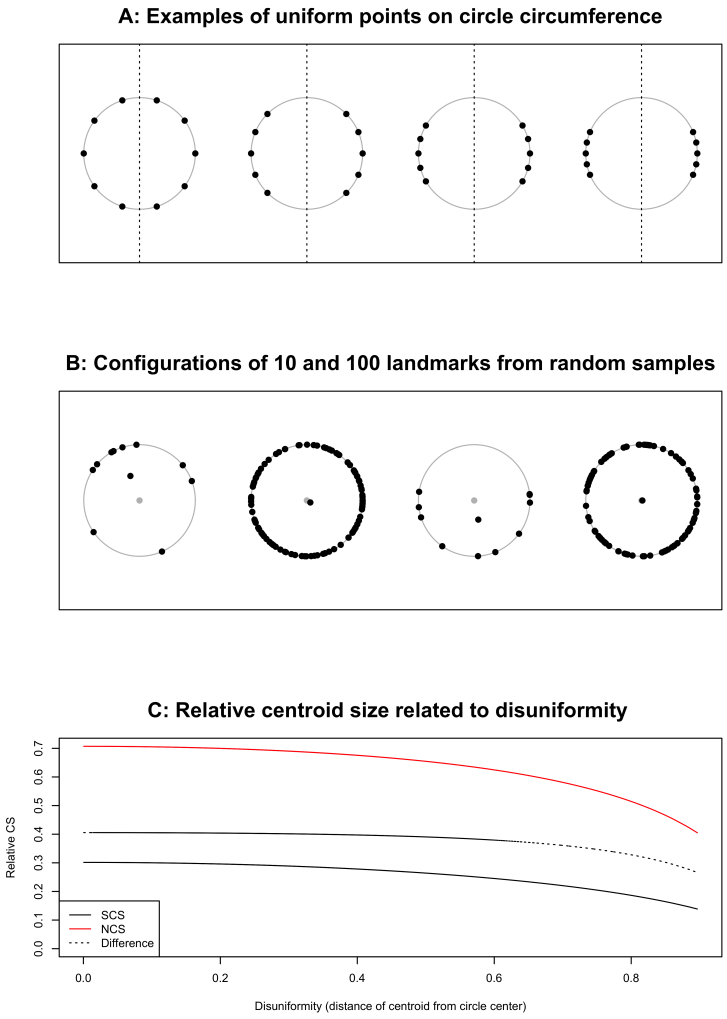
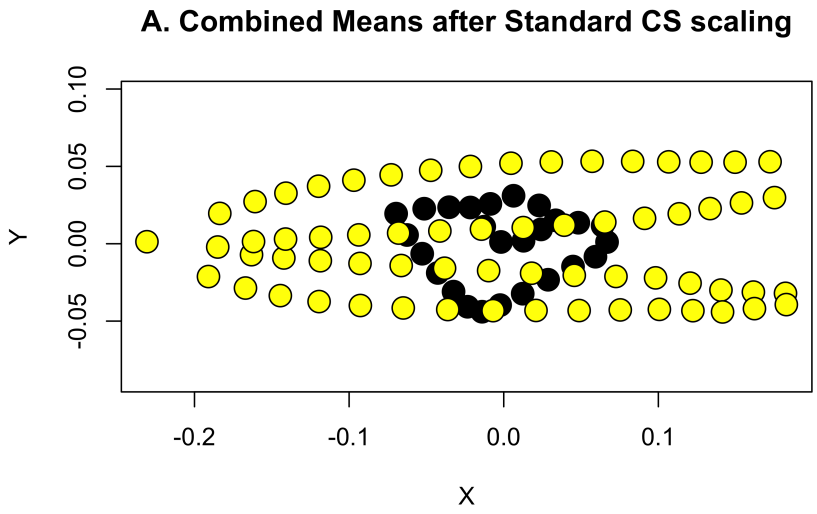


Figure 5: Examples of uniform distributions on circles (A) and examples of distributions of landmarks randomly drawn from 100,000 uniform points on a circle circumference (B), showing the issue of small samples. Repeating this sampling scheme for 10 landmarks over 100,000 simulation runs and comparing to a uniform distribution of 100 landmarks affects $CS^{'}$ in predictable ways (C).



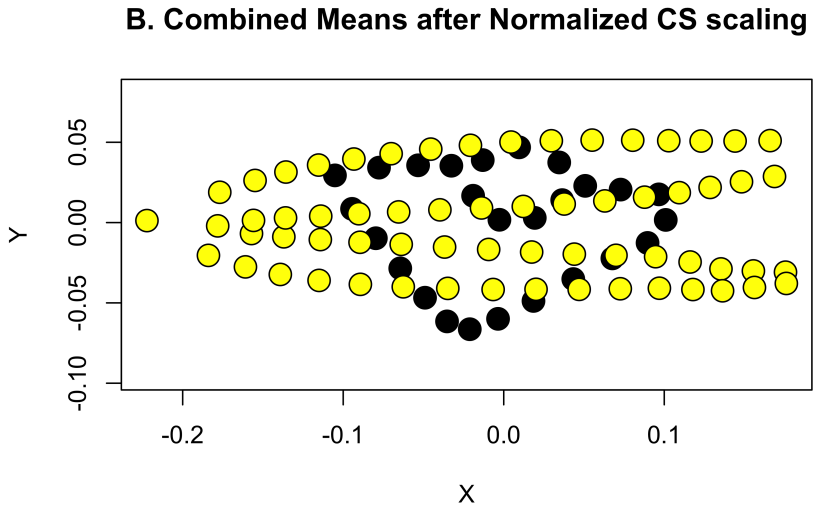


Figure 6: Means from combined configurations after SCS scaling (A) and NCS scaling (B). See Fig. 1 for other information.

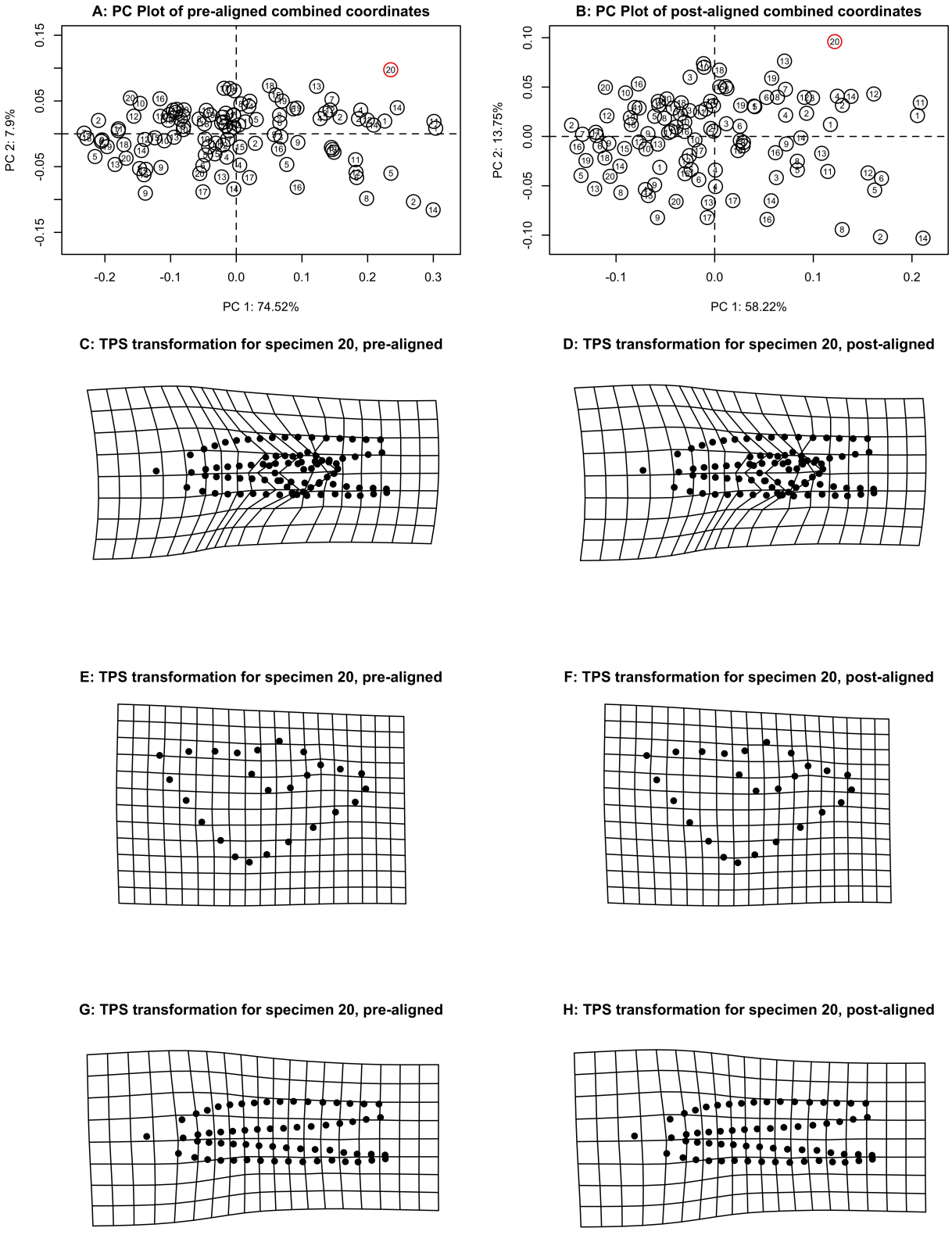


Figure 7: Principal component (PC) plots for combined configurations (A) and combined configurations subjected to GPA (B). Panels C, E, and G show TPS transformation grids for combined configurations, heads, and tails, respectively, for specimen 20 in the PC plot shown in A. Panels D, F, and H, correspondingly do the same for the plot in B. All TPS transformation grids are deformations of the means with respect to the first two PC scores.