Published by Associazione Teriologica Italiana
Hystrix, the Italian Journal of Mammalogy
Volume 24 (1): 67–73, 2013
Available online at:
http://www.italian-journal-of-mammalogy.it/article/view/6332/pdf
doi:10.4404/hystrix-24.1-6332

Research Article

Studying ontogenetic trajectories using resampling methods and landmark data

H. David Sheets[a,*], Miriam L. Zelditch[b]

[a]*Dept. of Physics, Canisius College, 2001 Main St, Buffalo, NY 14208, USA*
Dept. of Geology, SUNY at Buffalo, 411 Cooke Hall, Buffalo, NY 14260, USA
[b]*Museum of Paleontology, University of Michigan, Ann Arbor, Michigan 48109, USA*

Keywords:
ontogeny
shape
permutation
bootstrapping
resampling
MANCOVA

Article history:
Received: 20 June 2012
Accepted: 19 September 2012

Acknowledgements
We would like to thank Anna Loy, Philipp Mitteroecker and Andrea Cardini for helpful reviews of this manuscript.

Abstract

Comparative studies of ontogenies play a crucial role in the understanding of the processes of morphological diversification. These studies have benefited from the appearance of new mathematical and statistical tools, including geometric morphometrics, resampling statistics and general linear models. This paper presents an overview of how resampling methods may be applied to linear models of ontogenetic trajectories of landmark-based geometric morphometric data, to extract information about ontogeny. That information can be used to test hypotheses about the changes (or differences) in rate, direction, duration and starting point of ontogenetic trajectories that led to the observed patterns of morphological diversification.

Introduction

A central goal of evolutionary morphology is to explain the origin of morphological diversity. That diversity is now often termed "disparity" to distinguish it from the proliferation of species, a distinction that is important because the proliferation of species may not explain the proliferation of novel morphologies (e.g., Foote 1993; Adams et al. 2009). The proximate cause of disparity is evolutionary change in ontogeny; consequently, to understand the processes generating disparity we need to understand how ontogenies evolve (Zelditch et al., 2003; Adams and Nistri, 2010; Drake, 2011; Frederich and Vandewalle, 2011; Gerber, 2011; Ivanovic et al., 2011; Piras et al., 2011). Comparative studies of ontogenies play a critical role in such studies of disparity because they can discern which modifications of ontogeny are responsible for disparity, including the modifications that increase disparity, those that decrease it, and those that maintain it. Disparity itself often has an ontogeny because species may closely resemble each other during early states of morphogenesis, diverging thereafter or they may differ substantially early in development then come to resemble each other. Both these patterns have been detected empirically. For example, the disparity of both body shape and diet increase over ontogeny of some damselfishes (Frederich and Vandewalle, 2011) but body shape and diet show contrasting patterns in piranhas, with body shape disparity decreasing over ontogeny as the disparity of diet increases (Zelditch et al., 2003). Such decreases in disparity over ontogeny have been found in other groups as well, including body shape of crested newts (Ivanovic et al., 2011), and feet and interdigital webbing of European cave salamanders' feet (Adams and Nistri, 2010).

By combining comparative studies of ontogeny with analyses of disparity at two or more developmental stages, it is possible to test hypotheses about the developmental origins of disparity. When there are several modifications of ontogeny, we can ask how much each one, taken separately, would contribute to disparity of juveniles and adults and then how two or more interact with each other. The aim of this paper is to present methods for characterizing the modification of ontogeny, as revealed by morphometric data, and then methods for modeling the evolution of ontogeny to discern the impact of those modifications on disparity at two or more developmental stages.

Shape change: Ontogenetic change as a phenotypic trajectory

This paper focuses on ontogeny, but the analysis of ontogenetic change is a special case of a phenotypic trajectory. A phenotypic trajectory refers to the extent, direction and rate of any shape change in response to some factor, which could be geographic (e.g., latitude) or ecological (e.g., predation pressure). The concepts and methods discussed in context of phenotypic trajectories are thus readily adaptable to any context (see Adams and Collyer 2009, and in this volume). A variety of statistical tools can reveal how phenotypic trajectories differ; our objective in this paper is to discuss regression models and resampling techniques as applied to ontogenetic trajectories. We begin by applying resampling methods to a simple bivariate regression model, extending that to a multivariate regression for a single group and then to multiple groups by multivariate analysis of covariance (MANCOVA). Once MANCOVA establishes the statistical significance of differences in ontogenetic trajectories, a series of more specialized tests are used to characterize those differences.

Ontogenetic trajectories are complex to model because they comprise two distinct categories of variables: shape variables, and size and age variables. Whereas size and age are typically univariate measurements (scalar values), shape is multivariate, comprising multiple variables such as the coordinates of many landmarks. In this paper, we will typically discuss shape's dependence on size, but the same models extend to age-based analyses. The samples of shape and size may be taken continuously throughout ontogeny or at two or more times. In the first case, shape's dependence on size is modeled by a single line; in the second case its dependence may be modeled by several line seg-

*Corresponding author
Email address: sheets@canisius.edu (H. David Sheets)

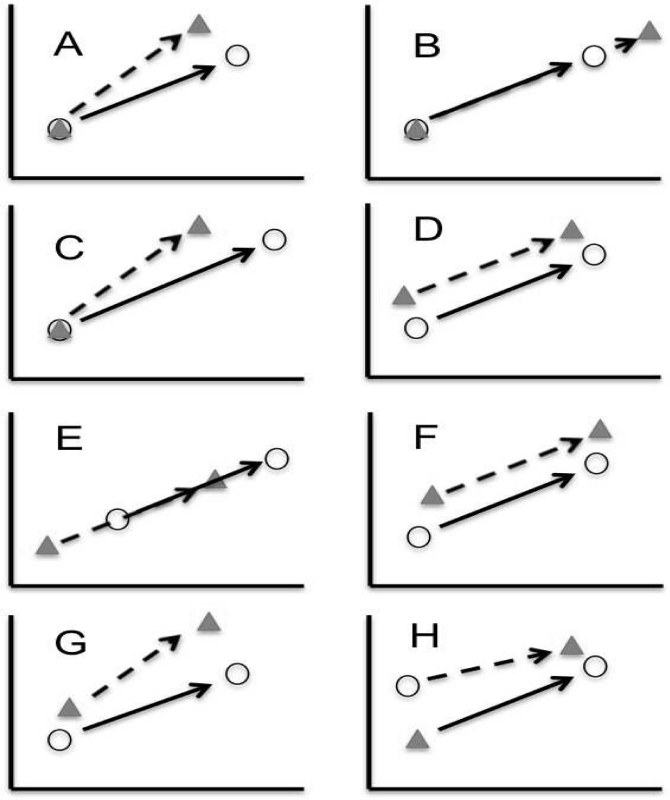

Figure 1 – Simple representation of a pairwise comparison of growth and development of two groups in a two dimensional morphospace. The circles and triangles represent the shapes of the two groups at the outset and end of the developmental stage, the arrows represent the ontogenetic trajectory. (A) Difference in the direction of ontogenetic trajectory, but equal magnitude of net change and a common shape at outset. (B) Overlapping trajectories with a common starting shape and direction, but unequal magnitude of net ontogenetic change. (C) Trajectories with differing directions and magnitudes. (D) Parallel trajectories with equal magnitudes of net change, with an elevation change in the outset shapes. (E) Overlapping parallel trajectories with a shift in outset shape along the direction of the trajectory and unequal magnitudes of change. (F) Parallel trajectories with both elevation changes and a shift along the trajectory of the shape at outset. (G) Equal length trajectories, showing divergence in shape over ontogeny. (H) Equal length trajectories showing convergence over ontogeny.

ments. This paper focuses on the analysis of a single linear segment because multiple segments may be analyzed by repeated application of the single segment models. The regression models used to analyze shape data may be readily adapted to discrete samples using dummy coding techniques. A linear model over a continuous covariate (age or size) suffices for both cases.

Fig. 1 shows a range of possible differences in growth and development of two organisms between two stages of development. The *ontogenetic trajectory* is the description of the change in shape from one stage to the next (Fig. 1). The trajectory itself may thus differ in direction, in length (magnitude) or in both. In addition to alterations in the trajectory, the *shape at the outset* of the trajectories may differ, due to alterations in development prior to the earliest stage. Differences in shape at this outset stage may be divided into *elevation* changes (Fig. 1D) and *shifts* in shape along an ontogenetic trajectory (Fig. 1E). This division of shape changes into two distinct categories at the outset becomes important when the trajectories share a common direction. Elevation changes (i.e., perpendicular to the common direction of the trajectory) produce parallel trajectories (Fig. 1D), while shifts along (i.e., parallel to) the trajectory produce overlapping trajectories (Fig. 1E), such that one group's juvenile resembles an older or younger age-class of the other group. In that case, their trajectories overlap but the age-classes differ in shape because of a shift in starting point along (parallel to) the trajectory. That can arise from differences in rate and/or duration of development prior to the youngest observed phase. A series of resampling tests based on linear models, and flowcharts illustrating a procedure for their use will be presented to address each of the possible differences in both the ontogenetic trajectory itself (direction and magnitude), differences in outset shape (elevation and shifts along a common trajectory) and duration of growth along the trajectory.

Resampling methods for testing a bivariate regression model

Classical parametric statistical methods use mathematical models of statistical distributions to calculate the distributions of test statistics. When the observed value is extreme relative to the distribution implied by the null hypothesis, it is possible to reject that null hypothesis at some calculated probability level. One alternative is to use Monte Carlo methods in which statistical models are fit to the data and then used as the bases for numerical simulations, which then can be used to determine the probability that the observed data was produced by a given null model (see Manly 1997 for examples). Resampling methods offer another approach. Resampling methods refer to a group of related methods (permutations, bootstrapping and jackknifing) that, not surprisingly, use resampling of the data to generate the distribution of a test statistic under the null hypothesis; the idea dates to Neyman (1923); Fisher (1935), and Pitman (1937) but practical application had to wait until inexpensive (and fast) computers became widely available (Efron, 1979; Efron and Tibshirani, 1998; Good, 2000, 2005).

The general process of constructing a resampling test consists of three distinct parts: (1) stating the null hypothesis, (2) determining what test statistic to use, and (3) deciding how to carry out the resampling. As a starting point, consider an ordinary bivariate linear regression model of the size of one trait (Y) relative to the size of another (X):

$$Y = MX + B + \varepsilon \quad (1)$$

M is the slope, B is the intercept and ε is a random error term (the residuals). Analytic approaches typically assume ε is an independently distributed random normal term with a mean of zero. Based on these assumptions, analytic statistics test the null hypothesis that there is no dependence of Y on X using test statistics such as the correlation coefficient R, or the estimate of the slope (M); the null is rejected if R is statistically significantly different from zero, or the confidence interval of M excludes zero. The value of R^2 also expresses the fraction of the variation in Y explained by X (which could, in principle, be used to test the null hypothesis). These analytic tests require algebraic models of the underlying distributions.

Resampling uses numerical randomization procedures to conduct the statistical tests. As noted, above, the first step is to state the hypothesis we want to test. It might seem that we have two, the first being that the correlation between Y and X is zero and the other that the slope is zero, but these are equivalent in that they assert that the model:

$$Y = B + \varepsilon \quad (2)$$

is equally effective at predicting Y as the original model which included X. The second model is called the "reduced model" because it omits a term (X) present in the full model. Whether we choose R, R^2 or M as our test statistic, the null model states that the *reduced* model will often produce a value of that statistic as extreme (i.e., as far from zero) as the *full* model (Eqn. 1). Because the omitted term is X, under the null hypothesis, the relationship between X and Y is not important. An important concept in the theory of resampling is that this makes X *exchangeable* under the null hypothesis (see discussions in Good 2000; Anderson 2001). That means that we could exchange the X value of any given specimen with that of any other specimen because, under the null hypothesis, this relationship is not important.

If X is exchangeable, the null model predicts that a randomly created version of the original data, in which the association between the X and Y values is randomized, will yield a similar distribution of values for any of our test statistics when the (full) regression model (Eqn. 1) is fit to the randomized data. Thus, if we create many such resampled versions of the data, we could generate a distribution of values of the test statistic under the null model. We can then use this distribution to determine the percentage of trials in which the observed value for that test statistic is as far from zero as the observed one. This is what we use

as our estimate of the p value. For example, if we are using R as our test statistic, and our observed value is 0.85 and a value this high never appeared in 999 trials, we would get a $p = \frac{1}{999+1} = 0.001$ or 0.1%. Note that we treat the original data as one possible resampling of the data; we resample 999 times and the 1000th value is the observed one – it therefore counts as a large value in the calculation of the p-value.

There are several different ways to resample the data that differ in two major respects. One is in how the data are selected and handled in the randomization, the other is in what exactly is permuted, the raw data X, or residuals (from either the full or reduced model). For hypothesis testing, permutation (or resampling without replacement) is thought to be the most effective approach, although permutation and bootstrap methods are thought to be asymptotically equivalent (Romano, 1989; Manly, 1997; Good, 2000). When permuting the data, the order of the X values is randomized and re-assigned to the Y values (see Manly 1997; Good 2000). When bootstrapping the data, which is resampling with replacement (Efron, 1979; Efron and Tibshirani, 1998), the individuals' values are randomly drawn, and each random draw is independent of all the others so a given individual's value may be used more than once or not at all. When jackknifing the data, a relatively small percentage of the specimens (from one specimen up to as many as 50% of the total) at a time are removed and the calculation is repeated. In addition to the distinction between the resampling procedure, methods differ according to what they permute (or bootstrap or jackknife). An alternative to permuting or bootstrapping the observed values (X) is to permute the residuals (ε) of the reduced model (Eqn. 2, which in this case implies that there is no slope, only a mean value and random variation around the mean). In this approach, the reduced model is fitted to the data and the residuals are computed and used in the permutation, as the null model implies that there is no ordering or relationship of the residuals relative to X or Y. Resampling residuals is thought to be more effective than simply permuting the observed (raw) variables (Anderson and ter Braak, 2003).

Example: Bivariate regression using different resampling methods

The following example is meant to show a range of different resampling methods applied to a simple bivariate regression model. If we start out with a set of measured values of the dependent and independent variables:

$$\begin{aligned} X =&[2.10\ 2.71\ 3.15\ 3.44\ 4.06\ 4.34\ 5.18\ 5.27 \\ &5.79\ 6.34\ 6.41\ 7.79] \end{aligned} \tag{3}$$

$$\begin{aligned} Y =&[6.54\ 7.69\ 9.10\ 9.15\ 10.86\ 11.47\ 13.46 \\ &13.67\ 14.60\ 16.00\ 16.14\ 19.40] \end{aligned} \tag{4}$$

fitting the full linear regression model, $Y = MX + B + \varepsilon$ we get estimates for the slope and intercept, and an R value

$$M = 2.2615 \qquad B = 1.6772 \qquad R = 0.9992 \tag{5}$$

We can then use the model to find the predicted values of y under the full model, and the residuals $\varepsilon = Y - Y_{predicted}$

$$\begin{aligned} Y_{predicted,full} =&[6.43\ 7.81\ 8.80\ 9.46\ 10.86\ 11.49\ 13.39 \\ &13.60\ 14.77\ 16.01\ 16.17\ 19.29] \end{aligned} \tag{6}$$

$$\begin{aligned} \varepsilon_{full} =&[0.11\ 0.12\ 0.30\ -0.31\ 0.00\ -0.02\ 0.07 \\ &0.08\ -0.17\ -0.02\ -0.03\ 0.11] \end{aligned} \tag{7}$$

For the reduced model (with no slope) $Y_{predicted,reduced} = B + \varepsilon$

$$Y_{predicted,reduced} = 12.34 \tag{8}$$

$$\begin{aligned} \varepsilon =&[-5.80\ -4.65\ -3.24\ -3.19\ -1.48\ -0.87 \\ &1.12\ 1.33\ 2.26\ 3.66\ 3.80\ 7.06] \end{aligned} \tag{9}$$

Based on these, we can see how the various types of resampled sets are formed. If we wanted to permute the variable itself, we would randomize the ordering of Y and regress this permutation set on X

$$\begin{aligned} Y_{perm,variables} =&[13.46\ 11.47\ 19.40\ 9.10\ 16.00\ 13.67 \\ &16.14\ 10.86\ 9.15\ 6.54\ 7.69\ 14.60] \end{aligned} \tag{10}$$

To permute residuals under the reduced model, we would permute them and add the permuted values to the predicted Y value ($Y_{predicted} = B$) under the reduced model

$$\begin{aligned} Y_{perm,residuals} =& [1.33\ 2.26\ 7.06\ -4.65\ 3.66\ -3.24 \\ &1.12\ -3.19\ -5.80\ 3.80\ -1.48\ -0.87] \\ &+ 12.34 \\ &= [13.67\ 14.60\ 19.40\ 7.69\ 16.00\ 9.10 \\ &13.46\ 9.15\ 6.54\ 16.14\ 10.86\ 11.47] \end{aligned} \tag{11}$$

A bootstrapping of the variables themselves might result in

$$\begin{aligned} Y_{bootstrap,variables} =&[16.00\ 9.10\ 19.40\ 10.86\ 9.10\ 9.15 \\ &13.67\ 11.47\ 10.86\ 16.00\ 13.67\ 13.46] \end{aligned} \tag{12}$$

Notice that in $Y_{bootstrap,variables}$, several values (16.00, 9.10, 10.86, 13.67) appear several times, while other values are omitted altogether. In contrast, in $Y_{permutation}$, residuals, each value in the original data set appears once and only once. If we bootstrap the residuals of the reduced model and add them to $Y_{predicted}$, we get

$$\begin{aligned} Y_{bootstrap,residuals} =& [7.06\ -3.19\ 3.66\ 3.66\ -1.48\ 1.12 \\ &-5.80\ -5.80\ 1.12\ 3.66\ 7.06\ -4.65] \\ &+ 12.34 \\ &= [19.40\ 9.15\ 16.00\ 16.00\ 10.86 \\ &13.46\ 6.54\ 6.54\ 13.46\ 16.00\ 19.40\ 7.69] \end{aligned} \tag{13}$$

and we can again see that some residual values (e.g., l7.06, 3.66, 1.12, -5.80) appear two or three times in the bootstrap set.

The original correlation was high, R = 0.992, and perhaps not surprisingly, in 9999 trials of each permutation and bootstrap, no resampled sets ever produced an R greater than or equal to 0.992. Since one of the 10000 (i.e., the original data) *did* have an R that large, the estimated p value is $\frac{1}{10000}$, or p = 0.0001. There are 12! (just over 479 million) possible permutations of the 12 values or residuals for this example, so 10000 trials does not come close to exhausting all possible combinations. Some authors urge using 1000 to 2000 trials in permutation tests (e.g., Manly 1997), but an alternative is to start with a relatively small number of trials and work upwards. If your p-value is 0.40 for 100 trials, running more trials to determine that p = 0.3874 is not productive. However, when the p-value is low relative to the desired α value, it is advisable to run several repetitions at 1000 to 2000 trials to see if that estimate is stable and reliable, increasing the number of trials until a stable estimate of p is obtained. Because resampling methods involve random processes, variance in the exact p-value obtained is expected. It may be necessary to run a relatively large number of trials to ensure that the variability in estimates of p are well below the desired α level. While the variation in p-value may seem worrisome and less precise than those from analytic tests, it is important to remember that analytic estimates of p-values are influenced by violations of the assumptions in the underlying analytic models, as well as by non-random sampling. Their apparent high precision may sometimes be illusionary.

Note that bootstrap and permutation methods assume that residuals are independently distributed, and that the data comprise a representative sample of the underlying population. These methods are not assumption free because they share some basic assumptions common to most statistical approaches.

Multivariate regression model

A wide range of methods have been used to capture information about the shapes of organisms. We focus here on the formalism of landmark-based geometric morphometrics (Bookstein, 1991; Dryden and Mardia, 1998; Adams et al., 2004; Zelditch et al., 2012), in which specimens are represented by a set of k landmarks measured in m dimensions; for 2D data, $m = 2$, and for 3D data, $m = 3$. One major advantage of landmark-based geometric morphometrics is the availability of a well characterized and robust distance metric, a univariate measure of the differences in shapes called a Procrustes distance.

We will assume that the data are superimposed by a Generalized Procrustes Analysis (GPA) so four degrees of freedom are used up when superimposing 2D landmarks and seven are used up when superimposing 3D landmarks. It is possible to include semilandmarks, i.e., points spaced along a curve or outline; these differ from landmarks in that they have only one degree of freedom per semilandmark for 2D data when semilandmark alignment procedures ("sliding") are used with semilandmark data (Sampson et al., 1996; Bookstein, 1997; Zelditch et al., 2012). This distinction becomes important when we consider estimating variance-covariance matrices later in this section.

In the case of shape data, the dependent variable Y is a vector quantity, denoted by $\mathbf{Y}$, each specimen is a row vector of k measurements per specimen. Our independent variable X is still a scalar, but the slope is now also a vector $\mathbf{M}$, as is the intercept $\mathbf{B}$. The error (residual) term $\mathbf{E}$ now also consists of a row vector of k values for each of the N specimens in the data set. This gives us a full model

$$\mathbf{Y} = \mathbf{M}X + \mathbf{B} + \mathbf{E} \tag{14}$$

and a reduced form

$$\mathbf{Y} = \mathbf{B} + \mathbf{E} \tag{15}$$

where $\mathbf{M}$ is the multivariate equivalent of the slope.

Like a univariate regression model, the significance of the full model may be estimated by a permutation test in which the residuals of the reduced model are permuted. The test statistic will be a version of an F-ratio because the F-ratio is the traditional statistic used in univariate analysis of variance (ANOVA and/or ANCOVA). There are many forms of F-ratios used in different experimental designs but all are ratios of sums of squares terms weighted by the degrees of freedom so they are akin to ratios of variances. In analytic models of multivariate data, the usual approach is to replace the sums of squares terms by a sum of squares and cross-products (SSCP) matrix. This is a major change relative to univariate data, not only because it requires a very large sample size to estimate the matrix reliably but because it requires a matrix inversion. For that inversion to be possible the variables must be linearly independent of one another and the degrees of freedom in the data must match the degrees of freedom in the measurements. But superimposition removes either four or seven degrees of freedom (and even more when semilandmark alignment is used) so the matrix is not of full rank. That is why partial warp scores (see Bookstein 1989) were typically used in multivariate statistical procedures that require the variance-covariance matrix to be invertible. Unfortunately, the matrix of partial warp scores is not of full rank when the data include semilandmarks. One approach is to reduce the dimensionality of the data using principal components, and to use the principal component scores in the analysis. That, however, does not solve the problem of estimating large variance-covariance matrices when the sample sizes are relatively small, and by "relative" here we mean relative to the number of landmarks plus semilandmarks.

Fortunately, there is another approach. We can work with *pseudo F-ratios* (Verdonschot and ter Braak, 1994; Legendre and Anderson, 1999), which are based on summed square distances of specimens about the mean rather than SSCP matrices. A distance metric, which for shape data is the Procrustes distance, is used to compute the sums of squares terms (SS), which are now scalars (simple distances) rather than SSCP matrices. For the simple regression model above, the F-ratio would be calculated as:

$$F = \frac{SS_{Model}/df_{Model}}{SS_{Residuals}/df_{Residuals}} \tag{16}$$

$$SS_{Model} = SS_{Total} - SS_{Residuals} \tag{17}$$

All sums of squares terms may be computed from a matrix of the pairwise distances between specimens (the outer product matrices, McArdle and Anderson 2001) because the sums of squares about the mean is proportional to the sums of squares between specimens. These methods were developed for a range of different types of statistical questions by Anderson and colleagues (Legendre and Anderson, 1999; McArdle and Anderson, 2001; Anderson and ter Braak, 2003) using a variety of different distance metrics, paralleling Goodall's (1991) derivation of the approach based specifically on Procrustes distance, adapted and generalized by Rohlf (2009) for permutation tests of regression models and MANCOVA designs. The availability of these pseudo F-tests (or Generalized Goodall's F-ratios) greatly speeds calculations in permutation tests, as well as having a number of other advantages (Anderson and ter Braak, 2003), plus the approach is readily adapted to a variety of experimental designs whether the data are univariate or multivariate. There are several different forms of Procrustes distances (see Dryden and Mardia 1998, or Zelditch et al. 2012), the form employed mostly commonly is properly called a *partial Procrustes distance*, in which the centroid size of all specimens is scaled to 1, and inference is carried out in the linear tangent space of the underlying curved space. The term "Procrustes distance" used hereafter refers to this partial Procrustes distance.

This set of ideas allows us to test the statistical significance of a linear regression model fitted to shape data; the pseudo F-ratio is determined for the full model and compared to the pseudo F-ratio derived from a large number of permutations of the residuals of the reduced model. From that we arrive at an estimated *p*-value for the pseudo-F statistic. This is precisely what we did when estimating the confidence interval for a test statistic in the univariate case.

Comparing ontogenies: Establishing evidence of differences in trajectories

To compare trajectories among two or more groups, the first step is to verify that there are statistically significant differences (of some kind) in the ontogenetic trajectories. Once that is established, we can go on to attempt to determine the nature of those differences. Simply computing the ontogenetic trajectories for each group and immediately doing pairwise comparisons of all the features of the trajectories rapidly leads to Bonferroni problems in the overall significance of the results (i.e., that there are statistically significant differences in the trajectories) because of the very large number of possible pairwise comparisons, which leads to an increased rate of false positive results if each test is done at the typical 5% alpha level. One can use a Bonferroni correction, carrying out each test at a lower alpha level to obtain overall results at the desired 5% alpha level, but another approach is a single overall test to establish overall significance at the desired alpha level. The first step is therefore a MANCOVA. To explain this, we introduce a factor A, the membership of each individual in a group, such as "species", which is a level in the factor A. In our full model $\mathbf{M}$ is now a function of A, as are the intercept terms:

$$\mathbf{Y} = \mathbf{M}(A)\mathbf{X} + \mathbf{B}(A) + \mathbf{E} \tag{18}$$

with the reduced model being:

$$\mathbf{Y} = \mathbf{M}\mathbf{X} + \mathbf{B}(A) + \mathbf{E} \tag{19}$$

This is sometimes called the common slopes model because it says that all groups share a common trajectory. The common slopes model itself has a reduced model with no slope at all:

$$\mathbf{Y} = \mathbf{B}(A) + \mathbf{E} \tag{20}$$

To test for statistically significant differences in the slopes $\mathbf{M}(A)$, we would form a pseudo F-ratio of the sums of squares (SS) explained by the model divided by the residual SS and then form permutation tests based on permutation of the residuals of the reduced model. Note that this permutation version of the F-ratio test does not assume equal variance at each landmark, nor does it assume that variances at each landmark are independent of one another. If this pseudo F-ratio is statistically significant at some desired level of confidence (based on an estimated distribution of pseudo F-ratio values obtained by a permutation or bootstrap test), it is then reasonable to proceed to a series of pairwise tests to understand the nature of the differences. Repeated use of the above procedure can determine which the levels of the group factor actually differed from one another. The tests discussed below can be used to determine the nature of those differences (rate of change, direction

of change or both). It is important to note that the common trajectory model is the null hypothesis in this procedure. The failure to reject the null may depend on sample size; the test is subject to some unknown rate of type II error, so accepting the null of common trajectory ($\mathbf{M}$) is not the same as a statistical proof that the trajectories are in fact common. Effect size in these systems (the magnitude of differences in slope, or in the variance explained) may also provide some insight, and should not be neglecting when examining the results of an F-test. Carefully structured tests based on geometric morphometric methods are thought to have high statistical power, and may detect statistically significant results when the differences in shape are too small to be of biological significance, particularly with relatively large sample sizes. Examination of the effect size involved in a comparison may thus be informative, both when the null is rejected, and when it is not.

Tests for differences in direction

Once a difference in the two vectors is known to exist, we need to determine if this is a difference in direction, in magnitude, or in both. Since we have two vectors, we can compute an angle between them,

$$\cos(\theta) = \frac{\mathbf{M_1} \cdot \mathbf{M_2}}{|\mathbf{M_1}||\mathbf{M_2}|} \tag{21}$$

where the numerator is the dot product of the two vectors and $|\mathbf{M}|$ is the magnitude of a vector $\mathbf{M}$. The angle between the two describes a difference in direction independent of the length (because the vectors are normalized to unit length), so the angle is a reasonable test statistic for differences in direction.

One null hypothesis addressing the need to determine if two ontogenetic trajectories are in the same direction may be stated as: *The observed angle between the trajectories is no larger than might be observed by randomly selecting two samples from within one of the groups*. That is, if we estimate the vector for each of two random samples drawn from a single population, and calculate the angle between them, the observed angle might not be large relative to the distribution of the angles under the null. To perform that test, we would calculate the vectors for each randomly drawn sample from a single group and compute the angle between them. A bootstrapping procedure can be used to estimate the range (or confidence interval) of angles that might appear within each sample (bootstrap methods allow for estimating confidence intervals from the range of variance within a sample via resampling). Bootstrapping is used here rather than permutation because there is no assumption of exchangeability here. Instead we are using the bootstrap to estimate the magnitude of variability in a derived measurement (the angle).

To compute the range of angles possible within a sample, the full regression model (eqn. 14) is fit to the data, and the residuals are calculated. Bootstrap samples are then produced by resampling the residuals with replacement to create two bootstrap sets, of the same sizes as the original data sets. To be conservative, both bootstrap samples of the smaller data set are limited to its sample size. The vectors are then calculated for each of the bootstrap sets and the angle between the bootstrap sets is determined. This is repeated for some large number of bootstrap sets for both groups to determine the confidence intervals of angles generated within each. The observed angle between the two groups may be judged statistically significant at the desired α if it lies beyond the $1 - \alpha$ confidence interval of both bootstrap sets. Failure to reject the null does not mean that the angle between the two trajectories actually *is* zero because the null result depends on the sample size and the unknown rate of type II error.

Tests based on distances between groups

Once the differences in ontogenetic trajectories have been identified, there are still some questions remaining about changes in ontogenies. One is whether the trajectories start at a common shape (Fig. 1B compared to 1E for example). Another is whether the adults are more, less or equally as different the juveniles are (Fig. 1G and 1H). Still another is whether one group undergoes a longer or shorter interval of net shape change than another (Fig. 1A compared to 1C, as one possibility). These questions can all be answered using differences in Procrustes distances between the means of groups, as shown in Fig. 2. Obtaining adequate estimates of the populations at these juvenile and adult stages is the difficult part of the process. The ideal situation is to have good collections of specimens at each stage, the second option is to attempt to estimate the mean shape and variation in the population at these stages based on the regression model and an estimate of the value of the independent variable X at each stage (see Frederich and Sheets 2009; Zelditch et al. 2012).

To test for differences in the mean shapes of two groups (at either juvenile or adult stage), the pseudo F-test based on the regression model discussed above may be used. In this pairwise test, the group membership of specimens is dummy coded as the independent variable X. Members of the first group are assigned an X value of $1/N_1$, members of the second group are assigned a value of -$1/N_2$. Other approaches to dummy coding are possible, this approach simply yields a mean value of zero for the dummy codes. Shape is then regressed on these dummy codes (X), and the permutation F-test is used to determine if the regression is statistically significant, indicating a difference in the mean juvenile shape. This procedure is equivalent to a permutation test based on Goodall's F, which has also been used to study ontogenies. The same test could be used to test for differences in adult shapes.

Confidence intervals for the Procrustes distance between means of two groups may be constructed by bootstrapping the residuals around the mean shape of each group. Both groups are resampled with replacement, and the distances between means recalculated for each bootstrap set. This allows for comparisons of shape differences from juvenile to adult shapes (i.e., the length of the ontogenetic vector), or from adult to adult ($\mathbf{A_{1-2}}$), or from juvenile to juvenile ($\mathbf{J_{1-2}}$), as shown in Fig. 1G, 1H and 2.

The net shape change during growth for two groups could be compared by comparing the distances from mean juvenile to mean adult shape for the two groups ($\delta = |\mathbf{A_{1-2}}| - |\mathbf{J_{1-2}}|$). Convergence on an adult form would imply the adult-to-adult distance for the two groups is smaller than the juvenile-to-juvenile distance (Fig. 1H). Conversely, divergence implies that the juveniles are more similar than the adults (Fig. 1G). It is possible to construct a bootstrap test of the observed differences in distances. For example, if we want to test the hypothesis that the distance from the mean of group A to the mean of group B (D_{AB}) is greater than the corresponding distance from groups C to E (D_{CE}), we could use the difference in distances, $\delta = D_{AB} - D_{CE}$ as our test statistic, measuring all distances in Procrustes units. We would then form a series of bootstrap sets of each of the sets A, B, C and E, bootstrapping within each, and then compute δ for each bootstrap set,

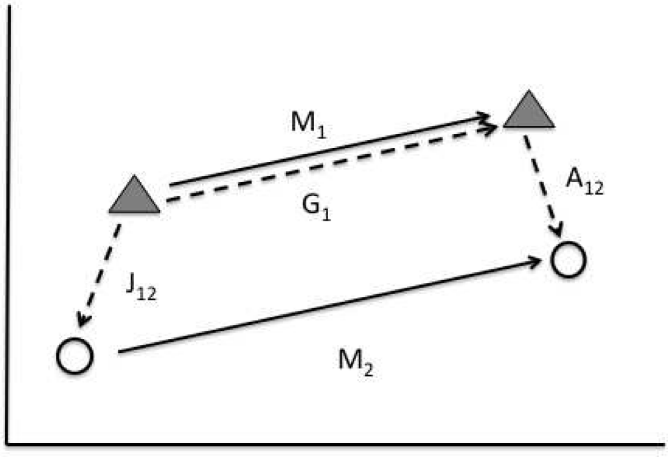

Figure 2 – Diagram of two ontogenetic trajectories with parallel ontogenetic trajectories lying along directions indicated by the multivariate regression slopes $\mathbf{M_1}$ and $\mathbf{M_2}$. The vectors $\mathbf{J_{12}}$ and $\mathbf{A_{12}}$ refer to the vector differences between the juveniles and adults of the two species.

generating a confidence interval on δ. If this interval excludes zero, then we can claim that δ is statistically significantly larger than zero.

Test of differences in elevation and shifts in starting position along the trajectory

A difference in elevation of two trajectories refers to a difference in juvenile shapes that is not along the ontogenetic trajectories, but rather perpendicular to it (Fig. 1D, 1F). If we have a common direction of the ontogenetic trajectory along the multivariate slope $\mathbf{M}$ and a difference vector between juvenile forms $\mathbf{J_{1-2}}$, then the elevation term would be the component of $\mathbf{J_{1-2}}$ perpendicular to $\mathbf{M}$ and the shift of the juvenile form along the trajectory would be the component of $\mathbf{J_{1-2}}$ parallel to $\mathbf{M}$ (Fig. 2). The parallel component is

$$\mathbf{J}_{parallel} = \frac{\mathbf{J_{1-2}} \cdot \mathbf{M}}{|\mathbf{M}|} \tag{22}$$

and

$$\mathbf{J}_{elevation} = \mathbf{J_{1-2}} - \mathbf{J}_{parallel} \tag{23}$$

If the $\mathbf{J}_{parallel}$ term is non-zero, then its dot product with $\mathbf{M}$ should always have the same sign, and should exclude zero, a hypothesis that can be tested via a bootstrap procedure. The magnitude of $\mathbf{J}_{elevation}$ could be tested computing the dot product of $\mathbf{J}_{elevation,bootstrap}$ derived from the bootstrapping procedure with the observed value of $\mathbf{J}_{elevation}$ to see if this dot product is also positive and excludes zero. Simple examination of the magnitudes (lengths) of these vectors would not necessarily be adequate, as distances are always positive. Random variation might generate small but non-zero values of these vectors, requiring the use of the dot product to detect random reversals in direction, which would not be consistent with a meaningful direction and magnitude of these components.

Overlapping trajectories

Overlapping trajectories would be a special case of parallel trajectories, but one in which the juvenile and/or adult shapes varied due to differences in rate along the trajectory or duration of shape change along the trajectory and/or shifts of the starting point along the trajectory (Fig. 1B, E). Overlapping trajectories would have a zero angle between them, and a non-significant difference in elevation, but might differ in magnitude of the ontogenetic trajectory and/or net shape change from juvenile to adult and/or exhibit shifts in juvenile shape along the trajectory.

Mitteroecker et al. (2005) looked for similar evidence of overlapping trajectories by fitting independent regression models to the specimens of each groups using the standard sum of squared residuals approach, but then used as a test criteria only the component of the residuals perpendicular to the predicted trajectory. This perpendicular component was then tested against a permutation of specimens among groups, testing the null hypothesis that group membership did not matter in predicting the perpendicular portion of the residuals, only the portion of the residuals along the trajectory as, specimens moving at different rates along the trajectory would be displaced parallel to the trajectory, not increasing the perpendicular error. If the null hypothesis is true, then the observed summed squared perpendicular errors would be consistent with the observed range of summed squared perpendicular errors generated by the permutation process.

Statistics derived from the parameters in regression models

In some situations it may be desirable to estimate derived statistics based on the results of a regression analysis. For example, if we want to compute the distance from the mean juvenile shape of one group to the mean juvenile form of a second group ($|\mathbf{J_{1-2}}|$), and use a bootstrap procedure to estimate a confidence interval on this distance, the ideal situation would be to have large sample of both groups of juveniles. In many situations, however, the researcher has a series of specimens sampled over wide range of sizes. As noted above, it is possible to estimate the predicted shape at a given juvenile size, and to use the residuals from the regression model to estimate variation around the mean juvenile shape. The distance between the means of two groups can then be estimated based on these bootstrap samples. In carrying out such an analysis, the within sample bootstrapping should be done on the residuals from the regression model, and that model should be refit to the predicted values to re-estimate the mean juvenile shape. That should be done at each iteration of the bootstrap to take the uncertainty of the regression into account.

Disparity

This approach can be used to estimate the uncertainty in derived statistic such as the disparity of a clade, which may be computed at both adult and juvenile states as was carried out in Zelditch et al. (2003). Disparity at any ontogenetic stage may be measured as:

$$Disparity = \frac{\sum_{i=1}^{m} d_i^2}{(m-1)} \tag{24}$$

where the sum is taken over all groups i out of m, and the distance d is the Procrustes distance from the mean of the i-th group to the mean of the group. Zelditch et al. (2003) examined disparity for juvenile and adult piranhas over several groups, using continuous ontogenetic series. To compare the juveniles and adults, the expected values for given sizes and the residuals were obtained from the regression model. Disparity was then calculated based on these predicted shapes, and confidence intervals for disparity were obtained by bootstrapping the residuals of the regression models, recalculating the regressions and re-estimating the disparity for the bootstrap sets. The bootstrapping procedure here incorporated the uncertainty in the regression model. The modeling procedure also allowed for creating hypothetical ontogenies, in which species were simulated to share common juvenile shapes, directions and/or rates, to see the impact of each parameter, singly and in combination, one the diversification of morphology. The simulations used bootstrapped residuals from the regression models employed to estimate the uncertainty in the simulated trajectories and resulting disparity values, in addition to the observed trajectories and disparities.

An approach to combining the tests

The individual tests presented here may be combined to determine the types of differences in ontogenetic trajectories present in a number of distinct groups of specimens. The sequence of tests shown in Fig. 3

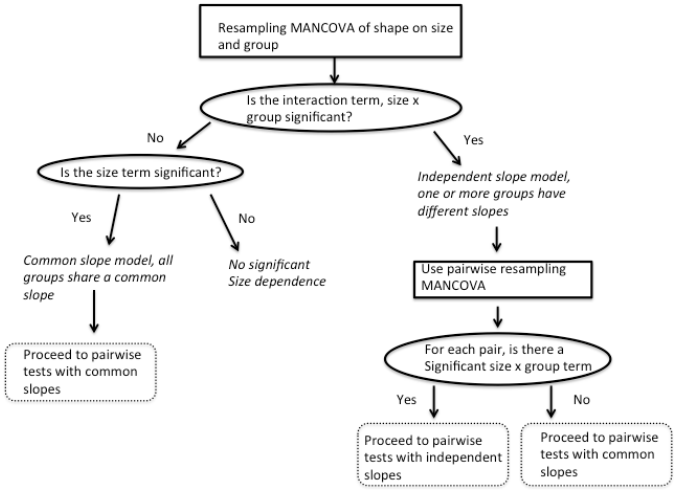

Figure 3 – Flowchart illustrating an approach to combining resampling tests to determine the types of differences in ontogenetic processes in two or more groups of specimens. Solid square boxes indicate a specific test, ellipse indicate decisions made based on the outcomes of the tests and the dashed boxes indicate what types of pairwise tests may be carried out to complete the analysis (Fig. 4 and 5).

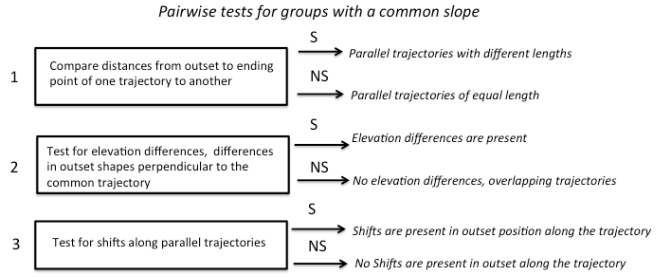

Figure 4 – The set of tests applicable to a pair of species with a shared multivariate slope, indicating parallel trajectories. The three tests may be carried out in any sequence, and need not all be used if they do not meet the goals of a study.

starts off with a MANCOVA of shape based on size and group, followed by a series of pairwise tests of the differences between groups (Fig. 4, 5). It is important to note that not all tests shown in these figures will be necessary for all analyses, the flowcharts are intended to be exhaustive in covering all possibilities. Some authors would chose to omit the pairwise MANCOVA step in favor of proceeding directly to the pairwise angle and trajectory length tests, on the grounds that the angle test adequately addresses the issue of pairwise differences in direction.

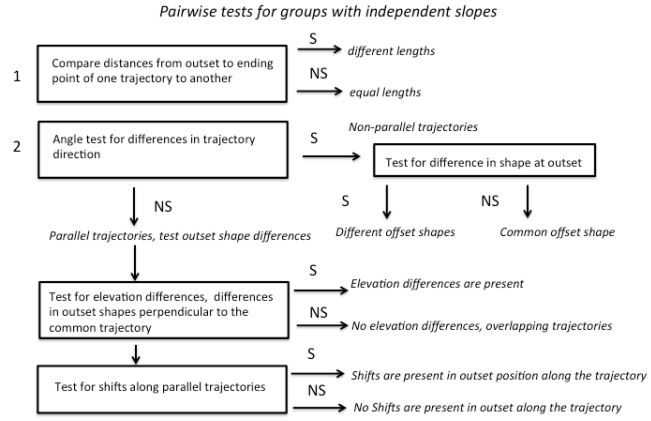

Figure 5 – The set of tests applicable to a pair of species with unequal multivariate slopes. The first and second test are independent of one another, but the result of the second test (differences in direction) does have a bearing on which of the remaining tests are applicable.

Conclusions

The combination of resampling methods, hypotheses based on general linear models and the well-established Procrustes distance as a measure of shape differences allows for a systematic and flexible approach to describing and testing how two or more ontogenetic trajectories differ. Most of the tests discussed here are available in specialized software such as the `tps` series (Rohlf, 2009) and `IMP` series (Sheets, 2001-2012; Zelditch et al., 2012). Customized bootstrapping and permutations methods are readily developed in R (see Good 2005), the `adonis` function in the `vegan` package (Oksanen et al., 2013) in R handles permutation MANCOVA, as does the `DistLM` program (Anderson, 2005). Several R scripts are either generally available or can be obtained by request from the authors that allow for testing many of these hypotheses about the evolution of ontogenies (Adams and Collyer, 2009; Gerber and Hopkins, 2011; Piras et al., 2011), making the approach outlined in this paper available to R users. The ongoing development of shape distance metrics, distance-based statistical tests, resampling methods and related software provide a steadily increasing set of analytic tools to examine morphological change in organisms, providing powerful methods for research in these areas.

References

Adams D.C., Collyer M.L., 2009. A general framework for the analysis of phenotypic trajectories in evolutionary studies. Evolution 63: 1143–1154.

Adams D.C., Nistri A., 2010. Ontogenetic convergence and evolution of foot morphology in European cave salamanders (Family: Plethodontidae). Bmc Evolutionary Biology 10: 216.

Adams D.C., Berns C.M., Kozak K.H., Wiens J.J., 2009. Are rates of species diversification correlated with rates of morphological evolution? Proceedings of the Royal Society B 276: 2729–2738.

Adams D.C., Rohlf F.J., Slice D.E., 2004. Geometric morphometrics: Ten years of progress following the "revolution". Italian Journal of Zoology 71: 5–16.

Anderson M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26: 32–46.

Anderson M.J., 2005. DistLM (software). Available from http://www.stat.auckland.ac.nz/~mja/Programs.htm#Mine [5 September 2012]

Anderson M.J., ter Braak C.J.F., 2003. Permutation tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73: 85–113.

Bookstein F.L., 1989. Principal warps: thing plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 11: 567–585.

Bookstein F.L., 1991. Morphometric tools for landmark data: Geometry and biology. Cambridge University Press, Cambridge.

Bookstein F.L., 1997. Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. Medical Image Analysis 1: 97–118.

Drake A.G., 2011. Dispelling dog dogma: an investigation of heterochrony in dogs using 3D geometric morphometric analysis of skull shape. Evolution & Development 13: 204–213.

Dryden I.L., Mardia K.V., 1998. Statistical shape analysis. Wiley, Chichester.

Efron B., 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM Review 21: 460–480.

Efron B., Tibshirani R.J., 1998. An Introduction to the Bootstrap. Chapman and Hall, Boca Raton.

Fisher R.A., 1935. The design of experiments. Oliver and Boyd, Edinburgh.

Foote M., 1993. Discordance and concordance between morphological and taxonomic diversity. Paleobiology 19: 185–204.

Frederich B., Sheets H.D., 2009. Evolution of ontogenetic allometry shaping giant species: a case study from the damselfish genus Dascyllus (Pomacentridae). Biological Journal of the Linnean Society 99: 99–117.

Frederich B., Vandewalle P., 2011. Bipartite life cycle of coral reef fishes promotes increasing shape disparity of the head skeleton during ontogeny: an example from damselfishes (Pomacentridae). Bmc Evolutionary Biology 11: 82.

Gerber S., 2011. Comparing the differential filling of morphospace and allometric space through time: the morphological and developmental dynamics of Early Jurassic ammonoids. Paleobiology 37: 369–382.

Gerber S., Hopkins M.J., 2011. Mosaic heterochrony and evolutionary modularity: The trilobite genus Zacanthopsis as a case study. Evolution 65: 3241–3252.

Good P.I., 2000. Permutation tests. Springer, New York.

Good P.I., 2005. Introduction to statistics through resampling methods and R/S-plus. Wiley, Hoboken.

Goodall C., 1991. Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society, Series B: Methodological 53: 285–339.

Ivanovic A., Cvijanovic M., Kalezic M.L., 2011. Ontogeny of body form and metamorphosis: insights from the crested newts. Journal of Zoology 283: 153–161.

Legendre P., Anderson M.J., 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs 69: 1-–24.

McArdle B.H., Anderson M.J., 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. Ecology 82: 290–297.

Manly B.F.J., 1997. Randomization, bootstrap and monte carlo methods in biology. Chapman and Hall, London.

Mitteroecker P., Gunz P., Bookstein F.L., 2005. Heterochrony and geometric morphometrics: a comparison of cranial growth in Pan paniscus versus Pan troglodytes. Evolution and Development 7: 244–258.

Neyman J., 1923. On the Application of Probability Theory to Agricultural Experiments. Statistical Science 5: 465–472

Oksanen J., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Stevens M.H.H., Wagner H., (2013). `vegan`: Community Ecology Package. R package version 2.0-6. http://CRAN.R-project.org/package=vegan

Piras P., Salvi D., Ferrar S., Maiorino L., Delfino M., Pedde L., Kotsakis T., 2011. The role of post-natal ontogeny in the evolution of phenotypic diversity in Podarcis lizards. Journal of Evolutionary Biology 24: 2705–2720.

Pitman E.J.G,. 1937. Significance tests which may be applied to samples from any populations. Journal of the Royal Statistical Society B 4: 119–130.

Rohlf F.J., 2009. tpsRegress 1.37. (software). Ecology and evolution. State University of New York at Stony Brook. Available from http://life.bio.sunysb.edu/morph/ [5 September 2012]

Romano J.P., 1989. Bootstrap and randomization tests of some nonparametric hypotheses. The Annals of Statistics 17: 141–159.

Sampson P.D., Bookstein F.L., Sheehan H., Bolson E.L., 1996. Eigenshape analysis of left ventricular outlines from contrast ventriculograms. In Marcus L.F., Corti M., Loy A., Naylor G.J.P., Slice D.E. (Eds.) Advances in Morphometrics. NATO ASI Series A: Life Science, New York. pp. 131–152

Sheets H.D., 2001-2012. IMP software series. Available from http://canisius.edu/~sheets. [5 September 2012]

Verdonschot P.E.M., ter Braak C.J.F., 1994. An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analyses with Monte Carlo permutation tests. Hydrobiologia 278: 251–266.

Zelditch M.L., Sheets H.D., Fink W.L., 2003. The ontogenetic dynamics of shape disparity. Paleobiology 29: 139–156.

Zelditch M.L., Swiderski D.L., Sheets H.D., 2012. Geometric morphometrics for biologists: a primer. Academic Press, London.

Associate Editor: A. Cardini