Systematic Biology Advance Access published August 9, 2011

DOI:10.1093/sysbio/syr061

Analysis of Ratios in Multivariate Morphometry

HANNES BAUR[1,*] AND CHRISTOPH LEUENBERGER[2]

[1]*Department of Invertebrates, Natural History Museum, Bernastrasse 15, 3005 Bern, Switzerland; and*
[2]*Département de mathématiques, Université de Fribourg, Chemin du Musée 3, 1705 Fribourg, Switzerland;*
[*]*Correspondence to be sent to: Department of Invertebrates, Natural History Museum, Bernastrasse 15, 3005 Bern, Switzerland; E-mail: hannes.baur@nmbe.ch.*

Received 15 November 2010; reviews returned 14 January 2011; accepted 17 March 2011
Associate Editor: Norman MacLeod

Abstract.—The analysis of ratios of body measurements is deeply ingrained in the taxonomic literature. Whether for plants or animals, certain ratios are commonly indicated in identification keys, diagnoses, and descriptions. They often provide the only means for separation of cryptic species that mostly lack distinguishing qualitative characters. Additionally, they provide an obvious way to study differences in body proportions, as ratios reflect geometric shape differences. However, when it comes to multivariate analysis of body measurements, for instance, with linear discriminant analysis (LDA) or principal component analysis (PCA), interpretation using body ratios is difficult. Both techniques are commonly applied for separating similar taxa or for exploring the structure of variation, respectively, and require standardized raw or log-transformed variables as input. Here, we develop statistical procedures for the analysis of body ratios in a consistent multivariate statistical framework. In particular, we present algorithms adapted to LDA and PCA that allow the interpretation of numerical results in terms of body proportions. We first introduce a method called the "LDA ratio extractor," which reveals the best ratios for separation of two or more groups with the help of discriminant analysis. We also provide measures for deciding how much of the total differences between individuals or groups of individuals is due to size and how much is due to shape. The second method, a graphical tool called the "PCA ratio spectrum," aims at the interpretation of principal components in terms of body ratios. Based on a similar idea, the "allometry ratio spectrum" is developed which can be used for studying the allometric behavior of ratios. Because size can be defined in different ways, we discuss several concepts of size. Central to this discussion is Jolicoeur's multivariate generalization of the allometry equation, a concept that was derived only with a heuristic argument. Here we present a statistical derivation of the allometric size vector using the method of least squares. The application of the above methods is extensively demonstrated using published data sets from parasitic wasps and rock crabs. [Allometry; Chalcidoidea; Hymenoptera; LDA ratio extractor; morphometry; multivariate statistics; PCA ratio spectrum.]

The use of ratios of measurements (i.e., of body proportions), has a long tradition and is deeply ingrained in morphometric taxonomy (Reyment et al. 1984; Winston 1999; Lestrel 2000; Schuh and Brower 2009). In many animal groups, the indication of such ratios is a standard of species descriptions, diagnoses, or identification keys (Mayr and Ashlock 1991). This is especially true for many arthropods, where ratios are a convenient means for distinguishing between morphologically similar species which often differ significantly in body proportions but not in qualitative characters. In certain insect groups, such as parasitic wasps, numerous ratios are routinely reported (e.g., Townes and Townes 1981; Kasparyan 1989; Noyes 2004; Horstmann 2009) and sometimes up to 30 ratios form the main body of a species description (see, e.g., Graham 1969, 1991). Often the use of ratios is rather implicit in descriptive terms, for instance, when leaves are described as being "narrow" or "broad," both attributes that could be translated into ratios without loss of information. In fact, botanists use numerous such terms for various plant parts that could be partly or wholly substituted by ratios (Stuessy 2009). Ratios are also used for phylogenetic analysis where they are treated as continuous characters (Thiele 1993; Wiens 2000; Rae 2002; Goloboff et al. 2006).

Besides tradition and ease of application, the widespread use of ratios is certainly related to a common way of looking at the shape of organisms. A taxonomist who notices similarity or dissimilarity in proportions of two specimens can always adequately translate them into a series of ratios. Any two individuals are then recognized as having the same shape (i.e., the same body proportions), when all measurements differ by a (positive) constant factor, for instance, when all of them are doubled. It does not matter if a head length to width ratio is, say, 2 : 4 mm or 4 : 8 mm, as long as the ratio (0.5) is the same, the shape (as captured by the ratio) is the same. The geometric shape expressed by ratios is thus invariant for a particular measure of size (Mosimann 1970).

Often it is useful to go one step further and analyze more than two linear distances in a single analysis with the help of multivariate statistical methods. Over the past decades, a wide array of tools has been developed in the field of multivariate morphometry (Reyment et al. 1984; Marcus 1990; Claude 2008). These methods help to unravel hidden population structure or to arrive at a better differentiation of groups, in other words, they give insights in the multivariate data structure that cannot be achieved solely by ratio analysis. Standard applications are principal component analysis (PCA) and Fisher's linear discriminant analysis (LDA), both with raw data (often transformed into logarithmic scale) as the primary input (see Pimentel 1979 for a readable account for biologists and Sorensen and Foottit 1992 for illustrative applications in insect systematics). Both methods aim to transform the original variables into a new system of coordinate axes, whereby most of the variance

is contained in the first two or three axes. Traditionally, the results are then presented as scatter plots. However, the geometric meaning of these plots differs from the one obtained by the analysis of body ratios (Bookstein 1989; Claude 2008).

For this reason, we present versions of the classical LDA and PCA algorithms that are directly adapted to body proportions. In particular, we develop tools that allow us to interpret the numerical results obtained by these multivariate analyses in terms of the body sizes and body proportions of the individuals in question. The first method, adapted to LDA and called the "LDA ratio extractor," allows the extraction of the ratios that are most informative for distinguishing between two or more groups. In this context, we also introduce a measure for deciding how much of the variation between individuals or groups of individuals is due to shape differences and how much is due to size differences. The second tool, called the "PCA ratio spectrum," allows the interpretation of principal components in terms of ratios. In a similar manner, the "allometry ratio spectrum" can be used to assess the extent of allometric behavior in ratios. Furthermore, we present several concepts of size and discuss their relation to multivariate allometry (Klingenberg 1996). Central to this discussion is allometric size (Jolicoeur 1963), a concept that was derived only heuristically. In the Appendix, we therefore provide a statistical derivation of Jolicoeur's allometric size vector using the method of least squares. Finally, the above methods are illustrated with a data set from parasitic wasps (Baur 2002) and a classic data set from rock crabs (Campbell and Mahon 1974). The former is ideally suited for our purpose as ratios are commonly used in the taxonomy of these wasps (see above). The latter is often used for testing new statistical methods; it is included here because of the strong allometric behavior of certain variables.

The mathematical framework, especially the definition of shape and size used in this paper, is adopted from the work of Mosimann (1970), Darroch and Mosimann (1985), Sampson and Siegel (1985), and Rao and Suryawanshi (1996) who has a long and acknowledged history in morphometry (see, e.g., Pimentel 1979; Reyment et al. 1984; Marcus 1990; Klingenberg 1996; Dryden and Mardia 1998; Richtsmeier et al. 2002; Claude 2008). The papers of Mosimann (1970) and Darroch and Mosimann (1985) established the theoretical foundation for the use of body ratios in multivariate analysis and thus provided an ideal starting point for our methods. Sampson and Siegel (1985) and Rao and Suryawanshi (1996) were more concerned with particular definitions of size and shape. In contrast to these authors, our focus is on interpretation of body proportions rather than mere size and shape. Of course, other concepts for the analysis of size and shape (e.g., Cadima and Jolliffe 1996; McCoy et al. 2006; Claude 2008; Hotz et al. 2010) or the analysis of ratios (e.g., Aitchison 1986 for compositional data) have been proposed, but these are, in our opinion, less suited in our context (see below).

METHODOLOGY

The methods presented below consist of a number of steps that are briefly itemized here. The data are first standardized and transformed into logarithms, then the shape space is defined and a suitable size vector chosen. Based on these steps, the best ratios for separation of groups are extracted using a new algorithm adapted to LDA, called the LDA ratio extractor. Associated with this method is a particular measure that allows us to compare the discriminatory power of size with that of shape. The second new tool, called the PCA ratio spectrum, allows us to interpret the axes of a PCA in terms of ratios. A related method, the allometry ratio spectrum, is suitable for examination of the allometric behavior of ratios. Computation of all examples was done with the R statistical software, version 2.11.1 (R Development Core Team 2010) (for obtaining data sets and R files for all methods presented here, see Supplementary Material section).

As mentioned in the introduction, the mathematical framework adopted here originates from Mosimann (1970) and followers. A statistical framework frequently used in the Earth Sciences is Aitchison's analysis of compositional data, also called simplicial analysis (Aitchison 1986; Pawlowsky-Glahn and Egozcue 2001). Typically, compositional data vectors have positive components that sum up to one: imagine, for instance, a rock composed of three minerals in proportions 20%, 50%, and 30%. The corresponding data points $(0.2, 0.5, 0.3)$ lie on a so-called simplex. The unit-sum constraint means a loss of 1 degree of freedom and requires special statistical tools, many of which have been developed by John Aitchison and his followers. We chose not to apply simplicial analysis to morphometric body ratios for two main reasons: First, ratios do not naturally satisfy the unit-sum constraint. Second, ratios have a complicated interrelationship not present in compositional data: the ratios a/b and b/c completely determine the ratio a/c. One could, alternatively, renormalize all body measurements to unit sum and thus obtain scale-free data on a simplex. This would free the path to simplicial analysis. However, it is not obvious to us how to extract statistical information about ratios from these renormalized data in a natural way. Also, our variants of LDA and PCA in Euclidean space would first have to be adapted to simplicial data, and it is not obvious how to do this, either. For these reasons, we preferred Mosimann's framework to that of Aitchison.

Standardizing the Data

For certain multivariate methods, it is important to standardize the data beforehand, otherwise, larger variables will dominate the analysis. As an example, let $\mathbf{u} = (u_1, \ldots, u_p)$ represent vectors of body measurements associated with N individuals of some animal population. It may happen that u_1, say, is many times larger than u_2 and u_3, and so the ratio u_2/u_3 will be largely dominated by the ratios u_1/u_2 and u_1/u_3. For this reason

the variables u_i should be transformed in a way that they are all in the same order of magnitude. A convenient way to achieve this is to divide each variable by its geometric population mean (Claude 2008). The transformed variables will be called y_i. They and their ratios vary around 1. In this scale, a value of $y_i = 1.2$, for example, means that the individual's corresponding body trait is 20% larger than the (geometric) average over the population (strictly speaking, this standardization is only crucial in PCA but has no impact on LDA).

Space of log-ratios.—Our interpretation of results from statistical analysis of shape will mainly take place in the space of ratios (or body proportions)

$$r_{ij} = y_i/y_j.$$

For p variables, there are in principle p^2 ratios; observe, however, that only $p(p-1)/2$ of these are informative and that even less, namely $p-1$, can vary freely.

The relations between ratios being of multiplicative nature, it is common in multivariate morphometry to pass to log-transformed values (Reyment et al. 1984; Klingenberg 1996; Claude 2008). This transformation allows the application of linear statistical methods and furthermore avoids some problems associated with the statistical analysis of ratios (see Hills 1978, in response to Atchley et al. 1976).

We thus denote $x_i = \log y_i$ and

$$d_{ij} = \log r_{ij} = \log(y_i/y_j) = x_i - x_j. \quad (1)$$

Following Aitchison (1983), we call the numbers d_{ij} *log-ratios*. Note that due to our standardization of the original data, the mean of the variables x_j is zero. Also, if $r_{ij} \approx 1$, we have

$$d_{ij} = \log\big(1 + (r_{ij} - 1)\big) \approx r_{ij} - 1$$

and thus the log-ratios roughly correspond to the deviation of the ratios from 100%.

Shape

As mentioned in the introduction, a ratio can be calculated from any two body measurements and be used to describe the form of a specimen. A ratio thus represents one way for defining shape (Claude 2008). Mosimann (1970) generalized this particular concept of shape for many measurements by posing the question, "When do two individuals have the same shape with respect to a finite number of measurements?". His definitions form the basis for our methods and are in the following formally introduced.

To the (standardized) body measurements $\mathbf{y} = (y_1, \ldots, y_p)^T$ of some individual, we would like to assign a set of numbers encapsulating the individual's body shape. We assume that these numbers can be calculated by formulas of the form $y_1^{b_1} y_2^{b_2} \cdot \ldots \cdot y_p^{b_p}$. As shape values should be invariant under scaling $\lambda\mathbf{y}$, the exponents must satisfy the shape restriction

$$b_1 + b_2 + \cdots + b_p = 0. \quad (2)$$

Passing to the log-values x_i, we define

$$\boldsymbol{\beta}(\mathbf{x}) = \log(y_1^{b_1} \cdot \ldots \cdot y_p^{b_p}) = \mathbf{b}^T\mathbf{x} \quad (3)$$

to be the shape function associated to the vector of coefficients $\mathbf{b} = (b_1, \ldots, b_p)$ subject to the shape restriction (2). We will also standardize $\mathbf{b}$ to length 1 ($\|\mathbf{b}\| = 1$). Geometrically, these constraints mean that $\mathbf{b}$ is a unit vector at right angles to the vector $1 = (1, \ldots, 1)^T$, that is, it lies in the $p-1$ dimensional subspace $1^\perp$ ("shape space") orthogonal to the vector 1. If

$$\mathbf{P} = \mathbf{I} - (11^T)/p \quad (4)$$

denotes the orthogonal projection onto the shape space $1^\perp$, then we calculate the shape values $(z_1, \ldots, z_p)$ according to

$$\mathbf{z} = \mathbf{P}\mathbf{x}. \quad (5)$$

The vector $\mathbf{b}$ represents a direction in shape space, and the shape function $\boldsymbol{\beta}(\mathbf{x})$ is the scalar product of $\mathbf{z}$ with the vector $\mathbf{b}$:

$$\boldsymbol{\beta}(\mathbf{x}) = \mathbf{b}^T\mathbf{x} = \mathbf{b}^T\mathbf{z}.$$

Log-ratios d_{ij} are represented by the *log-ratio vectors*

$$\mathbf{b}_{ij} = \mathbf{e}_i - \mathbf{e}_j, \quad (6)$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the i-th and j-th standard base vector in $\mathbb{R}^p$. We collect these vectors to a set $\mathcal{B} = \{\mathbf{b}_{ij}\}_{1 \le i < j \le p}$. The fact that there are many linearly independent subsets of $\mathcal{B}$ spanning 1^T reflects the interdependence of body ratios and poses a major problem for the interpretation of statistical results in terms of body proportions. We will address this problem below.

Size

Analogous to shape functions, a size function can be defined. We stipulate a size function to be of the form $y_1^{a_1} y_2^{a_2} \cdot \ldots \cdot y_p^{a_p}$, but this time the exponents fulfill the size restriction

$$a_1 + a_2 + \cdots + a_p = 1. \quad (7)$$

Thus, an individual with all body measurements doubled, say, will be twice as large. In terms of the log-values $\mathbf{x}$, we define

$$\boldsymbol{\alpha}(\mathbf{x}) = \log(y_1^{a_1} \cdot \ldots \cdot y_p^{a_p}) = \mathbf{a}^T\mathbf{x} \quad (8)$$

to be the size function corresponding to the size vector $\mathbf{a} = (a_1, \ldots, a_p)$. Three size vectors have been commonly proposed in the literature: Isometric size, allometric size, and shape-uncorrelated size, whose definitions are presented in the following. Shape-uncorrelated size is discussed here for the sake of completeness. In developing the methodology below, our focus will be on isometric and allometric size.

Isometric size.—The "democratic" way is to give equal weight to all body measurements. This is tantamount to the choice $\mathbf{a}_0 = (1/p)1$, and the size $\boldsymbol{\alpha}_0(\mathbf{x}) = \mathbf{a}_0^T\mathbf{x}$ is simply the arithmetic mean of $\mathbf{x}$. In many cases, the size $\boldsymbol{\alpha}_0(\mathbf{x})$ and the shape values $\mathbf{z}$ will show significant correlation over the population. This is a sign of the presence of allometry.

Allometric size.—Allometry was first observed by Cuvier and intensively studied by Huxley and Teissier for bivariate data (e.g., body weight vs. some body trait); see Gayon (2000) for a short history of allometry. A generalization to multivariate data sets was proposed by Jolicoeur (1963). He arrived at his definition of allometric size in a rather heuristic way, whereas we propose in the Appendix a statistical model that leads to Jolicoeur's generalization in a natural manner. One way to pass from the bivariate to the multivariate case is by putting forth the question: *Which is the measure of body size fitting optimally into the set of bivariate allometric power laws*

$$y_i = d_i \times (\text{body size})^{c_i}, \quad i = 1, \ldots, p,$$

for suitable coefficients d_i and exponents c_i? A mathematically more precise formulation is given in the Appendix. The answer to this question is the size function associated to the size vector $\mathbf{a}_J$ spanning the first principal component of the log-values $\mathbf{x}$, a fact that is proved in the Appendix by means of the least squares method. More precisely, $\mathbf{a}_J := \mathbf{a}_1/1^T\mathbf{a}_1$ where $\mathbf{a}_1$ is the unit eigenvector of the population covariance matrix $\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}^T)$ corresponding to the largest eigenvalue λ_1: $\boldsymbol{\Sigma}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$.

Shape-uncorrelated size.—A choice of size function that represents the other extreme to allometric size was proposed by Sampson and Siegel (1985) and by Rao and Suryawanshi (1996). Their size vector $\mathbf{a}_R$ has the property that size and shape over the population are uncorrelated. The shape-uncorrelated size vector is given by $\mathbf{a}_R := \boldsymbol{\Sigma}^{-1}1/1^T\boldsymbol{\Sigma}^{-1}1$.

The size vector $\mathbf{a}_R$ is harder to interpret geometrically than $\mathbf{a}_J$. An interpretation is offered in Rao and Suryawanshi (1996): A unit increase in shape-uncorrelated size represents the same average increase (or decrease) in all the variables $x_1, \ldots, x_p$. It is also proved in Rao and Suryawanshi (1996) that $\mathbf{a}_R^T\mathbf{x}$ is the only size function that is stochastically independent of shape if $\mathbf{x}$ has a multivariate normal distribution. This was already shown by Sampson and Siegel (1985) for linear size functions but it holds even true for nonlinear size functions.

The LDA Ratio Extractor: Selecting the Best Ratios with Discriminant Analysis

As mentioned above, LDA is a standard tool in multivariate morphometry. It often allows to distinguish most similar taxa but the numerical results obtained are then hard to interpret. Our aim is to adapt standard LDA in a way that its results admit a convenient interpretation in terms of the body proportions of the specimens under study. Our algorithm is recursive and the basic idea is as follows. In a first step, the ratio with the largest discriminating power is determined. Then a ratio is chosen that has maximal discriminating power but at the same time is as little correlated as possible to the first ratio. If needed, further ratios can be picked out in the same manner.

Suppose that the values $\mathbf{x}_1, \mathbf{x}_2$ stem from two distinct groups with mean $\mathbf{m}_1, \mathbf{m}_2$, and a common (nonsingular) within-groups covariance matrix $\boldsymbol{\Sigma}$. Then Fisher discriminant vector $\mathbf{w}$ is determined by

$$\mathbf{w} \propto \boldsymbol{\Sigma}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{9}$$

and $\|\mathbf{w}\| = 1$. The vector $\mathbf{w}$ is a mixture of size and shape.

Often taxonomists prefer to perform LDA purely within shape space $1^\perp$, that is to ignore the effects of size. Hence, the method is presented entirely in the shape space. The common within-groups covariance matrix of the shape values $\mathbf{z}_i = \mathbf{P}\mathbf{x}_i$, $i = 1, 2$, is given by $\boldsymbol{\Sigma}_1 = \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}$, which is symmetric and positive definite on the subspace $1^\perp$. Because it is singular in $\mathbb{R}^p$, its pseudoinverse must be used to perform the LDA. By singular value decomposition, there exists an orthogonal transformation matrix $\mathbf{O}$ in such a way that

$$\boldsymbol{\Lambda} := \mathbf{O}^T\boldsymbol{\Sigma}_1\mathbf{O} = \text{diag}(\sigma_1, \ldots, \sigma_{p-1}, 0).$$

Set $\boldsymbol{\Lambda}^+ = \text{diag}(\sigma_1^{-1}, \ldots, \sigma_{p-1}^{-1}, 0)$ and $\boldsymbol{\Sigma}_1^+ = \mathbf{O}\boldsymbol{\Lambda}^+\mathbf{O}^T$. The shape discrimination vector $\mathbf{w}_1$ is now determined by

$$\mathbf{w}_1 \propto \boldsymbol{\Sigma}_1^+\mathbf{P}(\mathbf{m}_1 - \mathbf{m}_2). \tag{10}$$

It is hard to interpret $\mathbf{w}_1$ in terms of body proportions because it is a mixture of ratios and, worse, can be written in infinitely many ways as a linear combination of log-ratio vectors (cf., formula 6) from set $\mathcal{B}$. In the next paragraph, we develop an algorithm that extracts the most informative body ratios for between-groups distinction.

Extracting ratios.—Let $\mathbf{x}$ denote the combined data set in which both groups $\mathbf{x}_1$ and $\mathbf{x}_2$ have been centered to 0 individually. Thus, $E(\mathbf{x}) = 0$ and $\text{var}(\mathbf{x}) = \boldsymbol{\Sigma}$. The dominant log-ratio vector from $\mathcal{B}$ with respect to discrimination between groups is the one that has the largest correlation with $\mathbf{w}_1$ in the data set $\mathbf{x}$. More precisely, we consider the correlation coefficients

$$c(\mathbf{b}_{ij}, \mathbf{w}_1) = \frac{|\text{cov}(\mathbf{b}_{ij}^T\mathbf{x}, \mathbf{w}_1^T\mathbf{x})|}{\sqrt{\text{var}(\mathbf{b}_{ij}^T\mathbf{x})\text{var}(\mathbf{w}_1^T\mathbf{x})}} = \frac{|\mathbf{b}_{ij}^T\boldsymbol{\Sigma}\mathbf{w}_1|}{\sqrt{\mathbf{b}_{ij}^T\boldsymbol{\Sigma}\mathbf{b}_{ij} \cdot \mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1}}$$

and set

$$\mathbf{b}_1 := \arg\max_{\mathbf{b}_{ij} \in \mathcal{B}} c(\mathbf{b}_{ij}, \mathbf{w}_1). \tag{11}$$

The discriminating power of a vector $\mathbf{v} \in \mathbb{R}^p$ can be measured by the standard distance $D(\mathbf{v})$, that is, the difference of the means of $\mathbf{v}^T\mathbf{x}_i$, $i=1,2$, divided by the common within-groups standard deviation:

$$D(\mathbf{v}) = \frac{|\mathbf{v}^T(\mathbf{m}_1 - \mathbf{m}_2)|}{(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v})^{1/2}}. \tag{12}$$

The term "standard distance" was introduced in Flury and Riedwyl (1986) (the square of D is sometimes called Rayleigh coefficient). Note that $\mathbf{w}_1$ is the vector in $1^\perp$ that maximizes $D(\mathbf{b})$ among all shape vectors $\mathbf{b} \in 1^\perp$. By (10) and because $\mathbf{P}\mathbf{w}_1 = \mathbf{w}_1$, we have for any $\mathbf{b} \in 1^\perp$:

$$\begin{aligned} c(\mathbf{b}, \mathbf{w}_1) &= \frac{|\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{w}_1|}{(\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b} \cdot \mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1)^{1/2}} = \frac{|\mathbf{b}^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^+\mathbf{P}(\mathbf{m}_1 - \mathbf{m}_2)|}{(\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b} \cdot \mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1)^{1/2}} \\ &= \frac{|\mathbf{b}^T\mathbf{P}(\mathbf{m}_1 - \mathbf{m}_2)|}{(\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b} \cdot \mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1)^{1/2}} = \frac{|(\mathbf{P}\mathbf{b})^T(\mathbf{m}_1 - \mathbf{m}_2)|}{(\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b} \cdot \mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1)^{1/2}} \\ &= \frac{1}{\sqrt{\mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1}} \cdot D(\mathbf{b}). \end{aligned}$$

Thus, we observe that $\mathbf{b}_1$ defined in (11) has the strongest discriminating power among all log-ratio vectors $\mathbf{b}_{ij} \in \mathcal{B}$. The highest possible standard distance for discrimination within size-and-shape space $\mathbb{R}^p$ is $D_{\text{tot}} := D(\mathbf{w})$, where the discrimination vector $\mathbf{w}$ is given by (9). It is necessary to list the values

$$D_{ij} := \frac{D(\mathbf{b}_{ij})}{D_{\text{tot}}}, \tag{13}$$

in order to get the magnitude of the discriminating power of each ratio. In this listing, the log-ratio $\mathbf{b}_{ij}$ with the second largest value D_{ij} is likely to be already largely explained by $\mathbf{b}_1$ due to the strong correlations between ratios. For this reason, we restrict the shape space $1^\perp$ to the subspace H_2 such that the values $\mathbf{b}^T\mathbf{x}$ for $\mathbf{b} \in H_2$ are uncorrelated to $\mathbf{b}_1^T\mathbf{x}$. It is easy to check that H_2 is orthogonal to the vector $\boldsymbol{\Sigma}\mathbf{b}_1$. Projection onto H_2 is given by the matrix

$$\mathbf{P}_2 = \mathbf{I} - \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T,$$

where $\mathbf{M}$ is the $p \times 2$-matrix $\mathbf{M} = [\mathbf{a}_0 | \boldsymbol{\Sigma}\mathbf{b}_1]$. Set $\boldsymbol{\Sigma}_2 = \mathbf{P}_2\boldsymbol{\Sigma}\mathbf{P}_2$ and calculate the (unit length) discrimination vector $\mathbf{w}_2$ according to

$$\mathbf{w}_2 \propto \boldsymbol{\Sigma}_2^+\mathbf{P}_2(\mathbf{m}_1 - \mathbf{m}_2),$$

where $\boldsymbol{\Sigma}_2^+$ is the pseudo-inverse of $\boldsymbol{\Sigma}_2$ (which has rank $p-2$). Now, let $\mathbf{b}_2$ be the log-ratio vector $\mathbf{b}_{ij}$ that shows largest correlation to $\mathbf{w}_2$. Iteration of this procedure leads to the following algorithm to compute the sequence of ratios $\mathbf{b}_i$, $i = 1, \ldots, p-1$:

1. Let $\mathbf{M}_1 = \mathbf{a}_0$ and initialize $k = 1$.
2. Set $\mathbf{P}_k = \mathbf{I} - \mathbf{M}_k(\mathbf{M}_k^T\mathbf{M}_k)^{-1}\mathbf{M}_k^T$ and $\boldsymbol{\Sigma}_k = \mathbf{P}_k\boldsymbol{\Sigma}\mathbf{P}_k$. Determine the pseudo-inverse $\boldsymbol{\Sigma}_k^+$ and set
$$\mathbf{w}_k = \boldsymbol{\Sigma}_k^+\mathbf{P}_k(\mathbf{m}_1 - \mathbf{m}_2).$$
3. Let $\mathbf{b}_k = \arg\max_{\mathbf{b}_{ij} \in \mathcal{B}} c(\mathbf{b}_{ij}, \mathbf{w}_k)$.
4. Add the column $\boldsymbol{\Sigma}\mathbf{b}_k$ to the matrix $\mathbf{M}_k$:
$$\mathbf{M}_{k+1} = [\mathbf{a}_0 | \boldsymbol{\Sigma}\mathbf{b}_1 | \ldots | \boldsymbol{\Sigma}\mathbf{b}_k].$$
5. Increase k by one unit (unless $i = p-1$), and continue at Step 2.

In practice, only a few iterations will be performed because the first two or three log-ratios $\mathbf{b}_1, \mathbf{b}_2, \ldots$ will already explain most of the discrimination between the two groups.

Extracting ratios for multiple groups.—Suppose we are given K groups (classes) $\mathbf{x}_1, \ldots, \mathbf{x}_K$ with means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ and a common within-groups covariance matrix $\boldsymbol{\Sigma}$. The between-groups covariance matrix is defined by

$$\mathbf{B} = \sum_{k=1}^{K} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T,$$

where $\mathbf{m}$ is the total mean and n_k is the number of individuals in each group. A frequently used criterion for discrimination in the multiple group case is

$$Q(\mathbf{v}) = \frac{\mathbf{v}^T\mathbf{B}\mathbf{v}}{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}.$$

The unit vector $\mathbf{v}_1$ maximizing $Q(\cdot)$ is the eigenvector of $\boldsymbol{\Sigma}^{-1}\mathbf{B}$ with largest eigenvalue. The generalization of our two-group algorithm explained above to the multiple group case is the following:

1. Let $\mathbf{M}_1 = \mathbf{a}_0$ and initialize $k = 1$.
2. Set $\mathbf{P}_k = \mathbf{I} - \mathbf{M}_k(\mathbf{M}_k^T\mathbf{M}_k)^{-1}\mathbf{M}_k^T$, $\boldsymbol{\Sigma}_k = \mathbf{P}_k\boldsymbol{\Sigma}\mathbf{P}_k$ and $\mathbf{B}_k = \mathbf{P}_k\mathbf{B}\mathbf{P}_k$. Determine the pseudo-inverse $\boldsymbol{\Sigma}_k^+$ and let $\mathbf{w}_k$ be the eigenvector of $\boldsymbol{\Sigma}_k^+\mathbf{B}$ with largest eigenvalue.
3. Determine
$$\mathbf{b}_k = \arg\max_{\mathbf{b}_{ij} \in \mathcal{B}} \frac{\mathbf{b}_{ij}^T\boldsymbol{\Sigma}\mathbf{w}_k}{\mathbf{b}_{ij}\boldsymbol{\Sigma}\mathbf{b}_{ij}}.$$
4. Add the column $\boldsymbol{\Sigma}\mathbf{b}_k$ to the matrix $\mathbf{M}_k$:
$$\mathbf{M}_{k+1} = [\mathbf{a}_0 | \boldsymbol{\Sigma}\mathbf{b}_1 | \ldots | \boldsymbol{\Sigma}\mathbf{b}_k].$$
5. Increase k by one unit and continue at Step 2.

The philosophy behind this algorithm is exactly the same as in the two-group case: First, we determine the linear discriminant $\mathbf{w}_k$ and choose the log-ratio vector $\mathbf{b}_{ij}$ with strongest correlation to $\mathbf{w}_k$. Then we project to a subspace of shape vectors that are uncorrelated to all log-ratio vectors that have already been chosen. Again, two or three iterations will be sufficient in practice.

Judging the influence of size.—As mentioned above, the LDA ratio extractor was developed in the shape space that is convenient for most circumstances. Sometimes,

however, it might be informative to know how well particular groups are separated in relation to size. In order to assess how much of the total separation is due to size, we define $D_{\text{size}} := D(\mathbf{a}_0)/D_{\text{tot}}$ and $D_{\text{shape}} = D(\mathbf{w}_1)/D_{\text{tot}}$. One can then view the number

$$\delta = \frac{D_{\text{size}}}{D_{size} + D_{\text{shape}}}, \tag{14}$$

as a measure of how well size discriminates in comparison with shape.

The PCA Ratio Spectrum: Interpreting Principal Components with Ratios

PCA is a very widely used method in multivariate statistics (Jolliffe 2004). In contrast to LDA, specimens are not assigned to different groups for a PCA but are treated as a single group. The resulting scatterplots can then be used to explore the structure of variation in this group. It might be the case that the pattern recovers groupings based on other sets of characters (qualitative morphology, molecular markers, etc.), which would give them additional weight. Usually, individual principal components are interpreted in terms of the original variables (see Jolicoeur and Mosimann 1960 and Manly 2005 for lucid examples). The method developed below allows an interpretation using ratios. The main ingredient of this method is a diagram that we call the PCA ratio spectrum. It allows the user to immediately read off the dominant ratios as well as their interrelationships (recall that ratios are always interdependent in a complex fashion as their number is larger than the degree of freedom in the data).

The technical details of this method and its theoretical justification are presented below. Let the random vector $\mathbf{x}$ with $E(\mathbf{x}) = 0$ and $\text{cov}(\mathbf{x}) = \mathbf{\Sigma}$ (assumed to be nonsingular) represent body measurements of a given population. The first principal components vector $\mathbf{u}_1 = (u_i)_{i=1,\ldots,p}$ of the shape values $\mathbf{z} = \mathbf{P}\mathbf{x}$ is the eigenvector of $\mathbf{\Sigma}_1 = \mathbf{P}\mathbf{\Sigma}\mathbf{P}$ corresponding to the largest eigenvalue λ_1 of $\mathbf{\Sigma}_1$:

$$\mathbf{\Sigma}_1\mathbf{u}_1 = \lambda_1 \cdot \mathbf{u}_1.$$

For a log-ratio vector $\mathbf{b}_{ij}$, we have

$$\text{cov}(\mathbf{b}_{ij}^T\mathbf{x}, \mathbf{u}_1^T\mathbf{x}) = \mathbf{b}_{ij}^T\mathbf{\Sigma}_1\mathbf{u}_1 = \lambda \cdot \mathbf{b}_{ij}^T\mathbf{u}_1 = \lambda \cdot (u_i - u_j). \tag{15}$$

This fact allows a simple graphical interpretation of the first principal component $\mathbf{u}$ in terms of body proportions: The numerical values (coefficients) of the components of $\mathbf{u}_1$ are drawn as points on the real line. We call this diagram the PCA ratio spectrum of the vector $\mathbf{u}_1$. To a pair of points u_i, u_j on the spectrum with a large difference corresponds a body proportion $\log(y_i/y_j)$ that contributes substantially to the first principal component; on the other hand, close points on the spectrum contribute little. The PCA ratio spectrum represents a mixture of all body proportions and shows how much each of them contributes to the variation in relation to the others. This can be illustrated with the example given in Figure 2b. As can be seen by their comparable separation in the spectrum, the ratios *gaster breadth:gaster length* and *postmarginal vein:tergum 7 length* have similar explaining power for the variance. On the other hand, the ratio *eye breadth:scape length* has no explanatory power because the corresponding points are very close in the spectrum.

If desired, the same procedure can be applied to the second and following principal components. Let us emphasize again that the method can only be applied in a statistically consistent manner when a PCA is performed within the shape space.

Statistical stability of the PCA ratio spectrum.—Sometimes it might be useful to test whether the PCA ratio spectrum is statistically stable. Instability occurs when the largest eigenvalue λ_1 is not sufficiently distinct from the smaller eigenvalues of $\mathbf{\Sigma}_1$, though this rarely might be the case in practice. In order to obtain confidence intervals for the points u_i on the PCA ratio spectrum we assume that the values $\mathbf{x}$ and hence $\mathbf{z}$ are normally distributed. More precisely: Let $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_n$ be a random sample created from a multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_1)$. Denote by $\hat{\mathbf{\Sigma}}_1$ the sample covariance matrix and by $\hat{\mathbf{u}}_1$ the standardized first principal components vector of $\hat{\mathbf{\Sigma}}_1$, pointing in the same half-space as $\mathbf{u}_1$. The sampling distribution of $\hat{\mathbf{u}}_1$ is complicated but Anderson has established its large-sample distribution (see theorem 13.5.1 in Anderson 2003). It follows from this result that for sufficiently large sample size n, the marginal distribution of the i-th component of the random vector $\hat{u}_i$ is approximatively normally distributed according to $\hat{u}_i \sim \mathcal{N}(u_i, \sigma_i^2)$ where

$$\sigma_i^2 = \frac{\lambda_1}{n}\sum_{k=2}^{p-1}\frac{\lambda_k}{(\lambda_1 - \lambda_k)^2}u_{i,k}^2. \tag{16}$$

Here, $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{p-1}$ are the positive eigenvalues of the matrix $\mathbf{\Sigma}_1$ (which has rank $p-1$) and $u_{i,k}$ are the elements of the matrix $\mathbf{U} = (\mathbf{u}_1|\ldots|\mathbf{u}_{p-1})$ formed by the corresponding standardized eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}$. (The eigenvector $\mathbf{u}_p$ corresponding to $\lambda_p = 0$ is proportional to the isometric size vector $\mathbf{a}_0$.) Graphically, we represent the 68% confidence intervals $[u_i - \sigma_i, u_i + \sigma_i]$ as perpendicular bars of length $2\sigma_i$ at the corresponding point u_i on the spectrum (Fig. 2b). If the interval lengths are not too large compared with the separation of the points on the spectrum—as is the case in Figure 2b—then the spectrum can be considered as statistically stable. Even when the normal assumption is violated, the confidence intervals still give some indication of the stability of the spectrum.

Alternatively, one can also sample the original values $\mathbf{z}$ directly from the empirical distribution and obtain similar intervals with a bootstrap. The latter was used for estimating the confidence intervals in Figure 2b.

The Allometry Ratio Spectrum: Assessing Allometric Behavior of Ratios

The idea of a ratio spectrum introduced above is also useful for extracting body ratios that show allometric behavior. For a given size vector (like $\mathbf{a}_0$ or $\mathbf{a}_J$), the body ratio that shows the most distinctive allometric growth can be interpreted as the one whose covariance with the body sizes $\mathbf{a}^T\mathbf{x}$ is maximal. We obtain

$$\mathrm{cov}(\mathbf{b}_{ij}^T\mathbf{x}, \mathbf{a}^T\mathbf{x}) = \mathbf{b}_{ij}^T\boldsymbol{\Sigma}\mathbf{a} = d_i - d_j,$$

where we have set $\boldsymbol{\Sigma}\mathbf{a} =: \mathbf{d} = (d_i)_{i=1,\ldots,p}$. Hence, exactly as in the preceding paragraph, the body proportions with strongest allometric growth along the size vector $\mathbf{a}$ can be read off the allometry ratio spectrum of $\mathbf{d}$. A reasonable choice of a size vector is Jolicoeur's size vector $\mathbf{a}_J$. In that case, we have $\mathbf{d} \propto \mathbf{a}_J$ and thus the allometric body proportions can be directly determined by the spectrum of $\mathbf{a}_J$. An illustration of such a spectrum is given in Figure 4.

RESULTS

Discriminating Species

As an illustration of how to apply the LDA ratio extractor, we revisit a statistical analysis from Baur (2002) where morphometric data from two species of parasitic wasps were examined, namely the species *Pteromalus albipennis* Walker, 1835 and *P. solidaginis* Graham and Gijswijt, 1991 from the *Pteromalus albipennis* group (Insecta: Hymenoptera: Chalcidoidea). The analysis is based on $p = 23$ characters (called "head breadth," "OOL," "eye height," etc.) measured on $n_1 = 32$ individuals from *P. albipennis* (Group 1) and $n_2 = 19$ individuals from *P. solidaginis* (Group 2), see Baur (2002) for a complete description. The common within-group variance is estimated by

$$\hat{\boldsymbol{\Sigma}} = \frac{n_1}{n_1 + n_2}\hat{\boldsymbol{\Sigma}}_1 + \frac{n_2}{n_1 + n_2}\hat{\boldsymbol{\Sigma}}_2,$$

where $\hat{\boldsymbol{\Sigma}}_1$, $\hat{\boldsymbol{\Sigma}}_1$ are the estimated covariance matrices of the two groups.

Before performing LDA, we would like to add a word of caution rarely mentioned in the textbooks: If the total number of individuals $n = n_1 + n_2$ is not distinctly larger than the number p of body traits, the results from an LDA can be completely spurious. The reason is that the dimension is large enough that a separating plane is likely to exist between the two groups even if the sample points are completely random. As a rule of thumb, one should always have $n > 2p + \sqrt{p}$. A theoretical justification of this rule is given in MacKay (2003, p. 490).

By applying the LDA ratio extractor introduced in the Methodology section, we obtain *OOL:gaster length* as the most discriminating ratio. We get $D_{\text{size}} = 0.064$ and $D_{\text{shape}} = 0.964$, hence $\delta = 0.063$ (cf., formula 14). Thus, discrimination between the groups stems mostly from shape differences. The next discriminating body ratio being as little correlated as possible with *OOL:gaster length* is *eye breadth:marginal vein*. Its standard distance D_{ij} (see formula 13) is 2.1 as compared with the standard distance $D_{ij} = 5.6$ for the first ratio. As can also be seen from the scatterplot in Figure 1a, the discriminating power as compared with the first ratio is already much lower. Figure 1b shows the next two ratios extracted from the algorithm, *funicle 1 length:propodeum length* and *scape length:postmarginal vein*, with standard distances $D_{ij} = 2.3$ and $D_{ij} = 1.7$, respectively. By looking at the plots in Figure 1a and b, one could be tempted to simply

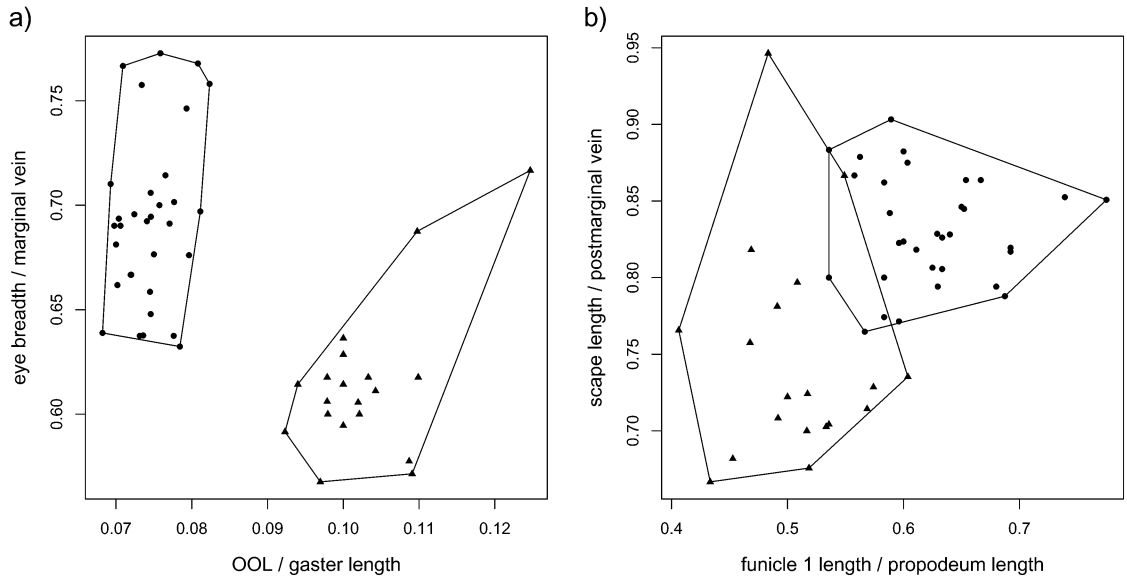

FIGURE 1. Scatter plots of the four most discriminating ratios for *Pteromalus albipennis* (dots) and *P. solidaginis* (triangles). Plot (a) shows first versus second ratio, plot (b) third versus fourth ratio.

combine the first *(OOL:gaster length)* with the third ratio *(funicle 1 length:propodeum length)* to arrive at an even better separation of groups. However, one should bear in mind that these ratios are highly correlated and therefore stand for more or less the same information.

Interpreting Principal Components

Figure 2a shows the results of a PCA on the same data set, but this time the two *Pteromalus* species were entered in the analysis as a single group. A PCA is always useful for examining the structure of variation in a single population, for instance, when it is difficult to assign specimens to different groups beforehand (Pimentel 1979; Reyment et al. 1984; Claude 2008). It can also give additional weight for groupings based on other features. In this case, the specimens in the scatterplot were labeled as either *P. albipennis* or *P. solidaginis* according to qualitative character differences, such as coloration or forewing pilosity, and host plant association (see Graham and Gijswijt 1991). As can be seen from Figure 2a, the first principal component is fully congruent with the separation of species. For the

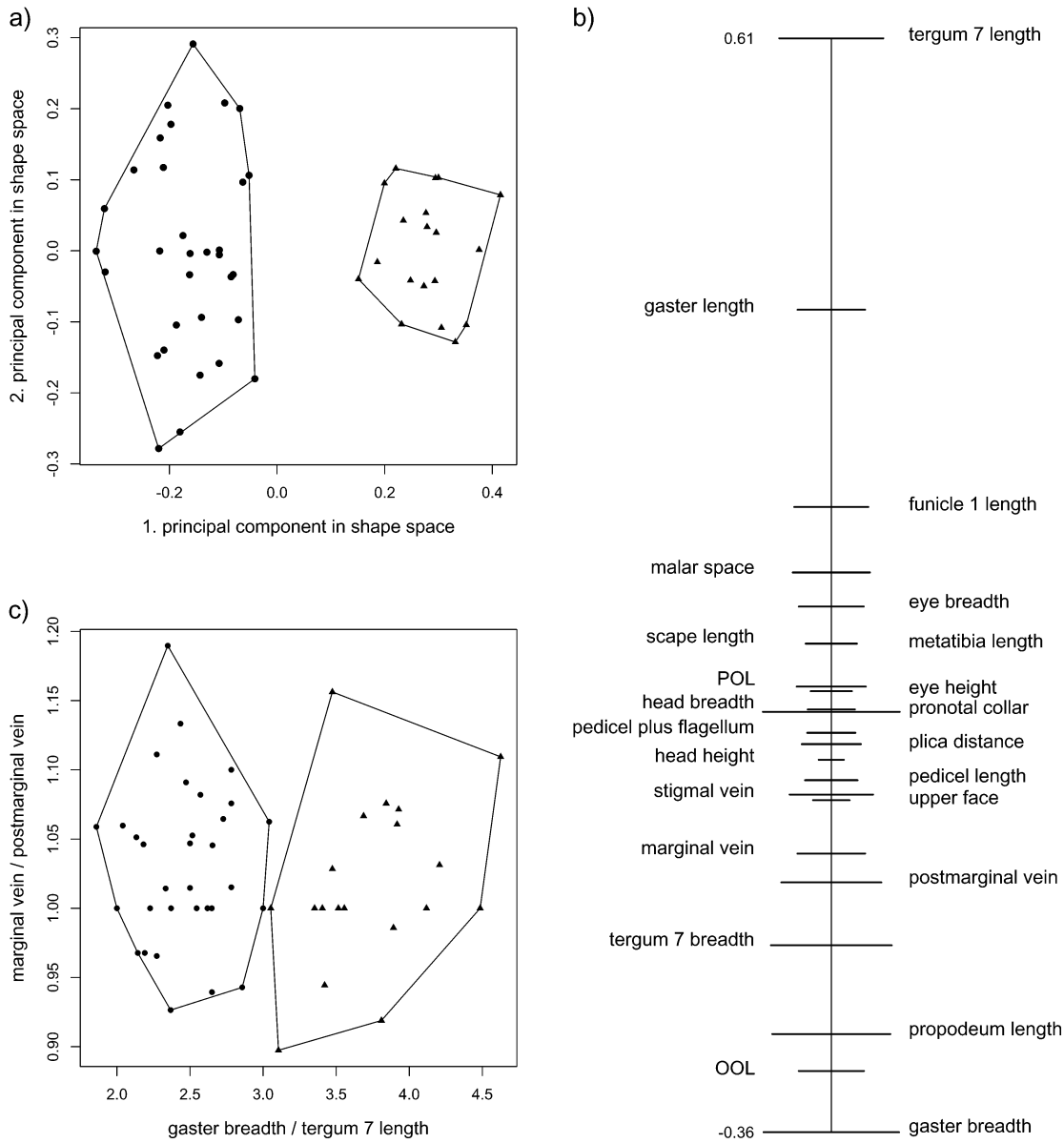

FIGURE 2. Application of the PCA ratio spectrum using the *Pteromalus* data, with *Pteromalus albipennis* (dots) and *P. solidaginis* (triangles). (a) Scatterplot of a principal component analysis (PCA) in shape space. (b) PCA ratio spectrum of the first principal component. The ratio formed from the extremal points (i.e., *gaster breadth:tergum 7 length*) explains a large part of the variation of the first component. In contrast, ratios formed from characters lying close to each other in the spectrum (e.g., *marginal vein:postmarginal vein*) explain very little. This is apparent in the scatterplot (c). Confidence intervals (horizontal bars in (b), see Methodology section) were estimated with a bootstrap.

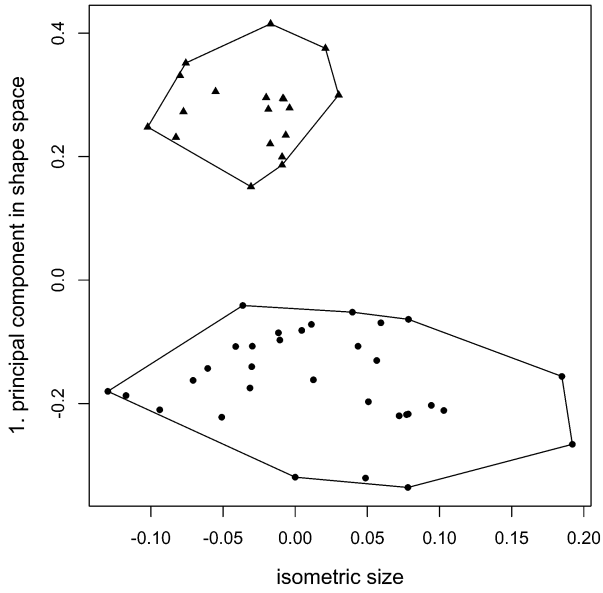

FIGURE 3. Scatterplot of isometric size versus first principal component in shape space for the *Pteromalus* data set, with *Pteromalus albipennis* (dots) and *P. solidaginis* (triangles). The mean size of *P. solidaginis* is obviously smaller but it still lies within the range of *Pteromalus albipennis*.

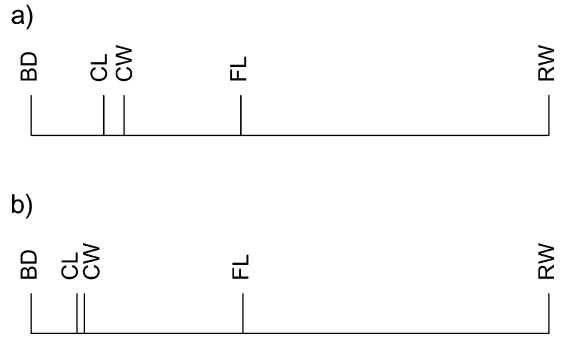

FIGURE 4. The allometry ratio spectrum for the *Leptograpsus variegatus* data set for blue type males (a) and for orange type males (b) respectively. The characters shown are carapace length (CL) and width (CW), width of frontal lobe (FL), rear width (RW), and body depth (BD) (see Results section). The bars do not represent confidence intervals here.

interpretation of this component, the PCA ratio spectrum is displayed in Figure 2b. Most of the variation is explained by ratios like *gaster breadth:tergum 7 length* that correspond to points lying at the opposite end of the spectrum. On the other hand, ratios formed from characters lying adjacent to each other in the spectrum, like *marginal vein:postmarginal vein*, explain very little. This is visualized in the scatterplot of the two ratios (Fig. 2c). Of course, also the ratio spectra of the second and third principal component could be drawn and sometimes this might be illuminating as well, for instance, for explaining the structure of variation within each species.

The above analysis exemplifies the use of our methodology in the shape space. Sometimes a researcher might be interested to examine differences in the size of the specimens, for instance, for investigating the influence of ecological parameters or different food regimes on populations (McCoy et al. 2006). Here, one could simply plot the isometric size axis (see Size section above) against the first principal component in shape space. From Figure 3 it is evident that the mean size of *Pteromalus solidaginis* is smaller, but that its range still lies within *P. albipennis*.

Assessing Allometry

We will illustrate the use of the allometry ratio spectrum on a classical data set of specimens of the purple rock crab *Leptograpsus variegatus* (Fabricius, 1793) (Crustacea: Brachyura: Grapsidae) from Western Australia (see Campbell and Mahon 1974). These occur in two color forms, blue and orange. Mahon collected 50 individuals from each color form and from each sex and made five body measurements: carapace length (CL) and width (CW), width of frontal lobe (FL), rear width (RW), and body depth (BD). We calculated the allometric size vectors $\mathbf{a}_J$ for the body measurements of the males of both the blue and the orange morph. Figure 4 shows the corresponding allometric ratio spectra for both morphs. As can be seen, the ratio BD:RW shows the largest allometric growth whereas for CL:CW allometry is negligible in both groups. Figure 5 confirms this conclusion: There we display a scatter plot for the orange type males of the isometric sizes versus the log-ratios of BD:RW and CL:CW, respectively. Whereas the first ratio (Fig. 5a) visibly has a strong correlation with isometric size, as is characteristic for allometry as explained in the Methodology section, this is much less the case for the second ratio (Fig. 5b).

It is useful to test allometry versus isometry, that is, to test the null hypothesis that $\mathbf{a}_J = \mathbf{a}_0$. Such a test, under the hypothesis of normality and relatively large sample size, was developed by Anderson (2003) (see section 11.6.2). Adapted to our situation, the P value of the null hypothesis is given by $\text{Prob}(\chi^2_{p-1} > \kappa)$ where the test value κ is determined by

$$\kappa = n(p\lambda_1 \mathbf{a}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{a}_0 + p\lambda_1^{-1} \mathbf{a}_0^T \boldsymbol{\Sigma} \mathbf{a}_0 - 2).$$

Here, $\boldsymbol{\Sigma}$ is the covariance matrix of the sample $\mathbf{x}$ of size n and λ_1 is its largest eigenvalue. For the male *Leptograpsus*, the P values are virtually zero for both color types, hence the null hypothesis that no allometry is present can safely be rejected.

DISCUSSION

As initially mentioned, a number of body measurements are commonly collected in taxonomic research. This mainly serves two purposes. First, the raw or

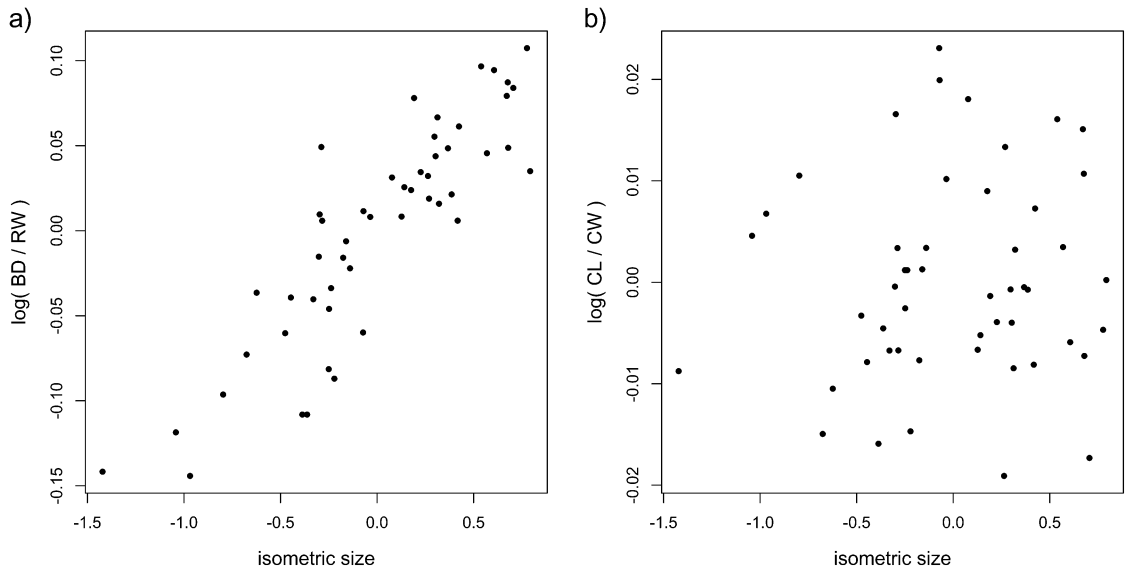

FIGURE 5. Scatter plots of isometric size versus log-ratios *body depth:rear width* (a) and *carapace length:width* (b) for the orange type males in the *Leptograpsus variegatus* data set.

log-transformed variables are entered in some kind of standard multivariate statistical analysis (MVA) for studying character variation and for discrimination of taxa. PCA and LDA are among the methods of choice in this respect and are the ones we refer to with MVA below. Second, the same measurements are integrated in descriptive works, but this time by calculating ratios (indeed, the numerical output from MVA would be far too awkward for inclusion in descriptions and identification keys). Of course, it would be most useful if, say, a discriminant function could be interpreted in terms of ratios that then could be directly used for a species description. One could, for instance, expect some guidelines for the choice of ratios. So far, this was not possible because the two kinds of analysis were not directly comparable (see below). Thus, ratio analysis usually adheres to certain standards established for a particular group, rather than following the insights gained from MVA. A case in point is the study of the *Encarsia meritoria* species complex (Insecta: Hymenoptera: Aphelinidae) by Polaszek et al. (2004), where some of the best ratios used for species discrimination were not even included in their elaborate PCA and LDA.

The incompatibility of MVA and ratio analysis results from the way, size and shape functions are defined for each method (see Fig. 6 for further details). However, the methods presented here, namely the newly developed LDA ratio extractor and the PCA ratio spectrum, solve these problems by using the same definitions for size and shape. Therefore, the results from MVA can now be interpreted in terms of ratios that, in turn, can be directly incorporated in a variety of descriptive taxonomic works. In fact, a more sophisticated use of ratios may be achieved, as is demonstrated by our application of the LDA ratio extractor to the data set from parasitic wasp species of the family Pteromalidae. Here, the best ratios found for separating the two *Pteromalus* species were *OOL (distance of lateral ocellus to eye margin):gaster (abdomen) length*, *funicle 1 (antenna) length:propodeum length*, etc. (see Results section and Fig. 1). These ratios relate characters from widely separated body parts and differ from those commonly used in the taxonomy of pteromalid wasps. For instance, in Graham (1969), still the standard reference in the field (Grissell and Schauff 1997), ratios are exclusively formed from characters lying adjacent to each other, like *eye height:breadth* or *thorax length:breadth* (see also Graham and Gijswijt 1991). Evidently, the variation of such ratios among specimens can—to a certain extent—be judged by eye. However, as demonstrated here, these ratios are apparently not the best ones for discrimination. It is of course very difficult if not impossible to judge by eye the discriminating power of ratios based on widely separated characters, a task that is best done analytically with the help of an algorithm such as the one presented in this paper.

The present methodology can thus easily be embedded in a consistent statistical frame work for the multivariate analysis of morphometric data. In particular, it allows us to interpret the results of a PCA and LDA entirely in terms of ratios, which themselves form the core information of most quantitative taxonomic works. The important point of the new methodology is to determine the shape values and to choose a particular size vector beforehand. For the size function, we mainly considered the isometric size vector $\mathbf{a}_0$, except for the allometry ratio spectrum, which relates to Jolicoeur's allometric size vector $\mathbf{a}_J$. Of course, other definitions of shape and size are possible (see Bookstein 1989 for a review). By using the "back-projection" method of Burnaby (1966), some authors (e.g., Klingenberg 1996; McCoy et al. 2006) choose to define their shape values by projecting the log-data $\mathbf{x}$ on the space orthogonal to

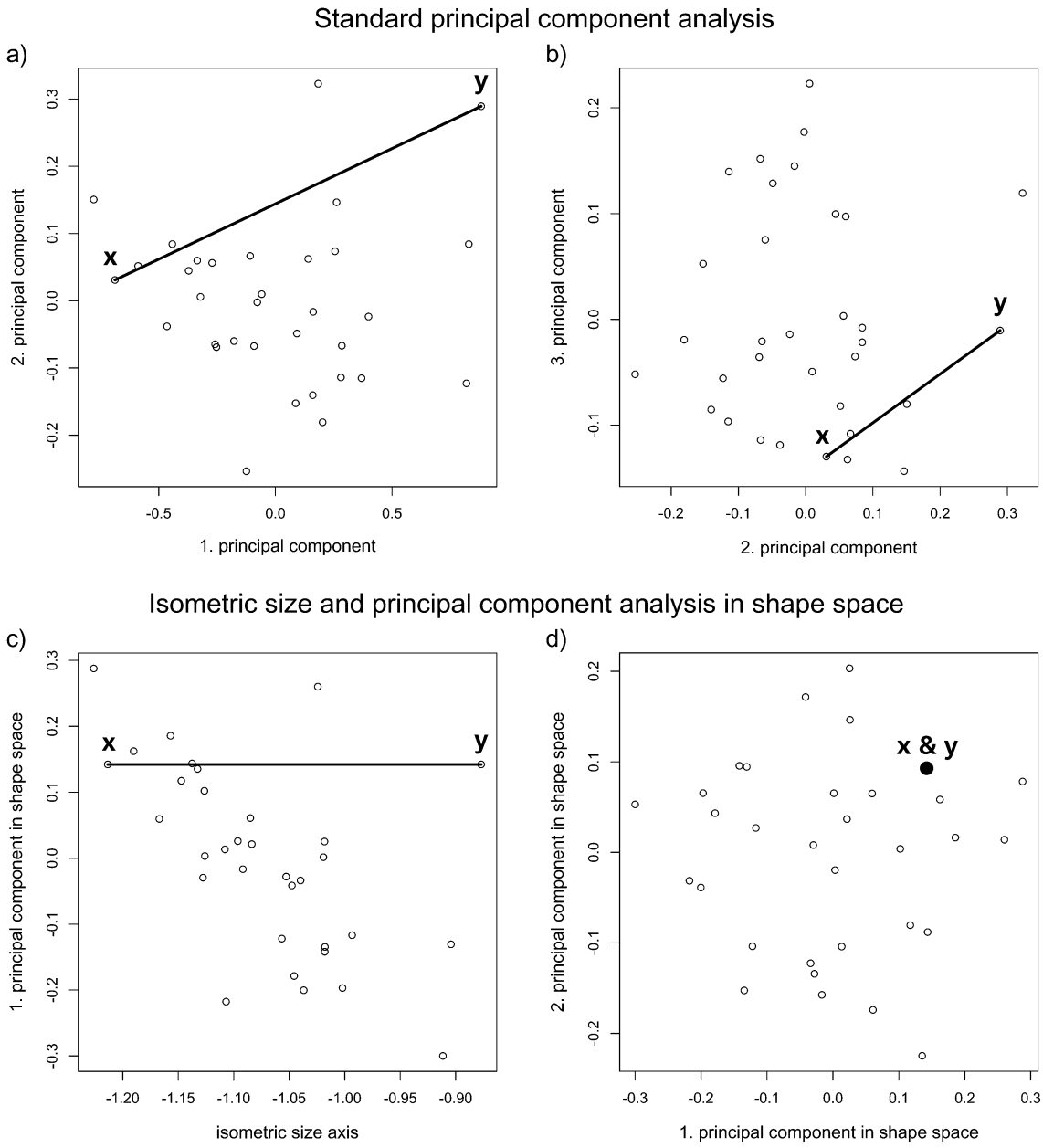

FIGURE 6. Scatterplots of principal component analyses (PCA) of a single species of *Pteromalus* (n = 32 specimens of *Pteromalus albipennis*, p = 23 variables of body measurements; data from Baur 2002), showing the effect of different definitions of size and shape. Specimen labeled y is a clone of specimen x but with all variables scaled by a factor of 1.4. The two specimens have therefore equal values for all their ratios and are only separated along the isometric size axis, as indicated by the line connecting x with y. (a) Scatterplot of first against second and (b) of second against third component respectively of a standard PCA on the covariance matrix of log-transformed data. The first component is considered as a general size measure because its coefficients have the same sign and are of similar magnitude for all variables. However, they are not exactly the same, thus the first component of a standard PCA is usually considered as the allometric size axis (Jolicoeur 1963; Claude 2008). The remaining components define the shape space in this analysis. Note that the line of isometry is not parallel to the first component, and, thus, reflects the different size measures. As a result, specimens x and y are also widely separated points in the shape space, although viewed from their body proportions they are identical. For (c) and (d) the same data were used, but here they were subjected to a PCA *after* removal of isometric size (for details of computation, see the Methodology section). Now, the line of isometry connecting x with y lies of course parallel to the isometric size axis (c). In the shape space (d) the two specimens form a single point, because only those specimens appear distinct which also differ in body proportions.

the allometric vector $\mathbf{a}_J$. The reason for this is to transform away shape effects related to allometric growth. According to this view, size is represented by the first, shape by all the following principal components of the log-data. It is, however, unclear how these shape values could be properly interpreted in terms of body proportions; in particular, no ratio-spectrum can be assigned in a mathematically consistent way to "shape" vectors

orthogonal to $\mathbf{a}_1$. Moreover, the allometric growth law in its bivariate or multivariate versions is just a convenient statistical model and by no means a "law of nature" (Gould 1966). In our opinion, allometry should rather be treated as a hypothesis to be tested *after* the size values are determined rather than be incorporated into the framework from the very beginning. We therefore prefer to analyze allometric variation with help of the allometry ratio spectrum, as demonstrated above (see Results section).

Our new methods are obviously rooted in the field of multivariate morphometrics (Reyment et al. 1984). The latter is occasionally dubbed traditional morphometrics (Marcus 1990), as opposed to "modern morphometrics" (Claude 2008) such as the analysis of landmarks (geometric morphometrics, Adams et al. 2004; Zelditch et al. 2004) or outlines (e.g., elliptic Fourier analysis, Lestrel 1989, 2000). The main reason why we stay within multivariate morphometrics is simply given by the nature of our data. Landmark and outline data are ideally suited for fixed objects, such as a skull or the body of a fish. For an insect with articulated extremities, those methods are of limited use unless one is willing to study the form of the head, thorax, or wings in separate analyses. This can and should be done. Nevertheless, it is often useful to include measurements from all over the body in a single analysis. For instance, a taxonomist trying to distinguish between two most similar species will be happy about any discriminating character. What if they are best separated by the ratio of, say, the length of the hind leg and the eye height? As we have shown above, it is here where methods of multivariate morphometrics, adapted for the analysis of ratios, could play a major role.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Acknowledgments

We thank Ian T. Jolliffe, Institute for Engineering, Computing & Mathematics, Exeter, for critical reading of the manuscript. Seraina Klopfstein, Swedish Museum of Natural History, Stockholm, and two reviewers also made numerous valuable comments and suggestions. We are finally grateful to Yvonne Kranz-Baltensperger, Christian Kropf and Elsa Obrecht, Natural History Museum, Bern, for discussion and useful corrections.

References

Adams D.C., Rohlf F.J., Slice D.E. 2004. Geometric morphometrics: ten years of progress following the "revolution". Ital. J. Zool. 71:5–16.
Aitchison J. 1983. Principal component analysis of compositional data. Biometrika. 70:57–65.
Aitchison J. 1986. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. London: Chapman and Hall.
Anderson T.W. 2003. An introduction to multivariate statistical analysis. 3rd ed. New York: Wiley.
Atchley W.R., Gaskins T.C., Anderson D. 1976. Statistical properties of ratios. I. Empirical results. Syst. Zool. 25:137–148.
Baur H. 2002. The power of multivariate statistical methods in the taxonomy of Pteromalidae (Hymenoptera: Chalcidoidea). In: Melika G., Thuróczy C., editors. Parasitic wasps: evolution, systematics, biodiversity and biological control. Budapest (Hungary): Agroinform. p. 73–81.
Bookstein F.L. 1989. "Size and shape": a comment on semantics. Syst. Zool. 38:173–180.
Burnaby T.P. 1966. Growth-invariant discriminant functions and generalized distances. Biometrics. 22:96–110.
Cadima J.F.C.L., Jolliffe I.T. 1996. Size- and shape-related principal component analysis. Biometrics. 52:710–716.
Campbell N.A., Mahon R.J. 1974. Multivariate study of variation in two species of rock crab of the genus *Leptograpsus*. Aust. J. Zool. 22:417–425.
Claude J. 2008. Morphometrics with R. Use R! New York: Springer.
Darroch J.N., Mosimann J.E. 1985. Canonical and principal components of shape. Biometrika. 72:241–252.
Dryden I.L., Mardia K.V. 1998. Statistical shape analysis. Chichester (UK): Wiley.
Flury B.K., Riedwyl H. 1986. Standard distance in univariate and multivariate analysis. Am. Stat. 40:249–251.
Gayon J. 2000. History of the concept of allometry. Am. Zool. 40: 748–758.
Goloboff P.A., Mattoni C.I., Quinteros A.S. 2006. Continuous characters analyzed as such. Cladistics. 22:589–601.
Gould S.J. 1966. Allometry and size in ontogeny and phylogeny. Biol. Rev. 41:587–640.
Graham M.W.R.d.V. 1969. The Pteromalidae of North-Western Europe. B. Brit. Mus. Nat. Hist. Entomol., Suppl. 16:1–908.
Graham M.W.R.d.V. 1991. A reclassification of the European Tetrastichinae (Hymenoptera: Eulophidae): revision of the remaining genera. Mem. Am. Entomol. Inst. 49:1–322.
Graham M.W.R.d.V., Gijswijt M.J. 1991. A new species of *Pteromalus* (Hymenoptera: Chalcidoidea) from France, associated with *Solidago virgaurea*. Entomol. Ber. Amst. 51:153–155.
Grissell E.E., Schauff M.E. 1997. A handbook of the families of Nearctic Chalcidoidea (Hymenoptera). 2nd ed . Washington (DC): Entomological Society of Washington.
Hills M. 1978. On ratios: a response to Atchley, Gaskins, and Anderson. Syst. Zool. 27:61–62.
Horstmann K. 2009. Revision of the western Palearctic species of *Dusona* Cameron (Hymenoptera: Ichneumonidae: Campopleginae). Spixiana. 32:45–110.
Hotz T., Huckemann S., Munk A., Gaffrey D., Sloboda B. 2010. Shape spaces for prealigned star-shaped objects: studying the growth of plants by principal components analysis. J. R. Stat. Soc. C Appl. Stat. 159:127–143.
Huxley J.S. 1932. Problems of relative growth (with an introduction by Frederick B. Churchill and an essay by Richard E. Strauss) (reprint edition). Baltimore (MD): The Johns Hopkins University Press.
Jolicoeur P. 1963. The multivariate generalization of the allometry equation. Biometrics. 19:497–499.
Jolicoeur P., Mosimann J.E. 1960. Size and shape variation in the painted turtle: a principal component analysis. Growth. 24:339–354.
Jolliffe I.T. 2004. Principal component analysis. 2nd ed. New York: Springer.
Kasparyan D.R. 1989. Ichneumonidae (Subfamily Tryphoninae), tribe Tryphonini. Fauna of the USSR, Hymenoptera 3. Leiden (the Netherlands): Brill.
Klingenberg C.P. 1996. Multivariate allometry. In: Marcus L.F., Corti M., Loy A., Naylor G.J.P., Slice D.E., editors. Advances in morphometrics. New York: Plenum Press. p. 23–49.
Lestrel P.E. 1989. A method for analyzing complex two-dimensional shapes: elliptic fourier functions. Am. J. Phys. Anthropol. 72:257–258.
Lestrel P.E. 2000. Morphometrics for the life sciences. Recent advances in human biology. Volume 7. Singapore: World Scientific.
MacKay D.J.C. 2003. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press.

Manly B.F.J. 2005. Multivariate staistical methods: a primer. 3rd ed. London: Chapman and Hall.
Marcus L.F. 1990. Traditional morphometrics. In: Rohlf F.J., Bookstein F.L., editors. Proceedings of the Michigan morphometrics workshop. Special publication 2. Ann Arbor (MI): University of Michigan Museum of Zoology. p. 77–122.
Mayr E., Ashlock P.D. 1991. Principles of systematic zoology. 2nd ed. New York: McGraw-Hill.
McCoy W., Bolker B.M., Osenberg C.W., Miner B.G., Vonesh J.R. 2006. Size correction: comparing morphological traits among populations and environments. Oecologia. 148:547–554.
Mosimann J.E. 1970. Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. J. Am. Stat. Assoc. 65:930–945.
Noyes J.S. 2004. Encyrtidae of Costa Rica (Hymenoptera: Chalcidoidea), 2: *Metaphycus* and related genera, parasitoids of scale insects (Coccoidea) and whiteflies (Aleyrodidae). Mem. Am. Entomol. Inst. 73:1–459.
Pawlowsky-Glahn V., Egozcue J.J. 2001. Geometric approach to statistical analysis on the simplex. Stoch. Environ. Res. Risk Assess. 15:384–398.
Pimentel R.A. 1979. Morphometrics: the multivariate analysis of morphological data. Dubuque (IA): Kendall/Hunt.
Polaszek A., Manzari S., Quicke D.L.J. 2004. Morphological and molecular taxonomic analysis of the *Encarsia meritoria* species-complex (Hymenoptera, Aphelinidae), parasitoids of whiteflies (Hemiptera, Aleyrodidae) of economic importance. Zool. Scripta. 33:403–421.
R Development Core Team. 2010. R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org.
Rae T.C. 2002. Scaling, polymorphism and cladistic analysis. In: MacLeod N., Forey P.L., editors. Morphology, shape and phylogeny. Systematic association special volume series 64. Boca Raton (FL): CRC Press. p. 45–52.
Rao C.R., Suryawanshi S. 1996. Statistical analysis of shape of objects based on landmark data. Proc. Natl. Acad. Sci. U.S.A. 93:12132–12136.
Reyment R.A., Blackith R.E., Campbell N.A. 1984. Multivariate morphometrics. 2nd ed. London: Academic Press.
Richtsmeier J., Deleon V.B., Lele S.R. 2002. The promise of geometric morphometrics. Yearb. Phys. Anthropol. 45:63–91.
Sampson P.D., Siegel A.F. 1985. The measure of "size" independent of "shape" for multivariate lognormal populations. J. Am. Stat. Assoc. 80:910–914.
Schuh R.T., Brower A.V.Z. 2009. Biological systematics: principles and applications. 2nd ed. Ithaca (NY): Cornell University Press.
Sorensen J.T., Foottit R. 1992. Ordination in the study of morphology, evolution and systematics of insects: applications and quantitative genetic rationals. Amsterdam: Elsevier.
Stuessy T.F. 2009. Plant taxonomy: the systematic evaluation of comparative data. 2nd ed. New York: Columbia University Press.
Thiele K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics. 9:275–304.
Townes H.K., Townes M. 1981. A revision of the Serphidae (Hymenoptera). Mem. Am. Entomol. Inst. 32:1–541.
Wiens J.J. 2000. Phylogenetic analysis of morphological data. Washington (DC): Smithsonian Institution Press.
Winston J.E. 1999. Describing species: practical taxonomic procedures for biologists. New York: Columbia University Press.
Zelditch M.L., Swiderski D.L., Sheets H.D., Fink W.L. 2004. Geometric morphometrics for biologists: a primer. Amsterdam: Elsevier.

APPENDIX

Statistical Derivation of the Allometric Size Vector

We would like to arrive at an estimation of the allometric size vector starting from a statistical model of the allometric growth hypothesis. Let $\mathbf{y}$ be the original data and $\alpha(\mathbf{y}) = \prod_{k=1}^{p} y_k^{a_k}$ a size function. According to Huxley (1932), each trait y_i when graphed against the individual's size $\alpha(\mathbf{y})$ should satisfy the power law

$$y_i = d_i \cdot \alpha(\mathbf{y})^{c_i}, \quad i = 1, \ldots, p. \tag{A1}$$

Here we consider d_i as positive random variables and c_i as constant coefficients. We shall use the approach of least squares to statistically estimate the coefficients a_i and c_i. Taking logarithms on both sides of (A1), we get

$$x_i = c_i \sum_{k=1}^{p} a_k x_k + \mu_i + \epsilon_i,$$

where $\mu_i = E(\log d_i)$ and $E(\epsilon_i) = 0$. In vector notation, this reads

$$\mathbf{x} = (\mathbf{a}^T\mathbf{x})\mathbf{c} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

with $E(\boldsymbol{\epsilon}) = 0$. Because $E(\mathbf{x}) = 0$, we conclude $\boldsymbol{\mu} = 0$. We estimate $\mathbf{a}$ and $\mathbf{c}$ in a way that the sum of squares is minimal:

$$(\hat{\mathbf{a}}, \hat{\mathbf{c}}) = \operatorname{argmin}_{\mathbf{a},\mathbf{c}} S(\mathbf{a}, \mathbf{c}),$$

where $S(\mathbf{a}, \mathbf{c}) = E\|\boldsymbol{\epsilon}\|^2$. We have

$$\begin{aligned} S(\mathbf{a}, \mathbf{c}) &= E\|\mathbf{x} - (\mathbf{a}^T\mathbf{x})\mathbf{c}\|^2 \\ &= E[\mathbf{x}^T\mathbf{x} + (\mathbf{a}^T\mathbf{x})^2(\mathbf{c}^T\mathbf{c}) - 2(\mathbf{a}^T\mathbf{x})(\mathbf{c}^T\mathbf{x})] \\ &= E(\mathbf{x}^T\mathbf{x}) + (\mathbf{c}^T\mathbf{c})\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} - 2\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{c}. \end{aligned}$$

Calculating vector derivatives with respect to $\mathbf{a}$ and $\mathbf{c}$ we get

$$\frac{\partial}{\partial \mathbf{a}} S(\mathbf{a}, \mathbf{c}) = 2(\mathbf{c}^T\mathbf{c})\boldsymbol{\Sigma}\mathbf{a} - 2\boldsymbol{\Sigma}\mathbf{c}$$

and

$$\frac{\partial}{\partial \mathbf{c}} S(\mathbf{a}, \mathbf{c}) = (\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})\mathbf{c} - \boldsymbol{\Sigma}\mathbf{a}.$$

Setting both equations equal to 0 and dropping the hats over $\hat{\mathbf{a}}$ and $\hat{\mathbf{c}}$, we arrive at the system of equations:

$$(\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})\mathbf{c} = \boldsymbol{\Sigma}\mathbf{a},$$
$$(\mathbf{c}^T\mathbf{c})\boldsymbol{\Sigma}\mathbf{a} = \boldsymbol{\Sigma}\mathbf{c}.$$

Multiplying the second equation from the left by $\boldsymbol{\Sigma}^{-1}$, solving for $\mathbf{c}$ and plugging the result into the first equation, one can see that $\mathbf{a}$ is an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue

$$\lambda = \frac{\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}}{\|\mathbf{a}\|^2}$$

and $\mathbf{c} = \mathbf{a}/\|\mathbf{a}\|^2$. Replacing these results in $S(\mathbf{a}, \mathbf{c})$ one gets:

$$S(\mathbf{a}, \mathbf{c}) = E(\mathbf{x}^T\mathbf{x}) - \lambda.$$

Evidently, this expression is minimal if λ is the largest eigenvalue of $\boldsymbol{\Sigma}$. Let $\mathbf{a}_1$ denote the unit vector representing the first principal component of the data $\mathbf{x}$. Imposing the size restriction, we arrive at the solution

$$\mathbf{a}_J = \mathbf{a}_1/1^T\mathbf{a}_1$$

and $\mathbf{c}_J = \mathbf{a}_J/\|\mathbf{a}_J\|^2$. Historically, Jolicoeur (1963) was the first to introduce a multivariate generalization of Huxley's allometric power law and he proposed our $\mathbf{a}_J$ as a measure of size (or rather $\mathbf{a}_1$ to be precise). He did not, however, give a statistical model to motivate his definition.