doi:10.1111/j.1558-5646.2009.00804.x

SIZE-CORRECTION AND PRINCIPAL COMPONENTS FOR INTERSPECIFIC COMPARATIVE STUDIES

Liam J. Revell[1,2,3]

[1] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138*

[2] *E-mail: lrevell@nescent.org*

Received February 25, 2009
Accepted July 21, 2009

Phylogenetic methods for the analysis of species data are widely used in evolutionary studies. However, preliminary data transformations and data reduction procedures (such as a size-correction and principal components analysis, PCA) are often performed without first correcting for nonindependence among the observations for species. In the present short comment and attached R and MATLAB code, I provide an overview of statistically correct procedures for phylogenetic size-correction and PCA. I also show that ignoring phylogeny in preliminary transformations can result in significantly elevated variance and type I error in our statistical estimators, even if subsequent analysis of the transformed data is performed using phylogenetic methods. This means that ignoring phylogeny during preliminary data transformations can possibly lead to spurious results in phylogenetic statistical analyses of species data.

KEY WORDS: **Data transformation, evolutionary regression, least squares, linear regression, PCA, phylogenetic comparative methods, principal components analysis.**

Phylogenetic methods for the statistical analysis of species data have become widely accepted in recent years (e.g., Cheverud et al. 1985; Felsenstein 1985; Grafen 1989; Harvey and Pagel 1991; Garland et al. 1992; Martins 1994; Hansen and Martins 1996; Hansen 1997; Butler et al. 2000; Hansen et al. 2008). This is because biologists are aware that data for species may be nonindependent due to shared history (Blomberg and Garland 2002; Freckleton et al. 2002; Blomberg et al. 2003; Garland et al. 2005; Revell et al. 2008). Although it has become rare to find published studies in which the authors choose to ignore phylogeny in their primary analysis, it is not uncommon to see phylogeny effectively bypassed during preliminary statistical data manipulations, such as size-correction and principal components analysis (PCA). However, both standard linear regression (the most common method of size-correction in evolutionary ecology) and PCA assume that the sample consists of independent datapoints—an assumption that is frequently violated by phylogenetic data from species (Harvey and Pagel 1991; Martins and Hansen 1997; Garland et al. 2005; but see Price 1997).

A relatively commonly used procedure for phylogenetic statistical transformation of interspecific data was provided by Garland et al. (1992). According to this procedure, we first calculate standardized phylogenetically independent contrasts (following Felsenstein 1985). To size-correct, we then regress the contrasts for each size-correlated variable on the contrasts for size (conducting regression "through the origin," i.e., without an intercept term; Garland et al. 1992) and compute the residuals as deviations between the predicted and estimated contrast values. To perform PCA, we first calculate the variance–covariance or correlation matrix from our contrasts (again, without intercept terms; e.g., Revell et al. 2007), then calculate its eigenvectors and eigenvalues, and, finally, compute PC scores for contrasts in the typical way (e.g., Ackerly and Donoghue 1998; Clobert et al. 1998).

[3] Current address: National Evolutionary Synthesis Center, 2024 W. Main St. A200, Durham, NC 27705

Although these methods should yield phylogenetically independent and size-corrected or rotated evolutionary differences, they have two main disadvantages. The first is that it is difficult or impossible to use these methods under non-Brownian motion (BM) models for the evolutionary process (although branch-length transformations can be used to mimic some other evolutionary processes, such as punctuational evolution; e.g., Martins and Garland 1991; Garland et al. 1993; Pagel 1994). The second disadvantage of size-correction and PCA on contrasts is that the residuals or scores that are returned are residual evolutionary differences between nodes rather than residual values for species (i.e., they are associated with internal nodes in the tree rather than with the species at the tips of the tree). This latter consideration might not be a problem in some studies (e.g., if the residuals are destined only to be analyzed by a linear regression and merely the slope and significance of this secondary analysis are of interest). However, many evolutionary analyses require the association of size-corrected or rotated phenotypic values with tip taxa in the tree. Types of analyses that might require transformed values for tip species include (but are not limited to): the estimation of phylogenetic signal (Freckleton et al. 2002; Blomberg et al. 2003); ancestral state reconstruction (Schluter et al. 1997); the phylogenetic analysis of variance (ANOVA) or analysis of covariance (ANCOVA) (Garland et al. 1993); and phylogenetic regression in which both slope and intercept (or prediction in the species space) are of concern (Rohlf 2001).

In the present short comment, I describe statistically correct procedures for phylogenetic size-correction and PCA. Although in some cases these procedures are developed elsewhere (e.g., Garland and Ives 2000; Rohlf 2001; Blomberg et al. 2003), they have not previously been presented in detail and in the explicit context of preliminary statistical data transformation. I also use numerical simulations to examine the consequences of ignoring phylogeny in preliminary transformations prior to a statistical analysis by standard phylogenetic methods, such as independent contrasts followed by linear regression (Felsenstein 1985).

The purpose of this comment is almost wholly utilitarian, and as such I have appended R (R Development Core Team 2008) and MATLAB (The Mathworks 2006) code for the analyses described herein. The inspiration for this comment comes from the frequent questions posed to me by evolutionary biologists and ecologists who are interested in phylogenetic methods for size-correction and PCA, but unsure how to proceed.

Methods

In the present section, I describe correct procedures for phylogenetic size-correction and PCA. Three primary disclaimers must accompany this material.

First, my presentation focuses on BM as a model for evolution. Various authors have already pointed out that BM is an inappropriate model under many circumstances (e.g., Felsenstein 1988; Hansen 1997; Butler and King 2004; Hansen et al. 2008), although it is a suitable model for genetic drift and some types of natural selection (e.g., O'Meara et al. 2006; Revell and Harmon 2008). In the present comment, I use BM as my model for the evolutionary process—however, as noted below, the methods described herein can theoretically be applied equally well using other evolutionary models (e.g., Pagel 1999; Butler and King 2004; Hansen et al. 2008).

Second, I focus on a single, very common method of size-correction in evolutionary studies. This is the method in which we obtain residuals from a least-squares regression of the size-dependent trait against body size (Gould 1966; Jolicoeur et al. 1984; e.g., Glossip and Losos 1997; Hulsey et al. 2007; Revell et al. 2007). However, size-correction can be performed by many methods, only one of which is detailed herein, and an extensive literature exists on procedures for size-correction (e.g., Humphries et al. 1981; Jolicoeur et al. 1984; Rohlf and Bookstein 1987; García-Berthou 2001; McCoy et al. 2006). My use of a linear regression for size-correction should not be considered an endorsement of this method. García-Berthou (2001) and Freckleton (2002, 2009) both caution against using size-corrected residuals as data for subsequent statistical analyses, and instead recommend that size be included as a covariate (García-Berthou 2001) or as an additional independent variable in a multiple regression (Freckleton 2002). Even when size-correction is desired, Rohlf and Bookstein (1987) have pointed out some of the reasons why a linear regression should not be our size-correction procedure of choice, particularly when data for size are collected with sampling error (as will be true of almost any evolutionary study). In most cases, it is possible to extend the presented phylogenetic approach to less commonly used statistical procedures for size-correction (e.g., shearing; Humphries et al. 1981).

Finally, third, although I present methods for phylogenetic size-correction and phylogenetic PCA, it is important for the reader to note that these procedures provide residuals and scores in the original, phylogenetically dependent, species space—not in a transformed phylogenetically independent space. This means that the residuals and scores from these analyses still probably need to be analyzed using phylogenetic methods! Why then, one might ask, should we bother to use phylogenetic methods for size-correction or data rotation? I show below, in section Examples, that phylogenetic size-correction and principal components provide estimates of the allometric coefficient and eigenstructure that will have lower variance relative to nonphylogenetic procedures, thus reducing type I error to its nominal level when residuals and scores are subsequently analyzed using phylogenetic methods. If phylogeny is instead ignored in these preliminary transformations, then variance and type I error of our statistical estimators and hypothesis tests can be substantially increased.

PHYLOGENETIC SIZE-CORRECTION

The procedure for conducting phylogenetic size-correction using the residuals from a least squares regression analysis, while controlling for nonindependence due to phylogenetic history, is straightforward.

First, we compute the matrix **C** that describes the expected covariances of our data due to phylogenetic relatedness and will provide the error structure in our linear model (Grafen 1989; Rohlf 2001). **C** is estimated from our tree and evolutionary model. In most phylogenetic analyses of continuously valued characters, the evolutionary model of choice is constant rate BM. Most of the procedures described herein can also be performed if the error structure, **C**, is obtained assuming a non-Brownian model of evolutionary change (Rohlf 2001; e.g., Pagel 1999; Butler and King 2004; Hansen et al. 2008).

The computation of **C** has been described in many prior studies (e.g., Hansen and Martins 1996; Martins and Hansen 1997; Rohlf 2001; O'Meara et al. 2006; Revell 2008; Revell and Harmon 2008). For *n* species, **C** is an $n \times n$ matrix containing, on its diagonal, values proportional to the expected variances for individual characters, and, on its off-diagonals, values proportional to the expected covariances between each trait value in different taxa due to the phylogeny. Under BM, in which variance among lineages is accumulated in a direct proportion to the time elapsed, each *i*, *j*th element of **C** (C_{ij}) is computed as the patristic distance from the root of the tree to the common ancestor of species *i* and *j*. If the phylogeny is ultrametric and contains only extant taxa, then $\mathbf{C} = t\mathbf{1}\mathbf{1}' - \frac{1}{2}\mathbf{P}$, where **1** is a $n \times 1$ column vector of 1.0 s, *t* is the total length of the tree, and **P** is a patristic distance matrix (Revell and Harmon 2008). For all of the analyses presented herein, the units of branch length of the tree used to calculate **C** are of no particular consequence, only their relative lengths are significant. (This is not true in all phylogenetic analyses of species data, e.g., Revell and Harmon 2008.)

To remove the effect of size from each of our variables, we next compute the least squares regression coefficients in the regression of our dependent variable on size, controlling for the phylogenetic correlations between our observations for species. To do this, we use the standard generalized least squares estimating equation

$$\mathbf{b} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}. \tag{1}$$

In this equation, **b** is a row vector containing our least squares estimates of the regression intercept and slope, and **y** is a vector containing the size-dependent morphological trait in all species. **X** is an $n \times 2$ matrix containing a column of ones and our data for size (Rohlf 2001; Rencher and Schaalje 2008).

We can then calculate the vector of residuals, **r**, using the equation

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{b}, \tag{2}$$

(Rohlf 2001). Although we used a phylogenetic regression to obtain these residuals, the reader should pay heed that, as noted above, the residuals in **r** are not phylogenetically independent! Size correction has been performed using the phylogeny to derive the error structure and thus properly estimate the linear regression model for size-correction, however the residuals from this analysis are in terms of the original species and should be subsequently analyzed by standard phylogenetic means (such as independent contrasts, PGLS, phylogenetic ANCOVA, etc.).

This procedure for size-correction can also be performed using independent contrasts with BM as our model for evolution (Felsenstein 1985). To do so, we first estimate the regression slope from bivariate regression through the origin performed on the independent contrasts for size and our dependent variable (Garland et al. 1992). We then fit the slope of that regression line to the species data through the phylogenetic means for both characters (see eq. 3, below). Finally, we calculate the residual deviations from this regression line in the standard way.

This is the same procedure that is described in Garland and Ives (2000; Blomberg et al. 2003), but is different from that in Garland et al. (1992) because the residuals are computed in the original space, rather than in the phylogenetically independent (contrasts) space, and thus a regression is performed through the phylogenetic mean (which corresponds to the maximum-likelihood estimate for the ancestral states at the root node of the tree; Schluter et al. 1997; Rohlf 2001) rather than through the origin. I consider the method based on generalized least squares and described above to be simpler (so long as we have an easy way to obtain **C**, for example, using the R package APE; Paradis et al. 2004, 2006; see Appendix), because it involves fewer steps. However, both the phylogenetic generalized least squares procedure and regression through the phylogenetic mean using the slope from contrasts regression should yield the same set of size-corrected residual values, given an evolutionary model of BM.

PHYLOGENETIC PRINCIPAL COMPONENTS ANALYSIS

The procedure for conducting phylogenetic PCA and obtaining scores for species in a rotated principal components space is related to the procedure for phylogenetic size-correction, above. Using the matrix **C** derived from the tree and our evolutionary model, as before, and an $n \times m$ data matrix **X** containing the data for *m* traits measured in *n* species, we first compute a vector containing the estimated ancestral states for each of our characters. This vector is also called the vector of "phylogenetic means," as previously mentioned, and can be estimated as follows:

$$\mathbf{a} = [(\mathbf{1}'\mathbf{C}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{C}^{-1}\mathbf{X}]'. \tag{3}$$

Equation (3) yields an $m \times 1$ column vector. We then estimate the evolutionary variance–covariance matrix for the m traits in our study. This is computed as follows:

$$\mathbf{R} = (n-1)^{-1}(\mathbf{X} - \mathbf{1a}')'\mathbf{C}^{-1}(\mathbf{X} - \mathbf{1a}'), \tag{4}$$

following Revell and Harmon (2008; also Revell and Collar 2009). In some instances we would like to perform PCA on the evolutionary correlation matrix, rather than the evolutionary variance–covariance matrix (Manly 2005). This is straightforward: we just calculate the elements of the correlation matrix $\mathbf{R}_{corr}$ as follows:

$$R_{corr,ij} = R_{ij}\Big/\sqrt{R_{ii}R_{jj}}, \tag{5}$$

and then use this matrix in subsequent analysis. Prior to computing scores, in this case, we must also standardize the values for species to have unit evolutionary variance. This is performed by dividing each element of **X** by the square-root of the corresponding diagonal element of **R**, i.e.,

$$X_{s,ij} = X_{ij}\Big/\sqrt{R_{jj}}, \tag{6}$$

in which X_{ij} is the value for the jth trait in the ith species, and $X_{s,ij}$ is the corresponding standardized value.

To perform PCA, we next obtain the diagonal matrix of eigenvalues, **D**, and the matrix, **V**, containing the eigenvectors of **R** in columns. **D** and **V** can be found by conventional means (and in many standard computational environments, such as MATLAB and R, see Appendix) or by singular value decomposition, such that $\mathbf{R} = \mathbf{VDV}^{-1}$. **V** and **D** contain the eigenvectors and eigenvalues, respectively, of the phylogenetic PCA. We can also compute scores in the original (species) space as follows:

$$\mathbf{S} = (\mathbf{X} - \mathbf{1a}')\mathbf{V}. \tag{7}$$

Here, **S** is an $n \times m$ matrix with the scores for n species and m principal components, in columns.

A key property of PC axes calculated in this way is that they are evolutionarily independent. This means that the phylogenetic correlation (i.e., the correlation of independent contrasts) between scores on each axis will be zero. This will not usually be true of PC axes computed ignoring phylogenetic nonindependence.

As with typical principal components, calculating component loadings can help in interpreting the principal component axes (but see Rencher 2002 for a criticism of loadings). Loadings are just the correlations between our phenotypic data and the principal component scores in the transformed space. As our principal component scores are nonindependent due to the phylogeny, as before, we should not ignore the tree in computing the correlations for our PC loadings. To compute our loadings while incorporating nonindependence due to the phylogeny, we must first compute the cross-covariance matrix, **K** as follows:

$$\mathbf{K} = (n-1)^{-1}(\mathbf{X} - \mathbf{1a}')'\mathbf{C}^{-1}\mathbf{S}. \tag{8}$$

Here, notably, we should subtract the phylogenetic mean from **X**, but we need not do so from **S** because the scores from our phylogenetic PCA should already be centered by our previous computations (i.e., they have phylogenetic means of 0.0).

We can then compute the loading of the ith trait on the jth PC axis using the following calculation:

$$L_{ij} = K_{ij}\Big/\sqrt{R_{ii}D_{jj}}. \tag{9}$$

This may actually be fairly important because Rohlf (2006) shows that the estimated correlation between two characters will be downwardly biased if the phylogeny is ignored. Because loadings are just correlation coefficients, as noted above, they may also be biased if they are calculated ignoring the phylogeny. As in the phylogenetic size-correction, PC scores are in terms of species and will still probably need to be analyzed phylogenetically in the subsequent statistical analyses.

We can also perform phylogenetic PCA using independent contrasts. In this case, we compute the uncentered variance–covariance or correlation matrix from our independent contrasts (Ackerly and Donoghue 1998; Clobert et al. 1998; Revell 2007; Revell et al. 2007). We then calculate the eigensystem for that matrix, and use the eigenvector matrix, **V**, and the set of phylogenetic means calculated using equation (3) to calculate principal component scores for species with equation (7), above.

Examples

PHYLOGENETIC SIZE-CORRECTION

For the first example, I generated size-correlated data for two characters on stochastic phylogenies. I used the following procedure. First I simulated 20,000 stochastic pure-birth phylogenetic trees, each containing $n = 100$ taxa. I then used BM as a model of evolution to simulate the evolution of three characters on each tree in two sets of simulations. In both sets of simulations, the first character was analogous to size. The other two characters were simulated to be highly correlated with size ($r = 0.95$), but to either (1) possess an expected residual regression coefficient of 0.00, or (2) possess an expected residual regression coefficient of 0.75. The mathematical details of these simulations are provided in the Appendix S1.

I then analyzed these data using two different procedures for size-correction. First, I size-corrected the data using ordinary

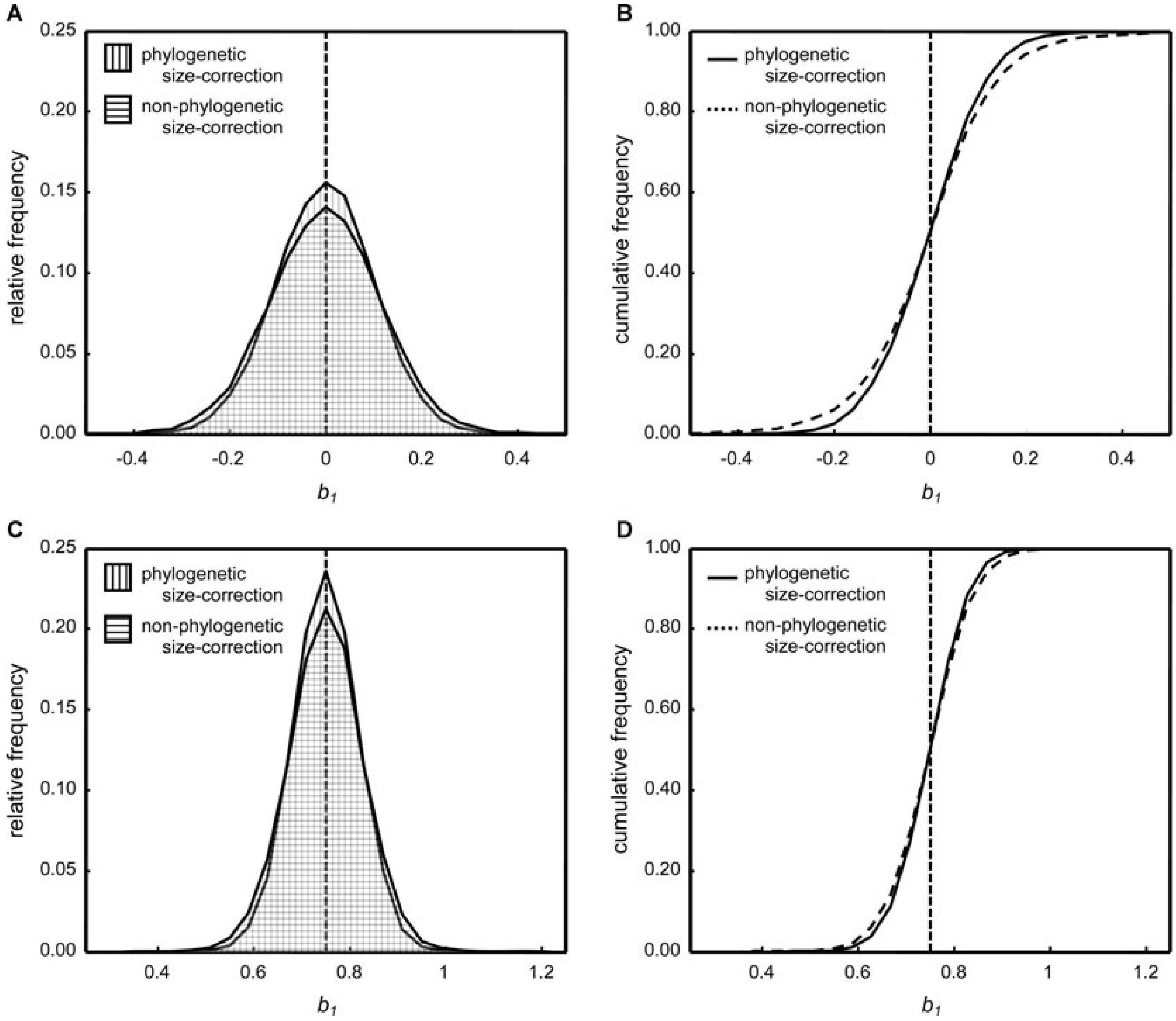

Figure 1. **(A) Relative frequency distribution of $\hat{b}_1$ (the estimated regression slope) from the regression of residuals from size-corrected data, with a generating residual regression slope of $\beta_1 = 0.0$. Vertical cross-hatching is for the distribution obtained when size-correction was performed using the phylogenetic method described herein, whereas horizontal cross-hatching is for the distribution of $\hat{b}_1$ obtained when preliminary size-correction was by ordinary least-squares regression (i.e., nonphylogenetic). In both cases, the subsequent evolutionary regression of size-corrected values was performed using phylogenetic generalized least-squares. (B) Cumulative frequency distribution of $\hat{b}_1$ for the same analyses as (A). The dashed line indicates the nonphylogenetic size-correction. (C) Distribution of $\hat{b}_1$ from the regression of residuals from size-corrected data, with a generating residual regression slope of $\beta_1 = 0.75$. (D) Cumulative distribution of $\hat{b}_1$ for the same analyses as (C). Phylogenetic and nonphylogenetic size-correction are indicated as in (A) and (B). $\hat{b}_1$ is an unbiased estimator of β_1 in both (A,B) and (C,D) regardless of the size-correction procedure, however the variance around the mean is approximately 33% larger when phylogeny is ignored during size-correction (horizontal cross-hatching; dashed line in cumulative distributions).**

least squares regression ignoring the phylogeny. Next, I size-corrected the data by calculating the residuals using the phylogenetic regression procedure described above. Then, I computed the regression line between each set of size-corrected variables using phylogenetic generalized least squares estimation (which will yield the same regression slope as the method of independent contrasts; Rohlf 2001). Thus, I performed size-correction using first a nonphylogenetic and then the analogous phylogenetic procedure described above, then I analyzed the residuals using a phylogenetic regression (Felsenstein 1985; Grafen 1989).

The relative frequency and cumulative frequency distributions of regression slopes estimated from each procedure on 20,000 simulated trees and datasets are shown in Figure 1. Because the differences between these distributions are subtle, in panels (A) and (C) I use line rather than bar graphs to more easily overlay the distributions obtained by each method. Both procedures yielded unbiased estimates of the regression slope from the phylogenetic regression on the size-corrected data, regardless of the generating conditions (Fig. 1A,C). However, phylogenetic size-correction using the procedure described in the Methods resulted in an approximately 25% decrease in the estimation variance around the regression slope (Fig. 1). When the generating residual correlation was 0.0, I also estimated the type I error probability associated with each method. I found that type I error was substantially greater than its nominal value of 0.05 when phylogeny was ignored in the size-correction procedure prior to statistical analysis (type I error $= 0.089$; P (true error ≤ 0.05) < 0.0001). By contrast, when I size-corrected using the phylogenetic method, type I error of the phylogenetic regression on the residuals was statistically indistinguishable from 0.05 (type I error $= 0.052$; P (true error ≤ 0.05) $= 0.153$). This shows that the consequence of ignoring phylogeny during size-correction is greater than simply obtaining an allometric coefficient with a high error. Rather, it can result in the substantial opportunity (nearly a doubling in the present analysis) of obtaining a spurious residual correlation between our size-corrected variables.

PHYLOGENETIC PRINCIPAL COMPONENTS ANALYSIS

For the second example, I generated data with known evolutionary covariance structure on stochastic phylogenies. In each of 20,000 simulations performed on the same trees as above, I simulated multivariate BM evolution. The generating variance–covariance structure for simulation was determined randomly by first drawing a matrix of $m = 4$ random orthonormal $m \times 1$ vectors, $\mathbf{V}$, and then a diagonal matrix $\mathbf{D}$ containing positive, evenly spaced eigenvalues, and computing $\mathbf{R} = \mathbf{VDV}^{-1}$. This procedure is guaranteed to yield a positive semidefinite covariance matrix, $\mathbf{R}$. The generating evolutionary covariance matrix, $\mathbf{R}$, must be positive semidefinite to be valid. I then analyzed all the datasets and trees using nonphylogenetic PCA and the procedure for phylogenetic PCA described above. The mathematical details of these simulations are provided in the Appendix S2.

PCA results in the calculation of eigenvalues, eigenvectors, component loadings, and scores for each analysis. To summarize the results of the 40,000 PCAs (20,000 nonphylogenetic and 20,000 phylogenetic, on the same datasets) I calculated two summary statistics from each analysis. First, I calculated the mean vector correlation between corresponding generating and estimated eigenvectors. Second, I calculated the mean correlation between the generating and estimated eigenvalues. Because I calculated each summary statistic twice for each dataset (once for the nonphylogenetic and once for the phylogenetic analysis), I measured the relative performance of each procedure in determining the generating eigensystem by then calculating the difference between the correlation from the phylogenetic and nonphylogenetic analyses. Good correspondence of the true and estimated eigenstructure should result in a high correlation between eigenvectors and eigenvalues, thus a positive difference indicates that the phylogenetic PCA was better at recovering the eigenstructure of the generating evolutionary variance–covariance matrix. The results from these analyses are given in Figure 2. I found that phylogenetic PCA was much more effective at recovering the generating eigenstructure, particularly the eigenvectors of that structure (Fig. 2A).

The scores and eigenstructure obtained from these analyses have various uses. For example, scores from multiple different PCAs on different sets of characters might subsequently be compared using a linear regression, or the scores from phylogenetic PCA might be used in the subsequent statistical analyses such as phylogenetic ANOVA or ANCOVA (Garland et al. 1993). Alternatively, the eigenstructure might be used to identify primary axes of diversification in the group (e.g., Schluter 1996; Revell et al. 2007). As such, I have not focused on any particular secondary analysis, and instead have concentrated on showing that the phylogenetic PCA procedure described herein recovers the eigenstructure of the generating evolutionary variance–covariance matrix with lower error.

Discussion

There is a growing consensus in evolutionary biology that phylogenetic information should not be ignored in the analysis of species data (e.g., Harvey and Pagel 1991; Martins and Hansen 1997; Rohlf 2001; Butler and King 2004; Garland et al. 2005; but see Price 1997). Various measures have been proposed to control for the statistical nonindependence of data from species (Cheverud et al. 1985; Felsenstein 1985; Grafen 1989; Harvey and Pagel 1991), and these procedures have become quite widely used in the evolutionary genetic, ecological, and anthropological literatures (e.g., Nunn and Barton 2001; Andrews et al. 2009; Pontzer and Kamilar 2009).

The most popular phylogenetic comparative method for more than 20 years has been the independent contrasts method of Felsenstein (1985). This method was developed based on the insight that although values for species related by a tree are nonindependent, the differences between species should not be (Felsenstein 1985). According to this method, we thus calculate the corrected differences between nodes (internal or tip) descending from each internal node of the tree. At the end of the procedure, we have a set of $n - 1$ contrasts for each trait that have no expected covariance due to the phylogeny, and thus can be treated as independent datapoints in a statistical analysis such as linear bivariate or multivariate regression (Garland et al. 1992).

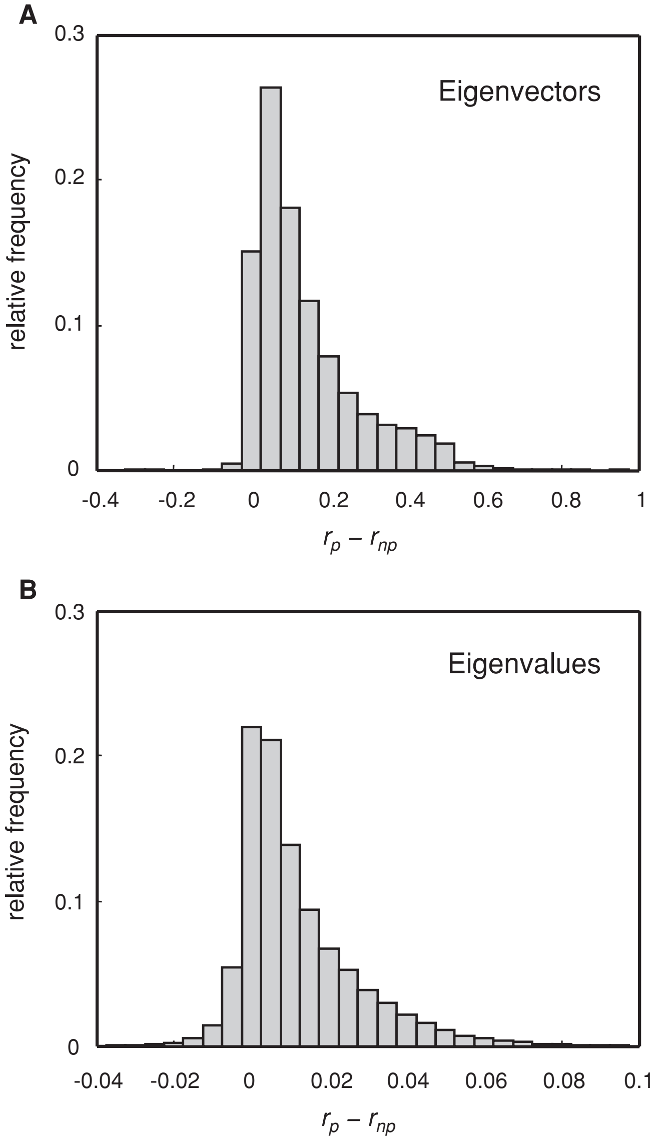

Figure 2. (A) Relative frequency distribution of $r_p - r_{np}$ for the eigenvectors. $r_p - r_{np}$ is the difference in the mean vector correlation between the estimated and generating eigenvectors in phylogenetic (*p*) and nonphylogenetic (*np*) principal components analyses. A positive $r_p - r_{np}$ thus indicates that the eigenvectors are better identified by the phylogenetic method. (B) Distribution of $r_p - r_{np}$ for the eigenvalues. Here, $r_p - r_{np}$ is the difference in the correlation between estimated and generating eigenvalues for phylogenetic (*p*) and nonphylogenetic (*np*) analyses.

However, the contrasts values represent differences between species, not species in the original space. As such, there is no immediately obvious way to use the values for contrasts in preliminary data transformations, such as size-correction by a linear regression or PCA, if we want to obtain residuals or scores for species from these procedures (but see Methods, above). Rightly or wrongly, it is thus very common for evolutionary ecologists and biologists to transform their data prior to analysis using nonphylogenetic means and thus without controlling for the statistical dependence of their observations due to the phylogeny.

In the present short comment, I provide a very simple overview of proper methods to incorporate phylogenetic information into preliminary data transformations prior to statistical analyses. The procedures yield residuals or scores in the original species space, but these are obtained while minimizing estimation error. As such, I show that the phylogenetic procedures will produce lower variance and type I error than is obtained if preliminary transformation is performed ignoring phylogeny.

Although the methods are general to multiple evolutionary models, I focus on a model of evolution by BM. Adjusting the methods presented herein to accommodate non-BM models of character change is conceptually straightforward, but might be easy or difficult in practice. For example, Pagel's (1999) λ statistic is a very useful phenomenological construct with which we can transform the error structure of our size-correction or PCA model (the matrix **C** in this article) to fit the observed covariance structure in our data using likelihood. After optimizing λ (e.g., Freckleton et al. 2002), we would then substitute the transformed matrix ($\mathbf{C}_\lambda$) for **C** in subsequent calculations (eqs. 1, 3, 4 and 8, herein). In the extreme case in which $\hat{\lambda} = 0.0$, this reduces our analysis to ordinary least squares regression (or nonphylogenetic PCA). An intriguing possibility is that we might find different values of λ in different steps of our analysis. In the Appendix S3, I illustrate size-correction under the same simulation conditions as the present study, but with $\lambda = 0.5$. Although this is a very simple non-Brownian evolutionary model, the principle is the same as for complex models (e.g., Hansen et al. 2008), although estimation is much more straightforward. As noted earlier, it is also possible to accommodate non-Brownian evolution using direct branch length transformations (such as log-transformation; e.g., Garland et al. 1992), rather than manipulation of the error structure of the model (e.g., Freckleton et al. 2002; Blomberg et al. 2003; Lavin et al. 2008; Revell and Harrison 2008).

In the case of more explicitly adaptive evolutionary models, for example the Ornstein–Uhlenbeck model of Hansen (1997), the error structure is more difficult to derive (e.g., Butler and King 2004). This is a very active area of research (e.g., Blomberg et al. 2003; Hansen et al. 2008; Lavin et al. 2008).

In the Appendix, I detail the procedures described in the text using R and MATLAB code. I am also distributing code for calculating the **C** matrix under a BM model from my website, and functions for the calculation of **C** under various evolutionary models already exist in R (Paradis et al. 2004; Paradis 2006). I hope that this article and the appended material are as useful as they are intended to be to evolutionary biologists, ecologists, geneticists, and evolutionary anthropologists, who are interested in performing statistical data-transformations using phylogenetic methods, but are presently unsure of precisely how to do so.

ACKNOWLEDGMENT

The author would like to thank D. Collar, T. Garland, L. Harmon, J. Losos, L. Mahler, C. Nunn, D. Posada, D. Rabosky, E. Rezende, and one anonymous reviewer for helpful comments on this manuscript.

LITERATURE CITED

Ackerly, D. D., and M. J. Donoghue. 1998. Leaf size, sapling allometry, and Corner's rules: phylogeny and correlated evolution in maples (*Acer*). Am. Nat. 152:767–791.

Andrews, C. B., S. A. Mackenzie, and T. R. Gregory. 2009. Genome size and wing parameters in passerine birds. Proc. R. Soc. Lond. B 276:55–61.

Blomberg, S. P., and T. Garland, Jr. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J. Evol. Biol. 15:899–910.

Blomberg, S. P., T. Garland, Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Butler, M. A., T. W. Schoener, and J. B. Losos. 2000. The relationship between sexual size dimorphism and habitat use in Greater Antillean *Anolis* lizards. Evolution 54:259–272.

Cheverud, J. M., M. M. Dow, and W. Leutenegger. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution 39:1335–1351.

Clobert, J., T. Garland, Jr., and R. Barbault. 1998. The evolution of demographic tactics in lizards: a test of some hypotheses concerning life history evolution. J. Evol. Biol. 11:329–364.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

———. 1988. Phylogenies and quantitative characters. Ann. Rev. Ecol. Syst. 19:445–471.

Freckleton, R. P. 2002. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. J. Anim. Ecol. 71:542–545.

———. 2009. The seven deadly sins of comparative analysis. J. Evol. Biol. 22:1367–1375.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160:712–726.

García-Berthou, E. 2001. On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. J. Anim. Ecol. 70:708–711.

Garland, T., Jr., and A. R. Ives. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346–364.

Garland, T., Jr., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Garland, T., Jr., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 42:265–292.

Garland, T., Jr., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. J. Exp. Biol. 208:3015–3035.

Glossip, D., and J. B. Losos. 1997. Ecological correlates of number of subdigital lamellae in anoles. Herpetologica 53:192–199.

Gould, S. J. 1966. Allometry and size in ontogeny and phylogeny. Biol. Rev. 41:587–638.

Grafen, A. 1989. The phylogenetic regression. Phil. Trans. R. Soc. Lond. B 326:119–157.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford Univ. Press, Oxford, UK.

Hulsey, C. D., M. C. Mims, and J. T. Streelman. 2007. Do constructional constraints influence cichlid craniofacial diversification? Proc. R. Soc. Lond. B 274:1867–1875.

Humphries, J. M., F. L. Bookstein, B. Chernoff, G. R. Smith, R. L. Elder, and S. G. Poss. 1981. Multivariate discrimination by shape in relation to size. Syst. Zool. 30:291–308.

Jolicoeur, P., P. Pirlot, G. Baron, and H. Stephan. 1984. Brain structure and correlation patterns in Insectivora, Chiroptera, and Primates. Syst. Zool. 33:14–29.

Lavin, S. R., W. H. Karasov, A. R. Ives, K. M. Middleton, and T. Garland, Jr. 2008. Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. Physiol. Biochem. Zool. 81:526–550.

Manly, B. F. J. 2005. Multivariate statistical methods: a primer, 3rd edn. Chapman & Hall / CRC, Boca Raton, FL.

Martins, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. Am. Nat. 144:193–209.

Martins, E. P., and T. Garland, Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.

McCoy, M. W., B. J. Bolker, C. W. Osenberg, B. G. Miner, and J. R. Vonesh. 2006. Size correction: comparing morphological traits among populations and environments. Oecologia 148:547–554.

Nunn, C. L., and R. A. Barton. 2001. Comparative methods for studying primate adaptation and allometry. Evol. Anthropol. 10:81–98.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37–45.

———. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paradis, E. 2006. Analysis of phylogenetics and evolution with R. Springer, New York, NY.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pontzer, H., and J. M. Kamilar. 2009. Great ranging associated with greater reproductive investment in mammals. Proc. Natl. Acad. Sci. USA 106:192–196.

Price, T. 1997. Correlated evolution and independent contrasts. Phil. Trans. R. Soc. Lond. B 352:519–529.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rencher, A. C. 2002. Methods of multivariate analysis, 2nd edn. John Wiley & Sons, Hoboken, NJ.

Rencher, A. C., and G. B. Schaalje. 2008. Linear models in statistics, 2nd edn. John Wiley & Sons, Hoboken, NJ.

Revell, L. J. 2007. Testing the genetic constraint hypothesis in a phylogenetic context: a simulation study. Evolution 61:2720–2727.

———. 2008. On the analysis of evolutionary change along single branches in a phylogeny. Am. Nat. 172:140–147.

Revell, L. J., and D. C. Collar. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. Evolution 63:1090–1100.

Revell, L. J., and L. J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res. 10:311–321.

Revell, L. J., and A. S. Harrison. 2008. PCCA: a program for phylogenetic canonical correlation analysis. Bioinformatics 24:1018–1020.

Revell, L. J., L. J. Harmon, R. B. Langerhans, and J. J. Kolbe. 2007. A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard *Anolis cristatellus*. Evol. Ecol. Res. 9:261–282.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

———. 2006. A comment on phylogenetic correction. Evolution 60:1509–1515.

Rohlf, F. J., and F. L. Bookstein. 1987. A comment on shearing as a method for “size correction”. Syst. Zool. 36:356–367.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.

Schluter, D., T. Price, A. Ø. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711.

The Mathworks,. 2006. Matlab: the language of technical computing (R2006a). The MathWorks Inc., Natick, MA.

Associate Editor: D. Posada

Appendix: Code for Data Transformations Described in the Methods

PHYLOGENETIC SIZE-CORRECTION

MATLAB code—The function below, phyl_resid.m, requires as input (1) an $n \times n$ matrix, **C**; (2) an $n \times 1$ vector **x** containing the sizes of each of the n species in the analysis; and (3) an $n \times m$ matrix **Y** containing the m size-correlated traits to be corrected using phylogenetic least squares regression. It returns an $n \times m$ matrix containing residuals in the original (species) space. Note that the species order in **C**, **x**, and **Y** is assumed to be identical.

```
% function R = phyl_resid(C, x, Y) computes R residuals given
% C, x, and Y
function R = phyl_resid(C, x, Y)
    % find out how many y variables and taxa we have
    [n,m] = size(Y);
    % prepare matrix containing size
    X = ones(n,1); X(:,2) = x;
    % now loop over those variables, each time calculating the
    % regression & residuals
    for i = 1:m
      % estimate beta
      beta = (X′*C^-1*X)^-1*(X′*C^-1*Y(:,i));
      % compute residuals and put in ith column of R
      R(:,i) = Y(:,i)-X*beta;
    end
% done
```

The astute reader (or MATLAB enthusiast) might note that the looped computations can actually be performed more succinctly using:

```
beta = (X′*C^-1*X)^-1*(X′*C^-1*Y); R = Y-X*beta;,
```

and I commend their careful reading, however I have presented the looped code for clarity and consistency with the description of the method in text.

R code—The function below, phyl_resid.R, requires as input **C**, the vector **x**, and the matrix **Y** containing the m size-correlated traits, as before. In R, the $n \times n$ matrix **C** under BM can be computed using the APE functions read.tree() and vcv.phylo() as follows:

```
> library(ape)
```

if the APE library has not already been loaded. Then:

```
> tree<-read.tree(“tree.file”);
> C<-vcv.phylo(tree);
```

in which tree.file is the name of the file containing the Newick format tree. If left blank, i.e.,

```
> tree<-read.tree(“”);
```

then the user will be prompted to enter his or her tree at the command line.

The rows and columns of **C** will be in the order of the tip labels from left to right in the Newick input tree. Because this order is unlikely to correspond to the order of taxa in **x** and **Y**, a useful trick to get the rows and columns of **C** into alphabetical or numerical order is as follows:

```
>C<-C[order(dimnames(C)[[1]]),order(dimnames(C)[[2]])].
```

Because the script specifically requires a matrix (not a data frame) to run, the user might want to use the following commands:

```
> x<-as.matrix(x);
> Y<-as.matrix(Y);
```

before analysis, if their data are indeed in data frame form.

Non-BM **C** can also be computed using function corMartins(), corPagel(), and corGrafen() in R using APE (Paradis 2006). Once we have obtained **C**, we load the function phyl_resid():

```
> source(“phyl_resid.R”);
```

and compute:

```
> r<-phyl_resid(C,x,Y);
```

The residuals are contained in the matrix, r. The function phyl_resid() is given below:

```
phyl_resid<-function(C,x,Y){
    # find out how many y variables and taxa we have
    m<-ncol(Y); n<-nrow(Y);
    # compute inverse of C
    invC<-solve(C);
    # create a matrix for residuals
    r<-matrix(,n,m);
    # prepare X matrix
    X<-matrix(,n,2); X[,1]<-1; X[,2]<-x;
    # now loop over those variables, each time calculating the
    # regression & residuals
    for(i in 1:m){
      # estimate beta
      beta<-
solve(t(X)%*%invC%*%X)%*%(t(X)%*%invC%*
%Y[,i]);
      # compute residuals
      r[,i]<-Y[,i]-X%*%beta;
    }
    phyl_resid<-r;
}
```

PHYLOGENETIC PRINCIPAL COMPONENTS ANALYSIS

MATLAB code—The function below, phyl_pca.m, takes as input: (1) the matrix, **C**, as above and in the text; (2) an $n \times m$ matrix **X** containing the data for m traits from n species; and (3) an optional string variable mode, which should be specified as 'corr' or 'cov' (the default is 'cov' if no mode is specified). As above, the ordering of **C** and **X** should be the same. It returns scores, eigenvalues, eigenvectors, and loadings, calculated as in the main text. It also calls the eigensystem sorting function sorteig.m, also provided below.

```
% function [S,Eval,Evec,L] = phyl_pca(C,X) performs PCA %
using the tree (C), data (X), and mode ('cov' or 'corr')
function [S,Eval,Evec,L] = phyl_pca(C,X,mode)
    % check mode (covariance or correlation matrix)
    if exist('mode','var')∼ = 1
      mode = 'cov';% default mode is 'cov'
    end
    % get m, n
    [n,m] = size(X);
    % first compute the vector of ancestral states
    one = ones(n,1); a = (one'*C^-1*one)^-1*(one'*C^-1*X)';
    % now compute the evolutionary VCV matrix, V
    V = (X-one*a')'*C^-1*(X-one*a')/(n-1);
    % if mode = = correlation matrix
    if strcmp(mode,'corr')
      % standardize X
      X = X./(one*sqrt(diag(V))');
      % change V to correlation matrix
      V = V./(sqrt(diag(V))*sqrt(diag(V))');
      % recalculate a
      a = (one'*C^-1*one)^-1*(one'*C^-1*X)';
    end
    % do eigenanalysis & sort
    [Evec,Eval] = eig(V); [Evec,Eval] = sorteig(Evec,Eval);
    % now compute scores in the original space
    S = (X-one*a')*Evec;
    % compute cross covariance matrix and loadings
    Ccv = (X-one*a')'*C^-1*S/(n-1);
    L = Ccv.*(diag(V)* diag(Eval)').^(-1/2);
% done
```

```
% function [V2,D2] = sorteig(V,D) sorts eigenvectors (V) by
% eigenvalues (D)
function [V2,D2] = sorteig(V,D)
    % make sorted diagonal matrix
    [d,index] = sort(diag(D),'descend'); D2 = diag(d);
    % use index to sort V
    V2 = V(:,index);
% done
```

R code– The function below, phyl_pca.R, requires as input **C** and **X**, as before, and mode. Unlike in phyl_pca.m, mode needs to be specified (as 'corr' or 'cov') for the function to run. In R, **C** under BM can easily be computed as described for phyl_resid.R, above. It returns a variable of the class phyl_pca with components $Eval (eigenvalues), $Evec (eigenvectors), $S (scores), and $L (PC loadings). The function is loaded and called using the following commands:

```
> source("phyl_pca.R");
> pca.results<-phyl_pca(C,X,mode);
```

As before, the user might also find the following command useful if the data in **X** are in data frame rather than matrix form:

```
> X<-as.matrix(X);
```

The function is as follows:

```
phyl_pca<-function(C,X,mode){
    # find out how many columns and taxa we have
    m<-ncol(X); n<-nrow(X);
    # compute inverse of C
    invC<-solve(C);
    # compute vector of ancestral states
    one<-matrix(1,n,1);
      a<-t(t(one)%*%invC%*%X)*sum(sum
(invC))^-1;
    # compute evolutionary VCV matrix
    V<-
t(X-one%*%t(a))%*%invC%*%(X-one%*%t(a))*(n-1)^-1;
    # if correlation matrix
    if(mode = = 'corr'){
      # standardize X
      X = X/(one%*%t(sqrt(diag(V))));
      # change V to correlation matrix
      V = V/(sqrt(diag(V))%*%t(sqrt(diag(V))));
```

```
        # recalculate a
        a<-t(t(one)%∗%invC%∗%X)∗sum(sum(invC))^-1;
      }
      # eigenanalyze
      es = eigen(V);
      result<-NULL; result$Eval<-diag(es$values);
      result$Evec
<-es$vectors;
      # compute scores in the species space
      result$S<-(X-one%∗%t(a))%∗%result$Evec;
      # compute cross covariance matrix
# and loadings
      Ccv<-t(X-one%∗%t(a))%∗%invC%∗%result$S/(n-1);
      result$L<-matrix(,m,m);
      for(i in 1:m)
        for(j in 1:m)
          result$L[i,j]<-Ccv[i,j]/sqrt(V[i,i]∗result$Eval[j,j]);
      phyl_pca<-result;
      # done
}
```

Supporting Information

The following supporting information is available for this article:

Appendix S1: Simulation code: Phylogenetic size-correction simulations and analysis written for MATLAB.
Appendix S2: Simulation code: Phylogenetic principal components simulations and analysis written for MATLAB.
Appendix S3: Size correction using Pagel's (1999) λ.

Supporting Information may be found in the online version of this article.
(This link will take you to the article abstract).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article. Additional results and discussion can be found in a document at http://www.repository.naturalis.nl/record/289893.