doi:10.1111/evo.12596

Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: What you shuffle matters

Dean C. Adams[1,2,3] and Michael L. Collyer[4]

[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011*

[2]*Department of Statistics, Iowa State University, Ames, Iowa 50011*

[3]*E-mail: dcadams@iastate.edu*

[4]*Department of Biology, Western Kentucky University, Bowling Green, Kentucky 42101*

Received November 5, 2014
Accepted December 16, 2014

Evaluating statistical trends in high-dimensional phenotypes poses challenges for comparative biologists, because the high-dimensionality of the trait data relative to the number of species can prohibit parametric tests from being computed. Recently, two comparative methods were proposed to circumvent this difficulty. One obtains phylogenetic independent contrasts for all variables, and statistically evaluates the linear model by permuting the phylogenetically independent contrasts (PICs) of the response data. The other uses a distance-based approach to obtain coefficients for generalized least squares models (*D*-PGLS), and subsequently permutes the original data to evaluate the model effects. Here, we show that permuting PICs is not equivalent to permuting the data prior to the analyses as in *D*-PGLS. We further explain why PICs are not the correct exchangeable units under the null hypothesis, and demonstrate that this misspecification of permutable units leads to inflated type I error rates of statistical tests. We then show that simply shuffling the original data and recalculating the independent contrasts with each iteration yields significance levels that correspond to those found using *D*-PGLS. Thus, while summary statistics from methods based on PICs and PGLS are the same, permuting PICs can lead to strikingly different inferential outcomes with respect to statistical and biological inferences.

KEY WORDS: **Geometric morphometrics, phylogenetic comparative method, phylogenetic generalized least squares, phylogenetic independent contrasts.**

The rapidly growing field of phylogenetic comparative biology provides a suite of analytical tools that enable the examination of trait evolution in a phylogenetic context, accounting for the non-independence of species traits due to shared evolutionary history (e.g., Felsenstein 1985; Grafen 1989; Hansen 1997; Garland and Ives 2000; Blomberg et al. 2003; O'Meara et al. 2006; Thomas et al. 2006; Revell and Collar 2009; Beaulieu et al. 2012). Seminal articles in this field (Felsenstein 1985; Harvey and Pagel 1991; Felsenstein 2004) focused on univariate traits, and thus macroevolutionary analyses of traits treated individually (or a few traits treated simultaneously) tend to dominate the literature (e.g., Garland et al. 1992; Ackerly and Donoghue 1998; Harmon et al. 2010; Mahler et al. 2010; Price et al. 2010; Valenzuela and Adams 2011; but see Rüber and Adams 2001; McPeek et al. 2008; Blankers et al. 2012). However, in the field of geometric morphometrics (GM; Bookstein 1991; Mitteroecker and Gunz 2009; Adams et al. 2013), an increasing number of studies have examined patterns of (multivariate) shape variation in a phylogenetic context (e.g., Rüber and Adams 2001; Bastir et al. 2010; Piras et al. 2010, 2013; Monteiro and Nogueira 2011; Klingenberg and Marugán-Lobón 2013; Monteiro 2013; Outomuro et al. 2013a,b; Polly et al. 2013; Sherratt et al. 2014). Because of the typical large number of shape variables in GM studies—which necessitates many observations in order for parametric multivariate analyses to have analytical solutions—many such studies rely on large phylogenies.

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

As a consequence, considerable effort has been devoted for developing nonparametric methods that merge components of the phylogenetic comparative toolkit with analytical tools for evaluating patterns in "high-dimensional" data, such as shape data, in which the number of shape variables can exceed the number of taxa examined (Klingenberg and Gidaszewski 2010; Klingenberg and Marugán-Lobón 2013; Adams 2014a,b,c; Adams and Felice 2014). For methods investigating the covariation of shape with other variables (i.e., phylogenetic regression), two approaches have been proposed to circumvent this large variable-to-small sample size challenge. One method (Klingenberg and Marugán-Lobón 2013) estimates the $n - 1$ phylogenetically independent contrasts (PICs) from every shape (dependent) variable and the independent variable (i.e., a covariate) for n taxa. A multivariate regression is then performed using Procrustes ANOVA (sensu Goodall 1991), in which the contrasts scores of the shape data are treated as the dependent variable and contrast scores of the covariate are used as the independent variable. Observed regression statistics (e.g., SS, F, R^2) are then calculated, and a sampling distribution of the F-statistic is generated by randomly shuffling the vectors of PIC scores for the shape data and recalculating the statistic from Procrustes ANOVA many times (Klingenberg and Marugán-Lobón 2013; e.g., see Figueirido et al. 2013; Klingenberg and Marugán-Lobón 2013; Santanta and Lofgren 2013; Martín-Serra et al. 2014). The second approach (Adams 2014a) is an extension of phylogenetic generalized least squares (PGLS) estimation of linear model coefficients. Here, phylogenetic transformations of both the shape data and independent variable are performed (sensu Garland and Ives 2000), and predicted values from a multivariate regression of shape on the independent variable are obtained. Summary regression statistics are then calculated from the distances among predicted values. However, in this case the empirical sampling distribution of the F-statistic is generated by randomly shuffling the vectors of shape values, performing again the phylogenetic transformation, and recalculating F-statistics from the multivariate regression in every random permutation (Adams 2014a).

Multivariate Generalizations: Similar Logic, Divergent Outcomes

Both of the procedures described above generalize well-established components of the phylogenetic comparative toolkit so that patterns in high-dimensional shape data may be assessed in light of phylogeny. The first method (hereafter termed as PIC_{rand}) is based on algorithmic estimation of independent contrasts, whereas the second (D-PGLS) is motivated through the algebra of generalized least squares. Further, because it is well known that phylogenetic independent contrasts and PGLS are equivalent in terms of ANOVA statistics (Garland and Ives 2000; Rohlf 2001; Blomberg et al. 2012), it is expected that PIC_{rand} and D-PGLS will produce equivalent sampling distributions for those statistics. Surprisingly however, when implemented on empirical datasets, the two procedures can lead to strikingly different outcomes with respect to the biological inferences that they suggest.

As a simple illustration of this, we examined patterns of evolutionary allometry in head shape across 42 species of *Plethodon* salamanders. Head shape (Fig. 1A) was obtained for each species using landmark-based GM (data from Maerz et al. 2006; Adams et al. 2007; Arif et al. 2007; Myers and Adams 2008; Adams 2010; Deitloff et al. 2013); typical adult body size was obtained from linear measurements (data from Adams and Church 2008, 2011). Using a time-calibrated multigene phylogeny for the genus (Wiens et al. 2006: Fig. 1B), we examined the relationship between head shape and body size in a phylogenetic context using both PIC_{rand} and D-PGLS (Fig. 1C displays an ordination of shape variation with the phylogeny superimposed). All analyses were performed in R 3.1.1 (R Core Team 2014) using the package *geomorph* (Adams and Otárola-Castillo 2013; Adams et al. 2014a) and routines written by the authors. For both analyses, $\alpha = 0.05$ was used as the level of significance for the tests.

As expected, all summary statistics (SS, F, R^2) were identical when estimated with PIC_{rand} and D-PGLS (Table 1). However, differences were found in the inferred significance levels, with D-PGLS identifying no trend between head shape and body size, while PIC_{rand} suggested a significant statistical relationship between them (Table 1). Because only the probability values varied, any differences between PIC_{rand} and D-PGLS must be in the manner in which the permutation tests were performed. Upon inspection, the discrepancy between the two approaches is obvious. With D-PGLS, the data are permuted first in every iteration of the analysis, prior to phylogenetic transformation and computation of statistics. For PIC_{rand}, the PICs are calculated only once, prior to all iterations in the analysis. All permutations and computation of statistics are then based on this single transformation. This difference in the order of operations is clearly important. Indeed, if one simply permutes the data prior to estimating the independent contrasts in every iteration of the analysis, the sampling distributions of ANOVA statistics (and hence P values) are the same for both PIC_{rand} and D-PGLS (Table 1). The conclusion from these findings is that despite the known mathematical relationship between independent contrasts and GLSs, the PIC_{rand} procedure is not equivalent to D-PGLS, because for inferential tests, the exchangeable units under the null hypothesis (i.e., the vectors that are randomized) are not the same between the two methods.

Exchangeable Units Under the Null Hypothesis

For permutation tests, exchangeable units under the null hypothesis are those elements that can be exchanged (i.e., permuted)

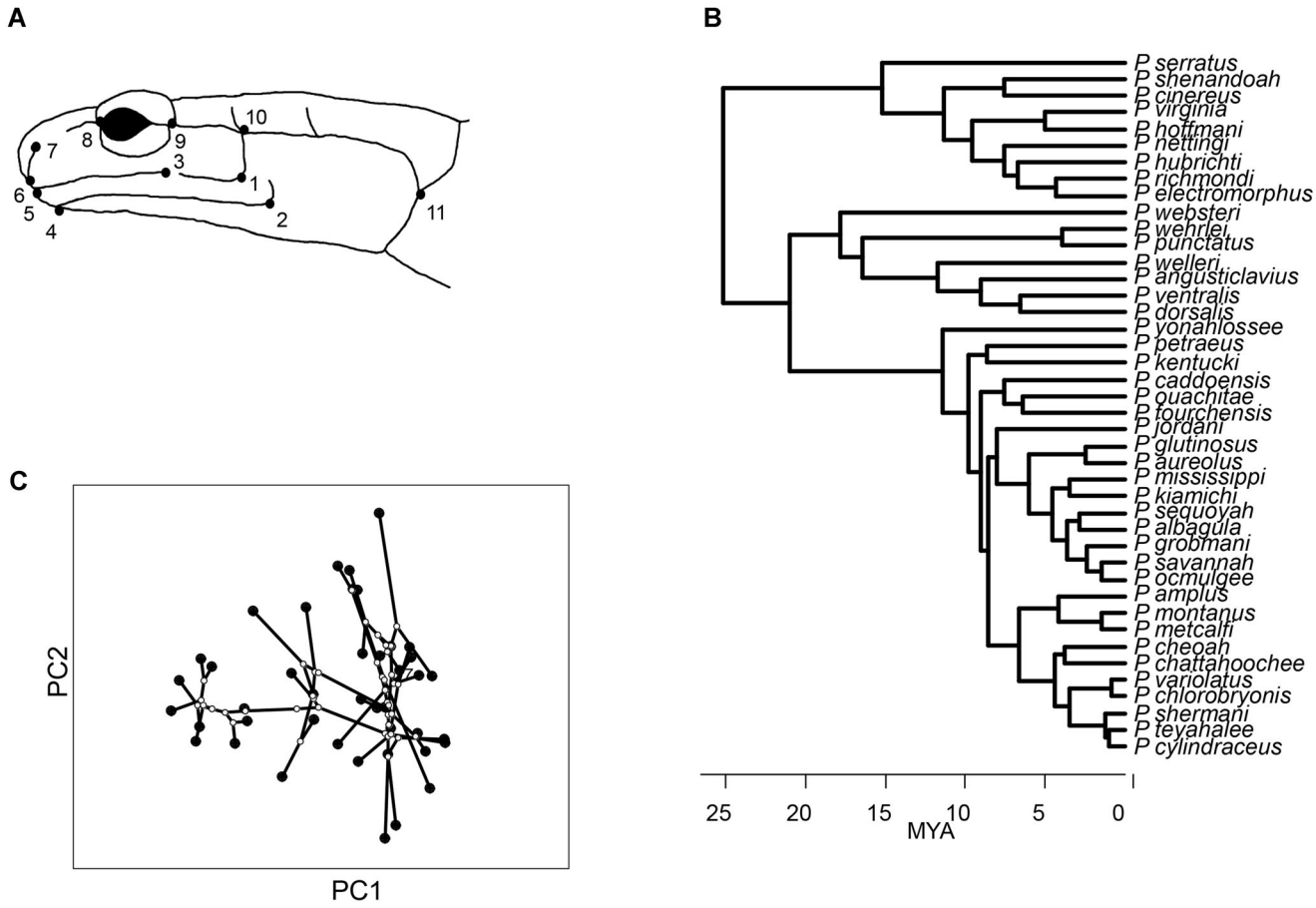

Figure 1. **(A) Positions of 11 anatomical landmarks used to quantify head shape in *Plethodon* salamanders (image from [Adams et al. 2007]). (B) Fossil-calibrated molecular phylogeny displaying the estimated phylogenetic relationships among the species of *Plethodon* examined here (from Wiens et al. 2006). (C) Plot of phylomorphospace for head shape, viewed as the first two principal component axes of tangent space (edges of phylogeny connect ancestral states and extant taxa [sensu Rohlf 2002]).**

Table 1. **Statistical results for empirical example in *Plethodon* salamanders, examining the relationship between head shape and body size (SVL) in a phylogenetic context. Head shape data and the phylogeny are shown in Figure 1.**

D-PGLS	df	*SS*	MS	*F*	R^2	P_{Yrand}	
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.221 NS	
Residual	40	0.0086976	0.0086976				
Total	41	0.0093562	0.0002174				
PIC	df	*SS*	MS	*F*	R^2	P_{PICrand}	P_{Yrand}
SVL	1	0.0006586	0.0006586	3.0288	0.07039	0.026	0.221 NS
Residual	40	0.0086976	0.0086976				
Total	41	0.0093562	0.0002174				

without altering the expected mean squares for the model under consideration (see Anderson and ter Braak 2003; Good 2005). For example, consider the linear model $\mathbf{Y} = \mathbf{1B} + \mathbf{E_1}$, where $\mathbf{Y}$ is an $n \times p$ matrix of shape variables, $\mathbf{1}$ is an $n \times 1$ vector of 1s, $\mathbf{B}$ is a $1 \times p$ vector of mean values for each shape variable, and $\mathbf{E}$ is an $n \times p$ matrix of residuals. The mean values are obtained from the general solution, $\mathbf{B} = \left(\mathbf{X^t X}\right)^{-1} \mathbf{X^t Y}$, where in this case $\mathbf{1}$ is used in place of $\mathbf{X}$. ($\mathbf{X}$ might also include one or more independent variables.) In this linear model, the residual row vectors of $\mathbf{E}$ (from the null model, $\mathbf{1}$) are the exchangeable units under the null hypothesis, such that $trace\,(\Delta\Sigma) = 0$, where the *trace* is the sum of diagonal elements (variable variances) found in the covariance matrix, and Δ indicates the difference in residuals found between null and regression models. In this equation, the expected mean squares of the regression, $\Delta\Sigma$, is estimated as $\Delta\hat{\Sigma} = (1/n)\,(\mathbf{E_X} - \mathbf{E_1})\,(\mathbf{E_X} - \mathbf{E_1})^{\mathbf{t}}$, where the subscripts of each

$\mathbf{E}$ indicate the corresponding model design matrices. Randomizing the vectors of $\mathbf{E_1}$ does not alter the residual sums of squares for the null model, but produces a distribution of possible outcomes under the null hypothesis. (For a comprehensive discussion of exchangeable units see Anderson and ter Braak 2003; Good 2005.) Specifically, if one shuffles the rows of $\mathbf{E_1}$ to produce a randomized matrix, $\mathbf{E_1^*}$, the residual sum of squares is

$$RSS = trace\left(\mathbf{E_1}\mathbf{E_1^t}\right) = trace\left(\mathbf{E_1^*}\left(\mathbf{E_1^*}\right)^{\mathbf{t}}\right). \quad (1)$$

Thus, each permutation produces alternative versions of $\mathbf{Y}$ under the null hypothesis from which sampling distributions of test statistics may be obtained. Note that for this simple model, shuffling row vectors in $\mathbf{Y}$ will provide an equivalent solution, as residuals are deviations from the overall mean (Anderson and ter Braak 2003; Collyer et al. 2015). An implicit assumption with this procedure is that the residuals obtained from $\mathbf{X}$ are independent, that is, the covariance matrix of the residuals is $\Sigma_{\mathbf{X}} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix.

For phylogenetic comparative methods, it is known that residuals are not independent, and that the expected covariance matrix is $\Sigma_{\mathbf{X}} = \sigma^2\mathbf{C}$, where $\mathbf{C}$ is an $n \times n$ phylogenetic covariance matrix. Thus, unlike the solution for ordinary least squares (OLS) coefficients above, GLS coefficients (Garland and Ives 2000; Rohlf 2001) are solved as

$$\mathbf{B} = \left(\mathbf{X^tC^{-1}X}\right)^{-1}\mathbf{X^tC^{-1}Y}. \quad (2)$$

The GLS model can thus be represented as

$$\mathbf{Y} = \mathbf{X}\left(\mathbf{X^tC^{-1}X}\right)^{-1}\mathbf{X^tC^{-1}Y} + \mathbf{E_X}. \quad (3)$$

Garland and Ives (2000) demonstrated that the OLS model can be transformed in one of the two ways to produce *SS* that are the same as those from GLS: via PIC or via calculating a phylogenetic transformation matrix (the latter of which is used in *D*-PGLS for high-dimensional data: Adams 2014a). Transformation via PIC solves coefficients for the linear model, $\mathbf{Y}_{PIC} = \mathbf{X}_{PIC}\mathbf{B} + \mathbf{E}_{PIC}$, where $\mathbf{Y}_{PIC}$ and $\mathbf{X}_{PIC}$ are matrices containing $n - 1$ PICs (rather than the n vectors that match the n taxa). $\mathbf{B}$ is solved via OLS computation, and the residuals of this model are found as $\mathbf{E}_{PIC} = \mathbf{Y}_{PIC} - \mathbf{X}_{PIC}\mathbf{B}$.

Alternatively, using the phylogenetic transformation of *D*-PGLS, equation (3) can be re-expressed as

$$\mathbf{PY} = \mathbf{P}\left(\mathbf{X^tX}\right)^{-1}\mathbf{X^tY} + \mathbf{PE_X} = \mathbf{P}\left[\left(\mathbf{X^tX}\right)^{-1}\mathbf{X^tY} + \mathbf{E_X}\right], \quad (4)$$

where $\mathbf{P}$ is a phylogenetic transformation matrix found as $\mathbf{P} = (\mathbf{UW}^{1/2}\mathbf{U}^t)^{-1}$, and $\mathbf{U}$ and $\mathbf{W}$ are the eigenvalues and eigenvectors of $\mathbf{C}$ (Garland and Ives 2000; see also Adams 2014a). Both PIC and *D*-PGLS produce unequivocally identical results in terms of *SS* for the observed regression, that is, $SS = trace\left(\Delta\mathbf{E}_{PIC}\Delta\mathbf{E}_{PIC}^{\mathbf{t}}\right) = trace\left(\Delta\mathbf{PE}\left(\Delta\mathbf{PE}\right)^{\mathbf{t}}\right)$. Furthermore, Garland and Ives (2000) also demonstrated that the expected covariance matrix of transformed residuals using *D*-PGLS is $\Sigma_{\mathbf{X}} = E\left\{\mathbf{PE_X}\left(\mathbf{PE_X}\right)^{\mathbf{t}}\right\} = \sigma_{\mathbf{X}}^2\mathbf{I}$, meaning the transformed residuals have the desirable property of independence. It might seem, therefore, intuitive that by replacing $\mathbf{Y}$ with $\mathbf{PY}$, and $\mathbf{E}$ with $\mathbf{PE}$, OLS regression can be performed, and either $\mathbf{PY}$ or $\mathbf{PE}$ of a null model can be shuffled in random permutations, as with OLS. This is the typical approach with PIC_{rand} as noted above.

However, neither the vectors of $\mathbf{PY}$ nor $\mathbf{PE}$ are the correct exchangeable units under the null hypothesis (as can be seen in eq. (4)). Because $\mathbf{P}$ is a phylogenetic transformation matrix, back-transformation to original values should be possible. In other words, $\mathbf{P^{-1}PE} = \mathbf{IE} = \mathbf{E}$. Randomizing $\mathbf{PE}$ confounds the phylogenetic correction and the error such that $\mathbf{P^{-1}}\left(\mathbf{PE}\right)^* = \mathbf{P^{-1}P^*E^*} \neq \mathbf{IE}^*$. (Or relatedly: $\mathbf{P^{-1}}\left(\mathbf{PY}\right)^* = \mathbf{P^{-1}P^*Y^*} \neq \mathbf{IY}^*$.) This has the undesirable property also that (building on eq. (1)), $\mathbf{P}\left(\mathbf{E}^*\left(\mathbf{E}^*\right)^{\mathbf{t}}\right) \neq \mathbf{P^*E^*}\left(\mathbf{P^*E^*}\right)^{\mathbf{t}}$, and thus, the original transformation of data to produce independent error cannot be expected to produce estimated error covariance matrixes that are $\hat{\Sigma}_{\mathbf{X}} \approx \sigma^2\mathbf{I}$ in subsequent random permutations. Importantly, the permutations of $\mathbf{E_{PIC}}$ (or $\mathbf{Y_{PIC}}$) suffer from this problem, and as such, the PIC_{rand} procedure may lead to undesirable statistical properties.

Indeed, this is the case. As a simple illustration, we performed a simulation that evaluated the type I error rate of PIC_{rand} and compared it to that of *D*-PGLS. Here, 100 datasets containing 32 species each were simulated from a normal distribution as $\sim N(0,1)$. For each dataset, the response data contained 10 random variables (representing a 10-dimensional shape), and the independent variable was represented by a single, continuous random variable. No covariation between $\mathbf{X}$ and $\mathbf{Y}$ was included in the simulation, and thus the datasets represented the null hypothesis of no relationship between $\mathbf{X}$ and $\mathbf{Y}$. Next, we generated 1000 random phylogenies of 32 species each, and obtained the significance of a phylogenetic regression of $\mathbf{Y}\sim\mathbf{X}$ using both the phylogenies PIC_{rand} and *D*-PGLS. The proportion of times (out of 1000) that each dataset was inferred to be significant (using $\alpha = 0.05$ as the criterion for significance) represented the type I error of the test.

As can be seen in Figure 2, the type I error rate for the *D*-PGLS procedure was at the nominal $\alpha = 0.05$. This confirmed previous findings demonstrating that *D*-PGLS displayed appropriate type I error rates under a wide range of conditions (the method also displays high power; Adams 2014a). By contrast, results from this simulation revealed that PIC_{rand} had unacceptably high type I error rates of nearly 30% (Fig. 2). This result highlights that independent contrasts of the response variables are not the correct exchangeable units under the null hypothesis, and that permuting them can lead to incorrect statistical and biological inferences. Others have shown via simulations that for univariate data and under some restricted conditions, PIC_{rand} can also have low

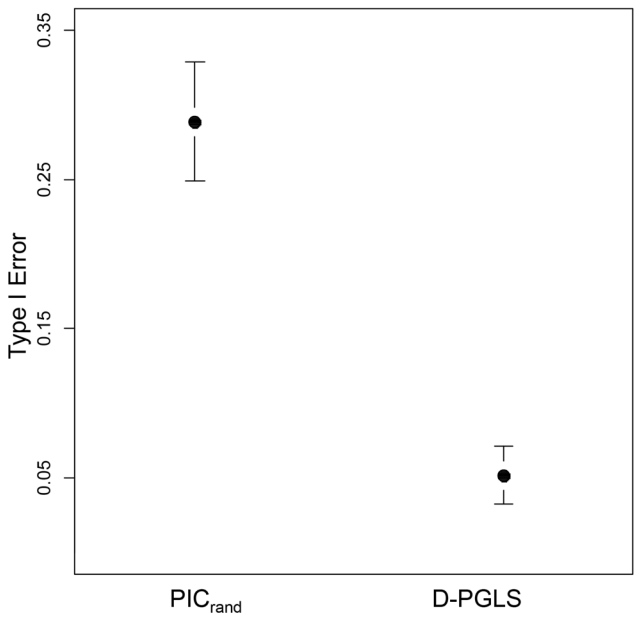

Figure 2. Simulation results displaying Type I error rates for PIC_{rand} and *D*-PGLS. Here, 100 datasets of 32 species were obtained from a normal distribution as $\sim N(0,1)$. For each dataset, the response data contained 10 random variables (representing a 10-dimensional shape), and the independent variable was represented by a single, continuous random variable. One thousand random phylogenies were then simulated, and the significance of each dataset was evaluated on each phylogeny. Type I error for each dataset was the proportion of times (out of 1000) that the phylogenetic regression was found to be significant (mean and 95% CI shown).

statistical power (Legendre and Desdevises 2009). Together, these findings demonstrate that randomizing PICs should be avoided as a method for generating sampling distributions of regression statistics.

Conclusions

In this note, we explored the relationship between two approaches that account for the phylogenetic nonindependence of traits in evolutionarily related taxa, and found that while summary statistics obtained from PIC_{rand} and *D*-PGLS were identical, significance-testing procedures differ. One method permutes independent contrasts, whereas the other permutes the original data (which is the same as permuting residuals of the null model). As a consequence of this difference, the two methods can obtain divergent statistical inferences that can subsequently lead to differing biological conclusions. We further show that PICs are not the correct exchangeable units under the null hypothesis. One undesirable consequence of this is that tests based on permuting independent contrasts have unacceptably high type I error rates (Fig. 2), whereas tests based on *D*-PGLS display type I error rates equal to α and have high statistical power (see Adams 2014a). From these results, we conclude that the PIC_{rand} procedure is not equivalent to *D*-PGLS, and that permuting PICs for assessing statistical significance can lead to incorrect biological inferences.

We were also able to determine that randomizing the original data rather than PICs, and recalculating PICs in every random permutation, not only produced the exact same summary statistics as *D*-PGLS, but also yielded equivalent random outcomes that produced identical sampling distributions of the *F*-statistics (i.e., *P* values of observed statistics were the same). Therefore, one possible adjustment to the PIC approach in order for it to have appropriate type I error rates is to simply not treat the PIC transformation as a one-time solution, but rather as an iterative solution (i.e., in each permutation shuffling the original data and recalculating the PICs). However, there is little appeal to this procedure as a statistical test, because the PIC algorithm would need to be repeated for every variable in every random permutation, thereby increasing computational time. For our *Plethodon* example, this resulted in an increase of over three times greater computation time than using *D*-PGLS, whether performed for 1000, 5000, or 10,000 random permutations. More importantly, *D*-PGLS is easily generalized to more complex models that are less straightforward to implement by using independent contrasts (Pennell and Harmon 2013). As a consequence, the comparisons of groups via ANOVA, or the assessment of several covariates via multiple regression, can all be accomplished efficiently using *D*-PGLS.

Interestingly, the issues we found with randomizing PICs bear similarity to those discussed with alternative permutation procedures for partial Mantel tests (Legendre 2000). With partial Mantel tests, one evaluates the correlation between two distance matrices, $\mathbf{D}_Y$ and $\mathbf{D}_X$, after accounting for information in a third matrix, $\mathbf{D}_P$. One implementation for obtaining the partial Mantel correlation is by first regressing both $\mathbf{D}_Y$ and $\mathbf{D}_X$ against $\mathbf{D}_P$ and finding the correlation between the corresponding residuals (Smouse et al. 1986). Although this—much like calculating PICs—allows a computationally easier solution to estimating the correlation between two distance matrices after accounting for a third, the inclination to use the residuals as exchangeable units under the null hypothesis yields inflated type I error rates (Legendre 2000; and confirmed in a phylogenetic context by Harmon and Glor 2010). Instead, a "matrix randomization" of $\mathbf{D}_Y$ followed by regression between the random $\mathbf{D}_Y$ and $\mathbf{D}_P$ in every permutation limits this maleficence—as did randomizing vectors of $\mathbf{Y}$ and recalculating PICs in this study—although partial Mantel tests are still prone to inferential errors if distributions of pairwise distances are skewed (Legendre 2000).

In essence, the point illustrated by our current demonstration is that computationally distinct methods may be used to obtain identical test statistics representing a given hypothesis, but how one implements the permutation test to assess statistical significance of these values is equally important and bears careful consideration as well. Randomizing PICs for null hypothesis tests is inappropriate. However, we feel it is still important to view PIC as a useful and meaningful descriptive transformation of the observed data that can be used to reveal the important contrasts associated with a phylogeny. PIC is nonetheless a transformation that should be decoupled from hypothesis testing, when such tests are based on randomizations. For this reason, we advocate that generating sampling distributions to evaluate statistics for phylogenetic comparative analyses should be performed using *D*-PGLS.

ACKNOWLEDGMENTS

We thank A. Cardini for identifying the discrepancy in significance levels between approaches. This work was sponsored in part by NSF grant DEB-1257287 to DCA.

DATA ARCHIVING

The doi for our data is 10.5061/dryad.2jv17.

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.

Adams, D. C. 2010. Parallel evolution of character displacement driven by competitive selection in terrestrial salamanders. BMC Evol. Biol. 10:1–10.

———. 2014a. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution 68:2675–2688.

———. 2014b. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. Syst. Biol. 63:166–177.

———. 2014c. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. Syst. Biol. 63:685–697.

Adams, D. C., and J. O. Church. 2008. Amphibians do not follow Bergmann's rule. Evolution 62:413–420.

Adams, D. C., and R. Felice. 2014. Assessing phylogenetic morphological integration and trait covariation in morphometric data using evolutionary covariance matrices. PLoS ONE 9:e94335.

———. 2011. The evolution of large-scale body size clines in Plethodon: evidence of heat-balance or species-specific artifact? Ecography 34:1067–1075.

Adams, D. C., and E. Otárola-Castillo. 2013. Geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods Ecol. Evol. 4:393–399.

Adams, D. C., M. E. West, and M. L. Collyer. 2007. Location-specific sympatric morphological divergence as a possible response to species interactions in West Virginia *Plethodon* salamander communities. J. Anim. Ecol. 76:289–295.

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

Adams, D. C., E. Otárola-Castillo, and E. Sherratt. 2014. Geomorph: software for geometric morphometric analyses. R package version 2.1.1. Available at http://CRAN.R-project.org/package=geomorph.

Anderson, M. J., and C. J. F. t. ter Braak. 2003. Permutation tests for multifactorial analysis of variance. J. Stat. Comput. Simul. 73:85–113.

Arif, S., D. C. Adams, and J. A. Wicknick. 2007. Bioclimatic modelling, morphology, and behaviour reveal alternative mechanisms regulating the distributions of two parapatric salamander species. Evol. Ecol. Res. 9:843–854.

Bastir, M., A. Rosas, C. B. Stringer, J. M. Cuétara, R. Kruszynski, G. W. Weber, C. F. Ross, and M. J. Ravosa. 2010. Effects of brain and facial size on basicranial form in human and primate evolution. J. Hum. Evol. 58:424–431.

Beaulieu, J. M., D. C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Blankers, T., D. C. Adams, and J. J. Wiens. 2012. Ecological radiation with limited morphological diversification in salamanders. J. Evol. Biol. 25:634–646.

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

Blomberg, S. P., J. G. Lefevre, J. A. Wells, and M. Waterhouse. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol. 61:382–391.

Bookstein, F. L. 1991. Morphometric tools for landmark data: geometry and biology. Cambridge Univ. Press, Cambridge, U.K.

Collyer, M. L., D. J. Sekora, and D. C. Adams. 2015. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. Heredity. doi:10.1038/hdy.2014.75.

Deitloff, J., J. D. Petersen, and D. C. Adams. 2013. Complex species interactions lead to unpredictable outcomes in *Plethodon*. Herpetologica 62:1–10.

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

———. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.

Garland, T. J., 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.

Figueirido, B., Z. J. Tseng, and A. Martín Serra. 2013. Skull shape evolution in durophagous carnivorans. Evolution. 67.

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

Good, P. 2005. Permutation, parametric, and bootstrap tests of hypotheses. Springer, New York.

Goodall, C. R. 1991. Procrustes methods in the statistical analysis of shape (with discussion and rejoinder). J. R. Stat. Soc. B 53:285–339.

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.

Harmon, L. J., J. B. Losos, T. J. Davies, R. G. Gillespie, J. L. Gittleman, W. B. Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Harmon, L. J., and R. E. Glor. 2010. Poor statistical performance of the Mantel test in phylogenetic comparative analyses. Evolution 64:2173–2178.

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

Klingenberg, C. P., and N. A. Gidaszewski. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Syst. Biol. 59:245–261.

Klingenberg, C. P., and J. Marugán-Lobón. 2013. Evolutionary covaration in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Syst. Biol. 62:591–610.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J. Stat. Comput. Simul. 67:37–73.

Legendre, P., and Y. Desdevises. 2009. Independent contrasts and regression through the origin. J. Theor. Biol. 259:727–743.

Maerz, J. C., E. M. Myers, and D. C. Adams. 2006. Trophic polymorphism in a terrestrial salamander. Evol. Ecol. Res. 8:23–35.

Mahler, D. L., L. J. Revell, R. E. Glor, and J. B. Losos. 2010. Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean Anoles. Evolution 64:2731–2745.

Martín-Serra, A., B. Figueirido, and P. Palmqvist. 2014. A three-dimensional analysis of the morphological evolution and locomotor behaviour of the carnivoran hind limb. BMC Evol. Biol. 14.

McPeek, M. A., L. Shen, J. Z. Torrey, and H. Farid. 2008. The tempo and mode of three-dimensional morphological evolution in male reproductive structures. Am. Nat. 171:E158–E178.

Mitteroecker, P., and P. Gunz. 2009. Advances in geometric morphometrics. Evol. Biol. 36:235–247.

Monteiro, L. R. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. Hystrix 24:25–32.

Monteiro, L. R. and M. R. Nogueira. 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. BMC Evol. Biol. 11:1–23.

Myers, E. M., and D. C. Adams. 2008. Morphology is decoupled from interspecific competition in *Plethodon* salamanders in the Shenandoah Mountains. Herpetologica 64:281–289.

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

Outomuro, D., D. C. Adams, and F. Johansson. 2013a. Evolution of wing shape in ornamented-winged damselflies. Evol. Biol. 40:300–309.

———. 2013b. Wing shape allometry and aerodynamics in calopterygid damselflies: a comparative approach. BMC Evol. Biol. 13:1–11.

Pennell, M. W., and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. N. Y. Acad. Sci. 1289:90–105.

Piras, P., L. Maiorino, P. Raia, F. Marcolini, D. Salvi, L. Vignoli, and T. Kotsakis. 2010. Functional and phylogenetic constraints in Rhinoceratinae craniodental morphology. Evol. Ecol. Res. 12:897–928.

Piras, P., L. Maiorino, L. Teresi, C. Meloro, F. Lucci, T. Kotsakis, and P. Raia. 2013. Bite of the Cats: Relationships between Functional Integration and Mechanical Performance as Revealed by Mandible Geometry. Syst. Biol. 62:878–900.

Price, S. A., P. C. Wainwright, D. R. Bellwood, E. Kazancioglu, D. C. Collar, and T. J. Near. 2010. Functional innovations and morphological diversification in parrotfish. Evolution 64:3057–3068.

Polly, P. D., A. M. Lawing, A.-C. Fabre, and A. Goswami. 2013. Phylogenetic principal components analysis and geometric morphometrics. Hystrix 24:1–9.

R Core Team. 2014. R: a language and environment for statistical computing. Version 3.1.1. R Foundation for Statistical Computing, Vienna, Austria. Available at http://cran.R-project.org.

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

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

———. 2002. Geometric morphometrics and phylogeny. Pp. 175–193 *in* N. MacLeod and P. L. Forey, eds. Morphology, shape and phylogeny. Taylor & Francis, Lond.

Rüber, L. and D. C. Adams. 2001. Evolutionary convergence of body shape and trophic morphology in cichlids from Lake Tanganyika. J. Evol. Biol. 14:325–332.

Santanta, S. E., and S. E. Lofgren. 2013. Does nasal echolocation influence the modularity of the mammal skull?. J. Evol. Biol 26:2520–2526.

Sherratt, E., D. J. Gower, C. P. Klingenberg, and M. Wilkinson. 2014. Evolution of cranial shape in caecilians (Amphibia: Gymnophiona). Evol. Biol. 41:528–545.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. 35:627–632.

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc. R. Soc. B 273:1619–1624.

Valenzuela, N. and D. C. Adams. 2011. Chromosome number and sex determination coevolve in turtles. Evolution 65:1808–1813.

Wiens, J. J., T. N. Engstrom, and P. T. Chippendale. 2006. Rapid diversification, incomplete isolation, and the "speciation clock" in North American salamanders (genus: Plethodon): testing the hybrid swarm hypothesis of rapid radiation. Evolution 60:2585–2603.

Associate Editor: M. Rosenberg
Handling Editor: R. Shaw