1-2018

Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations

Dean C. Adams
Iowa State University, dcadams@iastate.edu

Michael L. Collyer
Chatham University

Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations

Abstract

Recent years have seen increased interest in phylogenetic comparative analyses of multivariate datasets, but to date the varied proposed approaches have not been extensively examined. Here we review the mathematical properties required of any multivariate method, and specifically evaluate existing multivariate phylogenetic comparative methods in this context. Phylogenetic comparative methods based on the full multivariate likelihood are robust to levels of covariation among trait dimensions and are insensitive to the orientation of the dataset, but display increasing model misspecification as the number of trait dimensions increases. This is because the expected evolutionary covariance matrix (V) used in the likelihood calculations becomes more illconditioned as trait dimensionality increases, and as evolutionary models become more complex. Thus, these approaches are only appropriate for datasets with few traits and many species. Methods that summarize patterns across trait dimensions treated separately (e.g., SURFACE) incorrectly assume independence among trait dimensions, resulting in nearly a 100% model misspecification rate. Methods using pairwise composite likelihood are highly sensitive to levels of trait covariation, the orientation of the dataset, and the number of trait dimensions. The consequences of these debilitating deficiencies is that a user can arrive at differing statistical conclusions, and therefore biological inferences, simply from a dataspace rotation, like principal component analysis. By contrast, algebraic generalizations of the standard phylogenetic comparative toolkit that use the trace of covariance matrices are insensitive to levels of trait covariation, the number of trait dimensions, and the orientation of the dataset. Further, when appropriate permutation tests are used, these approaches display acceptable Type I error and statistical power. We conclude that methods summarizing information across trait dimensions, as well as pairwise composite likelihood methods should be avoided, while algebraic generalizations of the phylogenetic comparative toolkit provide a useful means of assessing macroevolutionary patterns in multivariate data. Finally, we discuss areas in which multivariate phylogenetic comparative methods are still in need of future development; namely highly multivariate Ornstein-Uhlenbeck models and approaches for multivariate evolutionary model comparisons.

Keywords

phylogenetic comparative methods, multivariate, high-dimensional data

Comments

This is a pre-copyedited, author-produced version of an article accepted for publication in Systematic Biology following peer review. The version of record, Dean C. Adams, Michael L. Collyer; Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations, *Systematic Biology*, Volume 67, Issue 1, January 2018, Pages 14–31 is available online at: doi: 10.1093/sysbio/syx055. Posted with permission.

MULTIVARIATE PHYLOGENETIC COMPARATIVE METHODS: EVALUATIONS, COMPARISONS, AND RECOMMENDATIONS

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

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

Department of Statistics, Iowa State University, Ames IA, USA

[2]*Department of Science, Chatham University, Pittsburgh, PA, USA*

[3]*Corresponding author email: dcadams@iastate.edu*

Short title: Comparative methods for high-dimensional data

Abstract

Recent years have seen increased interest in phylogenetic comparative analyses of multivariate datasets, but to date the varied proposed approaches have not been extensively examined. Here we review the mathematical properties required of any multivariate method, and specifically evaluate existing multivariate phylogenetic comparative methods in this context. Phylogenetic comparative methods based on the full multivariate likelihood are robust to levels of covariation among trait dimensions and are insensitive to the orientation of the dataset, but display increasing model misspecification as the number of trait dimensions increases. This is because the expected evolutionary covariance matrix (**V**) used in the likelihood calculations becomes more ill-conditioned as trait dimensionality increases, and as evolutionary models become more complex. Thus, these approaches are only appropriate for datasets with few traits and many species. Methods that summarize patterns across trait dimensions treated separately (e.g., *SURFACE*) incorrectly assume independence among trait dimensions, resulting in nearly a 100% model misspecification rate. Methods using pairwise composite likelihood are highly sensitive to levels of trait covariation, the orientation of the dataset, and the number of trait dimensions. The consequences of these debilitating deficiencies is that a user can arrive at differing statistical conclusions, and therefore biological inferences, simply from a dataspace rotation, like principal component analysis. By contrast, algebraic generalizations of the standard phylogenetic comparative toolkit that use the trace of covariance matrices are insensitive to levels of trait covariation, the number of trait dimensions, and the orientation of the dataset. Further, when appropriate permutation tests are used, these approaches display acceptable Type I error and statistical power. We conclude that methods summarizing information across trait dimensions, as well as pairwise composite likelihood methods should be avoided, while algebraic generalizations of the phylogenetic comparative toolkit provide a useful means of assessing macroevolutionary patterns in multivariate

data. Finally, we discuss areas in which multivariate phylogenetic comparative methods are still in need of future development; namely highly multivariate Ornstein-Uhlenbeck models and approaches for multivariate evolutionary model comparisons.

Keywords: phylogenetic comparative methods, multivariate, high-dimensional data

INTRODUCTION

Understanding patterns of trait evolution across sets of taxa requires accounting for the lack of independence among species due to shared evolutionary history (Felsenstein 1985). From this simple premise the burgeoning field of phylogenetic comparative biology has emerged, whose suite of statistical tools facilitate the evaluation of phenotypic patterns in a phylogenetic context to address a wide range of biological hypotheses. For example, phylogenetic comparative methods (PCMs) may be used to compare the rate of trait evolution among one or more sets of taxa or traits on a phylogeny (Garland 1992; O'Meara, et al. 2006; Thomas, et al. 2006; Revell and Harmon 2008; Adams 2013), and to distinguish among differing models that describe how trait variation accumulates (e.g., Brownian motion [BM] versus Ornstein-Uhlenbeck [OU] models: Hansen 1997; Beaulieu, et al. 2012). Other methods characterize the degree to which phenotypic traits display phylogenetic signal (Pagel 1999; Blomberg, et al. 2003), determine whether trait variation differs among groups of taxa (i.e., phylogenetic ANOVA: Garland, et al. 1993), or evaluate whether traits covary across the phylogeny (i.e., phylogenetic regression: Felsenstein 1985; Grafen 1989; Garland and Ives 2000). These and other methods provide evolutionary biologists with a panoply of analytical tools for testing hypotheses that describe the evolution of phenotypic diversity, and provide insight on the putative processes that have generated these macroevolutionary patterns (for a review see Pennell and Harmon 2013).

Presently, most analytical methods in the phylogenetic comparative toolkit model the evolution of a single trait across the phylogeny (Uyeda, et al. 2015). However, the last decade has seen increased interest in utilizing PCMs for examining phylogenetic patterns of trait evolution in multivariate datasets (e.g., Rüber and Adams 2001; Revell and Harmon 2008; Revell and Collar 2009; Bastir, et al. 2010; Monteiro and Nogueira 2011; Klingenberg and Marugán-Lobón 2013; Monteiro 2013; Outomuro, et al. 2013; Polly, et al. 2013; Sherratt, et al. 2014; Sherratt, et al.

2016). Several distinct approaches have been proposed for statistically evaluating multivariate trends in light of a phylogeny. One method uses likelihood ratio tests to evaluate the fit of the data to the phylogeny under differing models of trait evolution (Revell and Harmon 2008). This method relies on adequate estimation of model log-likelihood, which is possible with many taxa and few traits. However, when multivariate data sets have comparatively many trait variables or few taxa (a common occurrence with empirical data sets), estimating log-likelihoods is troublesome; thus, alternative methods for estimating log-likelihoods, or using test statistics that are correlated with log-likelihood estimates, have been proposed as log-likelihood surrogates.

One such approach evaluates evolutionary models via log-likelihood estimation across individual (univariate) trait dimensions treated separately and sums these to arrive at an overall hypothesis of the best-fitting evolutionary model for the data given a phylogeny (Ingram and Mahler 2013; Grundler and Rabosky 2014; Moen, et al. 2016). Another procedure uses test statistics based on traces of the same covariance matrices used for log-likelihood estimation (which are correlated with likelihood ratio test statistics) to evaluate macroevolutionary hypotheses in high-dimensional datasets (Adams 2014c; Adams 2014a; Adams 2014b; Adams and Felice 2014; Denton and Adams 2015). Finally, a recently introduced approach combines pairwise composite likelihood – a pseudo-likelihood estimated from all or a portion of possible pairwise combinations of trait variables - and phylogenetic simulation to compare the fit of the multivariate dataset to the phylogeny under a null and alternative hypothesis (Goolsby 2016). Strikingly, while all of these procedures have been developed to extend the phylogenetic comparative toolkit in various ways for the analysis of multivariate data, to date no study has compared their ability to accurately and reliably evaluate patterns in such multivariate dataspaces.

The purpose of this paper is to examine existing phylogenetic comparative approaches that evaluate trends in multivariate data in an effort to provide guidance for empiricists and to identify

areas ripe for future analytical development (findings summarized in Table 1). We first review the general properties required of any analytical method describing patterns in multivariate data, and describe how these properties are also applicable to phylogenetic comparative methods. We then review the procedures currently developed for characterizing multivariate patterns in a phylogenetic context, and use computer simulations to compare some of their properties under differing conditions. We find that even under simple conditions (e.g., Brownian motion for a small number of trait dimensions) approaches that summarize information across individual axes of the dataspace display high levels of model misspecification when comparing evolutionary models, which greatly limits their utility. Likewise, methods based on pairwise composite likelihood can arrive at differing statistical inferences based entirely on how the multivariate dataspace is oriented, rendering their conclusions arbitrary. By contrast, comparing the fit of differing evolutionary models using likelihood ratio tests or AIC scores do not suffer from these shortcomings, but display increased model misspecification as trait dimensionality increases. Finally, we find that log-likelihood correlates (statistics using the traces of covariance matrices) display none of these challenges, and thus appear appropriate for use on multivariate datasets for hypothesis testing under Brownian motion. We further find that when the correct permutation procedures are utilized, these approaches display acceptable statistical performance in terms of type I error and power, and are thus suitable for evolutionary hypothesis testing. We conclude that methods summarizing information across trait dimensions individually, as well as pairwise composite likelihood methods should be avoided, while using log-likelihood correlates based on algebraic generalizations of the phylogenetic comparative toolkit provide a useful means of assessing macroevolutionary patterns in multivariate data. Areas in need of additional theoretical development; namely the development of robust approaches for evaluating highly multivariate

Ornstein-Uhlenbeck models and methods for multivariate evolutionary model comparisons, are also discussed.

Necessary Characteristics of Analytical Methods for Multivariate Data

Here we consider the general geometric properties inherent to multivariate data, and the requirements of any analytical approach designed to evaluate patterns in such datasets. The goal of many analytical and statistical methods is to describe patterns of dispersion of species' trait values with respect to the hypothesis under investigation. This is the case for ordinary least squares (OLS) models such as analysis of variance and regression. Additionally, the hypothesis could incorporate some phylogenetic model for how trait variation is expected to accumulate over time. For instance, many phylogenetic comparative methods take into consideration the shared evolutionary history among species through the incorporation of the phylogenetic covariance matrix (**C**) when using generalized least squares (GLS) to estimate model parameters (see Rohlf 2006). For univariate data, such analytical methods describe the dispersion of species values along a number line, which represents a one-dimensional trait space. Likewise, multivariate analytical methods summarize patterns of dispersion among species in a multivariate trait space. The axes of this dataspace may represent a set of single-valued traits treated simultaneously (e.g., length, width, and height measures), a set of summary axes of the original variables (such as ordination scores), or the dimensions of a composite multi-dimensional trait such as shape derived from landmark-based morphometric methods (Bookstein 1991; Mitteroecker and Gunz 2009; Adams, et al. 2013). Figure 1a provides an example of a three-dimensional trait space in which the locations of 15 species in each of two categories are observed.

Both univariate and multivariate methods summarize dispersion in their respective datasets, but with multivariate data, additional mathematical properties must also be considered.

For instance, phenotypic traits are not independent, but can covary with one another. Hence, multivariate analytical methods must be capable of accurately summarizing the dispersion of species in this space regardless of the degree of covariation in the trait data. Additionally, the orientation of the dataspace should have no effect on statistical summaries obtained from the dataset. For example, rotating the data in Figure 1a to its principal component axes provides a different view of the multivariate dataspace (Fig. 1b), but the dispersion of points in the plot is exactly the same. Thus, statistical summaries of the data in either orientation must also be identical, so long as all trait dimensions that contain variation are included in the analysis. For Figures 1a and 1b this is indeed the case, as the summary parameter for multivariate ANOVA of the three-dimensional dataset is the same for both orientations (e.g., Pillai's trace = 0.65275 in this case). This property of rotation-invariance is well-known for multivariate OLS statistical methods in general (Mardia, et al. 1979; Rohlf 1999; Langsrud 2004), as linear models are invariant under linear transformations, including rigid rotations. Finally, rotation-invariance is not merely a matter of choice or convenience. For high-dimensional phenotypic data such as landmark-based morphometric shape data, rotation-invariance is in fact, essential, because there is no inherently 'natural' orientation in the data. That is, one orientation of the aligned landmark coordinates is as valid as any other, and each expresses the same information regarding patterns of multivariate shape disparity among specimens. As such, any orientation of the multivariate dataspace may be used as input in downstream statistical analyses, and summary test measures must therefore be insensitive to differing choices of orientation.

Importantly, and while not always considered, multivariate phylogenetic comparative methods must also conform to these fundamental geometric properties. Specifically, parameter estimates and statistical summaries must accurately characterize the evolutionary patterns of dispersion of species in the trait space regardless of the degree of covariation among trait

dimensions. Additionally, summary measures and statistical tests based on them should be invariant to the orientation of the multivariate dataspace, so long as all trait dimensions containing variation are treated simultaneously. For example, Figure 1C contains a phylomorphospace (sensu Sidlauskas 2008; Polly, et al. 2013) for a three-dimensional trait, displaying the dispersion of 16 species relative to their phylogenetic relationships. Figure 1d displays the same dataspace rotated to its principal axes. Clearly, the dispersion of species relative to their phylogenetic relationships remains unchanged irrespective of the orientation of the dataspace. Thus, any multivariate PCM summarizing patterns of trait dispersion relative to their phylogenetic relationships using the entire multivariate dataset must also exhibit rotation-invariance.

Conducting Phylogenetic Comparative Analyses on Multivariate Data

Phylogenetic comparative methods describe patterns of trait covariation by conditioning the data on the phylogeny under a particular model of evolutionary change (frequently Brownian motion: see Felsenstein 1973; Felsenstein 1981). This is accomplished via two steps: model estimation and model evaluation. First, most PCMs fit a GLS model to the data to describe the relationship:

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \varepsilon \qquad 1$$

where $\mathbf{Y}$ is a $N \times p$ matrix of trait values for the N species across p trait dimensions, and $\mathbf{X}$ is a $N \times k$ design matrix, which is frequently a column of ones, but may also contain one or more independent variables (e.g., for phylogenetic regression). The coefficients, are estimated via generalized least squares as, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\boldsymbol{t}}\mathbf{C}^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\boldsymbol{t}}\mathbf{C}^{-\mathbf{1}}\mathbf{Y}$, where $^{\boldsymbol{t}}$ and $^{-1}$ refer to matrix transposition and inversion, respectively, and $\mathbf{C}$ is an $N \times N$ phylogenetic covariance matrix. Fitted values are

estimated as, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and the residuals of the model (ε), found as, $\mathbf{Y} - E(\mathbf{Y}) = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, with the vectored form of the matrix of residuals having values assumed to be normally distributed as $\mathcal{N}(0, \mathbf{V})$, where **V** describes the lack of independence due to the phylogeny relative to the evolutionary model under consideration. The log-likelihood of the model (e.g., O'Meara, et al. 2006; Revell and Harmon 2008; Bartoszek, et al. 2012; Clavel, et al. 2015) may be estimated based on the equation:

$$-\frac{1}{2}((\mathbf{y} - E(\mathbf{y}))^{\boldsymbol{t}}\mathbf{V}^{-\mathbf{1}}(\mathbf{y} - E(\mathbf{y})) + log|\mathbf{V}| + Np \cdot log(2\pi)) \qquad 2$$

Here, **y** is a $Np \times 1$ column vector of trait values for the N species across p trait dimensions (found using the *vec* operator on the $N \times p$ data matrix **Y**), $E(\mathbf{y})$ is an $Np \times 1$ column vector of expected values (i.e., the vectorization of the matrix of phylogenetic means, $\mathbf{X}\hat{\beta}$), and **V** is a $Np \times Np$ expected covariance matrix for the evolutionary model under consideration (see Revell and Harmon 2008). For the case of a single Brownian motion model, **V** is found as: $\mathbf{V} = \mathbf{R} \otimes \mathbf{C}$, and represents the Kronecker product of an hypothesized $p \times p$ trait covariance matrix (typically called the rate matrix: **R**), with the $N \times N$ phylogenetic covariance matrix (**C**) as described above. For some evolutionary models, estimating **V** involves considering the evolutionary model and the uniqueness of multiple rates (for a general overview see Clavel, et al. 2015). In these cases, **R** matrices are typically empirically derived, and **V** represents the joint contribution of **C**, as estimated from the evolutionary model, and its influence on **R** (see Revell and Harmon 2008). For other evolutionary models (e.g., Ornstein-Uhlenbeck models), estimating **V** is considerably more complex (see Clavel, et al. 2015).

Alternatively, model coefficients may be obtained from phylogenetic independent contrasts (Felsenstein 1985), phylogenetic generalized least squares (Grafen 1989; Martins and Hansen 1997), or least squares estimation based on a phylogenetic transformation of the data (Garland and Ives 2000). Prior work has shown that these algebraic approaches lead to identical fitted values and covariance matrices, and thus provide alternative, but equivalent, implementations for fitting the data to the phylogeny given a model of evolutionary change (see Garland and Ives 2000; Rohlf 2001; Blomberg, et al. 2012).

Subsequent to the estimation of model coefficients, patterns of covariation in the response variables (**Y**) conditioned on the phylogeny can be statistically evaluated relative to the evolutionary hypothesis under investigation. Some multivariate PCMs use likelihood ratio tests (LRT) or indexing measures of penalized likelihood (such as Akaike information criterion, AIC) to accomplish this. For instance, LRT or AIC scores may be used to compare the fit of the data to the phylogeny under differing evolutionary models (e.g., Brownian motion versus Ornstein-Uhlenbeck models) to determine which provides the highest support (e.g., O'Meara, et al. 2006; Revell and Harmon 2008; Bartoszek, et al. 2012; Clavel, et al. 2015). In such cases, LRT and AIC are focused on evaluating different estimates of **V**, which represent differing evolutionary scenarios for how trait variation accumulates given a consistent statistical design (e.g., Revell and Harmon 2008).

Similarly, likelihood ratio tests may be used to evaluate the fit of the data to a set of independent variables in the design matrix (**X**), given the expected phylogenetic covariance, as described by phylogenetic regression. In these cases, the expected phylogenetic covariance under a particular evolutionary model (**V**) remains consistent, but the expected values from nested statistical models ($\mathbf{X}_i$) will vary. Using LRT to compare such models (e.g., **Y**~**X** versus **Y**~**1**, where **1** means a model including only an intercept), the phylogenetic covariance (**V**) remains

constant, and thus the LRT is simply the difference in scalars found from the first part of the log-likelihood equation (equation 2):

$$log\left(\frac{L(\mathbf{Y}|\mathbf{V},\mathbf{X}_F)}{L(\mathbf{Y}|\mathbf{V},\mathbf{X}_0)}\right) = -\frac{1}{2}\Big(\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_F)\big)^{\boldsymbol{t}}\mathbf{V}^{-1}\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_F)\big) + log|\mathbf{V}| + Np\cdot \log(2\pi)\Big) - \Big[-\frac{1}{2}\Big(\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_0)\big)^{\boldsymbol{t}}\mathbf{V}^{-1}\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_0)\big) + log|\mathbf{V}| + Np\cdot\log(2\pi)\Big)\Big] =$$

$$\frac{1}{2}\Big[\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_0)\big)^{\boldsymbol{t}}\mathbf{V}^{-1}\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_0)\big) - \big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_F)\big)^{\boldsymbol{t}}\mathbf{V}^{-1}\big(\mathbf{y}-E(\mathbf{y}|\mathbf{V},\mathbf{X}_F)\big)\Big]. \quad 3$$

Further, when $\mathbf{R}$ is a $p \times p$ identity matrix, the two times the LRT statistic is also the same as the sum of squares (*SS*) comprising the numerator of an *F*-statistic, calculated in analyses of variance (ANOVA) for multivariate data, based on the traces of estimated sums of squares and cross-products (SSCP) matrices (Anderson 2001):

$$2(\text{LRT}) = trace[(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_0))^t\mathbf{C}^{-1}(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_0)) - (\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_F))^t\mathbf{C}^{-1}(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_F))]. \quad 4$$

Likewise, if $\mathbf{R}$ is a $p \times p$ diagonal matrix with diagonal elements equal to the rank difference between $\mathbf{X}_0$ and $\mathbf{X}_F$ (the null and "full" design matrices, respectively), Δk, the two times the LRT statistic is the same as the trace of the estimated covariance matrix for $\mathbf{X}_F$ compared to $\mathbf{X}_0$:

$$2(\text{LRT}) = trace\big(\widehat{\boldsymbol{\Sigma}}_{\Delta k}\big) = \frac{1}{\Delta k}trace[(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_0))^t\mathbf{C}^{-1}(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_0)) - (\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_F))^t\mathbf{C}^{-1}(\mathbf{Y}-E(\mathbf{Y}|\mathbf{X}_F))]. \quad 5$$

In the event that **C** is singular, or just as a computationally efficient step, a phylogenetic transformation matrix, **P**, can be calculated from eigen-analysis of **C**, such that:

$$2(\text{LRT}) = trace\left(\widehat{\boldsymbol{\Sigma}}_{\Delta k}\right) = \frac{1}{\Delta k} trace[(\mathbf{PY} - E(\mathbf{PY}|\mathbf{PX}_0))^t(\mathbf{PY} - E(\mathbf{PY}|\mathbf{PX}_0)) - (\mathbf{PY} - E(\mathbf{PY}|\mathbf{PX}_F))^t(\mathbf{PY} - E(\mathbf{PY}|\mathbf{PX}_F))]. \quad 6$$

(Garland and Ives 2000; Adams2014b), meaning matrix inversion can be avoided altogether. Transforming data using phylogentically independent contrasts (Felsenstein 1985) also avoids matrix inversion. This association between LRT and alternative statistics using traces of covariance matrices is important, because other multivariate PCMs use the traces of the covariance matrices to evaluate the fit of the data to the phylogeny relative to a set of independent variables (e.g., Adams 2014b; Adams and Collyer 2015).

One challenge with likelihood approaches is that as the number of trait dimensions (p) increases, covariance matrices become more unstable. Also, as p approaches N they become singular, and as such, hypothesis tests based on estimating the log-likelihood (which includes inverting and finding the determinant of a singular **V** matrix) becomes computationally prohibitive (see Adams 2014c). By contrast, PCMs which utilize the trace of a covariance matrix are not as prohibitive under these conditions, because inverting singular matrices is not needed, as shown above.

Finally, it should be mentioned that while current multivariate PCMs largely follow the procedures described above, other analytical approaches for conditioning the data on the phylogeny, obtaining model parameters, and statistically evaluating evolutionary hypotheses could be envisioned. Such alternatives represent important avenues for future investigation, and are

briefly mentioned in the Discussion section. However, for the remainder of this article, we restrict our attention to those methods that have been formally described in the phylogenetic comparative literature, in an effort to determine how they perform under various conditions.

METHODS AND RESULTS

Simulation Approaches

To provide an assessment of the degree to which the different PCM approaches were capable of evaluating known patterns in multivariate datasets, we performed a series of computer simulations. For each simulation, 100 random phylogenies were generated, using both random-splits and pure-birth approaches (overall patterns from both methods were concordant, so only results from the former are shown: see Supplemental Material). On each phylogeny, multivariate phenotypic datasets were then simulated using a Brownian motion model of evolution. Datasets were simulated both with and without covariation among trait dimensions, ranging from complete trait independence ($cov_Y = 0.0$) to very high trait correlations ($cov_Y = 0.9$). Most simulations were performed using a 32 species phylogeny ($N = 32$) and eight trait dimensions ($p = 8$), resulting in a 4:1 $N:p$ ratio. However, some simulations were conducted across a broader range of species richness ($N = 32, 64, 128$) and a range of trait dimensionality ($p = 2$ - 32) to investigate their effects on PCM performance. All simulations were performed in R (R Core Team 2016), using the packages: *ape* (Paradis 2012), *geiger* (Pennell, et al. 2014), and *phytools* (Revell 2012). The multivariate PCM approaches evaluated were implemented using the packages: *geomorph* 3.0.3 (Adams, et al. 2016), *mvMORPH* (Clavel, et al. 2015), *mvSLOUCH* (Bartoszek, et al. 2012), *phylocurve* 2.0.6 (Goolsby 2015), and *SURFACE* (Ingram and Mahler 2013). Computer code and simulation results for all simulation experiments reported in this article may be found in the Supplemental Material on DRYAD (http://dx.doi.org/10.5061/dryad.29722).

As with any simulation study we acknowledge that the scenarios examined here are necessarily limited, and that their set of conditions do not represent the breadth of possible patterns displayed by empirical biological datasets. Nevertheless, valuable insights may still be obtained concerning the performance of multivariate PCMs even under these restricted conditions. Specifically, these simulations represent very simple evolutionary scenarios, with traits that evolve under a single Brownian motion model and with a specified (and common) degree of trait covariation between trait dimensions. However, if even under these conditions a particular multivariate PCM approach fails to reliably identify patterns in the data, that method has little hope of characterizing patterns under more realistic conditions, as may be found in datasets containing evolutionary outliers, displaying differing evolutionary rates among species, or that evolve under more complicated evolutionary models. Thus, the simulations implemented here may be used to establish a baseline performance for the various approaches that have thus far been proposed to evaluate patterns in multivariate datasets in light of their phylogenetic relationships.

Evaluating Multivariate PCM Patterns: Likelihood Ratio Tests and AIC Scores

One method for evaluating phylogenetic patterns in multivariate data is based on maximum likelihood. Here the fit of the multivariate data to the phylogeny may be obtained under differing evolutionary models (e.g., Brownian motion, Ornstein-Uhlenbeck, multi-rate Brownian motion models, etc.), and likelihood ratio tests or indexing measures of penalized likelihood, such as Akaike information criterion (AIC) scores, may be used to determine which alternative model provides the highest support. The procedure was originally proposed for evaluating rate shifts on the phylogeny in small numbers of univariate traits treated simultaneously (e.g., Revell and Harmon 2008; Revell and Collar 2009), but has recently been expanded to compare a wider class of evolutionary models (e.g., Bartoszek, et al. 2012).

One desirable attribute of this approach is that its summary measure (the multivariate log-likelihood, log$L$) is invariant to rotations of the multivariate dataspace. For instance, the multivariate log$L$ for the hypothetical example in Figure 1c and 1d is identical for both orientations of the dataspace (log$L$ = -20.0025). Additionally, the approach is robust to levels of covariation among trait dimensions. Figure 2a displays the correlation between log$L$ estimates obtained for a set of simulated datasets generated under conditions of increasing trait covariation, and those same datasets rotated to their principal axes. This value was 1.0 in all cases, confirming that log$L$ was both rotation-invariant and unaffected by increasing levels of trait covariation.

Nevertheless, evaluating evolutionary hypotheses in multivariate data using likelihood ratio tests or AIC scores does present some challenges. Specifically, these statistical approaches display increasing Type I error and increasing model misspecification as the number of trait dimensions (p) increases. This pattern is illustrated in Figure 2b, where the type I error rate of likelihood ratio tests increases with p (see also Fig. 2 of Adams 2014c). Likewise, using AIC comparisons, the percent model misspecification also increases with increasing trait dimensionality (Fig. 2b). The reason for this pattern is that the calculation of the likelihood (and subsequently its AIC) requires estimating both the inverse and the determinant of expected evolutionary covariance matrix (**V**), and this matrix becomes more ill-conditioned as the number of trait dimensions increases. This pattern is an embodiment of the well-known 'curse of dimensionality' (Bellman 1957) inherent in many multivariate methods (whereby adding variables increases the sparseness of dataspaces, rendering classification and prediction models insufficient for the available data). Additionally, this problem is expected to become more acute as the evolutionary model under examination becomes more complex. In such cases, **V** is likely to become increasingly ill-conditioned, leading to further computational instability. Additionally, when $p \geq N$, the likelihood of the data given the model cannot be calculated, because **V** will be

singular and its inverse cannot be computed (see Adams 2014c). Therefore, as the phenotypic dataset under investigation becomes more highly multivariate, evaluating the fit of alternative models using LRT and AIC becomes increasingly difficult. As a consequence, model evaluation using test measures such as LRT and AIC scores, when based on unstable log*L* estimates, do not provide a general analytical solution for the evaluation of phylogenetic comparative trends in high-dimensional multivariate datasets.

Evaluating Multivariate PCM Patterns: Individual-Axis Methods

Because of the challenges of evaluating evolutionary models in multivariate data, some implementations use data simplification. For instance, comparisons of the fit of the data to the phylogeny under differing evolutionary models may be based on only subset of summary axes, such as the first few principal components (e.g., Monteiro and Nogueira 2011; Monteiro 2013). However, as cogently pointed out by Uyeda, et al. (2015), this approach is positively misleading, and displays is a high degree of model misspecification. Specifically, for data generated under a BM process, the first few principal component dimensions incorrectly provide strong support for more complex OU models (see Fig. 1 in Uyeda, et al. 2015). Therefore, using only a few principal component axes in place of the full multivariate dataset for the purposes of completing the algebra is unlikely to result in meaningful macroevolutionary inferences, because such inferences will be biased towards identifying more complex evolutionary models than are actually present.

Alternatively, some approaches assume independence among trait dimensions, and for each trait dimension estimate the fit of the data to the phylogeny under differing evolutionary models separately. They then use the individual log*L* estimates across trait dimensions to obtain summary measures (Σlog*L* and AIC) for subsequent model comparisons (e.g., Ingram and Mahler 2013; Grundler and Rabosky 2014; Moen, et al. 2016). Unfortunately, these approaches are

conceptually flawed, because it is mathematically impossible for the multivariate trait dimensions to be independent under different evolutionary models simultaneously. For instance, principal component axes are uncorrelated in the phylogenetically-naïve multivariate dataspace, but *are* correlated evolutionarily, because the evolutionary covariance matrix (**R**) for principal component scores contains non-zero off-diagonal elements. Thus, summing likelihood values across dimensions will yield incorrect values (for the example in Fig. 1: $\Sigma \log L$ = - 21.16605 instead of $\log L$ = -20.0025). Furthermore, even when using phylogenetic principal component analysis, the PPCA axes are only uncorrelated under Brownian motion; for all other evolutionary models, the evolutionary rate matrix will contain non-zero correlations, meaning that summing across trait dimensions will result in incorrect $\log L$ estimates for all other evolutionary models.

The consequences of obtaining incorrect $\Sigma \log L$ and AIC estimates is that increased model misspecification can occur. This is clearly demonstrated using the simulated BM datasets above, as shown in Figure 2c. Here, 95% of the datasets simulated under Brownian motion were inferred to display two or more OU optima when using one such individual axis method (SURFACE: Ingram and Mahler 2013). This extremely high level of model misspecification demonstrates that individual axis methods do not provide a robust approach for evaluating phylogenetic comparative trends in multivariate datasets, and should be avoided.

Evaluating Multivariate PCM Patterns: Pairwise Composite Likelihood

One recent PCM approach uses a pseudolikelihood score based on pairwise composite likelihood (PCL: Goolsby 2016). Here the fit of the data to the phylogeny under both a null and an alternative model are found for pairs of trait dimensions, which are then summed across all pairs to arrive at a pseudolikelihood score for the multivariate dataset under each model. The difference in PCL scores for the two models is then calculated, and phylogenetic simulations (sensu Boettiger,

et al. 2012) are performed to obtain a distribution of possible test values to assess significance. Currently, neither the properties of PCL nor the statistical consequences of using it as a surrogate for the actual multivariate log*L* have been fully investigated.

Using the simulated datasets above, we found that the PCL score suffers from several debilitating properties. First, there is not a one-to-one correspondence between the multivariate log*L* and PCL (Fig. 3a), and as the degree of trait covariation increased, the correlation between to two decreased precipitously. Also, PCL values were not perfectly correlated when the same datasets were examined in different orientations (Fig. 3b: for additional results see Supplemental Material). Finally, as the number of trait dimensions increased, the correlation between PCL estimates obtained for the same data oriented in different directions decreased (Fig. 3c).

In addition, statistical inferences based on PCL are arbitrarily affected by levels of trait covariation and the orientation of the dataset. For instance, comparing the fit of two alternative models (BM versus OU) for the simulated datasets above revealed low levels of model misspecification in one orientation, but high levels of support for the incorrect (OU) model when data were rotated to a different direction (Fig. 4a). Further, the pattern was more acute as levels of trait covariation increased. Additionally, comparisons of two alternative rate matrix models (BM1 vs. BMM) are similarly affected. To demonstrate this we simulated multivariate datasets as above, but with two subclades that differed in their evolutionary rates and did so in a reciprocal manner (following the example in Goolsby, 2016: pg. 859). Again we found that that levels of trait covariation and the orientation of the multivariate dataspace had a large influence on statistical estimates from PCL (Fig. 4b). Finally, PCL approaches approximating phylogenetic regression suffer from similar issues. To demonstrate this we simulated data as above, but with the addition that the covariance between **Y** and **X** was set at 0.3. Again we found that tests based on PCL were

highly sensitive to the degree of covariation among trait dimensions, with a strong increase in support for the incorrect model as trait covariation increased (Fig. 4c).

Taken together, these results demonstrate that pairwise composite likelihood is not an accurate representation of the multivariate log*L* it is intended to represent, it is rotation-dependent, dimension-dependent, and is adversely affected by increasing levels of covariation among trait dimensions. In other words, PCL is sensitive to the very characteristics commonly found in the high-dimensional datasets for which it was proposed. Further, tests based on PCL are adversely affected by these undesirable properties, where differing statistical conclusions, and thus biological inferences, may be obtained for the same dataset based entirely on arbitrary input decisions made by the user. Such pathologies were observed for all PCL methods evaluated. From this it is clear that PCL-based methods yield unpredictable and uninterpretable results, and as such these approaches should be avoided in macroevolutionary studies of multivariate datasets.

Evaluating Multivariate PCM Patterns: Algebraic Generalizations

One approach to multivariate phylogenetic comparative methods is not based on estimating maximum likelihood, but instead uses test statistics based on traces of either sums of squares and cross-products matrices or covariance matrices (Adams 2014c; Adams 2014a; Adams 2014b; Adams and Felice 2014; Denton and Adams 2015). These approaches circumvent the computational issues of estimating log*L* while still retaining the components necessary for conducting statistical evaluations based on surrogates for LRT (see "Conducting Phylogenetic Comparative Analyses on Multivariate Data" section above). The methods provide summary statistics that represent algebraic extensions of test measures commonly utilized to evaluate phylogenetic patterns for univariate datasets: K_{mult} for phylogenetic signal, σ^2_{mult} for net evolutionary rates, sums of squares (*SS*) for phylogenetic regression and ANOVA models, and r_{PLS}

for the covariation between sets of variables, which are evaluated with empirically-generated probability distributions to assess statistical significance. Typically, this is accomplished using permutation procedures, where the rows (objects) of the data matrix are permuted in some fashion, and relative to the design matrix for the hypothesis under investigation. For comparisons of net evolutionary rates (σ^2_{mult}), both permutation and phylogenetic simulations were suggested (see Adams 2014c), though the latter is typically used.

One important property of these approaches is that their multivariate test statistics are rotation-invariant and insensitive to levels of covariation among trait dimensions. Using the simulated datasets above, we found a perfect correlation between summary test measures obtained from different orientations of the dataset, regardless of the degree of trait covariation. Likewise, levels of statistical inference for all permutation-based testing procedures were also identical under these conditions (Table 2). Using phylogenetic simulations to evaluate net evolutionary rates displayed slightly lower correlations as compared to using permutation methods (r = 0.93 to 0.99 versus r = 1.00), and subsequent investigations revealed that permutation tests for comparing net evolutionary rates displayed appropriate Type I error and statistical power as well (see Appendix). We therefore recommend that future empirical studies evaluating net evolutionary rates for high-dimensional datasets use permutation tests for statistical evaluation.

Algebraic Generalizations PCMs: Statistical Performance

In terms of statistical performance, prior investigations demonstrated that tests based on summary measures from algebraic extensions of PCMs displayed appropriate Type I error rates and reasonable statistical power (e.g., Adams 2014c; Adams 2014b). A recent study largely confirmed these earlier findings (Goolsby 2016), but found elevated Type I error rates for two approaches: phylogenetic partial least squares (PPLS) and phylogenetic generalized least squares

(PGLS). Because of this discrepancy we re-evaluated the Type I error of all methods using simulated datasets generated as described above. Additionally, we evaluated the Type I error of PGLS using a large set of empirically-generated chronograms from the OpenTree database (Hinchliff, et al. 2015) available on DateLife (O'Meara, et al. 2016). A total of 104 empirical chronograms containing between 32 and 512 species were used, and on each we simulated 1000 datasets at differing levels of trait dimension (p = 2, 8, 16, 32), as described above.

For virtually all simulation conditions, we found that algebraic generalizations of PCMs displayed appropriate Type I error, including phylogenetic partial least squares (Table 3). This result differed from that of Goolsby (2016), and is explained by a difference in how the permutation tests were performed. Earlier studies permuted the original trait values (following Adams and Felice 2014), which resulted in elevated Type I error rates (see Goolsby 2016). However, results reported here are based on permuting the phylogenetically-transformed data (*sensu* Garland and Ives 2000), which represent the correct exchangeable units under the null hypothesis for PPLS (demonstrated mathematically in the Appendix; for a general description of exchangeable units see: Collyer, et al. 2015). Thus, when the correct exchangeable units are permuted, PPLS does in fact display appropriate statistical properties.

For PGLS, appropriate Type I error rates were obtained when using data simulated on random-splits trees (Table 3: replicating results of Adams 2014b), but were slightly elevated when utilizing data simulated on pure-birth trees (as per results in Goolsby 2016). However, when data simulated on actual empirical phylogenies was examined, PGLS displayed stable and appropriate Type I error rates near the nominal $\alpha = 0.05$ (Table 3). Thus, through consilience one may conclude that PGLS does display appropriate Type I error, and that some statistical property of simulated pure-birth trees, and not PGLS, was responsible for the aberrant results.

Indeed, this appears to be the case. Examining the condition number obtained from phylogenetic covariance matrices revealed an increase in condition number with the number of species in the phylogeny, but this pattern was much steeper for pure-birth trees as compared to both random-splits trees and empirically-generated phylogenies (Fig. 5a). Further, because the condition number is a numerical measure of how stable a covariance matrix is under operations such as matrix inversion, larger condition numbers represent more ill-conditioned matrices, which can result in less stable estimates from down-stream algebraic operations (see Belsley, et al. 2004). As such, phylogenetic covariance matrices from pure-birth phylogenies were less stable than those obtained from empirical data or other simulation procedures, and could adversely affect PGLS computations. Indeed, we found that the condition numbers were significantly higher in those simulations displaying significant effects when using PGLS as compared to those not displaying significant effects (Fig. 5b; $F_{1,998} = 115.06$, $P < 0.0001$: $\log(\bar{k}_{sig}) = 8.41$; $\log(\bar{k}_{non-sig}) = 7.17$). This confirmed that pure-birth phylogenies displayed poor mathematical properties and were ill-conditioned for downstream analyses, resulting in the spurious Type I error rates. Additional work is needed to fully evaluate the consequences of using pure-birth phylogenies to examine other phylogenetic comparative methods in this context.

Algebraic Generalizations PCMs: Sampling Distributions of Trait Covariance Matrices

One potential concern with algebraic extensions of PCMs is that their permutation procedures do not behave as expected when compared to parametric methods. Chiefly, two criteria should be essential for these permutation procedures. First, the trait covariance matrix of a null model should be approximately constant through all permutations of the permutation procedure. Second, pertaining to linear models, the sampling distribution of trait covariance matrices (of fitted

values) for evaluated models is expected to follow a Wishart (1928) distribution (the parametric sampling distribution for data sampled from a multivariate normal distribution). Non-parametric test alternatives should produce empirical sampling distributions of covariance matrices similar to a Wishart distribution.

For the first criterion, null model covariance matrices are held constant through all permutations for analyses of K_{mult} for phylogenetic signal and for σ^2_{mult} for net evolutionary rates, as each permutation iteration randomizes the joint phenotypic-phylogenetic covariances but does not alter the resulting covariance matrices (Adams 2014a, b). Likewise, the phylogenetically transformed within-sample covariance matrices for r_{PLS} are also constant across permutations, although the cross-covariances between samples are randomized in each permutation. For PGLS, trait covariance matrices are constant across permutations for single-factor models, though with multiple covariates (beyond the scope of the current work), this pattern is more complex, owing to the type of sums of squares and cross-products (SSCP) calculated and how permutations are performed. We have performed simulations (see below) showing that using sequential SSCP (adding factors to models, sequentially) and randomized residual permutation procedures (RRPP) comes close to preserving null model covariance matrices, producing an isotropic distribution of covariance matrices centered on the observed model covariance matrix. However, further research is needed to understand the implications of SSCP choice and using full randomization of data vectors rather than RRPP.

For the second criterion, one may evaluate the sampling distribution of random covariance matrices produced by permutation relative to what is expected under a Wishart distribution, by comparing the two in a principal coordinate space defined by the Riemannian distance based on the relative eigenvalues of pairwise comparisons of covariance matrices (Mitteroecker and Bookstein 2009; see also Forstner and Moonen 1999). We posit that the empirical results from

RRPP can be compared to results from a Wishart distribution, with the expectation that the two should produce similar isotropic (spherical) scatter in the plots of principal coordinates. Using the simulation procedure above, we generated 100 datasets based on a simulated random-splits phylogeny ($N = 100$), a Brownian motion model of evolution, and a four-dimensional trait ($p = 4$), with a covariance between **Y** and a single independent variable (**X**) equal to 0.3. For each dataset 1000 iterations of a permutation procedure for PGLS were performed, from which the $p \times p$ trait covariance was obtained. Additionally, we generated the same number of random covariance matrices from a Wishart distribution, conditioned on N and p above. We obtained a measure of sphericity (eccentricity, *sensu* Turner, et al. 2010) for each sampling distribution (here, 0.0 is spherical while 1.0 is a linear trend in the covariance pattern). We found that the permutation procedure from PGLS produced isotropic sampling distributions of covariance matrices similar to, and more spherical than, those expected by sampling a Wishart distribution Fig. 6). The conclusion from this finding is that while PGLS using permutation does not utilize log$L$ estimates from sampled covariance matrices for test statistics, the method nevertheless retains appropriate sampling distributions of the covariance matrices produced by RRPP.

DISCUSSION

The State of Multivariate Phylogenetic Comparative Methods

The question posed in the introduction of this article was: How should phylogenetic comparative analyses of multivariate data be performed? Recent years have seen increased interest in the analysis of multivariate datasets in a phylogenetic context, and numerous approaches have been proposed to evaluate phylogenetic hypotheses in multivariate datasets. However, to date no study has compared the ability of these approaches to reliably assess patterns of evolutionary dispersion in such multivariate dataspaces. Here we provide the first comparative analysis of

existing multivariate phylogenetic comparative methods, examining not only their ability to make reliable statistical inferences, but also their adherence to the geometric properties required of any multivariate method. From these perspectives we found widely varying performance across the proposed approaches. As such, the answer to how one should conduct multivariate PCMs depends upon the evolutionary hypothesis one wishes to consider (see Table 1).

First, if one is interested in characterizing the degree of phylogenetic signal in multivariate datasets, this may be accomplished effectively using K_{mult} (Adams 2014a): the algebraic generalization of *Kappa* (Blomberg, et al. 2003). The approach is invariant to rotations of the multivariate dataspace, and is robust to levels of trait covariation and the number of trait dimensions (Table 1). Further, statistical tests based on this measure display appropriate Type I error (Table 2) and high statistical power (shown previously). Finally, as with *Kappa*, this measure provides a constant expected value under Brownian motion ($K_{mult} = 1.0$) against which the relative degree of phylogenetic signal may be described. Thus, researchers interested in the degree of phylogenetic signal in multivariate datasets have an appropriate tool for such investigations.

Second, for macroevolutionary hypotheses that evaluate the degree of evolutionary covariation between dependent and independent variables, the multivariate equivalents of phylogenetic regression (PGLS) and evolutionary correlation methods can be used. Specifically, such hypotheses may be examined properly using algebraic generalizations of PGLS and PPLS (Adams 2014b; Adams and Felice 2014; Adams and Collyer 2015). As with K_{mult}, these methods are rotation-invariant, are robust to differing levels of trait covariation, and are robust to the number of trait dimensions (Table 2). Further, when appropriate permutation procedures are utilized, statistical tests based on these approaches display appropriate Type I error rates (Table 3) and statistical power (shown previously). Their implementations are also flexible, as the multivariate PGLS approach is capable of performing phylogenetic ANOVA, phylogenetic

regression, and phylogenetic factorial models. Thus, with these approaches a considerable number of evolutionary hypotheses may be reliably examined in multivariate data and in a phylogenetic context. By contrast, the alternative procedure proposed for evaluating these hypotheses (PCL: Goolsby 2016) was shown to be sensitive to trait covariation and dataspace orientation (Fig. 3). The consequence of these deficiencies is that PCL can arrive at different statistical conclusions for the same dataset (Fig. 4). Therefore, we recommend that PCL should not be used to investigate the degree of evolutionary covariation between traits, and that instead algebraic extensions of PGLS and PPLS be utilized for this purpose.

Unfortunately, with respect to comparing alternative evolutionary models for describing patterns of trait evolution in multivariate datasets (e.g., BM versus OU), the situation is not so positive. First, simplifying the multivariate dataspace to a single summary axis is not a solution, as the first few principal component axes display a bias towards more complex evolutionary models, even when the data were generated under Brownian motion (see Uyeda, et al. 2015). Thus, analyses comparing evolutionary models based on the first, or even the first few principal components (*sensu* Monteiro and Nogueira 2011) are likely to provide incorrect support for OU or early burst models, thereby yielding unreliable results. Likewise, methods that assume independence across trait dimensions also do not provide a solution. These methods display extreme levels of model over-fitting and model misspecification. In the example shown here, the *SURFACE* method (Ingram and Mahler 2013) inferred multiple phenotypic optima in over 95% of the Brownian motion datasets examined (Fig. 2), implying that complex OU models were incorrectly preferred over the correct BM model. Thus, this approach, and others that make the same assumption of independence (e.g., Grundler and Rabosky 2014; Moen, et al. 2016), should not be used for macroevolutionary inference. Additionally, comparisons of evolutionary models using pairwise composite likelihood (Goolsby 2016) also do not yield meaningful biological

inferences. As shown above, PCL is sensitive to trait covariation and dataspace orientation (Fig. 3), and comparisons between evolutionary models (e.g,. BM vs. OU) can arrive at different statistical conclusions for the same dataset (Fig. 4). Thus, the method is unreliable, and results based on PCL depend almost entirely on arbitrary decisions of the user. Based on these findings, we recommend that single axis methods, methods that summarize across trait dimensions (e.g., *SURFACE*), and methods based on pairwise composite likelihood - all methods that are surrogates for estimating log$L$ for LRTs - should be avoided in future macroevolutionary studies.

On the other hand, multivariate phylogenetic comparative methods based on log-likelihood (when estimable) are rotation-invariant, and are robust to levels of trait covariation. Further, a previous study showed that for a small number of traits (p = 4) and a large number of taxa (N = 100), comparisons of evolutionary rate models (using LRT) display only slightly elevated Type I error rates (Revell and Harmon 2008). However, as the number of variables (p) increases, the Type I error rate of these procedures also increases (Adams 2014c; this study). Additionally, with only a moderate number of trait dimensions, LRT and AIC-based approaches were shown to display high levels of model misspecification that can exceed 50% (Fig. 2). As discussed above, the reason is that calculation of the likelihood requires estimating both the inverse and the determinant of expected evolutionary covariance matrix (**V**), and this matrix becomes more ill-conditioned as the number of trait dimensions increases, and as the evolutionary model under examination becomes more complex. Thus, while these methods are fully multivariate, they are only reliable when there is a large ratio of species to variables (i.e., a high $N{:}p$ ratio). How large the $N{:}p$ ratio must be to maintain acceptable levels model misspecification will depend upon the complexity of the models being compared, and is a question that requires further investigation.

In fact, the only current approach that provides a robust means of comparing evolutionary models for multivariate datasets are algebraic extensions of univariate methods for comparing net

evolutionary rates between groups of taxa or sets of traits under Brownian motion (Adams 2014c; Denton and Adams 2015). As above, these methods are rotation-invariant, are robust to levels of trait covariation, and display appropriate Type I error and statistical power (e.g., Table 3). We appreciate that comparisons of net evolutionary rates under BM represents a very restricted set of the possible evolutionary models of interest to macroevolutionary biologists; particularly when compared to the panoply of models that may be evaluated with univariate datasets. Nonetheless, the results of our investigation lead us to the conclusion that all other currently-available methods for multivariate evolutionary model comparison fail to display appropriate properties that facilitate such analyses for high-dimensional datasets and to make reliable inferences. We also fully recognize that this conclusion is rather disappointing, particularly because of the intense interest in evaluating multivariate trends relative to alternative evolutionary models that may have generated those patterns. However, while this may be seen as a macroevolutionary "inconvenient truth", it is nevertheless a conclusion supported by the evidence. As such we echo the plea of Uyeda, et al. (2015) albeit in modified form: "These results highlight the need for truly multivariate phylogenetic comparative methods [for the comparison of evolutionary models]." (Uyeda, et al. 2015; pg. 677).

Conclusions and Prospectus

So what is the prospectus for the future, and how might comparisons of fully multivariate models for distinct evolutionary scenarios (e.g., BM vs. OU) be accomplished? While we do not provide a full analytical solution to this dilemma in this article, our investigation provides essential insight on the properties that future multivariate phylogenetic comparative methods must display. First and foremost, any new multivariate method for macroevolutionary inference must adhere to the geometric properties of multivariate dataspaces: they must be robust to differing levels of trait

covariation, and must be rotation-invariant. Any newly proposed method whose inferences differ with increasing levels of trait covariation is not up to the task (see Fig. 4), and any method that is rotation-dependent will result in arbitrary outcomes (e.g., PCL). Next, once these geometric properties are satisfied, any new approach must have appropriate statistical properties; namely Type I error and power. Third, a truly multivariate approach should be robust when implemented on highly multivariate datasets; otherwise the approach will be restricted to a small number of trait dimensions, and will not provide a solution for highly dimensional multivariate datasets. Finally, we urge researchers proposing potential approaches to thoroughly investigate all of these properties of their new methods, as all are crucially important in determining whether new procedures are robust analytical alternatives that move the field towards a fully multivariate solution.

In considering the varied approaches for performing multivariate phylogenetic comparative analyses, several avenues forward may be envisioned to alleviate the challenges our study has identified. First, future research could focus on alternative methods of model estimation. That is, one could envision other approaches for conditioning patterns of trait covariation on the phylogeny, and from this obtaining a sampling distribution of possible covariance matrices conditioned on the phylogenetic non-independence among taxa for statistical evaluation. This would represent an important direction of future research. Second, one could focus on model evaluation by envisioning alternative approaches that avert the computational problems associated with ill-conditioned covariance matrices. For instance, if likelihood ratio tests or other testing procedures can avoid inverting ill-conditioned **V** matrices, a reliance on traces of covariance matrices could avert computational problems. In this vein we suggest it would be fruitful to reconsider how LRT statistics that compare different evolutionary models are estimated, rather

than reconsidering the log-likelihoods that comprise them. A possible solution could target finding stable forms of **V** matrices via eigen-analysis.

To consider this option, we suggest that equation 3 could be rewritten for a putative estimate model covariance matrix ($\widehat{\mathbf{V}}$) and null model covariance matrix ($\mathbf{V_0}$) as:

$$\left(\frac{L(\widehat{\mathbf{V}}|\mathbf{X}_0)}{L(\mathbf{V_0}|\mathbf{X}_0)}\right) = -\frac{1}{2}\Big[\Big(\big(\mathbf{y} - E(\mathbf{y}|\widehat{\mathbf{V}}, \mathbf{X}_0)\big)^{\boldsymbol{t}}\, \widehat{\mathbf{V}}^{-\mathbf{1}} \big(\mathbf{y} - E(\mathbf{y}|\widehat{\mathbf{V}}, \mathbf{X}_0)\big) + log|\widehat{\mathbf{V}}|\Big)$$

$$- \Big(\big(\mathbf{y} - E(\mathbf{y}|\mathbf{V_0}, \mathbf{X}_0)\big)^{\boldsymbol{t}} \mathbf{V}^{-\mathbf{1}} \big(\mathbf{y} - E(\mathbf{y}|\mathbf{V_0}, \mathbf{X}_0)\big) + log|\mathbf{V_0}|\Big)\Big]$$

$$= \frac{1}{2}\Big[\big(\mathbf{y} - E(\mathbf{y}|\mathbf{V_0}, \mathbf{X}_0)\big)^{\boldsymbol{t}} \mathbf{V_0}^{-\mathbf{1}} \big(\mathbf{y} - E(\mathbf{y}|\mathbf{V_0}, \mathbf{X}_0)\big) - \big(\mathbf{y} - E(\mathbf{y}|\widehat{\mathbf{V}}, \mathbf{X}_0)\big)^{\boldsymbol{t}} \widehat{\mathbf{V}}^{-\mathbf{1}} \big(\mathbf{y} - E(\mathbf{y}|\widehat{\mathbf{V}}, \mathbf{X}_0)\big)\Big] +$$

$$\frac{p}{2} log\left(\frac{trace(\mathbf{V_0})}{trace(\widehat{\mathbf{V}})}\right). \qquad 7$$

The latter component of this equation uses traces of the evolutionary covariance matrices (**V**), and takes advantage of the inequality of arithmetic and geometric means; i.e., $\frac{1}{p} trace(\mathbf{V}) \geq |\mathbf{V}|^{1/p}$ (this may be utilized if and only if **V** is symmetric, and thus, the trace is the same as the sum of eigen values found from eigen-analysis of **V**). Additionally, recognizing that if one performed eigen-analysis on each of the **V** matrices, the sum of positive eigen-values could replace the trace of the original matrices. Thus the latter part could be rewritten as: $\frac{p}{2} log\left(\frac{\sum_{i=1}^{k_0} \lambda_i}{\sum_{j=1}^{k_1} \lambda_j}\right)$, where k_0 and k_1 refer to the ranks of $\mathbf{V_0}$ and $\widehat{\mathbf{V}}$, respectively. Furthermore, the phylogenetic residuals, $\mathbf{y} - E(\mathbf{y}|\mathbf{V}, \mathbf{X}_0)$, can be estimated by utilizing a phylogenetic projection matrix to avoid matrix inversion (Garland and Ives 2000; Adams 2014b) in the estimation of **V**; i.e., $\mathbf{y} - E(\mathbf{y}|\mathbf{V}, \mathbf{X}_0) = vec[\mathbf{Y} - E(\mathbf{PY}|\mathbf{PX}_0)]$. These residuals can then be projected onto the eigen-vectors of **V** to solve the former part of equation 4, substituting **V** with the $k \times k$ diagonal matrix of positive eigen-

values, **Λ**, in each case. This approach would achieve comparing evolutionary models in appropriately dimensioned subspaces of covariance matrices, rotated to their major axes of covariance. However, it should be recognized that the LRT statistic would summarize both scale and rotational differences of **R** matrices, and further theoretical development would be needed to decompose these attributes (*sensu* Revell and Harmon 2008). From our perspective the development of model evaluation procedures that are robust to ill-conditioned covariance matrices (such as **V**) represents an important avenue for future consideration.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.29722.

Funding

This work was sponsored in part by National Science Foundation grants DEB-1556379 (to DCA) and DEB-1737895 (to MLC).

Acknowledgements

We thank E. Baken, B. Juarez, E. Sherratt, and N. Valenzuela for comments on drafts of this manuscript. The comments of E. Goolsby, N. MacLeod, D. Polly, and two anonymous reviewers greatly improved this work.

Appendix

1. Demonstration that phylogenetically-transformed data are the correct exchangeable units for phylogenetic partial least squares: PPLS

Identifying the correct exchangeable units under the null hypothesis is essential for any permutation procedure (Anderson and Braak 2003). As shown by Adams and Collyer (2015), choosing the incorrect exchangeable units can have dire consequences, such as inflating Type I error rates. For PCMs, data transformations are often used as an analytical step, which can make permutation procedures challenging. For example, OLS models can be represented as $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$; where $\mathbf{Y}$ is a species (N) × trait (p) matrix of phenotypic values, $\mathbf{X}$ is an $N \times k$ design matrix for the k linear model parameters, $\hat{\boldsymbol{\beta}}$ is a $k \times p$ matrix of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $N \times p$ matrix of residuals. The row vectors of $\boldsymbol{\varepsilon}$ from a null model are the exchangeable units under the null hypothesis for a model whose design contains the same parameters of $\mathbf{X}$, plus additional parameters for the effect that is tested. The method of D-PGLS (Adams 2014b) involves calculating an $N \times N$ phylogenetic transformation matrix, $\mathbf{P}$, to facilitate OLS estimation of parameters during permutation procedures, but the exchangeable units are unchanged; i.e., $\mathbf{P(Y)} = \mathbf{P}(\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon})$. As long as the phylogenetic transformation is applied to every random permutation of $\boldsymbol{\varepsilon}$, D-PGLS has appropriate Type I error rates (Adams and Collyer 2015). Transforming the data once – i.e., obtaining $\mathbf{PY}$, $\mathbf{PX}\hat{\boldsymbol{\beta}}$, and $\mathbf{PE}$ – followed by randomizing either $\mathbf{PE}$ or $\mathbf{PY}$ (the latter often performed with "full" randomization of data), fails to randomize exchangeable units under the null hypothesis and influences statistical errors for PGLS models. That is to say, it inherently randomizes the phylogenetic covariances among species, in addition to the model error (see Adams and Collyer 2015 for further details.)

However, while PPLS displays some apparent similarities to D-PGLS, this similarity could

inadvertently obscure proper detection of exchangeable units for null hypothesis testing. Concerning two-block partial least squares (PLS) analysis (Rohlf and Corti 2000), the correlation between two (centered) matrices with N rows in the same order, $\mathbf{Y_1}$, and $\mathbf{Y_2}$, with p_1 and p_2 phenotypic traits, respectively, is calculated to measure the level of phenotypic integration between data sets. (Note, these matrices must be "centered" by subtracting trait means.) Singular value decomposition (SVD) on the $p_1 \times p_2$ cross-traits covariance matrix, calculated as $N^{-1}\mathbf{Y_1^t Y_2}$, where the superscript, ^{t}, means matrix transposition, along with projection of each data set onto corresponding singular vectors, is used to calculate Pearson product-moment correlations as measures of integration. The null hypothesis is that the correlation is asymptotically 0, although the number of species and traits influences the expected value under the null hypothesis (Adams and Collyer 2016). The permutation procedure for testing integration randomizes the row vectors of either $\mathbf{Y_1}$ or $\mathbf{Y_2}$ in each random permutation, in order to calculate random versions of the correlation coefficients. This procedure has the property that $N^{-1}(\mathbf{Y_1^*})^{\boldsymbol{t}}\ (\mathbf{Y_1^*}) = N^{-1}\mathbf{Y_1^t Y_1}$ in every random permutation, where $\mathbf{Y_1^*}$ is a randomized version of $\mathbf{Y_1}$. Thus, the within-set covariance matrix remains constant through every random permutation, suggesting – as with D-PGLS – that this procedure exchanges the correct units under the null hypothesis. (Note, $N^{-1}\mathbf{E^t E}$ is the trait by trait error covariance matrix in D-PGLS, which is constant in every random permutation. If the model design only contains an intercept, this is the same as $N^{-1}\mathbf{Y_1^t Y_1}$.)

In order to account for phylogenetic relatedness in PPLS, the same phylogenetic transformation matrix for D-PGLS is used on both $\mathbf{Y_1}$, and $\mathbf{Y_2}$, prior to performing SVD; i.e., the cross-traits covariance matrix is calculated as $N^{-1}(\mathbf{PY_1})^{\mathbf{t}}(\mathbf{PY_2})$**.** At first glance, it might seem appropriate to randomize either $\mathbf{Y_1}$ and $\mathbf{Y_2}$ in every random permutation, prior to transformation. However, shuffling the row vectors of either $\mathbf{Y_1}$ or $\mathbf{Y_2}$ does not preserve within-set covariance matrices; *i.e.*,

$N^{-1}(\mathbf{PY_1^*})^t(\mathbf{PY_1^*}) \neq N^{-1}\left(\mathbf{PY_1}\right)^t\left(\mathbf{PY_1}\right)$. Rather, performing the transformation once and randomizing the transformed values results in a constant within-set trait covariance matrix in every random permutation; *i.e.*, $N^{-1}((\mathbf{PY_1})^*)^t((\mathbf{PY_1})^*) = N^{-1}\left(\mathbf{PY_1}\right)^t\left(\mathbf{PY_1}\right)$.

It is important to realize here that the null hypothesis test targets the covariances between data sets rather than the difference in model parameters, as in D-PGLS. The phylogenetic transformation is merely a method to adjust values prior to measuring covariation; the correct exchangeable units maintain this transformation. This subtlety was not appreciated by Adams and Felice (2014), prior to the discussion of appropriate exchangeable units by Adams and Collyer (2015), and lead to the elevated type I error rates reported by Goolsby (2016). However, when the correct exchangeable units are utilized, the approach does in fact have appropriate type I error rates.

2. Evaluation of statistical properties of permutation tests for evaluating net evolutionary rates

As described in the text, comparisons of net evolutionary rates are typically accomplished via phylogenetic simulation, where evolutionary rate matrices for the set of traits is used as an input covariance matrix for generating sets of data under those conditions (Adams 2014c; Denton and Adams 2015). However, Adams (2014c) also mentioned that permutation procedures are commonly utilized to assess phylogenetic patterns in data. Here we evaluate the statistical properties of this new procedure.

Simulation protocol: First, for each simulation run, 1000 random-splits phylogenies containing 32 species each were generated, and taxa were divided equally into two groups. Multivariate data were then simulated on each phylogeny using a Brownian motion model of evolution. Trait dimensionality was varied across simulation runs (p = 2,8,16, 32). For Type I error simulations, a

single input covariance matrix was used, where the diagonal elements were set to 1.0 for all trait dimensions, and the covariation among trait dimensions was set to one of three values depending on simulation conditions (Y_{cov} = 0.0, 0.5, 0.9). For power simulations, the input covariance matrix for the first group was set as described above, but for the second group the diagonal elements of the input covariance matrix were set to: 2.0 or 4.0. Tests comparing one-rate and two-rate models were then performed using these datasets, as well as the datasets rotated to their principal axes. Simulations were performed in R using the packages *geomorph, geiger,* and *phylocurve*.

Results. Simulations revealed that the method attained appropriate Type I error rates at the nominal value of $\alpha = 0.05$ (Figure A1). This was consistent across a range of trait dimensionality as well as the degree of covariation among trait dimensions. Additionally, power increased as the true difference in net evolutionary rates increased, and this pattern was more acute for greater numbers of trait dimensions (Figure A1). Finally, results were identical when datasets were rotated to their principal axes, demonstrating that the permutation procedure is rotation-invariant. Overall these patterns confirm that permutation-based approaches for comparing net evolutionary rates display appropriate statistical properties across a wide range of conditions.

References

Adams DC. 2013. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. Syst. Biol. 62:181-192.

Adams DC. 2014a. A generalized Kappa statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. Syst. Biol. 63:685-697.

Adams DC. 2014b. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution 68:2675-2688.

Adams DC. 2014c. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. Syst. Biol. 63:166-177.

Adams DC, Collyer M, Sherratt E. 2016. geomorph 3.0.3: Software for geometric morphometric analyses. R package version 3.0.3. http://CRAN.R-project.org/package=geomorph.

Adams DC, Collyer ML. 2015. Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: what you shuffle matters. Evolution 69:823-829.

Adams DC, Collyer ML. 2016. On the comparison of the strength of morphological integration across morphometric datasets. Evolution 70:2623-2631.

Adams DC, Felice RN. 2014. Assessing trait covariation and morphological integration on phylogenies using evolutionary covariance matrices. PLoS ONE 9:e94335.

Adams DC, Rohlf FJ, Slice DE. 2013. A field comes of age: Geometric morphometrics in the 21st century. Hystrix 24:7-14.

Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26:32-46.

Anderson MJ, Braak CJFt. 2003. Permutation tests for multi-factorial analysis of variance. J. Stat. Comput. Simul. 73:85-113.

Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF. 2012. A phylogenetic comparative method for studying multivariate adaptation. J. Theor. Biol. 314:204-215.

Bastir M, Rosas A, Stringer CB, Cuétara JM, Kruszynski R, Weber GW, Ross CF, Ravosa MJ. 2010. Effects of brain and facial size on basicranial form in human and primate evolution. J. Hum. Evol. 58:424–431.

Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution 66:2369-2383.

Bellman R. 1957. Dynamic Programming. Princeton, NJ. USA, Princeton University Press.

Belsley DA, Kuh E, Welsch RE. 2004. Regression diagnostics: Identifying influential data and sources of collinearity. Hoboken, New Jersey, John Wiley and Sons.

Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717-745.

Blomberg SP, Lefevre JG, Wells JA, Waterhouse M. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol. 61:382-391.

Boettiger C, Coop G, Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 67:2240-2251.

Bookstein FL. 1991. Morphometric tools for landmark data: geometry and biology. Cambridge, Cambridge University Press.

Clavel J, Escarguel G, Merceron G. 2015. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311-1319.

Collyer ML, Sekora DJ, Adams DC. 2015. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. Heredity 115:357-365.

Denton JSS, Adams DC. 2015. A new phylogenetic test for comparing multiple high-dimensional evolutionary rates suggests interplay of evolutionary rates and modularity in lanternfishes (Myctophiformes; Myctophidae). Evolution 69:2425-2440.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Gen. 25:471-492.

Felsenstein J. 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution 35:1229-1242.

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

Forstner W, Moonen B. 1999. A Metric for Covariance Matrices. In: Krumm F, Schwarze VS editors. Quo vadis geode sia... ?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday. Stuttgart, Stuttgart University Press, p. 113-128.

Garland T, Jr., Dickerman AW, Janis CM, Jones JA. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 43:265-292.

Garland TJ. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140:2104-2111.

Garland TJ, Ives AR. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346-364.

Goolsby EW. 2015. Phylogenetic comparative methods for evaluating the evolutionary history of function-valued traits. Syst. Biol. 64:568–578.

Goolsby EW. 2016. Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of "distance-based" methods. Syst. Biol. 65:852-870.

Grafen A. 1989. The phylogenetic regression. Phil. Trans. Roy. Soc. London B. 326:119-157.

Grundler MC, Rabosky DL. 2014. Trophic divergence despite morphological convergence in a continental radiation of snakes. Proc. Roy. Soc. B. 281:20140413.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341-1351.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, *et al.* 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. U.S.A. 112:12764–12769.

Ingram T, Mahler DL. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol. Evol. 4:416-425.

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

Langsrud O. 2004. The geometrical interpretation of statistical tests in multivariate linear regression. Stat. Papers 5:111-122.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate Analysis. London, Academic Press.

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646-667.

Mitteroecker P, Bookstein FL. 2009. The ontogenetic trajectory of the phenotypic covariance matrix, with examples from craniofacial shape in rats and humans. Evolution 63:727-737.

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

Moen DS, Morlon H, Wiens JJ. 2016. Testing convergence versus history: Convergence dominates phenotypic evolution for over 150 million years in frogs. Syst. Biol. 65:146-160.

Monteiro LR. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. Hystrix 24:25-32.

Monteiro LR, Nogueira MR. 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. BMC Evol. Biol. 11:1-23.

O'Meara BC, C.Ane, Sanderson MJ, Wainwright PC. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922-933.

O'Meara BC, Eastman J, Heath T, Wright A, Schliep K, Chamberlain S, Midford P, Harmon L, Brown J, Pennell M, *et al.* 2016. DateLife. Version 0.23. 10.5281/zenodo.56803.

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

Pagel MD. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Paradis E. 2012. Analyses of Phylogenetics and Evolution with R. 2nd ed. New York, Springer

Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. 2014. Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 15:2216–2218.

Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. New York Acad. Sci. 1289:90-105.

Polly DP, Lawing AM, Fabre A, Goswami A. 2013. Phylogenetic principal components analysis and geometric morphometrics. Hystrix 24:33-41.

R Core Team. 2016. R: a language and environment for statistical computing. Version 3.3.1. http://cran.R-project.org. R Foundation for Statistical Computing, Vienna.

Revell LJ. 2012. Phytools: An R package for phylogenetic comparative biology (and other things) Methods Ecol. Evol. 3:217-223.

Revell LJ, Collar DC. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. Evolution 63:1090-1100.

Revell LJ, Harmon LJ. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res. 10:311-331.

Rohlf FJ. 1999. Shape statistics: Procrustes superimpositions and tangent spaces. J. Classific. 16:197-223.

Rohlf FJ. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143-2160.

Rohlf FJ. 2006. A comment on phylogenetic correction. Evolution 60:1509-1515.

Rohlf FJ, Corti M. 2000. The use of partial least-squares to study covariation in shape. Syst. Biol. 49:740-753.

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

Sherratt E, Alejandrino A, Kraemer AC, Serb JM, Adams DC. 2016. Trends in the sand: directional evolution in the shell shape of recessing scallops (Bivalvia: Pectinidae). Evolution 70:2061-2073.

Sherratt E, Gower DJ, Klingenberg CP, Wilkinson M. 2014. Evolution of cranial shape in caecilians (Amphibia: Gymnophiona). Evol. Biol. 41:528-545.

Sidlauskas B. 2008. Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. Evolution 62:3135-3156.

Thomas GH, Freckleton RP, Székely. T. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc. Roy. Soc. B. 273:1619-1624.

Turner TF, Collyer ML, Krabbenhoft TJ. 2010. A general hypothesis□testing framework for stable isotope ratios in ecological studies. Ecology 91:2227-2233.

Uyeda JC, Caetano DS, Pennell MW. 2015. Comparative analysis of principal components can be misleading. Syst. Biol. 64:677-689.

Wishart J. 1928. The generalised product moment distribution in samples from a normal multivariate population. Biometrika 20:32–52.

Table 1. Summarization of the efficacy of current phylogenetic comparative methods for evaluating macroevolutionary patterns in highly multivariate datasets. Methods not applicable for a particular hypothesis are designated as '—'. The different approaches are abbreviated as follows: multivariate log-likelihood ($\log L_{Mult}$), multivariate log-likelihood from subset of trait dimensions ($\log L_{Subset}$), summation of log-likelihood across dimensions ($\Sigma \log L$), pairwise composite likelihood (PCL), and multivariate generalizations of the algebra of univariate PCMs (MultG).

Analysis Type	$\log L_{Mult}$	$\log L_{Subset}$	$\Sigma \log L$	PCL	MultG
Phylogenetic Signal	—	—	—	—	Yes
Phylogenetic ANOVA	—	—	—	—	Yes
Phylogenetic Regression	—	—	—	No (1-3)	Yes
Phylogenetic Covariation (blocks of variables)	—	—	—	No (1-3)	Yes
Comparing Evolutionary Models: BM vs BM_{Mult}	Limited (6)	No (5)	No (4)	No (1-3)	Limited (7)
Comparing Evolutionary Models: BM vs OU, etc.	Limited (6)	No (5)	No (4)	No (1-3)	—

1. Method is orientation-dependent: high model misspecification.
2. Method is covariation-dependent.
3. Method is dependent on number of variables.
4. Method incorrectly assumes trait independence: high model misspecification.
5. Method has high model misspecification.
6. Method limited to a small numbers of traits: has high model misspecification otherwise.
7. Method limited to net evolutionary rate comparisons only

Table 2. Results from statistical simulations for datasets generated on 32 species phylogenies using differing levels of covariation among trait dimensions (Y_{cov}). The table displays the correlation between summary test measures obtained for the 100 datasets in each of two orientations, followed by the correlation between significance levels of tests based on each approach.

	$Y_{cov} = 0.0$	$Y_{cov} = 0.5$	$Y_{cov} = 0.9$			$Y_{cov} = 0.0$	$Y_{cov} = 0.5$	$Y_{cov} = 0.9$
K_{mult}	1.00	1.00	1.00		P_{perm}	1.00	1.00	1.00
σ^2_{mult}	1.00	1.00	1.00		P_{sim}	0.99	0.98	0.93
					P_{perm}	1.00	1.00	1.00
SS_{PGLS}	1.00	1.00	1.00		P_{perm}	1.00	1.00	1.00
r_{PLS}	1.00	1.00	1.00		P_{perm}	1.00	1.00	1.00

Table 3. Results from simulations evaluating the Type I error of PCMs based on algebraic generalizations to a multivariate context. For PGLS, type I error was also evaluated on data simulated on 104 empirically-generated chronograms.

	Random-Splits Trees					Pure-Birth Trees			
σ^2_{mult}	$p = 2$	$p = 8$	$p = 16$	$p = 32$	σ^2_{mult}	$p = 2$	$p = 8$	$p = 16$	$p = 32$
$Y_{cov} = 0.0$	0.046	0.03	0.016	0.004	$Y_{cov} = 0.0$	0.038	0.023	0.012	0.014
$Y_{cov} = 0.5$	0.041	0.028	0.045	0.029	$Y_{cov} = 0.5$	0.049	0.045	0.059	0.045
$Y_{cov} = 0.9$	0.049	0.05	0.048	0.045	$Y_{cov} = 0.9$	0.05	0.036	0.048	0.053
K_{mult}	$p = 2$	$p = 8$	$p = 16$	$p = 32$	K_{mult}	$p = 2$	$p = 8$	$p = 16$	$p = 32$
$Y_{cov} = 0.0$	0.056	0.048	0.049	0.044	$Y_{cov} = 0.0$	0.055	0.044	0.056	0.05
$Y_{cov} = 0.5$	0.053	0.048	0.051	0.045	$Y_{cov} = 0.5$	0.05	0.057	0.046	0.036
$Y_{cov} = 0.9$	0.043	0.046	0.05	0.05	$Y_{cov} = 0.9$	0.048	0.057	0.062	0.061
PGLS	$p = 2$	$p = 8$	$p = 16$	$p = 32$	PGLS	$p = 2$	$p = 8$	$p = 16$	$p = 32$
$Y_{cov} = 00$	0.05	0.029	0.019	0.004	$Y_{cov} = 0.0$	0.087	0.161	0.149	0.186
$Y_{cov} = 0.5$	0.066	0.042	0.047	0.043	$Y_{cov} = 0.5$	0.102	0.097	0.131	0.129
$Y_{cov} = 0.9$	0.046	0.058	0.057	0.047	$Y_{cov} = 0.9$	0.103	0.116	0.099	0.092
Empirical Trees	0.058	0.061	0.059	0.054					
PPLS	$P = 2$	$P = 8$	$P = 16$	$P = 32$	PPLS	$P = 2$	$P = 8$	$P = 16$	$P = 32$
$Y_{cov} = 0.0$	0.049	0.054	0.037	0.043	$Y_{cov} = 0.0$	0.042	0.048	0.042	0.043

$Y_{cov} = 0.5$	0.042	0.05	0.044	0.048		$Y_{cov} = 0.5$	0.06	0.05	0.045	0.045
$Y_{cov} = 0.9$	0.054	0.041	0.051	0.047		$Y_{cov} = 0.9$	0.051	0.062	0.056	0.049

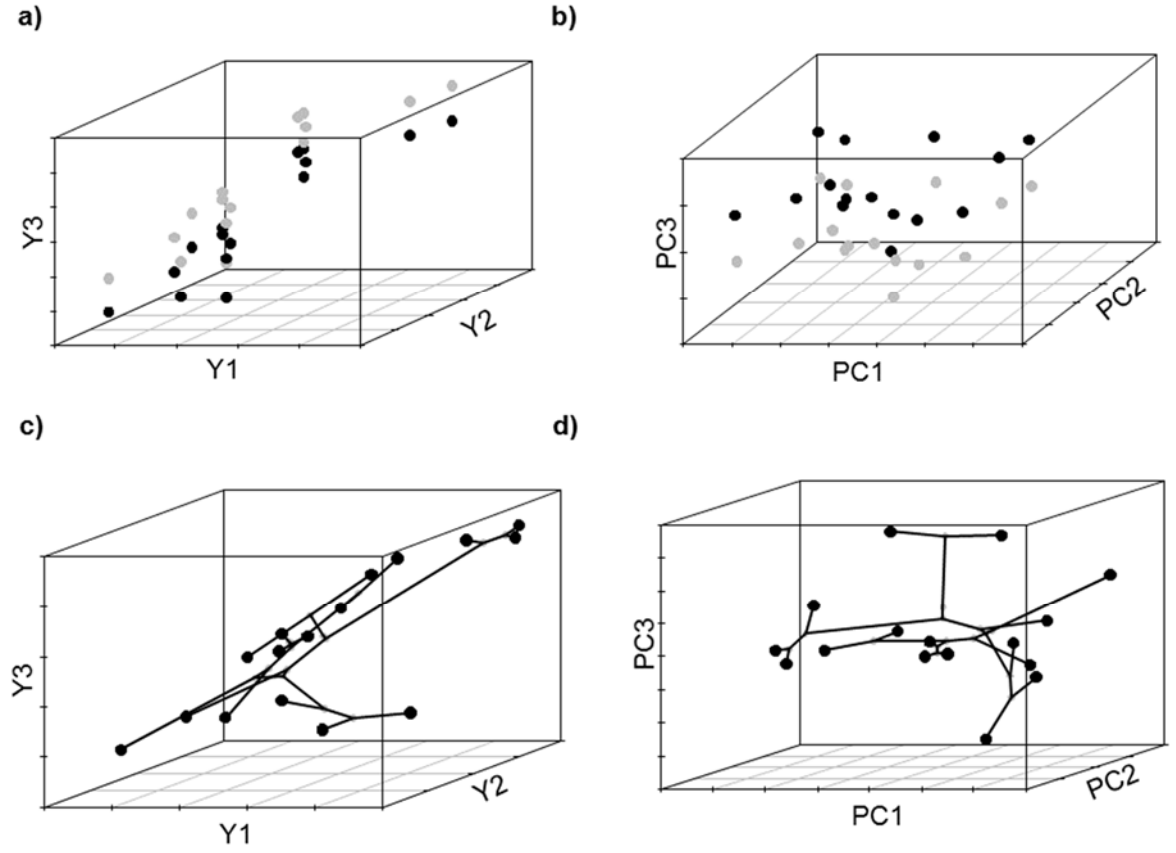

Figure 1. Graphical representation of various three-dimensional phenotype spaces. a) Plot of the means of 30 hypothetical species in a three-dimensional space, where 15 inhabit islands (dark symbols) and 15 inhabit continental locations (light symbols). b) The same three-dimensional dataspace rotated to its principal axes. c) Phylomorphospace of 16 hypothetical species for three-dimensional data with their phylogeny superimposed. d) The same phylomorphospace rotated to its principal axes.

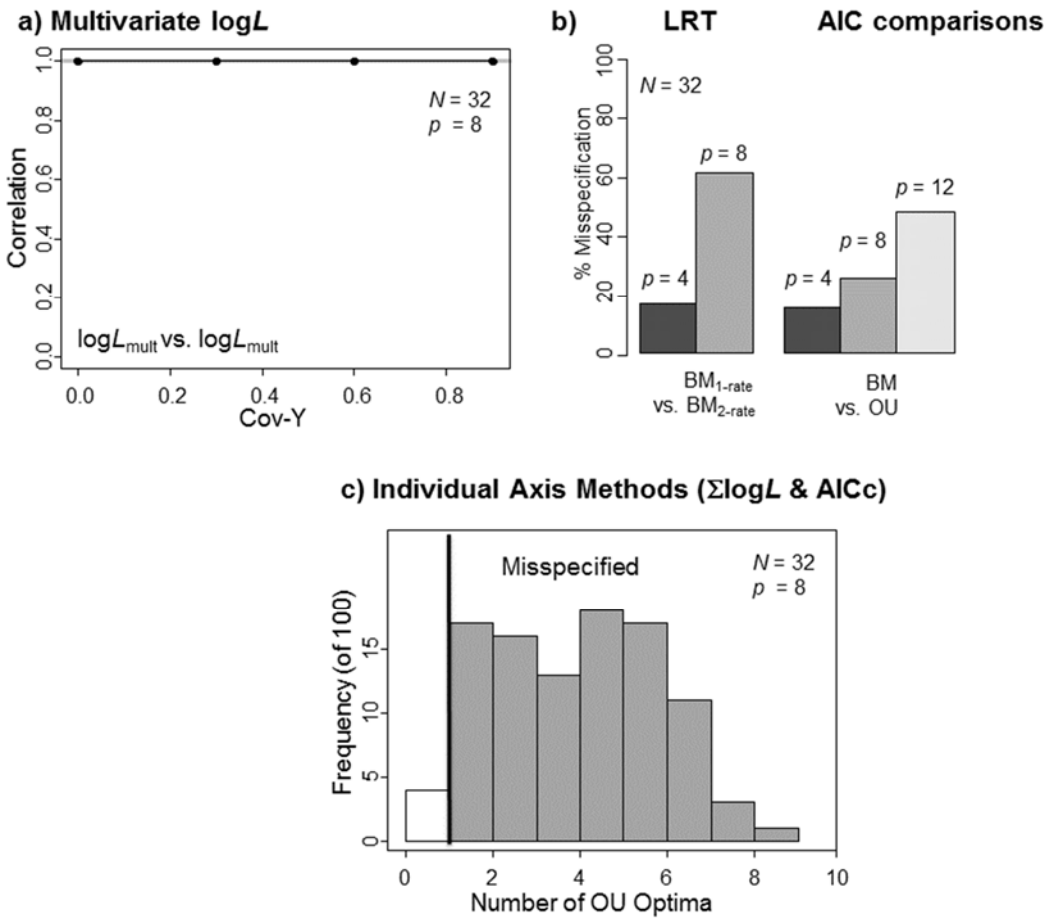

Figure 2. Results from statistical simulations evaluating a) the correlation between multivariate log*L* and multivariate log*L* for the same simulated datasets rotated to a different orientation. b) Percent model misspecification based on comparisons of evolutionary models, where data were simulated under a single-rate Brownian motion model. Comparisons of BM1 versus BMM were evaluated using likelihood ratio tests, while comparisons of BM versus OU models were accomplished using AIC. Results obtained using the *mvMORPH* package (see Supplemental Material for results using *mvSLOUCH*). c) Model misspecification of individual axis methods using (SURFACE) AIC model comparisons. Models inferring two or more optima were considered model misspecification (shown in gray), as input data were simulated under Brownian motion.

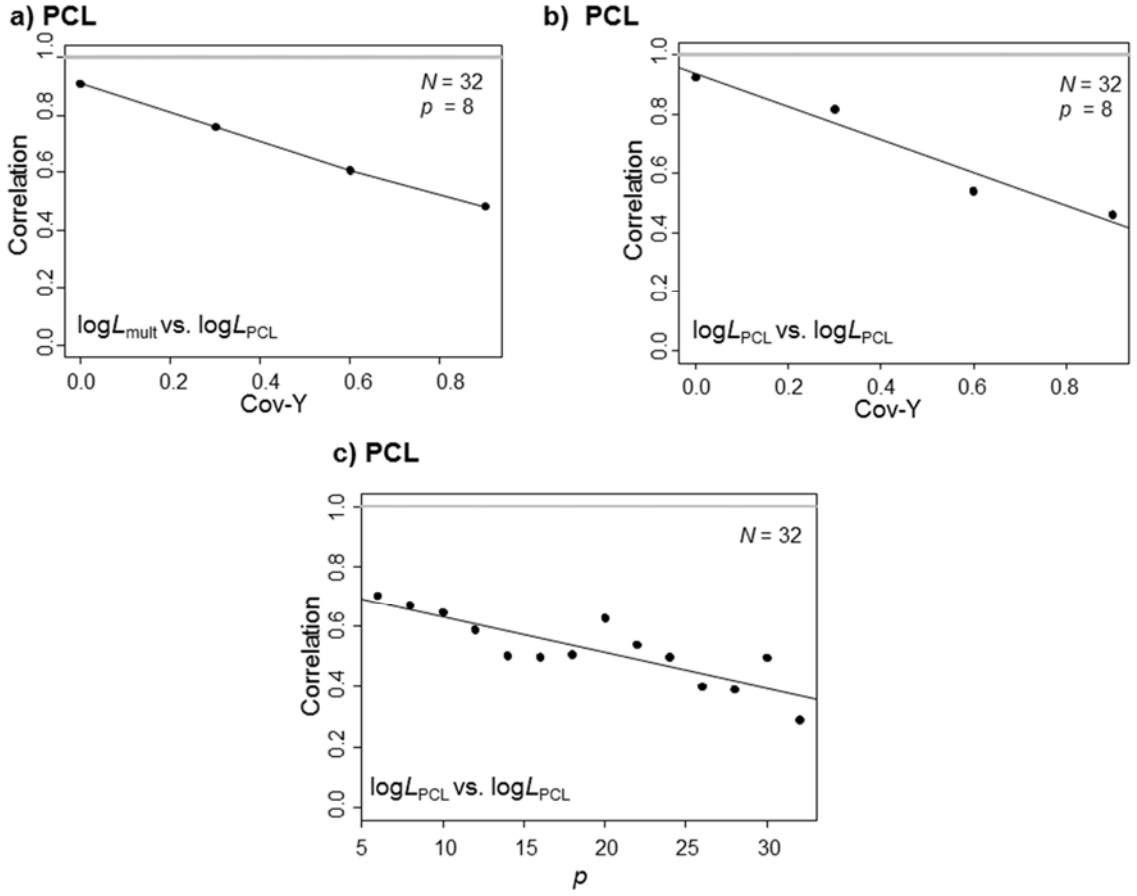

Figure 3. Results from statistical simulations for datasets generated on 32 species phylogenies using differing levels of covariation among trait dimensions (a & b) and differing numbers of trait dimensions (c). For each simulation condition, 100 phylogenies and 100 simulated datasets were generated (see text). a) Correlation between multivariate log*L* and PCL for the same datasets. b) Correlation between PCL and PCL for the same datasets rotated to a different orientation. c) Correlation between PCL and PCL for the same datasets rotated to a different orientation as the number of trait dimensions increase.

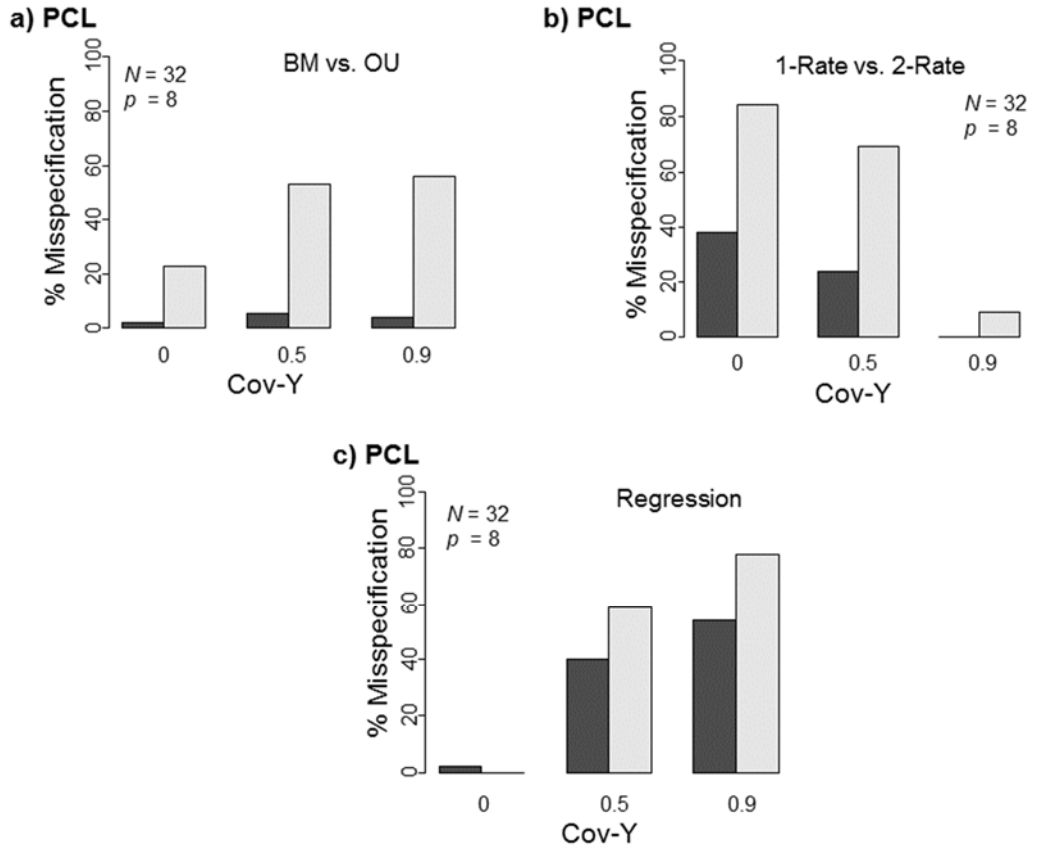

Figure 4. Results from statistical tests using PCL for datasets generated on 32 species phylogenies using differing levels of covariation among trait dimensions (X-axis). For each analysis, 100 phylogenies and 100 simulated datasets were generated, and results are reported for the same datasets in two different orientations of the multivariate dataspace (black and gray bars). The percent of model mis-specification is shown for three examples: a) comparisons of a BM (correct) model versus OU (incorrect) model, b) comparisons of a two evolutionary rate (correct) model versus a one evolutionary rate (incorrect) model, and c) comparisons of a regression (correct) model with a null model lacking the covariate (incorrect).

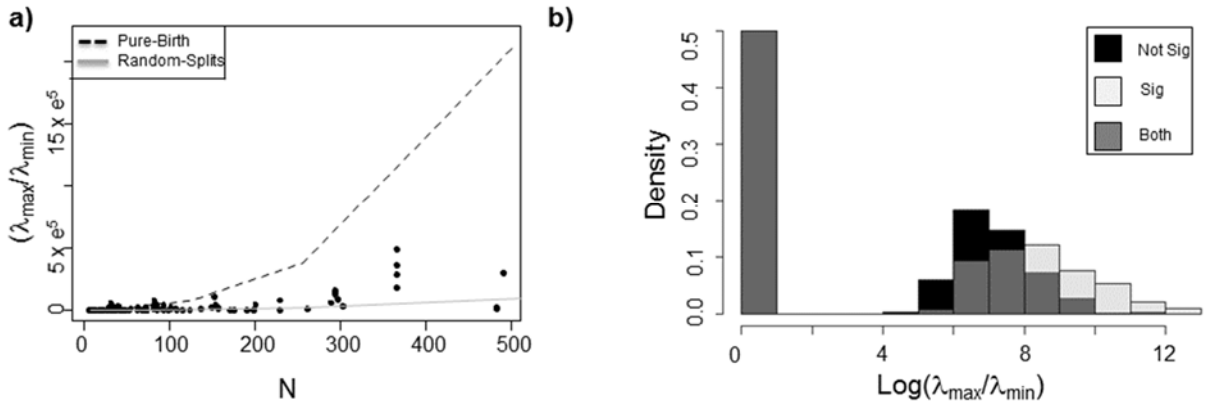

Figure 5. a) Condition number of phylogenetic covariance matrices at differing levels of sample size (*N*). The condition number is a numerical measure of how stable a covariance matrix is under operations such as matrix inversion. Larger condition numbers represent more ill-conditioned matrices, which can result in less stable estimates from down-stream algebraic operations. Values from 104 empirically-generated phylogenies are shown as black dots. The dashed line represents the mean value for 500 pure-birth trees simulated at each level of sample size, while the solid line represents the mean of 500 random-splits phylogenies for the same sample sizes. b) Distribution of condition numbers of simulated pure-birth phylogenies for nonsignificant (black) and significant (light gray) datasets obtained from and tested on those phylogenies using PGLS (N = 32, p = 32). Dark gray bars represent the overlap of the two distributions.

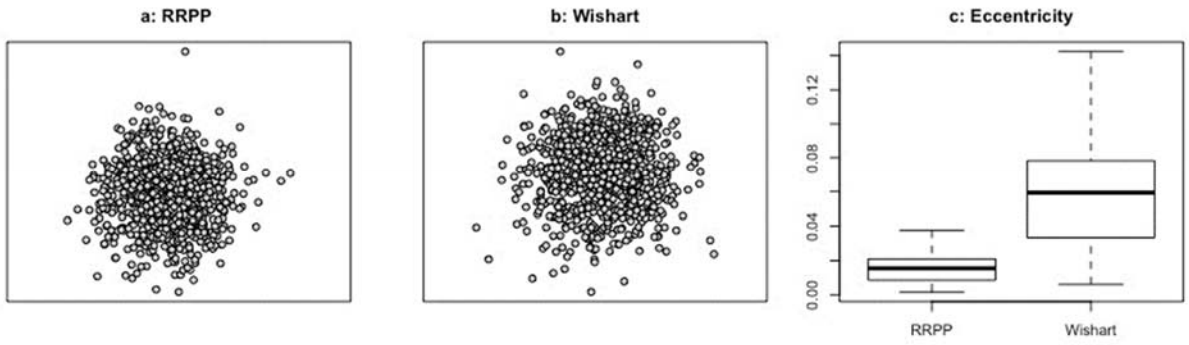

Figure 6. Results from a sampling experiment to compare covariance matrix distributions. a) A sampling distribution generated from 1,000 permutations using D-PGLS for 100 taxa, four dependent variables, and 1 independent variable. b) A second sampling distribution generated from 1,000 samplings from a Wishart distribution. In both a) and b) the two-dimensional ordinations are from principal coordinates (axes not labeled) of Riemannian distances among covariance matrices. c) Boxplots of the eccentricities of 100 sampling iterations, repeating the process summarized in a) and b). Interquartile ranges are shown as boxes, with bolded lines representing medians. Fences extend to maximum and minimum values within 1.5 times the interquartile range (no outliers were found).

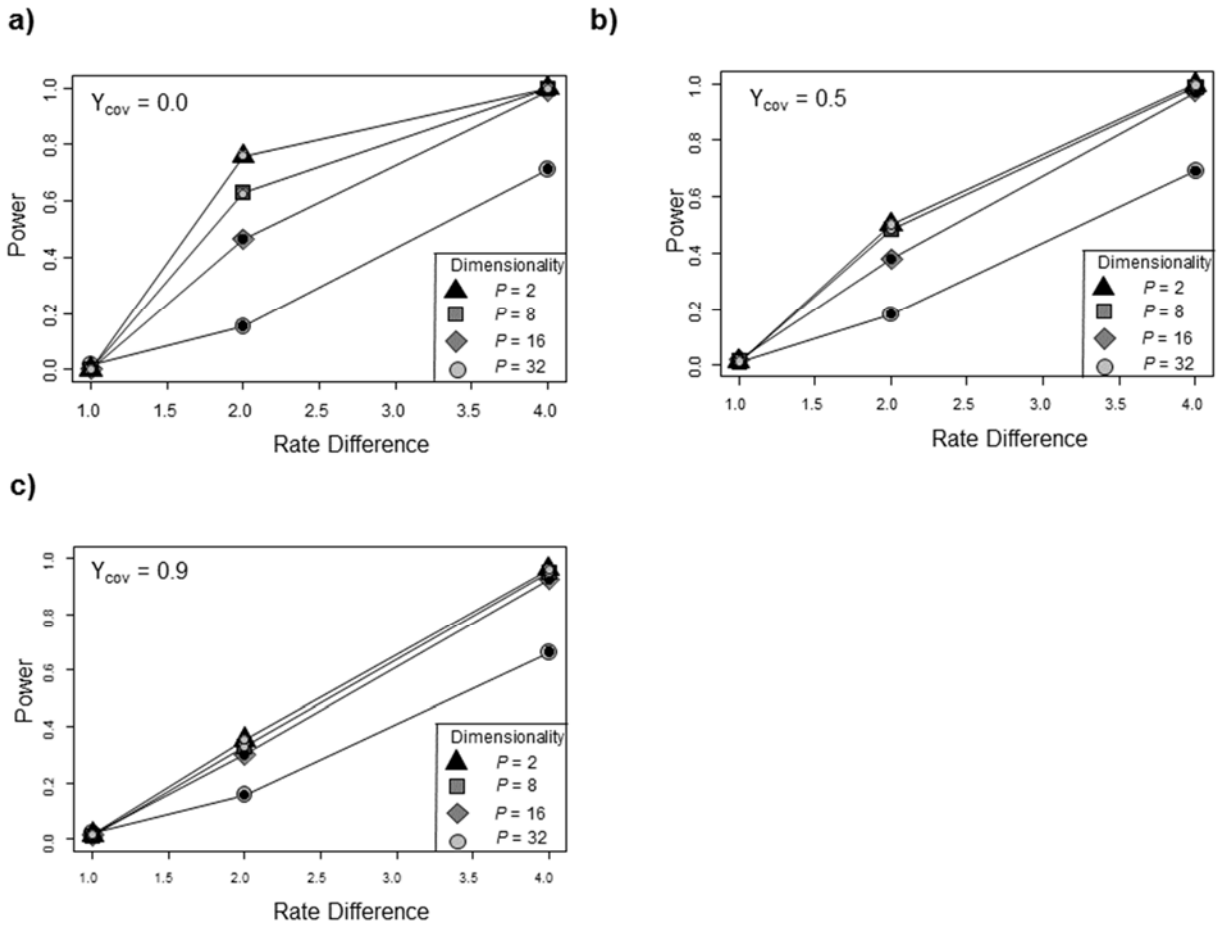

Figure A1. Simulation results evaluating the statistical power of permutation-based hypothesis testing procedures for comparing net evolutionary rates. Data were simulated on random-splits phylogenies containing 32 taxa, and using: a) no covariation among trait dimensions, b) moderate levels of covariation among trait dimensions, and c) high levels of covariation among trait dimensions.