Ecology, 97(2), 2016, pp. 286–293

Improving phylogenetic regression under complex evolutionary models

Florent Mazel,[1,2,6] T. Jonathan Davies,[3,4] Damien Georges,[1,2] Sébastien Lavergne,[1,2]
Wilfried Thuiller,[1,2] and Pedro R. Peres-Neto[5]

[1]*Laboratoire d'Écologie Alpine (LECA), University of Grenoble Alpes, F-38000, Grenoble, France*
[2]*Laboratoire d'Écologie Alpine (LECA), CNRS, F-38000, Grenoble, France*
[3]*Department of Biology, McGill University, 1205, Avenue Docteur Penfield, Montreal, Quebec, Canada*
[4]*African Centre for DNA Barcoding, University of Johannesburg, APK Campus, PO Box 524, Auckland Park 2006, Johannesburg, South Africa*
[5]*Canada Research Chair in Spatial Modelling and Biodiversity, Département des sciences biologiques, Université du Québec à Montréal, Montreal, Quebec H3C 3P8, Canada*

Abstract. Phylogenetic Generalized Least Square (PGLS) is the tool of choice among phylogenetic comparative methods to measure the correlation between species features such as morphological and life-history traits or niche characteristics. In its usual form, it assumes that the residual variation follows a homogenous model of evolution across the branches of the phylogenetic tree. Since a homogenous model of evolution is unlikely to be realistic in nature, we explored the robustness of the phylogenetic regression when this assumption is violated. We did so by simulating a set of traits under various heterogeneous models of evolution, and evaluating the statistical performance (type I error [the percentage of tests based on samples that incorrectly rejected a true null hypothesis] and power [the percentage of tests that correctly rejected a false null hypothesis]) of classical phylogenetic regression. We found that PGLS has good power but unacceptable type I error rates. This finding is important since this method has been increasingly used in comparative analyses over the last decade. To address this issue, we propose a simple solution based on transforming the underlying variance–covariance matrix to adjust for model heterogeneity within PGLS. We suggest that heterogeneous rates of evolution might be particularly prevalent in large phylogenetic trees, while most current approaches assume a homogenous rate of evolution. Our analysis demonstrates that overlooking rate heterogeneity can result in inflated type I errors, thus misleading comparative analyses. We show that it is possible to correct for this bias even when the underlying model of evolution is not known a priori.

Key words: comparative methods; non-stationarity; Phylogenetic Generalized Least Square; statistical performance.

Introduction

Comparative methods are among the key tools for understanding ecological and evolutionary processes (Felsenstein 1985, Harvey and Pagel 1991) and are used to test hypotheses about the correlated evolution of traits (e.g., Pearman et al. 2014). Since species share common ancestry, they should not be considered statistically independent units, thus traditional statistical methods such as Ordinary Least Square (OLS) regression are not appropriate for analyzing comparative data. When analyzed by OLS, the two major issues that arise from shared evolution are increased type I error when traits are uncorrelated with each other and reduced precision in parameter estimation when traits are correlated with each other (Revell 2010). Therefore, interspecific trait data should be analyzed within a phylogenetic framework (Harvey and Pagel 1991, Freckleton et al. 2002, Revell 2010).

Multiple methods have been proposed to account for phylogenetic nonindependence of species when

Manuscript received 15 January 2015; revised 16 September 2015; accepted 22 September 2015. Corresponding Editor: P. de Valpine.

[6] E-mail: flo.mazel@gmail.com

regressing two (or more) continuous or categorical traits (Felsenstein 1985, Grafen 1989, Maddison 1990, Lynch 1991, Garland et al. 1992, Martins and Hansen 1997, Diniz-Filho et al. 1998, Freckleton et al. 2002, Paradis and Claude 2002, Ives and Garland 2010). The phylogenetic regression based on a generalized least square, where the inverse of the phylogenetic covariance matrix is used as weights, is perhaps now the most widely adopted (Grafen 1989, Martins and Hansen 1997): it is a generalization of phylogenetic independent contrasts (Rohlf 2001) originally proposed by Felsenstein (1985) and a particular case of general linear models (Rencher and Schaalje 2007). Phylogenetic regression assumes that the model residual error ε is distributed according to $\sigma_s^2\mathbf{C}$ where and σ_s^2 represents the residual variance and $\mathbf{C}$ is an $n \times n$ matrix (n is the number of tips, i.e., species in most cases) describing the evolutionary relationships among species (i.e., a phylogenetic covariance matrix with diagonal elements estimated as the total branch length between each tip and the root, and off-diagonal elements estimated as the evolutionary time shared by each species pair).

In its simplest form, PGLS assumes a Brownian motion (BM; Edwards and Cavalli-Sforza 1964) model of evolution with a single rate, σ^2. Nevertheless, PGLS is highly flexible and it can be extended to alternative evolutionary models (Martins and Hansen 1997). For example, recent PGLS implementations have incorporated tree transformation models that capture different modes of evolution (e.g., early vs. late trait diversification, continuous vs. punctual evolution [Pagel 1997, 1999, Freckleton et al. 2002, Revell 2010]) or selective regimes (e.g., Ornstein-Uhlenbeck [OU] models [Hansen 1997, Butler and Kings 2004, Lavin et al. 2008]). This model flexibility has been key in reducing type I errors due to model misspecification, one of major issues in comparative biology (Freckleton 2009).

Despite recent advances, however, current PGLS implementations still assume that the tempo and mode of evolution remain constant across the phylogenetic tree (although they allow for rate variation over time), whereas it is likely that both are highly heterogeneous (Simpson 1944, Gould 2002), particularly in the case of large phylogenetic trees (O'Meara 2012). The construction of very large trees (Jetz et al. 2012) hand in hand with the increased availability of corresponding large trait data sets (e.g., Wilman et al. 2014) has generated an increasing need to consider more complex models of evolution within the regression framework. Heterogeneous trait evolution, where trait evolution has been markedly different across multiple clades is a potential source of bias that has been largely overlooked. If two traits evolved under complex models of evolution, standard PGLS (i.e., assuming a single rate of evolution) may not be appropriate. Since it has been demonstrated that an incorrectly defined variance–covariance (VCV) matrix in PGLS increases the type I error rate (Revell 2010) for simple homogeneous models, it naturally follows that for large comparative phylogenetic data sets, where evolutionary processes are likely heterogeneous, there will also be an increase in type I error rates and/or reduced statistical power.

One potential solution arises from the development of heterogeneous models of evolution, which allow the fit of highly complex VCV matrices. Variation in evolutionary rates across the phylogenetic tree can be modeled, for instance, with a heterogeneous BM process where σ^2 varies across the phylogenetic tree (O'Meara et al. 2006, Thomas et al. 2006). Similarly, heterogeneous OU models with multiple optima, strength of selection, and evolutionary rates have been proposed to model adaptive peaks (Ingram and Mahler 2013). These models have been successfully applied in the literature to investigate how a single trait has evolved across the phylogeny (e.g., to study the evolution of a clade life form [Adams et al. 2009, Boucher et al. 2012]) but they are not yet incorporated into the toolbox of comparative analyses of trait correlation. Transforming the phylogenetic tree according to a heterogeneous model of trait evolution fitted with the data and using this transformed tree to derive a new VCV matrix could help in increasing the flexibility of the PGLS framework, but this potential has remained unexplored.

Here, using simulations, we first explore the performance (type I error rate and statistical power) of PGLS under models of heterogeneous trait evolution. We simulate multiple models of trait evolution and cross correlations among traits, and contrast power and type I error rates. We show that complex models of evolution lead to inflated type I error rates, but PGLS is able to handle such complexities when the correct VCV matrices are known. We then propose an implementation of PGLS that has valid type I error rates even under models with large rate heterogeneity where the underlying model of evolution is not known a priori.

Methods and Results

Simulating traits under complex evolutionary models

Simulation model.—Our simulations considered two traits (X and Y) generated by the following basic equation, setting a to zero (following Revell 2010):

$$\mathrm{Y} = a + \beta \mathrm{X} + \varepsilon. \qquad (1)$$

We defined evolutionary models by manipulating the phylogenetic covariance structure and rates of evolution (see Homogeneous and Heterogeneous models of trait evolution subsection) and simulate X and the residuals error ε assuming normally distributed values $N(0, \sigma^2)$; where σ^2 represents the instantaneous rate of evolution that was set depending on the specific scenarios of homogenous (one single evolutionary rate) and heterogeneous (multiple changes in evolutionary rates) models of trait evolution (see Homogeneous and

Heterogeneous models of trait evolution subsection). Type I error for a given method was assessed by simulating data with β = 0 while statistical power was assessed by simulating data with β = 1 (Appendix S1). In both cases, we tested the null hypothesis H_0: β = 0 and reported the percentage of simulations in which H_0 was rejected with an alpha level of 0.05 (i.e., type I error rate as number of rejections when β = 0 and statistical power as the number of rejections when β = 1). For β = 0, *Y* = ε and the evolution of *X* and *Y* are simulated independently. For β = 1, the evolutionary model for *Y* was a function of the two evolutionary processes generating *X* and ε. We consider three different scenarios representing different evolutionary models for *X* and ε (Appendix S1). We generated *X* and ε under (1) the same evolutionary model, (2) different models, or (3) assuming that only ε followed an evolutionary model while *X* was drawn from a normal distribution (with mean of 0 and a standard deviation of 1).

Phylogenetic trees.—To make sure our results were not just representative of a particular phylogenetic topology, we ran all analyses on two very different phylogenetic topologies of 128 species each (rescaled so that their total depth equaled 1). One tree was completely balanced, whereas the second one was obtained using a pure birth process leading to a more realistic unbalanced tree (see Appendix S2).

Homogeneous models of trait evolution.—We considered three classic models (Brownian motion [BM, Edwards and Cavalli-Sforza 1964], Ornstein-Uhlenbeck [OU; Hansen 1997]) and the lambda [λ] tree transformation [Pagel 1999]). For BM, the change in species traits over time was expressed as

$$dX(t) = \sigma dB(t) \tag{2}$$

where $dX(t)$ is the change in trait *X* over time period *dt*. The parameter σ measures the rate of evolution, while the term $B(t)$ is random noise ~ $N(0, dt)$.

For an OU process, the change in species traits over time was expressed as

$$dX(t) = \alpha[\theta - X(t)]dt + \sigma dB(t) \tag{3}$$

where θ represents the mean trait value (often interpreted as the trait optimum) and alpha measures the rate of decay of trait similarity through time (often interpreted as the intensity of stabilizing selection). When $\alpha = 0$, the OU model simplifies to a BM model (see Eq. 2).

The λ tree transformation model simply rescales the phylogenetic tree before applying the classic BM model. In our case, λ is the multiplier of internal branches and we considered values between 1 (no transformation: the trait evolved under a classical BM) and 0 (the tree is a star phylogeny and the trait has no "phylogenetic signal", i.e., related species do not tend to share similar trait values). Continuous characters (representing *X* and ε) under each model had a starting value of 0 at the root of the tree and were evolved tipward according to each model.

Heterogeneous models of trait evolution.—For heterogeneous BM models, we simulated traits with two different rates of evolution occurring in different parts of the phylogenetic tree. In the simplest case, one sub-clade evolved with a rate $\sigma^2 = 1$ while the other sub-clade evolved with one of the following σ^2: 1/1000, 1/100, 1/10, 1/4, 1/2 or 3/4. In this case we simulated a single rate shift near the root of the phylogenetic trees so that the two major sub-clades of the phylogeny evolved under different σ^2 (see Appendix S2). We also generated traits evolving under multiple rate shifts (3, 5, and 9) within each of the two major clades, but we kept the total number of rate values restricted to two (see Appendix S2).

For heterogeneous OU models, we simulated optima shifts and kept σ^2 and α constant at 1 and 0.5, respectively. The initial OU regime started with an optimal value of θ (either 1, 2, 3 or 4) and then shifts to $-\theta$ at the same node as described for BM rate shifts (see Appendix S2).

For heterogeneous λ-transformed models, we simulated a single shift occurring near the root of the tree (i.e., separating two major clades A and B, see Appendix S3) by multiplying the internal branch length of clade A by λ, keeping the internal branches lengths of clade B unchanged. The resulting sub-clades therefore differed in their root-to-tip distances and tip-to-internal branch length ratios. If we had used a single σ^2 for this transformed tree, the trait evolution of the two sub-clades would thus have different rates and phylogenetic signal. Because we were interested here in the differences in phylogenetic signal only, we rescaled the transformed branches of clade A so that all species in the complete tree had the same root-to-tip distance (same overall evolutionary rate) but differed in their ratio of tip-to-internal branch length (see Appendix S3). For λ, we restricted our analysis to the balanced tree (see Phylogenetic trees subsection) because it is not straightforward to retain tree ultrametricity for more complex tree topologies, potentially confounding comparisons between our alternative models. As for the homogenously evolved traits, traits had a starting value of 0 at the root of the tree and were evolved tipward according to each model and evolutionary rates σ^2 that were clade dependent.

Result 1: Assessing power and type 1 error rates of the phylogenetic regression

We fitted two classical linear regression models to each of the different simulated data sets (scenarios). We first fitted an OLS regression, for which we tested the significance of the slope with a *t* test using *n*

− 2 = 126 df. Note that a t test was used given its common usage in the comparative analysis literature, though a likelihood test contrasting the slope and the intercept-only models could have been equally applied. Implementation of the latter could be the object of future studies. Second, we fitted a PGLS (using the pgls function in the caper R package, Orme et al. 2013) that also simultaneously optimized a single λ value for the residuals of the model (λ = ML in the pgls function, $\text{PGLS}_{\text{global\_}\lambda}$ hereafter). The λ parameter (Pagel 1999) is a multiplier for the off-diagonal elements (i.e., the internal phylogenetic branches) of the VCV matrix and usually varies between 0 and 1. If $\lambda = 1$, the VCV is left unchanged and the PGLS assumes a BM while if $\lambda = 0$, the VCV is a diagonal matrix and the PGLS reduces to an OLS. Any value between 0 and 1 indicates that the phylogenetic strength in the residuals is reduced in contrast to a BM. As such, the optimization of λ in PGLS allows encompassing both classical OLS and PGLS. As for OLS, we then tested the significance of the slope with a t test with $n - 2 - 1 = 125$ df (λ is an additional parameter).

Power analysis

First, we simulated X and ε independently from each other to generate Y ($Y = \beta X + \varepsilon$ with $\beta = 1$).

Under this scenario, both OLS and $\text{PGLS}_{\text{global\_}\lambda}$ had good statistical power (i.e., they both detected the simulated correlation) for all models of evolution tested (Table 1 and Appendix S4).

Type 1 error analysis

Second, we simulated Y and X independently from each other ($Y = \beta X + \varepsilon$ with $\beta = 0$, so $\varepsilon = Y$), and evaluated the percentage of simulations where a correlation was (incorrectly) detected (i.e., false positives). The type I error of a valid test should equal the alpha value selected when assessing test significance (5% here). Results differed across the three simulated scenarios (Table 1 and Appendices S5 and S6) (1) when X was simulated without phylogenetic signal and Y followed different models of trait evolution (i.e., heterogeneous BM and OU models), both OLS and PGLS had correct type I error rates (Appendix S6: Fig. S1), (2) when X and Y followed different models of trait evolution (e.g., X followed a heterogeneous BM model and Y followed a heterogeneous OU model), OLS showed inflated type I error while $\text{PGLS}_{\text{global\_}\lambda}$ still performed well (i.e., it had correct type I error rates, Fig. S2 in Appendix S6), and (3) when X and Y followed the same heterogeneous model of trait evolution, results were more mixed. When we simulated BM with heterogeneous Pagel λ, all methods had an inflated type I error, which covaried with the heterogeneity in the strength of the phylogenetic signal (see Fig. 1). When we simulated

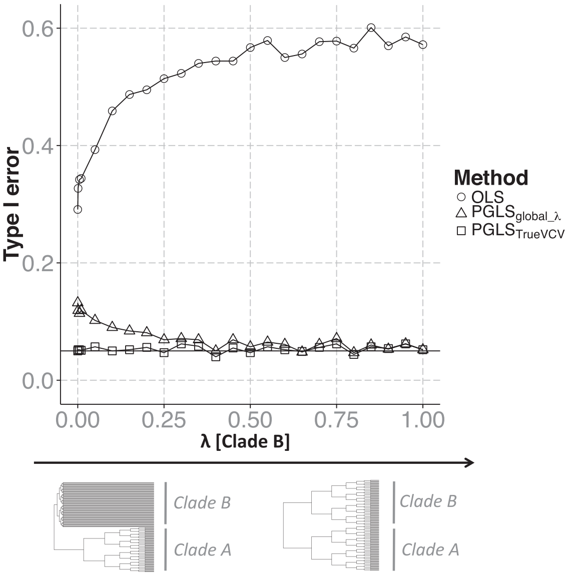

FIG. 1. Effect of heterogeneous phylogenetic signal on type I error. Effect of variation in phylogenetic signal heterogeneity on Type I error rates for the different comparative methods: classical ordinary least square (OLS), phylogenetic generalized least square (PGLS) that jointly optimizes a single λ value for the residuals together with parameter estimates ($\text{PGLS}_{\text{global\_}\lambda}$), and a PGLS that uses the true variance–covariance (VCV) matrix ($\text{PGLS}_{\text{TrueVCV}}$). The x-axis represents λ of clade B (λ of clade A is set to one). Plotted below the x-axis are the corresponding transformed trees for a homogeneous signal (λ[Clade A] = λ[Clade B] = 1) and a heterogeneous signal (λ[Clade A] = 1; λ[Clade B] = 0.01). The type I error represents the percentage of simulation that detected a significant correlation at the 5% level (1000 simulations) between the two traits, which is expected to be 5% for a valid method (black horizontal line).

TABLE 1. The statistical performance (type I error [the percentage of tests based on samples that incorrectly rejected a true null hypothesis] and power [the percentage of tests that correctly rejected a false null hypothesis]) of classical phylogenetic generalized least square (PGLS; optimized for a single λ) under different simulation scenarios.

Statistical performance	Residuals & X	Only X (residuals are normally distributed)	Only residuals (X is normally distributed)
Type 1 error	KO	OK	OK
Power	OK	OK	OK

Notes: Complex evolutionary models were used to simulate either X and the residuals of Y (column 1), only X (column 2), or only the residuals (column 3). KO indicates reduced power or inflated type I error, OK indicates good power and correct type I errors.

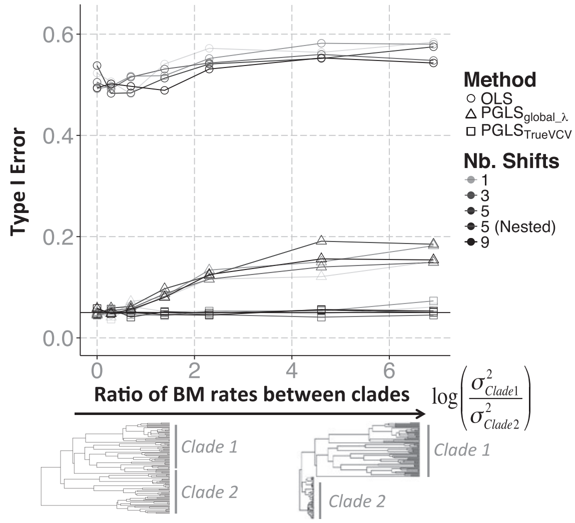

FIG. 2. Effect of heterogeneous rate of trait evolution on type I error. Effect of variation in evolutionary rate heterogeneity (i.e., the ratio of rate evolution between clades) on Type I error rates for different comparative methods (see legend of Fig. 1). Different models of rate heterogeneity are presented (i.e., from 1 to 9 rate shifts; BM rate is Brownian motion rate). For the simplest case (one single rate shift), we plotted below the x-axis the corresponding transformed trees for a homogeneous rate (σ^2 [Clade 1] = σ^2 [Clade 2] = 1) and a heterogeneous signal (σ^2 [Clade 1] = 1; σ^2 [Clade 2] = 0.01). The type I error represents the percentage of simulation that detected a significant correlation at the 5% level (1000 simulations) between the two traits, which is expected to be 5% for a valid method (black horizontal line).

BM models with heterogeneous rates of evolution (different σ^2 across the tree), type I error rates were also inflated, and varied with the strength of the evolutionary rate variation, but were only weakly influenced by the number of rate shifts (Fig. 2). We did not evaluate type I error rates under a heterogeneous OU model of evolution for reasons described in the *Discussion: PGLS and hidden selective trends*. Unbalanced and balanced trees gave qualitatively similar results (see Fig. 2 and Appendix S6: Fig. S3, respectively); for brevity, we report only values for the unbalanced tree in the main text.

Providing the correct VCV to PGLS led to correct type I error rates in all cases (Fig. 2 and Appendix S6: Fig. S3). However, this is obviously not a viable solution for most empirical studies, since the true VCV is not known a priori, but it nonetheless shows that PGLS is able to deal with complex models of evolution when they are correctly estimated.

RESULT 2: A SOLUTION TO CORRECT FOR INFLATED TYPE 1 ERROR

Our simulations showed that PGLS had inflated type I error rates under heterogeneous BM but that it could theoretically handle such models when a correct VCV was provided.

To correct for inflated type I errors when the correct VCV is not known, we studied the statistical performance of the following three-step procedure: (1) fit heterogeneous BM models of trait evolution to the raw OLS residuals, (2) use this fit to modify the VCV matrix used in the standard PGLS, and (3) apply a significance criterion (see *Modified method for significance testing*) that allows for proper inference (i.e., correct type I error) associated to the two initial fits that themselves involve one statistical test each. For clarity, note that the correction in step 3 is completely independent from the issues of inflated type I errors in comparative analysis involving multiple rates of evolution (the focus of this paper). However, because our proposed framework is based on a two-step procedure, one would incur additional type I errors if a multiple testing criterion was not applied.

We tested two approaches to detect rate shifts on the OLS residuals: the auteur Bayesian approach (Eastman et al. 2011; implemented in the R package geiger) and the trait medusa approach (Thomas and Freckleton 2012; implemented in the R package motmot). As auteur and trait medusa yielded similar results, but auteur was much faster, we present here the detailed procedure and results for auteur ($PGLS_{auteur}$ hereafter). The details and results of the procedure involving trait medusa are provided in Appendix S7. We provide R code and example data of both procedures as supplements (Supplements 1 and 2).

Detecting evolutionary rate shifts with auteur

Auteur (Eastman et al. 2011) is a recently developed Bayesian approach to model evolutionary rate heterogeneity along a phylogeny. It uses a reversible jump Markov chain Monte Carlo procedure to sample a distribution of multi-rate BM models in inverse proportion to their poorness of fit. In the optimization procedure, the Markov chain jumps from models of differing complexities (i.e., number and position of the shifts in the phylogenetic tree).

We sampled parameters every 500 generations for a total of 20 000 generations. We removed 25% of the sampled parameters as burn-in to obtain marginalized distributions of relative rates for each branch of the tree.

Using fitted multi-rate models in standard PGLS

We then used the parameter estimates from auteur to rescale the phylogenetic tree using the function rescale in the geiger package. The rescaled tree was then used in a standard PGLS procedure ($PGLS_{auteur,}$ function pgls in the R package caper) as previously described. We then tested the significance of the slope with a t test with $n - 2$ df ($PGLS_{auteur}$, see Fig. 3). As described previously, this stepwise procedure may be expected to have a slight inflated type I error rate

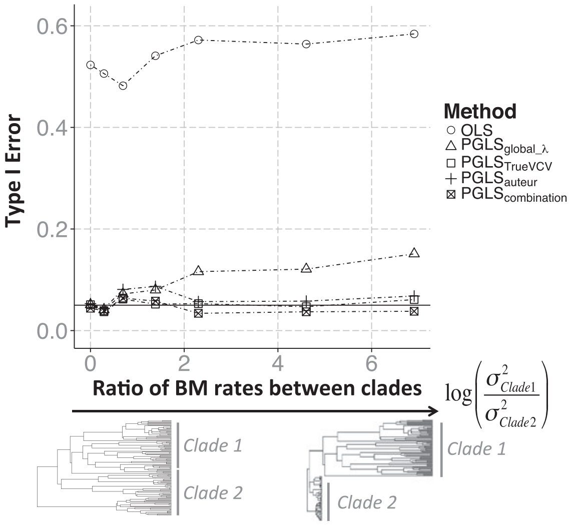

FIG. 3. Type I error rates for modified PGLS. Comparison of type I error rate for classical (OLS, $PGLS_{global\_\lambda}$, and $PGLS_{TrueVCV}$) and modified ($PGLS_{auteur}$ and $PGLS_{combination}$) comparative methods as a function of evolutionary rate heterogeneity between clades. We show here the result for the simplest model of rate heterogeneity (i.e., one single rate shift) and plot below the *x*-axis the corresponding transformed trees for a homogeneous rate (σ^2 [Clade 1] = σ^2 [Clade 2] = 1) and a heterogeneous rate (σ^2 [Clade 1] = 1; σ^2 [Clade 2] = 0.01). The type I error represents the percentage of simulation that detected a significant correlation at the 5% level (1000 simulations) between the two traits, which is expected to be 5% for a valid method (black horizontal line).

because it represents a two-step procedure (Fig. 3), and each step has independent errors (the same reasoning applies to the $PGLS_{trait\ medusa}$ procedure, Appendix S7). Step 1 estimates the evolutionary model of the OLS residuals, and step 2 fits PGLS using the estimated VCV matrix in step 1. As before, because two statistical tests were involved here, we applied a modified significance testing criterion.

Modified method for significance testing

Statistical testing when using a two-step procedure is likely to have inflated incorrect family-wise type I error. This has been previously recognized by ter Braak et al. (2012) when testing for the links between trait to environmental variation. Their procedure was based on a two-step procedure in which the links between species trait variation and species distributions, and the links between environmental variation and species distributions were both tested. They adopted a rejection criterion for establishing the significance of the links between environment and trait variation in which the largest probability of the two tests involved needed to be below the pre-established alpha-level (i.e., 0.05). Retaining the largest *P* value is equivalent to conducting only one statistical test instead of two, assuring correct type I errors. Following ter Braak et al. (2012), we thus considered the relationship between *Y* and *X* ($PGLS_{combination}$) and retained the largest *P* value between the $PGLS_{global\_\lambda}$ and $PGLS_{auteur}$. This approach produced correct type I error rate in all cases (Fig. 3).

DISCUSSION

With the increasing availability of well-resolved phylogenetic trees, PGLS has become routinely employed in the analysis of interspecific data over the past decades (Felsenstein 1985, Grafen 1989, Martins and Hansen 1997, Freckleton et al. 2002). By assuming an explicit model of evolution, PGLS contrasts with some other approaches, for example, nonparametric eigenvector decomposition (Diniz-Filho et al. 1998, Freckleton et al. 2011). While it has been argued that the inclusion of an explicit evolutionary model within PGLS allows for increased efficiency of estimation, decreased variance in parameter estimates, and decreased Type I errors (Freckleton et al. 2011), misspecifying the evolutionary model may have important consequences for hypotheses testing (Revell 2010; though not for parameter estimation; see, e.g., Rohlf 2006). One solution is to simultaneously estimate the parameters of the PGLS model and the model of evolution of the residuals. For some simple models of trait evolution, in which both the tempo and mode of evolution remain constant across the phylogenetic tree, it is possible to adjust the PGLS model residuals using Pagel's (1999) lambda tree transformations (Freckleton 2002). However, such simple models of evolution are likely rare, particularly for large trees (O'Meara 2012), which have become widely used in the comparative literature with the increasingly availability of large scale mega-phylogenies including several thousand species. Here we evaluated the performance (power and type I errors) of PGLS methods, under more complex evolutionary scenarios. We show that most developed methods around PGLS have good power, but unacceptably high type I errors under some scenarios with heterogeneous evolutionary rates. Nonetheless, PGLS methods perform well when the VCV matrix is estimated properly.

X and Y follow different models of evolution

When there is no phylogenetic signal in the independent variable (i.e., *X* was normally distributed and independent from phylogeny) but the residuals (of *Y*) follow a heterogeneous model of evolution (either BM or OU), all methods (OLS and PGLS) showed correct type I error rate. Similar results have been reported for homogeneous BM models (see Revell 2010). However, when we simulated a heterogeneous BM in the independent variable (*X*) and the dependent variable (*Y*) followed a heterogeneous OU, OLS shows inflated type I error rates, but PGLS still performs well.

X and Y follow the same model of evolution

When we simulated *X* and *Y* with the same heterogeneous model of evolution all classical methods (OLS and PGLS) showed inflated type I error rates. Nevertheless PGLS is theoretically able to handle such bias (Martins and Hansen 1997), and we demonstrated empirically that providing the correct VCV leads to valid test of correlated evolution with appropriate type I error rates.

Because knowing the correct VCV transformation a priori is difficult for traits evolving under complex evolutionary models, we implemented a simple approach that allowed us to estimate the model of evolution on the residuals of the OLS, and used this to transform the VCV matrix for PGLS. Here we used two methods (a Bayesian approach and a bootstrap approach in conjunction with an algorithm for estimating multiple rate shifts) to identify the appropriate model for transforming the VCV matrix, though we acknowledge that alternative methods are available (see Revell et al. [2012] for an example). We show that assuming the transformed VCV matrix, PGLS had appropriate type I errors after correcting for the two-step model selection procedure (ter Braak et al. 2012). However, we caution that one of the methods used is sensitive to overfitting (a known problem with the medusa algorithm assuming the Akaike information criterion adjusted for sample size [AIC_c] stopping criterion) and that a wrongly defined VCV matrix can lead to type I error rates as high as for raw OLS models. It is crucial, therefore, to use a methodology that correctly assigns rate shifts, or to test overfitting by parametric bootstrapping (Boettiger et al. 2012).

It is possible that more efficient strategies will become soon available for fitting PGLS under heterogeneous models of evolution, nonetheless we have shown that our approach works well, and is reasonably transparent. Due to computational constrains, it was not feasible to explore the full range of possible evolutionary models, tree shapes, and trees sizes (the combinations of which are effectively infinite), but we see no reason why our approach should fail under different conditions except when clade sizes for which rate changes have occurred are too small to accurately infer the correct evolutionary model.

Last, we note that we do not provide a solution for the heterogeneous phylogenetic signal scenario; however, future advances that allow the fit of heterogeneous phylogenetic signal to the tree could be easily implemented within our framework.

PGLS and hidden selective trends

We did not evaluate type I error rates under a heterogeneous OU model of evolution. Under this scenario, even for $\beta = 0$ (i.e., no expected correlation between *X* and *Y*), the two traits might be significantly correlated if both have followed the same selective trends (i.e., optima and selection strength shifts). In this case, we are no longer considering type I error rates, but rather the power of the method, and it is not straightforward to derive expectations if the two traits follow slightly different models. For example, one could imagine that the two traits show the same general trends but with slightly different optima, strength of selection, or location of shifts. One interesting idea would be to simulate traits with different OU parameters and positions and plot the percentage of detected correlations against the amount of difference between the two models using a version of PGLS that takes into account the drift part of the OU model. It might also be informative to explore models where traits coevolve, or the evolution of one is driven by the evolution of the other, causing consistent time lags between their evolutionary shifts. It would then be possible to compare the likelihood of such a model with another in which traits evolve independently. This would avoid the direct correlation of extant species traits. Finally, in the present study we have assumed a parametric method for estimating changes in evolutionary rates in the sense that the procedures here estimate variance–covariance phylogenetic matrices based on families of known models of evolution. Another potential solution, not explored here, would be to use an iterative weighted least-square implementation (e.g., Björck 1996) in order to estimate an appropriate variance–covariance structure that would make model residuals independent.

Taken together, we have shown that currently implemented phylogenetic comparative methods have unacceptable type I error rate when species' traits evolve under heterogeneous models of evolution. We proposed a flexible solution based on PGLS and showed that it had correct type I error. Our framework is potentially extendible to most complex evolutionary scenarios.

Acknowledgments

F. Mazel would like to thank Rafael Wüest for stimulating discussions on the PGLS approach. The research leading to these results had received funding from the European Research Council under the European Community's Seven Framework Programme FP7/2007-2013 Grant Agreement no. 281422 (TEEMBIO). F. Mazel, D. Georges, S. Lavergne, and W. Thuiller belong to the Laboratoire d'Écologie Alpine, which is part of Labex OSUG@2020 (ANR10 LABX56). All computations performed using the CiGRI platform of the CIMENT infrastructure, which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA) and France-Grille. T. J. Davies and P. R. Perez-Neto were supported by a team grant from the Fonds de recherche du Québec – Nature et technologies.

Literature Cited

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

Björck, Å. 1996. Numerical methods for least squares problems. SIAM, Philadelphia, Pennsylvania, USA.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Boucher, F. C., W. Thuiller, C. Roquet, R. Douzet, S. Aubert, N. Alvarez, and S. Lavergne. 2012. Reconstructing the origins of high-alpine niches and cushion life form in the genus Androsace sl (Primulaceae). Evolution 66:1255–1268.

Butler, M. A., and A. A. Kings. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. American Naturalist 164:683–695.

David Orme, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac and Will Pearse 2013. caper: Comparative Analyses of Phylogenetics and Evolution in R. http://CRAN.R-project.org/package=caper

Diniz-Filho, J. A. F., C. E. R. de Sant'Ana, and L. M. Bini. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262.

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Pages 67–76 *in* V. H. Heywood, and J. McNeill, editors. Phenetic and phylogenetic classification. The Systematics Association, London.

Felsenstein, J.. 1985. Phylogenies and the comparative method. American Naturalist 125:1–15.

Freckleton, R. 2009. The seven deadly sins of comparative analysis. Journal of Evolutionary Biology 22:1367–1375.

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

Freckleton, R. P., N. Cooper, and W. Jetz. 2011. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. American Naturalist 178:E10–E17.

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

Gould, S. J.. 2002. The structure of evolutionary theory. Harvard University Press, The publisher is Harvard University Press, Cambridge, MA.

Grafen, A.. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London B 326:119–157.

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

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

Ingram, T., and D. Mahler. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods in Ecology and Evolution 4:416–425.

Ives, A., and T. Garland. 2010. Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59:9–26.

Jetz, W., G. Thomas, J. Joy, and A. Ø. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

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

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.

Maddison, W. P.. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:223–539–557.

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

O'Meara, B. 2012. Evolutionary inferences from phylogenies: a review of methods. Annual Review of Ecology, Evolution, and Systematics 43:267–285.

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

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. Zoologica Scripta 26:331–348.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paradis, E., and J. Claude. 2002. Analysis of comparative data using generalized estimating equations. Journal of Theoretical Biology 218:175–185.

Pearman, P., S. Lavergne, C. Roquet, R. Wuest, N. E. Zimmermann, and W. Thuiller. 2014. Phylogenetic patterns of climatic, habitat and trophic niches in a European avian assemblage. Global Ecology and Biogeography 23:414–424.

Rencher, A. C., and G. B. Schaalje. 2007. Linear models in statistics. John Wiley & Sons Inc, Hoboken, New Jersey, USA.

Revell, L. 2010. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1:319–329.

Revell, L., D. Mahler, P. Peres-Neto, and B. Redelings. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution 66:135–146.

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

Rohlf, F. 2006. A comment on phylogenetic correction. Evolution 60:1509–1515.

Simpson, G. G.. 1944. Tempo and mode in evolution. Columbia University Press, Colombia University Press (NY).

ter Braak, C., A. Cormont, and S. Dray. 2012. Improved testing of species traits-environment relationships in the fourth-corner problem. Ecology 93:1525–1526.

Thomas, G. H., and R. P. Freckleton. 2012. MOTMOT: models of trait macroevolution on trees. Methods in Ecology and Evolution 3:145–151.

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proceedings Biological Sciences / The Royal Society 273:1619–1624.

Wilman, H., J. Belmaker, J. Simpson, C. de la Rosa, M. M. Rivadeneira, and W. Jetz. 2014. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. Ecology 95:2027.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1890/15-0086.1/suppinfo