PREPRINT

This preprint is a PDF of a manuscript that has been accepted for publication in an ESA journal. It is the final version that was uploaded and approved by the author(s). While the paper has been through the usual rigorous peer review process of ESA journals, it has not been copy-edited, nor have the graphics and tables been modified for final publication. Also note that the paper may refer to online Appendices and/or Supplements that are not yet available. We have posted this preliminary version of the manuscript online in the interest of making the scientific findings available for distribution and citation as quickly as possible following acceptance. However, readers should be aware that the final, published version will look different from this version and may also have some differences in content.

The doi for this manuscript and the correct format for citing the paper are given at the top of the online (html) abstract.

Once the final published version of this paper is posted online, it will replace the preliminary version at the specified doi.
Improving phylogenetic regression under complex evolutionary models

Florent Mazel, T. Jonathan Davies, Damien Georges, Sébastien Lavergne, Wilfried Thuiller and Pedro R. Peres-Neto

Running title: PGLS under complex models of evolution

Author affiliations:

Florent Mazel. Univ. Grenoble Alpes, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France; CNRS, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France, flo.mazel@gmail.com

Jonathan T. Davies. Department of Biology, McGill University, 1205, Avenue Docteur Penfield, Montreal, Quebec, Canada. African Centre for DNA Barcoding, University of Johannesburg, APK Campus, PO Box 524, Auckland Park 2006, Johannesburg, South Africa; j.davies@mcgill.ca

Damien Georges. Univ. Grenoble Alpes, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France ; CNRS, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France. damien.georges2@gmail.com

Sébastien Lavergne. Univ. Grenoble Alpes, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France ; CNRS, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France. sebastien.lavergne@ujf-grenoble.fr

Wilfried Thuiller. Univ. Grenoble Alpes, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France ; CNRS, Laboratoire d’Écologie Alpine (LECA), F-38000 Grenoble, France. wilfried.thuiller@ujf-grenoble.fr

Pedro R. Peres-Neto. Canada Research Chair in Spatial Modelling and Biodiversity, Département des sciences biologiques, Université du Québec à Montréal, Montréal, QC H3C 3P8, Canada; peres-neto.pedro@uqam.ca

Abstract
Phylogenetic Generalised 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 % of tests based on samples that incorrectly rejected a true null hypothesis - and power - the % 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, whilst 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.

**Keywords.** Comparative methods, PGLS, non-stationarity, statistical performance, Phylogenetic Generalised Least Square.

**Introduction**

Comparative methods are among the key tools for understanding ecological and evolutionary processes (Felsenstein 1985; Harvey & 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 (Revel 2010). Therefore, interspecific trait data should be analysed within a phylogenetic framework (Harvey and Pagel 1991, Freckleton et al. 2002, Revell 2010).

Multiple methods have been proposed to account for phylogenetic non-independence of species when 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 generalised least square, where the inverse of the phylogenetic covariance matrix is used as weights, is perhaps now the most widely adopted (Grafen 1989; Martins & Hansen 1997): it is a generalisation of phylogenetic independent contrasts (Rohlf 2001) originally proposed by Felsenstein (1985) and a particular case of general linear models (Rencher & Schaalje, 2007). Phylogenetic regression assumes that the model residual error $\epsilon$ is distributed according to $\sigma^2_C$ where and $\sigma^2_c$ represents the residual variance and $C$ is a $n \times n$ 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 & Cavalli-Sforza 1964) model of evolution with a single rate, $\sigma^2$. 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, Freckleton et al. 2002; Revell 2010) or selective regimes (e.g. Ornstein–Uhlenbeck [OU] models, Hansen 1997; Butler & 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 (Jetz et al. 2012) hand in hand with the increased availability of corresponding large trait datasets (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 VCV matrix in PGLS increases the type I error rate (Revell 2010) for simple homogeneous models, it naturally follows that for large comparative phylogenetic datasets, 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 $\sigma^2$ 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 be proposed to
model adaptive peaks (Beaulieu 2013; Ingram & 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 is 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 & 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 Revel 2010):

\[ Y = a + \beta X + \varepsilon \]  
(Eqn. 1)

We defined evolutionary models by manipulating the phylogenetic covariance structure and rates of evolution (see below) and simulate X and the residuals error \( \varepsilon \) assuming normally distributed values \( N(0, \sigma^2) \); where \( \sigma^2 \) 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 below). Type I error for a given method was assessed by simulating data with $\beta=0$ while statistical power was assessed by simulating data with $\beta=1$ (Appendix A). In both cases, we tested the null hypothesis $H_0: \beta=0$ 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 $\beta=0$ and statistical power as the number of rejections when $\beta=1$). For $\beta=0$, $Y=\epsilon$ and the evolution of $X$ and $Y$ are simulated independently. For $\beta=1$, the evolutionary model for $Y$ was a function of the two evolutionary processes generating $X$ and $\epsilon$. We consider three different scenarios representing different evolutionary models for $X$ and $\epsilon$ (Appendix A). We generated $X$ and $\epsilon$ under (1) the same evolutionary model, (2) different models or (3) assuming that only $\epsilon$ followed an evolutionary model while $X$ was drawn from a normal distribution (with mean of 0 and a standard deviation of one).

**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 one). 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 B).

**Homogeneous models of trait evolution**

We considered three classic models (Brownian motion (BM, Edwards & Cavalli-Sforza 1964), Ornstein–Uhlenbeck (OU, Hansen 1997) and the lambda ($\lambda$) tree transformation (Pagel 1999)). For BM, the change in species traits over time was expressed as:

$$dX(t) = \sigma dB(t) \quad (Eq. \ 2)$$

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

For an OU process, the change in species traits over time was expressed as:
\[ dX(t) = \alpha [\theta - X(t)] dt + \sigma dB(t) \quad \text{(Eq. 3)} \]

where \( \theta \) 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 \( \lambda \) tree transformation model simply rescales the phylogenetic tree before applying the classic BM model. In our case, \( \lambda \) 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 value). Continuous characters (representing \( X \) and \( \epsilon \)) 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 \( \sigma^2 \): 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 \( \sigma^2 \) (see Appendix B). 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 B).

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

For heterogeneous \( \lambda \)-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 C) by multiplying the internal branch length of clade A by \( \lambda \), 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.
176 If we had used a single $\sigma^2$ 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 C). For $\lambda$, we restricted our analysis to the balanced tree (see above) 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 $\sigma^2$ that were clade dependent.

187 Result 1: assessing power and type I error rates of the phylogenetic regression.

188 We fitted two classical linear regression models to each of the different simulated datasets (scenarios). We first fitted an OLS regression, for which we tested the significance of the slope with a t-test using n-2=126 d.f. 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 had 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) that also simultaneously optimized a single $\lambda$ value for the residuals of the model ($\lambda$="ML" in the pgls function, ‘PGLSglobal_\lambda’ hereafter). The $\lambda$ 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 $\lambda$ 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 d.f. \((\lambda)\) is an additional parameter).

**Power analysis**

First, we simulated \(X\) and \(\varepsilon\) independently from each other to generate \(Y\) \((Y = \beta X + \varepsilon\) with \(\beta=1\)). Under this scenario, both OLS and PGLS had good statistical power (i.e. they both detected the simulated correlation) for all models of evolution tested (Table 1 and Appendix D).

**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 see Appendix E-F): (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 (Fig. F1 in Appendix F); (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 PGLS still performed well (i.e., it had correct type I error rates, Fig. F2 in Appendix F); (3) When \(X\) and \(Y\) followed the same heterogeneous model of trait evolution results were more mixed. When we simulated BM with heterogeneous Pagel \(\lambda\), all methods had an inflated type I error, which covaried with the heterogeneity in the strength of the phylogenetic signal (see Figure 1). When we simulated BM models with heterogeneous rates of evolution (different \(\sigma^2\) 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 section *PGLS and hidden selective trends*. Unbalanced and balanced tree gave qualitatively similar results (see Fig. 2
and Fig. F3 in Appendix F, 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 Fig. F3 in Appendix F). However, this is obviously not a viable solution for most empirical studies since the true VCV is not known \textit{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 \textit{Modified method for significance testing} below) 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 \textit{auteur} bayesian approach (Eastman \textit{et al.} 2011, implemented in the R package \textit{geiger}) and the \textit{trait medusa} approach (Thomas and Freckleton 2012, implemented in the R package \textit{motmot}). As \textit{auteur} and \textit{trait medusa} yielded similar results but \textit{auteur} being much faster, we present in the main text the detailed procedure and results for \textit{auteur} (PGLS\textsubscript{auteur} hereafter). The details and results of the
procedure involving *trait medusa* are provided in Appendix G. We provide R code and example data of both procedures as a supplement (Supplements 1 & 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 described above. We then tested the significance of the slope with a t-test with n-2 d.f. (*PGLS_{auteur}* see Fig. 3). As described above, this stepwise procedure may be expected to have a slight inflated type I error rate 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 G). 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\textsubscript{combination}) and retained the largest p-value
between the PGLS\textsubscript{global}, and PGLS\textsubscript{auteur}. This approach produced correct type I error rate in all
cases (Fig. 3).

Discussion

With the increasingly 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, non-parametric 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 AICc 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 clades 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 B=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 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 co-evolve, 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.

Acknowledgements:

FM 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). FM, DG, SL and WT 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 (https://ciment.ujf-grenoble.fr), which is supported by the
Rhône-Alpes region (GRANT CPER07_13 CIRA) and France-Grille (http://www.france-grilles.fr).

TJD and PRPN were supported by a team grant from the Fonds de recherche du Québec – Nature et technologies.

References


List of Supplementary material

Appendix A. Simulation workflow
Appendix B. Phylogenetic trees and associated regime shifts
Appendix C. λ tree transformation
Appendix D. Power Analysis
Appendix E. Type 1 error Analysis workflow
Appendix F. Type 1 error Analysis results
Appendix G. Trait Medusa approach: methods & results
Supplement 1. R code Example
Supplement 2. Data Example
Table 1. Table 1 summarizes the statistical performance (type I error - the % of tests based on samples that incorrectly rejected a true null hypothesis - and power - the % of tests that correctly rejected a false null hypothesis) of classical PGLS (optimized for a single $\lambda$) under different simulation scenarios. 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.

<table>
<thead>
<tr>
<th>Complex Evolutionary model in:</th>
<th>Residuals &amp; $X$</th>
<th>Only X (residuals are normally distributed)</th>
<th>Only residuals (X is normally distributed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical performance:</td>
<td>Type I error</td>
<td>Power</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KO</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
</tbody>
</table>


**Figure 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 OLS, PGLS that jointly optimises a single $\lambda$ value for the residuals together with parameter estimates (‘PGLS$^{\text{global}}_{\lambda}$’) and a PGLS that uses the true VCV matrix (‘PGLS$^{\text{TrueVCV}}$’). The X-axis represents $\lambda$ of clade B ($\lambda$ of clade A is set to one). Plotted below the X-axis are the corresponding transformed trees for a homogeneous signal ($\lambda[\text{Clade A]} = \lambda[\text{Clade B]} = 1$) and a heterogeneous signal ($\lambda[\text{Clade A]} = 1; \lambda[\text{Clade B]} = 0.01$). 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).

**Figure 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 Figure 1). Different models of rate heterogeneity are presented (i.e. from 1 to 9 rate shifts). For the simplest case (one single rate shift), we plotted below the X-axis the corresponding transformed trees for a homogeneous rate ($\sigma^2[\text{Clade 1]} = \sigma^2[\text{Clade 2]} = 1$) and a heterogeneous signal ($\sigma^2[\text{Clade 1]} = 1; \sigma^2[\text{Clade 2]} = 0.01$). 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).

**Figure 3. Type I error rates for modified PGLS.** Comparison of type I error rate for classical (OLS, PGLS$^{\text{global}}_{\lambda}$, and PGLS$^{\text{TrueVCV}}$) and modified (PGLS$^{\text{auteur}}$ and PGLS$^{\text{combination}}$) 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 ($\sigma^2[\text{Clade 1]} = \sigma^2[\text{Clade 2]} = 1$) and a heterogeneous rate ($\sigma^2[\text{Clade 1]} = 1; \sigma^2[\text{Clade 2]} = 0.01$). The type I error represents the
509 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).
Type I Error

Method
- OLS
- PGLS_{global, \lambda}
- PGLS_{TrueVCV}

Nb. Shifts
- 1
- 3
- 5
- 5 (Nested)
- 9

Ratio of BM rates between clades

\[ \log \left( \frac{\sigma^2_{\text{Clade 1}}}{\sigma^2_{\text{Clade 2}}} \right) \]
Type I Error

Ratio of BM rates between clades

\[ \log \left( \frac{\sigma^2_{\text{Clade}1}}{\sigma^2_{\text{Clade}2}} \right) \]

Method
- OLS
- PGLS_{global, \lambda}
- PGLS_{TrueVCV}
- PGLS_{auteur}
- PGLS_{combination}

Clade 1

Clade 2

Clade 1

Clade 2