A General Method for Simultaneously Accounting for Phylogenetic and Species Sampling Uncertainty via Rubin's Rules in Comparative Analysis

Shinichi Nakagawa, Pierre de Villemereuil

Syst. Biol. 0(0):1–11, 2019

DOI:10.1093/sysbio/syy089

A General Method for Simultaneously Accounting for Phylogenetic and Species Sampling Uncertainty via Rubin's Rules in Comparative Analysis

SHINICHI NAKAGAWA[1,2,*] AND PIERRE DE VILLEMEREUIL[3]

[1]*Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia;* [2]*Diabetes and Metabolism Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; and* [3]*CEFE, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France*
Shinichi Nakagawa and Pierre de Villemereuil contributed equally to this article.
[*]*Correspondence to be sent to: Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia;*
E-mail: s.nakagawa@unsw.edu.au.

Received 18 June 2018; reviews returned 6 December 2018; accepted 15 December 2018
Associate Editor: Luke Harmon

Abstract.—Phylogenetic comparative methods (PCMs), especially ones based on linear models, have played a central role in understanding species' trait evolution. These methods, however, usually assume that phylogenetic trees are known without error or uncertainty, but this assumption is most likely incorrect. So far, Markov chain Monte Carlo (MCMC)-based Bayesian methods have mainly been deployed to account for such "phylogenetic uncertainty" in PCMs. Herein, we propose an approach with which phylogenetic uncertainty is incorporated in a simple, readily implementable and reliable manner. Our approach uses Rubin's rules, which are an integral part of a standard multiple imputation procedure, often employed to recover missing data. We see true phylogenetic trees as missing data under this approach. Further, unmeasured species in comparative data (i.e., missing trait data) can be seen as another source of uncertainty in PCMs because arbitrary sampling of species in a given taxon or "species sampling uncertainty" can affect estimation in PCMs. Using two simulation studies, we show our method can account for phylogenetic uncertainty under many different scenarios (e.g., uncertainty in topology and branch lengths) and, at the same time, it can handle missing trait data (i.e., species sampling uncertainty). A unique property of the multiple imputation procedure is that an index, named "relative efficiency," could be used to quantify the number of trees required for incorporating phylogenetic uncertainty. Thus, by using the relative efficiency, we show the required tree number is surprisingly small (~50 trees). However, the most notable advantage of our method is that it could be combined seamlessly with PCMs that utilize multiple imputation to handle simultaneously phylogenetic uncertainty (i.e., missing true trees) and species sampling uncertainty (i.e., missing trait data) in PCMs. [Bayesian statistics; comparative analysis; data augmentation; information theory; model averaging; phylogenetics.]

Phylogenetic comparative methods (PCMs) have been playing a central role in investigating trait evolution across species (reviewed in Garamszegi 2014). The most popular methods in comparative biology are based on linear regression such as independent contrasts (Felsenstein 1985), phylogenetic generalized least squares (PGLS; Grafen 1989), or phylogenetic (generalized) linear mixed models (Lynch 1991; Hadfield and Nakagawa 2010). PCMs also include methods estimating lineage diversification (O'Meara 2012; Pennell and Harmon 2013). When one phylogenetic tree is used in analysis, all these methods assume that the phylogeny of organisms is known without error.

However, no phylogenetic trees (or hypotheses) are known without error. Errors come in the form of uncertainty in branch length, topology, and also in the model of assumed character evolution (Cooper et al. 2016; Cornwell and Nakagawa 2017). Researchers have been investigating the impact of these types of uncertainty on statistical inference (e.g., Diaz-Uriarte and Garland 1996; Symonds 2002). These studies generally suggest the importance of incorporating "phylogenetic uncertainty" in PCMs; note that by using one tree, point estimates (e.g., regression coefficients) are not necessarily biased (Stone 2011), but uncertainty estimates [e.g., standard error or confidence intervals (CI)] are not accurate. Therefore, a number of methods have been proposed to include phylogenetic uncertainty (e.g., Losos 1994; Martins 1996; Huelsenbeck et al. 2000; Housworth and Martins 2001; Rangel et al. 2015). Among these methods, probably the best one is to use Bayesian Markov Chain Mote Carlo (MCMC; Huelsenbeck et al. 2000; Huelsenbeck and Rannala 2003; de Villemereuil et al. 2012); the Bayesian MCMC methods utilize phylogenetic trees sampled from posterior tree set obtained from Bayesian phylogenetic tree estimation programs such as BEAST (Drummond and Rambaut 2007) and MrBayes (Ronquist and Huelsenbeck 2003).

Nonetheless, these methods are not always met with enthusiasm in the evolutionary biology community (cf. Pagel et al. 2004; Pagel and Meade 2006). Difficulties we see are 2-fold: (1) currently, few easy-to-use implementations for such Bayesian MCMC methods are widely available, at least, for regression-based PCMs (but see Hadfield 2010; de Villemereuil et al. 2012); and (2) even if implemented, Bayesian MCMC-based analysis may take a long time to process many phylogenetic trees (e.g., see figure 6 in de Villemereuil et al. 2012). More recently, Garamszegi and Mundry (2014) have proposed a readily implementable frequentist solution, which employs model averaging with Akaike information criterion (AIC) in PGLS incorporating many phylogenetic trees (see also Mahler et al. 2010). Such a method overcomes the aforementioned difficulties. However, Garamszegi and Mundry (2014) acknowledge the lack of theoretical basis for this proposal, and that theoretical

or simulation-based confirmation of their method is necessary.

Herein, we propose another solution to account for phylogenetic uncertainty. Our method is simple, generally applicable, and, what is more, it is fairly reliable and readily implementable (see below). Also, it is firmly based on missing data theory (reviewed in Little and Rubin 2002), and utilizes Rubin's rules, which have been proposed as a part of the multiple imputation procedure (Rubin 1987). Evolutionary biologists and ecologists have just recently recognized the usefulness of techniques based on missing data theory (reviewed in Nakagawa and Freckleton 2008; Nakagawa 2015). Also, the importance of these missing-data methods has been discussed in the phylogenetic comparative literature (e.g., Garamszegi and Moller 2011; de Villemereuil and Nakagawa 2014). Notably, multiple imputation has been successfully employed in a number of comparative studies to handle missing data (e.g., Fisher et al. 2003; Gonzalez-Suarez et al. 2012; Liker et al. 2014; Pollux et al. 2014). Yet, so far, nobody seems to have made a use of Rubin's rules to deal with phylogenetic uncertainty. We note that Martins' work (1996) is conceptually very similar to the proposed method in terms of incorporating uncertainty due to fitting "incorrect" trees (see below), but we do not advocate the use of randomly generated trees (Symonds 2002; see also Rangel et al. 2015).

Paterno et al. (2018) recently discussed three main sources of uncertainty which affect PCMs: (1) phylogenetic uncertainty, (2) species sampling uncertainty, which can be seen as a missing-data problem (because one can see unsampled species as missing data; Nakagawa and Freckleton 2008), and (3) data uncertainty, which include measurement error and within-species variation (see also Rangel et al. 2015; Cooper et al. 2016; Cornwell and Nakagawa 2017). Once we could show Rubin's rules can be used for accounting for phylogenetic uncertainty, there is a highly practical possibility that we could seamlessly combine multiple imputation with PCMs to handle missing trait data, thus, addressing species sampling uncertainty simultaneously. There are two ways of imputing missing phenotypic data. The one is that we directly use a phylogenetic correlation (variance-covariance) matrix in the multiple imputation process (e.g., Bruggeman et al. 2009; Goolsby et al. 2017; see below for more details). The other is that we employ (phylogenetic) eigenvectors from a phylogenetic correlation (or variance-covariance) matrix (Penone et al. 2014). These two approaches, surprisingly, have never been systematically compared in terms of performance in augmenting missing comparative data.

Below, we first describe Rubin's rules associated with multiple imputation, and explain the rationale and potential advantages of our proposed method. Then, we conduct two simulation studies: (1) using 12 phylogenetic trees covering different taxa, we compare the performance of our proposed method to other methods such as methods using only one phylogenetic tree and the AIC-based method; and (2) we test how the proposed method can perform with different degrees and types of missing data, when used with the two types of multiple imputation methods (i.e., the one using a phylogenetic correlation matrix and the other phylogenetic eigenvectors).

Multiple Imputation and Rubin's Rules

Multiple imputation is a three-step process: imputing data, analyzing imputed data and pooling results. In the first step, m copies of "complete" datasets are generated from an incomplete original dataset. Popular techniques for the imputation steps use EM/EMB (expectation maximization with bootstrap) and MCMC algorithms, both of which are implemented in R packages such as Amelia (Honaker et al. 2011), mice (van Buuren and Groothuis-Oudshoorn 2011) and mi (Su et al. 2011); for more details regarding the algorithms, see Schafer (1997), Enders (2010), and van Buuren (2012). In the second step (analysis), we run separate statistical analyses on m datasets. In the final step (pooling), we use Rubin's rules (see below) to aggregate m sets of results to produce parameter estimates along with their uncertainty (the overall process is illustrated in Fig. 1).

As an example of applying this three-step process to PCMs, let us first assume that we have complete data for species traits (i.e., ignoring top-left side of Fig. 1). Then, what remains missing is the "true phylogenetic tree"; note that this is the central reason for us using (a part of) multiple imputation to account for phylogenetic uncertainty. Currently, a standard approach to creating candidate trees is to use Bayesian phylogenetic methods, as mentioned above, such as BEAST and MrBayes, which yield a posterior distribution of phylogenetic trees (for a guidance on building phylogenetic trees, see Garamszegi and Gonzalez-Voyer 2014). Alternatively, we can use published Bayesian tree sets as in Jetz et al. (2012) for birds, and Arnold et al. (2010) for primates. We consider this tree generation stage as our imputation step (the first step). The second step can be conducted using any frequentist or Bayesian statistical procedures including PCMs, such as independent contrasts, PGLS, and phylogenetic mixed models. Say, we will run PGLS with m randomly sampled phylogenetic trees from a Bayesian posterior tree set, which will result in m sets of results. Then, by combining these result sets via Rubin's rules (the final step), we will have integrated phylogenetic uncertainty in our estimates from PGLS.

Rubin's rules are a set of formulas for combining multiple statistical results, and they are as follows (Rubin 1987). With m imputations, parameters (e.g., regression coefficients) can be estimated as the average $\bar{b}$ over the estimates yielded by each imputed dataset b_j:

$$\bar{b} = \frac{1}{m}\sum_{j=1}^{m} b_j. \quad (1)$$

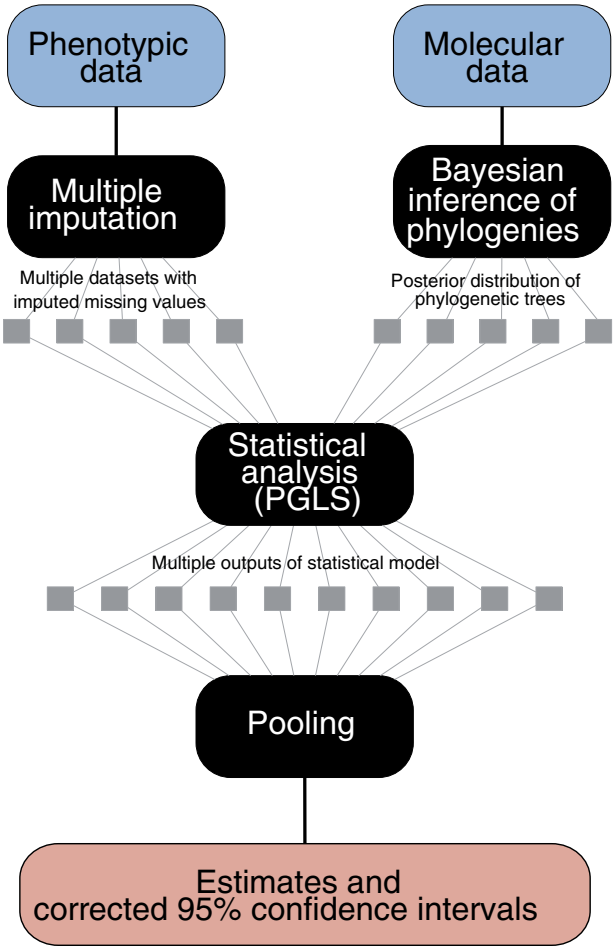

FIGURE 1. A conceptual diagram of our proposed method for accounting for phylogenetic and species sampling uncertainty in PCMs (e.g., PGLS) using multiple imputation and pooling (i.e., Rubin's rules).

The within-imputation variance (V_W) is calculated from the standard error (se) associated with b_j:

$$V_W = \frac{1}{m}\sum_{j=1}^{m} se_j^2. \quad (2)$$

The between-imputation variance (V_B) and the total variance (V_T: $\sqrt{V_T}$ is the overall standard error for $\bar{b}$) are as follows:

$$V_B = \frac{1}{m-1}\sum_{j=1}^{m}(b_j-\bar{b})^2, \quad (3)$$

$$V_T = V_W + V_B + \frac{V_B}{m}. \quad (4)$$

A quantity named the "fraction of missing information" and noted γ can be computed for each parameter:

$$\gamma = \left(1+\frac{1}{m}\right)\left(\frac{V_B}{V_T}\right). \quad (5)$$

This parameter conveniently represents the proportion of our parameter uncertainty that is arising from the multiple imputation process, that is, it represents the proportion of uncertainty due to using different trees.

From this parameter γ, we can obtain statistical significance and CI (with a particular α level) for our averaged point estimate, $\bar{b}$ based on t distributions with the degrees of freedom (ν) of the following:

$$\nu = (m-1)\frac{1}{\gamma^2}, \quad (6)$$

$$t_\nu = \frac{\bar{b}}{\sqrt{V_T}}, \quad (7)$$

$$100(1-\alpha)\%\ \text{CI} = \bar{b} \pm t_{\nu,(1-\alpha/2)}\sqrt{V_T}. \quad (8)$$

This computation, however, assumes a very large sample size, n (which is the length of data when no data are missing; Rubin and Schenker 1986; Rubin 1987). Barnard and Rubin (1999) proposed the following adjustment in the degrees of freedom (cf. Lipsitz et al. 2002):

$$\nu^* = \left(\frac{1}{\nu}+\frac{1}{\nu_{\text{obs}}}\right)^{-1}, \quad (9)$$

$$\nu_{\text{obs}} = (1-\gamma)\left(\frac{n-k+1}{n-k+3}\right)(n-k). \quad (10)$$

where ν^* is the corrected degrees of freedom corresponding to our averaged point estimate $\bar{b}$, especially suitable when sample size, n is small. The degrees of freedom, ν_{obs} denotes the observed degrees of freedom and k is the number of the parameters estimated in the model. In the next section, we will compare the performance of both ν (hereafter denoted "original df") and ν^* (hereafter denoted "corrected df").

Once we have an estimate of the corrected degrees of freedom, we can obtain a refined estimate of the fraction of missing information, γ^* for each parameter:

$$\gamma^* = \gamma + \frac{2}{(\nu^*+3)V_T}. \quad (11)$$

Finally, a useful measure of the multiple imputation process is named "relative efficiency" (ε), which represents the efficiency of the multiple imputation process, compared to the case of m being infinite. It ranges from 0 to 1 (1 being efficiency identical to a case with infinite m) and can be obtained as follows:

$$\varepsilon = \left(1+\frac{\gamma^*}{m}\right)^{-1}. \quad (12)$$

Relative efficiency represents the efficacy of multiple imputation process, compared to the case of m being infinite. In other words, this number can be used to assess how many imputations (m) are needed to account for uncertainty due to missing data. In our case, relative efficiency can indicate how many phylogenetic trees we should use for analysis (typically, the number of required

trees to account for phylogenetic uncertainty is chosen arbitrarily). Notably, to achieve fairly high relative efficiency, the required number of m is surprisingly low, even when the fraction of missing information is relatively large. For example, with $\gamma^* = 0.5$ and $m = 5$, relative efficiency is 90.91%, while it is 95.24% when $\gamma^* = 0.5$ and $m = 10$. Rubin's (1987) initial recommendation of m was low (3–10) probably due to computational limitation at that time, but current thinking is to use much larger m, aiming at over 99% relative efficiency (e.g., Graham et al. 2007; von Hippel 2009; Nakagawa 2015). We obtain a relative efficiency value (ε) for every parameter, and such values vary among parameters. For assessing efficiency of a model, we will use the relative efficiency ($\tilde{\varepsilon}$) that is obtained from the largest value of the fraction of missing information, following McKnight et al. (2007); that is:

$$\tilde{\varepsilon} = \left(1 + \frac{\max(\gamma*)}{m}\right)^{-1} = \min(\varepsilon) \qquad (13)$$

where the maximum and minimum are taken over all k parameters of the model. Another practical implication of the relative efficiency is that, for example, a low relative efficiency (due to a small m) would mean associated CI remain wider than a higher relative efficiency could achieve (i.e., larger m; this can be seen in the term, V_B/m in Equation 4). These equations are shown here mainly to illustrate the philosophy behind the process (again, summarized in Fig. 1). We can easily automate calculations involving the above formulae with currently available R packages for multiple imputation such as mice (reviewed in Nakagawa and Freckleton 2011; see also Penone et al. 2014).

Simulation Studies

Incorporating Phylogenetic Uncertainty as Missing Data

In order to assess the overall quality of our new method and compare it to existing ones, we performed a simulation study using 12 "maximum likelihood" trees extracted from TreeBASE (the number of tips ranging from 67 to 174; www.treebase.org, see Supplementary Table S1). We simulated datasets in which a variable y was linearly predicted from a variable x, with an intercept of 5 and a slope of 2. The error structure of this relationship was constrained by the phylogenetic tree chosen among the 12 trees (hereafter called the "true tree"), following a Brownian motion model. Different residual standard deviations were used ($\sigma = 2$, 5, 10, or 15). From the true tree, a distribution of trees was created by altering branch lengths and topology to artificially reproduce, while controlling precision, the kinds of tree variability obtained in a Bayesian posterior distribution of trees. To alter branch lengths, random noise drawn from a uniform distribution centered around 0 was added to the true value. The maximum level of that noise varied between 0% (no branch length noise), 10%, 20%, 40%, 70%, or 90% of the true branch length. To alter topology, we randomly "swapped" branches belonging to a focal clade to a sister clade. To choose the branch to swap, a tip was chosen at random, and a "threshold" was chosen from a uniform distribution with the thresholds of [0.1, 1]. The node just below this threshold in the path from the tip chosen to the root was swapped. We used several levels of topological noise (no swaps, i.e., no topological noise, or 1, 2, 5, 10, 20, 30 swaps in the tree). To construct the distribution of trees, the probability of each swap was set to 0.5. For each set of parameters (true tree, level of branch noise, level of topological noise), we constructed a distribution of 100 trees and replicated the analysis 100 times. This resulted in 2016 conditions, hence 201,600 different analyzes. Using the simulated phenotypes and tree distributions, we compared PGLS using the true tree or two types of consensus trees (majority rule or consensus), with both multiple PGLS with pooling of the results using AIC averaging (as in Garamszegi and Mundry 2014) and pooling with Rubin's rules as described above (either using the original degrees of freedom, df, or the corrected df as in Equation 9).

The accuracy of the intercept and slope were only slightly influenced by the different parameters (Table 1 and Supplementary Figs. S1–S3). On the contrary, the estimation of the residual standard deviation depended strongly on the method used (as well as, trivially, the true parameter sigma, and to a far lesser extent, all of the other parameters, see Table 1). Notably, the estimation of residual standard deviation was biased upward for the two methods using consensus trees (strict or majority rule, see Supplementary Figs. S1–S3).

The coverage of the CI for the slope was heavily influenced by the method used and more marginally by other parameters (except the true parameter sigma which had negligible influence, Table 1). The coverage was correctly calibrated when using the true tree (True PGLS, Fig. 2) and heavily mis-calibrated when using consensus trees (strict and majority rule consensus PGLS, Fig. 2). Accounting for uncertainty yielded better-calibrated coverages. AIC averaging was the closest to correct calibration. It was, however, slightly but consistently too liberal (Fig. 2). Using Rubin's rule yielded conservative coverages. Contrary to AIC averaging, the coverage was sensitive to the level of branch length and/or topological noise, decreasing when the noise increased (thus being even more conservative, Fig. 2).

In order to assess the behavior of the proposed method using Rubin's rules to account for phylogenetic uncertainty, we also conducted a study using different sample size for the trees (T = 10, 20, 50, or 100) and computed the relative efficiency as shown in Equation (13). This analysis revealed two interesting patterns (Fig. 3). First, no efficiency lower than 0.90 was recorded for a total of 806,400 simulated datasets, even for a sample size of trees as low as T = 10. Second, the relative

TABLE 1. Variance partitioning using a linear model to model the distribution of the inferred parameters, confidence interval coverage and efficiency

Parameter estimation	Model R^2	Parameter contribution to R^2					
		True tree	Method	Sigma	Branch length noise	Topology noise	Number of trees
Intercept	0.0075	0.51	0.018	0.29	0.062	0.12	–
Slope	0.007	0.8	0.041	0.027	0.11	0.026	–
Residual St. Dev.	0.79	0.043	0.3	0.66	0.00015	0.0017	–
CI coverage							
Slope	0.66	0.013	0.98	3.4×10^{-5}	0.0019	0.0055	–
Efficiency analysis							
Efficiency	0.71	0.023	–	1.9×10^{-7}	0.37	0.037	0.58

The total R^2 of the linear model is given, followed by the relative contribution (i.e. relative Pratt's measure; Pratt 1987) from each parameter to the total R^2. Relative contributions sum up to 1. "Number of trees"ž was available only for the study of efficiency.

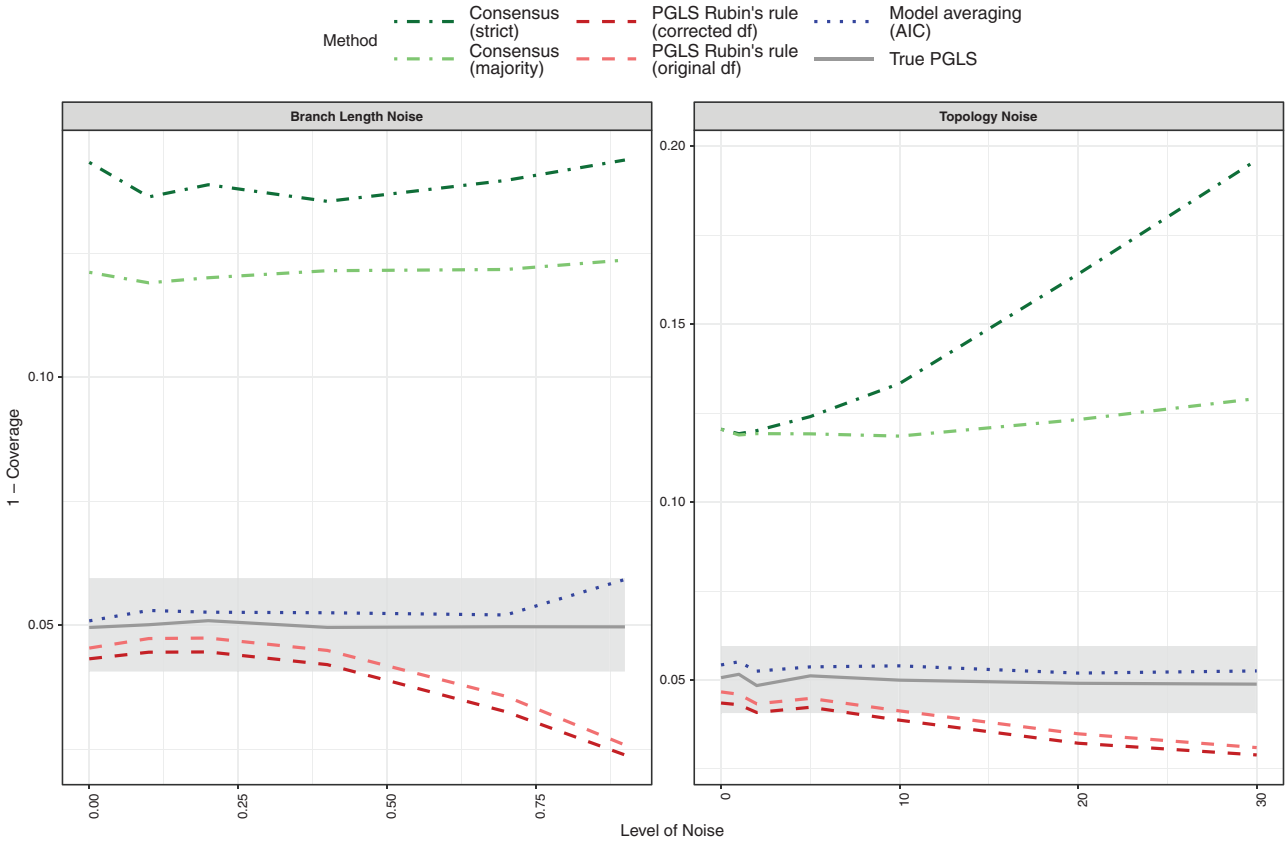

FIGURE 2. Complementary of the coverage (1—coverage) for 95% CI for the different estimation methods against the two types of noise (left: branch length noise, right: topological noise). Grey area is the zone of nonsignificance for a binomial test with a true probability of 0.05 (i.e., expected complementary coverage).

efficiency depended strongly on the number of trees used (Fig. 3 and Table 1). It also depended on the level of branch length noise, and to a lesser extent, on the level of topological noise (Fig. 3 and Table 1), as well as, even more marginally, on the nature of the true tree (Table 1). Third, in order to reach a relative efficiency over 0.99, on average, only 50 trees were necessary even with high levels of branch length and topological noise. With 100 trees, the relative efficiency was always over 0.99.

Incorporating Both Phylogenetic Uncertainty and Missing Trait Data

We then investigated the possibility to combine the ability of multiple imputation to account simultaneously for phylogenetic uncertainty and missing phenotypic values. To do so, we conducted a study with parameters fixed to the following values: the residual standard deviation σ was set to 5, the branch length noise to 20% and topological noise to 2 swaps. For simulated

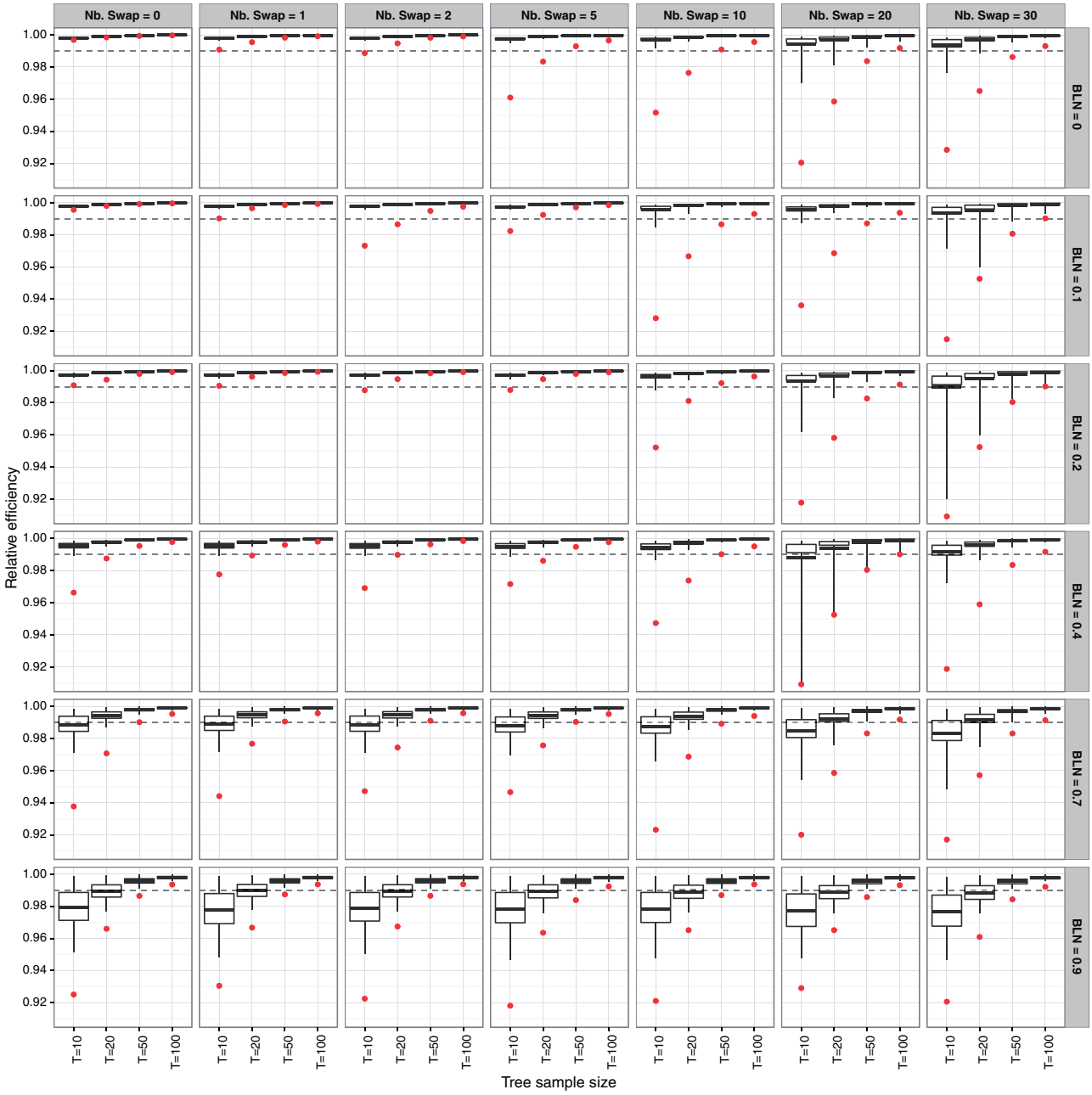

FIGURE 3. Relative efficiency distribution for different tree sample size (T) and different levels of branch length noise (BLN) and topological noise (Nb. Swap). The boxes depict the 50% interquantile interval, the whiskers depict the 95% interquantile interval, and the horizontal bar is the average estimate. The lower dot is the minimal relative efficiency yielded during the simulations.

data according to these parameters, we deleted records of phenotypic values at various proportions (10%, 30%, and 50%) and according to three mechanisms inspired from Penone et al. (2014): values were missing completely at random (MCAR), missing at random according to the environmental variable (MARvar) or missing at random according to the phylogeny (MARphylo). For more details of missing data mechanisms (e.g., MCAR, MAR), see Little and Rubin (2002, see also Nakagawa and Freckleton 2008;). The multiple imputation of the missing phenotypic values were handled using two different methods: on the one hand, we used an R implementation of the method PhyloPars (Bruggeman et al. 2009), called Rphylopars (Goolsby et al. 2017), to impute the missing values according to both the phylogeny and environmental (nonmissing) data (hereafter, the matrix method). On the other hand, we used the method described in Penone et al. (2014) using the information contained in phylogenetic eigenvectors (Diniz et al. 1998; see also Guenard et al. 2013) to impute the missing vales (hereafter, the eigenvector method).

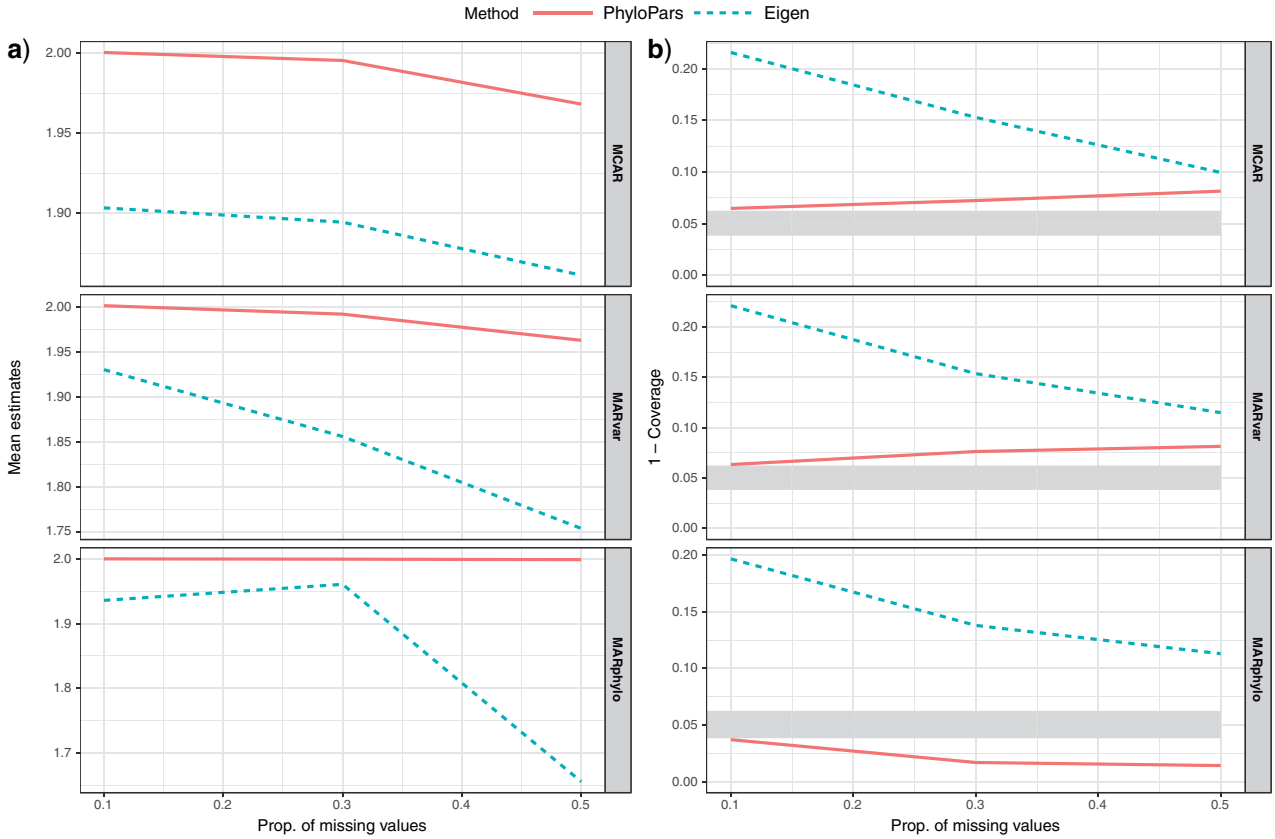

FIGURE 4. Estimate of the slope (a) and the coverage (1−coverage) of its associated CI (b) for the two methods of multiple imputation of missing phenotypic values (PhyloPars and Eigenvectors) according to the proportion of missing values in the data and mechanism of missing values: MCAR; MARvar, missing at random according to the environmental variable: MARphylo, missing at random according to the phylogeny. Grey area in B is the zone of nonsignificance for a binomial test with a true probability of 0.05 (i.e., expected complementary coverage).

The results of our simulations show that the matrix method (RphyloPars) yielded estimates with little bias (Fig. 4a, especially when missing values are missing according to the phylogeny, MARphylo), while using eigenvectors resulted in a stronger bias, strongly increasing with the proportion of missing values. Overall, the level of bias strongly depended on the characteristics of the true tree and the method used, and only slightly on the rate of missing values (Table 2). Coverage analysis of the CI (Fig. 4b) show that the matrix method is slightly too liberal when values are MCAR or missing at random according to the environmental variable (MARvar), but slightly conservative when they are missing at random according to the phylogeny (MARphylo). By contrast, the eigenvector method produced a coverage too liberal to be useful, although, interestingly, decreasing with the proportion of missing values. Overall, the coverage depended mostly the true tree and method used, and only marginally on the mechanism and rate of missing values (Table 2). The strong influence of the true tree on the estimate and its coverage is mainly driven by a strong instability of the eigenvector method regarding a particular tree (Tree #11 in Supplementary Fig. S4 and Table S1). Removing this tree from the analysis does not qualitatively impact the results shown in Fig. 4. However, this example makes an interesting point about the eigenvector method being potentially very sensitive to the nature of a phylogenetic tree.

DISCUSSION

The aim of this article is to introduce a simple and readily implementable method (i.e., Rubin's rubles) to account for phylogenetic uncertainty in PCMs. More practically, we explored the use of Rubin's rules simultaneously handling phylogenetic uncertainty and species sampling uncertainty (i.e., missing trait data; see Paterno et al. 2018). Via a simulation study using a simple PGLS, we compared the proposed method using Rubin's rules with other existing methods across different levels of branch length and topological noise, and we also assessed the number of trees required to accurately account for phylogenetic uncertainty. Further, we tested the practicality of our method to handle missing trait data under different imputation procedures and missing-data mechanisms. Four main results have emerged from our simulation study.

TABLE 2. Variance partitioning using a linear model to model the distribution of the inferred slope and confidence interval coverage in the simulation study on missing values

Parameter estimation	Model R^2	Parameter contribution to R^2			
		True tree	Method	Mechanism	Proportion of missing
Slope	0.39	0.41	0.43	0.0078	0.15
CI coverage					
Slope	0.65	0.33	0.59	0.026	0.056

The total R^2 of the linear model is given, followed by the relative contribution (i.e. relative Pratt's measure; Pratt 1987) from each parameter to the total R^2. Relative contributions sum up to 1.

First, in terms of error rate, methods ignoring phylogenetic uncertainty performed poorly and had a bad coverage for the slope CI. These findings are concordant with the previous work by de Villemereuil et al. (2012) comparing different methods. Both our proposed methods using Rubin's rule and the AIC-based method were much closer to the expected results using a PGLS with the true tree. Hence, using a consensus tree (either being a strict consensus or a majority rule based one) will yield too narrow CI, meaning that any test framework linked to it (e.g., slope significance testing) will yield an uncontrolled type I error rate.

The second main result is that the behavior of the methods accounting for phylogenetic uncertainty differed between them and depends on the level of phylogenetic noise in the tree distribution. Whereas the AIC-based method was consistently slightly too liberal, our proposed method using Rubin's rule was, by contrast, slightly conservative. The method assuming infinite sample size ("original df") was less conservative than the method correcting for small sample size ("corrected df"). This conservative behavior depended on the level of noise: our proposed method became more conservative as the level of phylogenetic noise increased. The AIC-based method was, on the contrary, less sensitive to the level of noise.

The third main result is that the number of phylogenetic trees needed to correct for phylogenetic uncertainty is surprisingly low. The required number of trees is far less than 1000 (as in Garamszegi and Mundry 2014), and probably less than 100 (as in de Villemereuil et al. 2012). It is likely to be a matter of dozens. In our simulation, sets of 50 randomly selected trees achieved almost always over 99% relative efficiency; in other words, using 50 trees should be almost as good as using an infinite number of trees. For low to medium levels of noise, even a sample size as low as 10 trees almost always yielded over 99% relative efficiency. As a whole, we recommend the use of over 50 phylogenetic trees in a PCM to account for phylogenetic uncertainty. However, for any given analysis and tree set, we recommend checking the number of trees needed to reach a relative efficiency of 99% (Nakagawa 2015). In practice, indeed, the required number of trees required to achieve high efficiency will strongly depend on the phenotypic data (e.g., phylogenetic signal), the complexity of the model and the variability in the tree estimates (e.g., strong topological and branch length uncertainty). We note that the statistical literature has discussed other criteria apart from the relative efficiency to determine how many imputations one requires (see Graham et al. 2007; White et al. 2011).

As mentioned, the AIC-based method (Garamszegi and Mundry 2014) accounted for phylogenetic uncertainty performed well, although with slightly liberal CIs. Therefore, the AIC-based method is definitely an option to correct for phylogenetic uncertainty. The method based on Rubin's rules (or multiple imputation), despite being slightly conservative, has the advantage of being a theoretical founded, yet simple method (we note that being conservative is probably preferred to being slightly liberal). This is, given that the imputation step is "proper," which is the case here as long as the trees come from a Bayesian posterior distribution and the estimates are maximum likelihood estimators (e.g., BEAST/PGLS combination; for the definition on proper multiple imputation, see Rubin 1987; Nielsen 2003). However, there is another clear benefit of using the proposed method.

This leads to our fourth point, that is, multiple imputation can simultaneously handle missing trait data (species sampling uncertainty) and phylogenetic uncertainty in a comparative dataset. Especially, using the matrix method (PhyloPars; Bruggeman et al. 2009; implemented as Rphylopars by Goolsby et al. 2017) to account for missing phenotypic values, while accounting for the phylogenetic uncertainty at the same time, yields estimate with little bias on the slope and almost calibrated coverage of the CI. Using the eigenvector method, as suggested in Penone et al. (2014) does not seem to yield satisfying results, however. The sensitivity of the matrix method (Rphylopars) to the rate and mechanism of missing data was relatively small, suggesting that the method should perform fairly well in many different circumstances. An exception to this is that when missing values are missing at random according to the phylogeny, the matrix method is slightly too conservative, while it is slightly too liberal for the two other missing-data mechanisms we tested here. Given the pervasive nature of missing data, we suggest multiple imputation may be useful for virtually every comparative dataset (Nakagawa and Freckleton 2008; Garamszegi and Moller 2011). Note that Rphylopars is intended to produce point estimate

of the missing phenotypic value with standard errors, which can be used to produce multiple imputation as we did. However, this process might not conserve all the properties of the multiple imputation model (e.g., it might slightly decreased covariance between species in the multiple imputation). Work is being conducted on a more proper multiple imputation method using a matrix method for missing values in the context of phylogenetic comparative analysis (S. Blomberg, personal communication, see also the package in development at https://github.com/pdrhlik/phylomice). We provide implementations of our method using R (https://github.com/devillemereuil/SimulTrees).

It is notable that the procedure known as "data augmentation" can also be used for dealing with missing data instead of multiple imputation. The term data augmentation is used in a number of ways in the statistical literature, but here we follow the usage by McKnight et al. (2007); that is, in this procedure, uncertainty of missing data is incorporated into parameter estimates during analysis (see the original usage of this term as in Tanner and Wing 1987). A data augmentation procedure is implemented, for instance, in MCMCglmm (Hadfield 2010). However, there is one disadvantage to data augmentation, which does not affect multiple imputation. Data augmentation assumes the use of just identified or overidentified models (Enders and Bandalos 2001; Enders 2010). That is, a particular model (for imputation) includes enough or more predictor variables, so that missing values can be recovered accurately from these predictors. In contrast, because multiple imputation separates the steps of data imputation and analysis, we do not need to clutter a statistical model for analysis (i.e., the analysis step) with many variables, which assist in recovering missing values (known as auxiliary variables; Enders 2010; Nakagawa 2015). Technically speaking, auxiliary variables are supported to make missing values to fulfill the assumption of missing at random, MAR (Little and Rubin 2002). In a multiple imputation procedure, we need add auxiliary variables only to a statistical model for imputation (i.e., the imputation step). For example, known data on species body size can be used during the imputation step to help impute missing data on species longevity, given the strong correlation between the two. However, because multiple imputation separates imputation and analysis, body size does not need to be a part of the final model. The use of multiple imputation probably has wider applications over data augmentation. Most importantly, to integrate phylogenetic uncertainty in a comparative dataset with missing data, one just needs to conduct extra imputations—for example, more m as in Equations (1–4)—to include the adequate number of trees, which can be measured by the efficiency index as in Equation (13).

Another notable point is that although we focused on the application of Rubin's rules on PGLS in this paper, its application clearly goes beyond trait evolution models like PGLS. The use of Rubin's rules should also be useful for models investigating lineage diversification (O'Meara 2012; Pennell and Harmon 2013). For example, one should be able to incorporate phylogenetic uncertainty into the estimation of birth–death (speciation–extinction) parameters from phylogenetic trees. Likewise, one could integrate both phylogenetic and species sampling uncertainty into such parameter estimation from trait-dependent diversification models.

In conclusion, the method using Rubin's rules is readily usable for all comparative biologists. Clearly, the use of multiple imputation used with the matrix method is extremely useful not only for imputing missing trait data, but also for integrating phylogenetic uncertainty, even simultaneously, as we have shown above. We expect such a simultaneous use of these two aspects of multiple imputation to be common in phylogenetic comparative analyses in the near future.

Funding

S.N. was supported by an ARC Future Fellowship (FT130100268) and an ARC Discovery grant (DP180100818).

Acknowledgments

We thank Will Cornwell, Travis Ingram, Losia Lagisz, Alistair Senior, and Simon Blomberg for comments, which have improved the manuscript. We also thank Eric Goolsby who provided help with Rphylopars.

References

Arnold C., Matthews L.J., Nunn C.L. 2010. The 10k trees website: a new online resource for primate phylogeny. Evol. Anthropol. 19:114–118.
Barnard J., Rubin D.B. 1999. Small-sample degrees of freedom with multiple imputation. Biometrika 86:948–955.
Bruggeman J., Heringa J., Brandt B.W. 2009. Phylopars: estimation of missing parameter values using phylogeny. Nucleic Acids Res. 37:W179–W184.
Cooper N, Thomas G.H., FitzJohn R.G. 2016. Shedding light on the "dark side" of phylogenetic comparative methods. Methods Ecol. Evol. 7:693–699.
Cornwell W, Nakagawa S. 2017. Phylogenetic comparative methods. Curr. Biol. 27:R333–R336.
de Villemereuil P., Nakagawa S. 2014. General quantitative genetic methods for comparative biology. In: Garamszegi L.Z., editor. Modern phylogenetic comparative methods and their application in evolutionary biology. Berlin, Heidelberg: Springer. p. 287–303.
de Villemereuil P., Wells J.A., Edwards R.D., Blomberg S.P. 2012. Bayesian models for comparative analysis integrating phylogenetic uncertainty. BMC Evol. Biol. 12.
Diaz-Uriarte R., Garland T. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Syst. Biol. 45:27–47.
Diniz J.A.F., De Sant'ana C.E.R., Bini L.M. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution, 52:1247–1262.
Drummond A.J., Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7.
Enders C.K. 2010. Applied missing data analysis. New York: Guilford Press.
Enders C.K., Bandalos D.L. 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Struct. Equ. Modeling 8:430–457.
Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Fisher D.O., Blomberg S.P., Owens I.P.F. 2003. Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proc. R. Soc. Lond. B 270:1801–1808.
Garamszegi L., Gonzalez-Voyer A. 2014. Working with the tree of life in comparative studies: how to build and tailor phylogenies to interspecific datasets. In: Garamszegi L.Z., editor. Modern phylogenetic comparative methods and their application in evolutionary biology. Berlin, Heidelberg: Springer, p. 19–48.
Garamszegi L., Mundry R. 2014. Multimodel-inference in comparative analyses. In: Garamszegi L.Z., editor. Modern phylogenetic comparative methods and their application in evolutionary biology. Berlin, Heidelberg: Springer, p. 305–331.
Garamszegi L.Z. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology. New York: Springer, p. pages cm.
Garamszegi L.Z., Moller A.P. 2011. Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst. Biol. 60:876–880.
Gonzalez-Suarez M., Lucas P.M., Revilla E. 2012. Biases in comparative analyses of extinction risk: mind the gap. J. Anim. Ecol. 81:1211–1222.
Goolsby E.W., Bruggeman J., Ane C. 2017. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. Methods Ecol. Evol. 8:22–27.
Grafen A. 1989. The phylogenetic regression. Philos. T. Roy. Soc. B 326:119–157.
Graham J.W., Olchowski A.E., Gilreath T.D. 2007. How many imputations are really needed? – some practical clarifications of multiple imputation theory. Prev. Sci. 8:206–213.
Guenard G., Legendre P., Peres-Neto P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. Methods Ecol. Evol. 4:1120–1131.
Hadfield J. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J. Stat. Softw. 33:1–22.
Hadfield J.D., Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J. Evol. Biol. 23:494–508.
Honaker J., King G., Blackwell M. 2011. Amelia ii: a program for missing data. J. Stat. Softw. 45:1–47.
Housworth E.A., Martins E.P. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. Syst. Biol. 50:628–639.
Huelsenbeck J.P., Rannala B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57:1237–1247.
Huelsenbeck J.P., Rannala B., Masly J.P. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288:2349–2350.
Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature 491:444–448.
Liker A., Freckleton R.P., Szekely T. 2014. Divorce and infidelity are associated with skewed adult sex ratios in birds. Curr. Biol. 24:880–884.
Lipsitz S.R., Parzen M., Zhao L.P. 2002. A degrees-of-freedom approximation in multiple imputation. J. Stat. Comput. Sim. 72:309–318.
Little R.J.A., Rubin D.B. 2002. Statistical analysis with missing data. 2nd ed. Hoboken (NJ): Wiley.
Losos J.B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. Syst. Biol. 43:117–123.
Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.
Martins E.P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50:12–22.
McKnight P.E., McKnight K.M., Sidani S., Figueredo A.J. 2007. Missing data: a gentle introduction. New York (NY): The Guilford Press.
Nakagawa S. 2015. Missing data: mechanisms, methods and messages In: Fox G.A., Negrete-Yankelevich S., Sosa V.J., editors. Ecological statistics. Oxford: Oxford University Press. p. 81–105.
Nakagawa S., Freckleton R. 2011. Model averaging, missing data and multiple imputation: a case study for behavioural ecology. Behav. Ecol. Sociobiol. 65:103–116.
Nakagawa S., Freckleton R.P. 2008. Missing inaction: the dangers of ignoring missing data. Trends Ecol. Evol. 23:592–596.
Nielsen S.F. 2003. Proper and improper multiple imputation. Int. Stat. Rev. 71:593–607.
O'Meara B.C. 2012. Evolutionary inferences from phylogenies: a review of methods. Annu. Rev. Ecol. Evol. Syst. 43:267–285.
Pagel M., Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167:808–825.
Pagel M., Meade A., Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. 53:673–684.
Paterno G.B., Penone C., Werner G.D. 2018. Sensiphy: an R-package for sensitivity analysis in phylogenetic comparative methods. Methods Ecol. Evol. 9:1461–1467.
Pennell M.W., Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. N. Y. Acad. Sci. 1289: 90–105.
Penone C., Davidson A.D., Shoemaker K.T., Di Marco M., Rondinini C., Brooks T.M., Young B.E., Graham C.H., Costa G.C. 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol. Evol. 5:961–970.
Pollux B.J.A., Meredith R.W., Springer M.S., Garland T., Reznick D.N. 2014. The evolution of the placenta drives a shift in sexual selection in livebearing fish. Nature 513:233–236.
Pratt J.W. 1987. Dividing the indivisible: using simple symmetry to partition variance explained. Proceedings of the Second International Tampere Conference in Statistics; 1987; Department of Mathematical Sciences, University of Tampere. p. 245–260.
Rangel T.F., Colwell R.K., Graves G.R., Fucikova K., Rahbek C., Diniz J.A.F. 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution 69:1301–1312.
Ronquist F, Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Rubin D.B. 1987. Multiple imputation for nonresponse in surveys. New York (NY): J. Wiley & Sons.
Rubin D.B., Schenker N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Am. Stat. Assoc. 81:366–374.
Schafer J.L. 1997. Analysis of incomplete multivariate data. London: Chapman & Hall.
Stone E.A. 2011. Why the phylogenetic regression appears robust to tree misspecification. Syst. Biol. 60:245–260.
Su Y.S., Gelman A., Hill J., Yajima M. 2011. Multiple imputation with diagnostics (mi) in r: opening windows into the black box. J. Stat. Softw. 45:1–31.
Symonds M.R.E. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst. Biol. 51:541–553.
Tanner M.A., Wing H.W. 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82:528–540.
van Buuren S. 2012. Flexible imputation of missing data. Boca Raton (FL): CRC Press.
van Buuren S., Groothuis-Oudshoorn K. 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45:1–67.
von Hippel P.T. 2009. How to impute interactions, squares and other transformed variables. Sociol Methodol. 39:265–291.
White I.R., Royston P., Wood A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30:377–399.