Exploring Dependent Variable Selection in Phylogenetic Eigenvector Regression

Zheng-Lin Chen, Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, 100875, China

*Correspondence author:
Deng-Ke Niu, College of Life Sciences, Beijing Normal University, Beijing, 100875, China.
dkniu@bnu.edu.cn, dengkeniu@hotmail.com

ORCID:
Deng-Ke Niu, https://orcid.org/0000-0003-4503-2658

Abstract

Phylogenetic autocorrelation among samples is a critical problem in evolutionary analysis across multiple species. To address this challenge, methods like phylogenetic eigenvector regression (PVR) have been developed. This study investigates the impact of variable selection on PVR outcomes in evolutionary biology research. Through simulations and model comparisons, we explore how different selections of dependent and independent variables influence phylogenetic eigenvector selection and correlation outcomes. Our results reveal significant discrepancies in phylogenetic eigenvector selection and correlation outcomes when swapping dependent and independent variables. R^2 has been identified as an effective criterion for selecting variables in PVR analysis. Additionally, we compare the performance of three non-sequential phylogenetic eigenvector selection methods. The first is to select phylogenetic eigenvectors significantly related to the response variable. The second and third methods involve forward stepwise regression, and their model fit is assessed by minimizing the Akaike Information Criterion value and by minimizing Moran's I of model residuals, respectively. We find no significant difference between the first and third methods, but both are superior over the second method. Overall, our findings highlight the critical role of variable selection, particularly based on R^2, in enhancing the efficacy of PVR analysis and improving correlation analyses in evolutionary biology.

Keywords

Phylogenetic autocorrelation, spatial autocorrelation, phylogenetic eigenvector regression (PVR), dependent and independent variable selection, correlation analysis, linear regression model,

phylogenetic signal, R^2, forward stepwise regression

Introduction

In ecological data analysis, a commonly observed and intriguing phenomenon is the tendency for samples closer in distance to exhibit more similar traits. This positive spatial autocorrelation, where adjacent observations show similar data values, is often visually represented on maps (Legendre 1993; Cottenie 2005; Diniz-Filho and Bini 2005; Dale and Fortin 2009). Similarly, in studies of biological evolution involving cross-species data, researchers highlight a phenomenon called phylogenetic autocorrelation. This phenomenon occurs when species that are closer on a phylogenetic tree tend to share more similar features (Cheverud et al. 1985; Felsenstein 1985). Such forms of sample autocorrelation, whether originating from spatial or phylogenetic relationships, challenge the assumptions of conventional statistical methods. These methods typically assume independence among observations or data points. Failure to consider sample autocorrelation and solely relying on conventional statistical approaches when analyzing such data may lead to inaccurate conclusions (Felsenstein 1985; Garamszegi 2014).

To address the issue of interdependence among samples, various methods have been proposed (Felsenstein 1985; Grafen 1989; Pagel 1999; Keitt et al. 2002; Dormann et al. 2007; Peres-Neto and Legendre 2010; Garamszegi 2014). While some of these methods may have been initially designed for one specific fields, they share underlying principles that can be applied across various disciplines. Originally, researchers sought to tackle phylogenetic autocorrelation among samples using autoregressive models (Cheverud et al. 1985). Autoregressive models decompose studied traits into components influenced by phylogeny (P) and components reflecting specific independent evolution (S) (Harvey et al. 1995; Westoby et al. 1995; Desdevises et al. 2003), mirroring the approach of autoregressive models in spatial statistics for addressing spatial autocorrelation (Doreian 1980; Dow et al. 1982; He et al. 2003). Phylogenetic eigenvector regression (PVR) builds upon Cheverud's model and is widely employed in evolutionary biology (Diniz-Filho et al. 1998). PVR entails principal coordinate analysis on a double-centered phylogenetic matrix ($n \times n$), generating $n - 1$ orthogonal phylogenetic eigenvectors. These eigenvectors represent species' positions on the phylogenetic tree and can be integrated as covariates into the original model, such as ordinary linear regression models (Desdevises et al. 2003; Diniz-Filho et al. 2011; Peres-Neto et al. 2012). PVR's adaptability in incorporating covariates to address phylogenetic

autocorrelation enables seamless integration with various statistical techniques, thereby extending the applicability of conventional statistical approaches to a phylogenetic framework. This adaptability has led to widespread use of PVR in evolutionary analyses, including evaluating the contribution of phylogenetic structure to trait variance (Diniz-Filho and Nabout 2009; LeRoy et al. 2020; Gao et al. 2023), examining non-stationarity in trait evolution rates (Diniz-Filho et al. 2019), exploring trait correlations (Diniz-Filho et al. 2009), and imputing missing trait data within a phylogenetic context (Debastiani et al. 2021).

While the PVR method offers flexibility and validity, researchers have pinpointed several associated issues, with the main concern revolving around the selection of phylogenetic eigenvectors. Theoretically, including all phylogenetic eigenvectors as covariates in the original model is essential to fully encompass phylogenetic information. However, this approach may lead to the complete explanation of variation in the dependent variable Y solely by phylogenetic information, hindering the detection of relationships between traits X and Y (Rohlf 2001). Conversely, omitting certain phylogenetic eigenvectors from the model results in the loss of pertinent phylogenetic tree information. Fortunately, studies have indicated that not all phylogenetic eigenvectors are indispensable, as some contribute minimally to the analysis. A judiciously selected subset of phylogenetic eigenvectors is often adequate to capture the requisite phylogenetic information (Peres-Neto et al. 2005; Griffith and Peres-Neto 2006; Diniz-Filho et al. 2009; Diniz-Filho et al. 2012). The authors of such studies have proposed several methods for selecting phylogenetic eigenvectors, including sequential selection and three non-sequential methods: 1) choosing phylogenetic eigenvectors significantly correlated with the trait under investigation, 2) utilizing stepwise regression to identify a set of phylogenetic eigenvectors minimizing the Akaike Information Criterion (AIC) value, and 3) selecting a set of phylogenetic eigenvectors minimizing the autocorrelation of residuals (S-component). Comparative analyses of these methods have demonstrated that selecting phylogenetic eigenvectors to minimize the autocorrelation of residuals (S-component) effectively reduces the autocorrelation of residuals while necessitating only a small number of phylogenetic eigenvectors. Consequently, it is recommended to use the set of phylogenetic eigenvectors minimizing the autocorrelation of residuals. This approach is commonly adopted in subsequent research as well (Burns et al. 2015; Divíšek et al. 2018; Lanuza et al. 2023).

Effectively selecting phylogenetic eigenvectors can significantly improve the performance of PVR, thereby enhancing the method's efficacy. However, another critical issue arises in eigenvector selection concerning the designation of dependent and

independent variables. Previous research on phylogenetic generalized least squares has highlighted that improper designation of these variables can lead to substantial statistical errors (Chen et al. 2023). When analyzing the correlation between two variables, with one choice of independent and dependent variables, traits may appear significantly correlated, whereas with another selection, they may seem unrelated. Alternatively, traits may exhibit positive correlation under one configuration and negative correlation under another. These potential scenarios represent mutually exclusive outcomes of the model's correlation results. We are concerned that in PVR analysis, variations in the designation of dependent and independent variables may result in different eigenvector selections, consequently influencing subsequent tests of trait correlation and potentially leading to significant errors.

We constructed simulated data sets (X_1 and X_2) to select phylogenetic eigenvectors, comparing differences between different selections of independent and dependent variables and evaluating their influence on accurately predicting trait correlations. Our analysis revealed conflicting correlation results due to different selections of dependent and independent variables, exhibit a significant occurrence. To accurately depict evolutionary correlations, we established a gold standard based on traits' changes per unit length along the phylogenetic tree branches. Comparing models based on R^2, phylogenetic signal strength, and AIC values, we found that larger R^2 models provided more accurate predictions. After determining variables based on R^2, we examined various methods for selecting phylogenetic eigenvectors' effects on PVR model fit.

Materials and Methods

Simulated data

Firstly, we constructed a phylogenetic tree comprising 100 species (n = 100) and simulated the evolution of two traits (X_1 and X_2) on this tree. The relationship between the traits is defined as $X_2 = X_1 + \varepsilon$, where ε represents the noise term. By manipulating the variance of the noise term ε, we adjusted the strength of the correlation between X_1 and X_2. Our simulation design draws inspiration from Revell (2010) and encompasses four distinct scenarios:

1. In the "BM & BM + Norm" scenario, trait X_1 evolves along the phylogenetic tree following a Brownian motion model (variance set to 1), while ε follows a normal distribution with a mean of 0 and a variance of d. Values of d used are 10^{-4}, 0.01, 1, 4, 16, 64, 256, and 1024.

2. In the "Norm & Norm + BM" scenario, trait X_1 evolves under a normal distribution with a mean of 0 and a variance of 1, while ε evolves along the phylogenetic tree following a Brownian motion model with a variance of d. The same range of d values is utilized: 10^{-4}, 0.01, 1, 4, 16, 64, 256, and 1024.

3. In the "BM1 & BM1 + BM2" scenario, trait X_1 evolves according to a Brownian motion model (variance set to 1) on the phylogenetic tree, while ε evolves under a Brownian motion model with a variance of d. Again, the values of d range from 10^{-4} to 1024.

4. The "Norm1 & Norm1 + Norm2" scenario involves trait X_1 following a normal distribution with a mean of 0 and a variance of 1, while ε follows a normal distribution with a mean of 0 and a variance of d. The values of d used are consistent with the previous scenarios: 10^{-4}, 0.01, 1, 4, 16, 64, 256, and 1024.

For each value of d within each scenario, we conducted 1000 repeated simulations, resulting in a total of 3.2×10^4 simulations.

Phylogenetic eigenvectors selection methods

We begin with Cheverud's autoregressive model (Cheverud et al. 1985), expressed as $T = P + S$, where T represents the trait under study, P represents the portion influenced by phylogenetic relationships, and S represents the portion influenced by other traits. The PVR method first conducts principal coordinate analysis on the double-centered phylogenetic matrix ($n \times n$), decomposing it into $n - 1$ phylogenetic eigenvectors. Then, k (usually, $k < n - 1$) eigenvectors are chosen based on the methods described in the following paragraphs and added as covariates to the original model. By substituting the dependent variable Y and the phylogenetic eigenvectors V for T and P, respectively, we can derive the ordinary linear model, $Y = V\beta + \varepsilon$, here Y represents the trait as the dependent variable, V represents the selected k phylogenetic eigenvectors, β represents the regression coefficients corresponding to the k phylogenetic eigenvectors, and ε represents S, which is the part explained by another explanatory variable.

Diniz-Filho et al. (2012) proposed two approaches for selecting phylogenetic eigenvectors: the sequential approach and the non-sequential approach. In the sequential approach, the $n - 1$ phylogenetic eigenvectors obtained from the decomposition of the phylogenetic matrix are initially sorted based on their eigenvalues from largest to smallest (EV_1, EV_2, ..., EV_{n-1}). Then, new eigenvectors are successively added to the selected set of

phylogenetic eigenvectors in descending order until the Moran's I value of ε equals 0, indicating no significant autocorrelation. Moran's I, originally a measure of spatial autocorrelation, has been adapted to measure phylogenetic autocorrelation, ranging from −1 (complete negative correlation) to +1 (complete positive correlation), with 0 indicating no phylogenetic autocorrelation, thus representing randomness.

The non-sequential approach comprises three methods. The first method involves selecting phylogenetic eigenvectors significantly related with the response variable (ESRRV). Diniz-Filho et al. identified significant linearly related phylogenetic eigenvectors through regression models without normality tests for variables (Diniz-Filho and Nabout 2009; Diniz-Filho et al. 2012). In this study, normality tests are first performed on both the phylogenetic eigenvectors and the response variable. If both conform to a normal distribution, Pearson correlation tests are employed; otherwise, Spearman rank correlation tests are utilized. The second and third methods involve stepwise regression, including forward, backward, and bidirectional selections. In this study, we utilize forward selection, beginning with fewer variables and sequentially adding one at a time until the model fit no longer improves with the addition of a new variable. In the second method, model fit is assessed by minimizing the AIC, termed mAIC. In the third method, model fit is evaluated by minimizing the Moran's I of model residuals, termed mMorI. The third method aligns with the sequential selection method in its aim but does not necessarily require the selected phylogenetic eigenvectors to be continuous. The sequential approach is essentially a special case of the third method. Therefore, we concentrated on these three non-sequential selection methods.

The selected phylogenetic eigenvectors, obtained through different methods, are incorporated as covariates into a simple linear regression model: $y = \sum_{i=1}^{k} a_i * ev_i + b * x + \epsilon$, where $\{ev_1, ev_2, \dots, ev_k\}$ represents the set of selected phylogenetic eigenvectors. Here, y denotes the dependent variable trait, x denotes the independent variable trait, and ϵ represents the residual term. The correlation between x and y can be inferred based on the sign and *p*-value of the regression coefficient b (significantly positive, significantly negative, or not significant).

Gold Standard for trait correlations in each simulation

In our study, we simulated the evolution of traits along a phylogenetic tree, obtaining the traits values at each node, Subsequently, we calculated the changes $\Delta X_1/L$ and $\Delta X_2/L$ per

unit time on each branch, where L represents the length of the corresponding branch. The changes on each branch are independent, allowing us to utilize conventional correlation tests to analyze the correlation between $\Delta X_1/L$ and $\Delta X_2/L$. If both $\Delta X_1/L$ and $\Delta X_2/L$ follow a normal distribution, we will perform Pearson correlation analysis; otherwise, we utilize the Spearman rank test.

By examining the relationship between $\Delta X_1/L$ and $\Delta X_2/L$, we capture the true evolutionary relationship between the two traits X_1 and X_2. Consequently, we can assess the model's performance by comparing the predicted trait correlations from various models with the actual evolutionary relationship between X_1 and X_2.

Results

For convenience, we denote V_{X_1} as the set of phylogenetic eigenvectors selected when X_1 is the dependent variable, and V_{X_2} as the set when X_2 is the dependent variable. $OLS_{X_1 \sim X_2 + V_{X_1}}$ represents the ordinary least squares linear regression model with X_1 as the dependent variable, X_2 as the independent variable, and V_{X_1} as the covariates. Similarly, $OLS_{X_2 \sim X_1 + V_{X_2}}$ denotes the model with X_2 as the dependent variable, X_1 as the independent variable, and V_{X_2} as the covariates. Both $OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$ are used to assess the relationship between X_1 and X_2.

Impact of dependent and independent variable selection on phylogenetic eigenvector selection

For a given set of simulated data X_1 and X_2, we derived two distinct sets of phylogenetic eigenvectors, V_{X_1} and V_{X_2}. When these sets differ, it suggests that the choice of dependent and independent variables influences the selection of phylogenetic eigenvectors. We analyze the frequency of such discrepancies when switching dependent and independent variables across simulations (Figure 1). With each set of simulated data comprising 8000 results, Figure 1 reveals that over half of the selected phylogenetic eigenvector sets differ when swapping dependent and independent variables. Specifically, the proportion of discrepancies in phylogenetic eigenvector selection by the ESRRV method is 81.6%, while for the mMorI method, it stands at 70.6%. Notably, the mAIC method exhibits the highest discrepancy rate at 93.1%. These findings underscore the significant impact of variable selection on phylogenetic eigenvector choice.

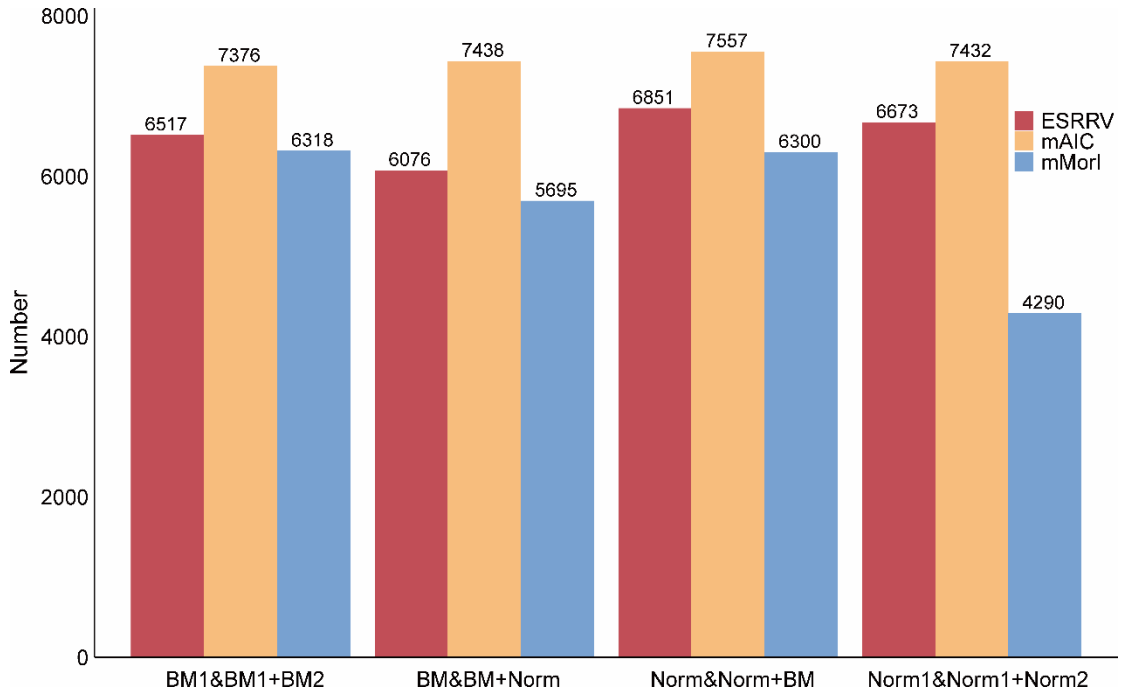

Figure 1. Dependence of phylogenetic eigenvector selection on variable exchange. This figure illustrates the count of divergent selections of phylogenetic eigenvector sets resulting from interchanging dependent and independent variables (denoted as V_{X_1} and V_{X_2}) across various phylogenetic eigenvector selection methodologies. Each scenerio encompasses 8000 simulations. Please refer to the Materials and Methods section for details regarding the three phylogenetic eigenvector selection methods (ESRRV, mAIC, and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Impact of dependent and independent variable selection on correlation results

The above findings reveal that across all four simulation scenarios (BM&BM+Norm, Norm&Norm+BM, BM1&BM1+BM2, and Norm1&Norm1+Norm2), the three phylogenetic eigenvector selection methods (ESRRV, mMorI, and mAIC) may produce disparate sets of phylogenetic eigenvectors due to variations in the selection of dependent variables. Consequently, this variability can lead to conflicting correlation outcomes. Across the 32,000 simulations, employing the ESRRV method resulted in 3935 instances of inconsistent correlation results for PVR analyses when switching dependent and independent variables. On the other hand, when using the mAIC method, there were 2393 instances of inconsistent correlation results, and with the mMorI method, there were 2108 instances of inconsistency. Notably, the proportion of conflicting correlation results attributed to the selection of dependent variables is approximately 10%, which is considerable. This underscores the significance of variable selection in correlation analysis. Should the selection of dependent

and independent variables influence the detection of trait correlations without researchers' awareness, the resulting correlation conclusions could be unreliable.

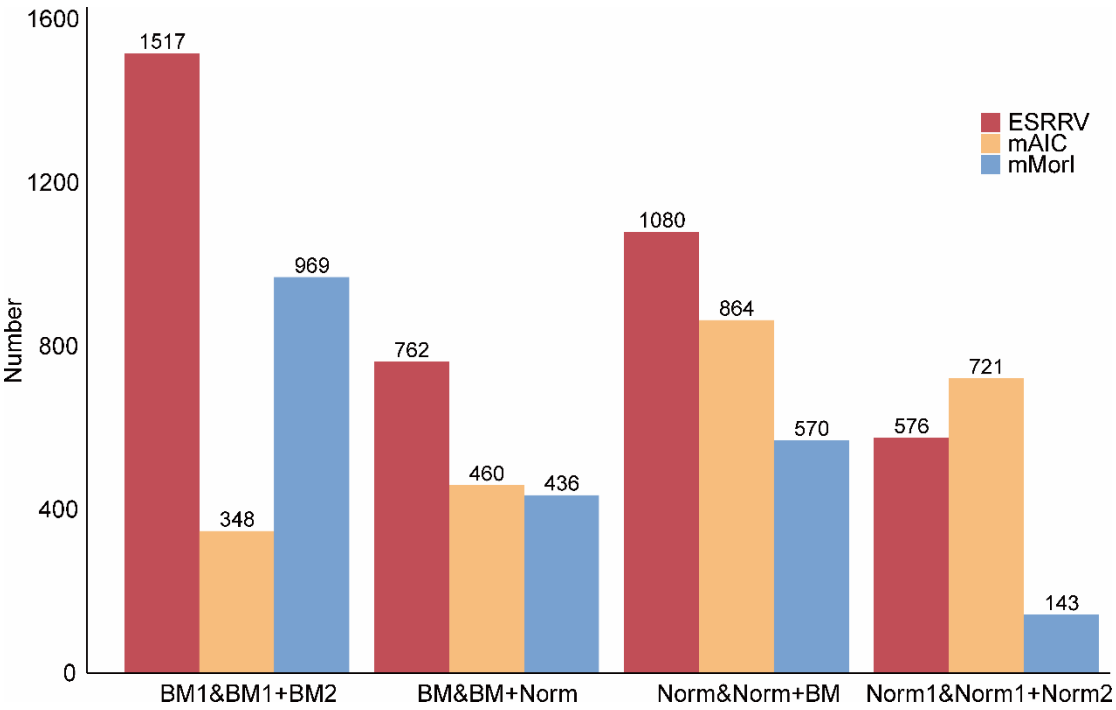

Figure 2. Impact of variable exchange on correlation results in PVR models. The figure depicts the number of conflicting correlation results resulting from switching dependent and independent variables in PVR models. This switch leads to inconsistent sets of phylogenetic eigenvectors used in PVR analysis, thus affecting the detection of trait correlations within PVR models. Conflicting correlation results occur when switching dependent and independent variables, leading to either one significant positive and one negative correlations, or one significant correlation and the other nonsignificant. Each scenarios encompasses 8000 simulated results. Please refer to the Materials and Methods section for details regarding the three phylogenetic eigenvector selection methods (ESRRV, mAIC, and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Impact of dependent/independent variable selection and phylogenetic eigenvector selection methods on PVR model accuracy

We first established a gold standard for model evaluation by calculating the unit change values $\Delta X_1/L$ and $\Delta X_2/L$ for each branch of the phylogenetic trees, conduct correlation tests on them. The correlation test results are regarded as the true relationship between traits X_1 and X_2 (significant positive correlation, significant negative correlation, or nonsignificant correlation), which can be used to evaluate competing phylogenetic models.

Then, we conducted a thorough examination of correlation results obtained by swapping dependent and independent variables in the PVR models ($OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$).

We examined the frequencies of consistent and both correct results, consistent and both incorrect results, and instances where only one correlation result is accurate. Our findings demonstrate that when swapping dependent and independent variables, the ESRRV and mMorI methods consistently yield more accurate correlation results across all four simulation scenarios compared to the mAIC method (Figure 3). To address conflicting outcomes resulting from this switch, we will devise a criterion to identify the correct result for datasets lacking a gold standard, recognizing that empirical data often lack such a reference point. Here, we first define a best scenario and a worst scenario, representing the most accurate and the least accurate identification of correct results. Apart from instances categorized as both correct, all correct results in cases where only one correlation is accurate can be accurately selected in the best scenario. Conversely, no correct result in cases where only one correlation is accurate can be identified in the worst scenario. A comparison between the mAIC method's correct predictions in the best scenario and the ESRRV and mMorI methods' correct predictions in the worst scenario (Table 1) underscores the superior performance of the ESRRV and mMorI methods. Consequently, subsequent analyses, focusing on selecting accurate results in instances of inconsistent correlation outcomes and comparing post-optimization performance across different methods, primarily concentrate on the ESRRV and mMorI methods.

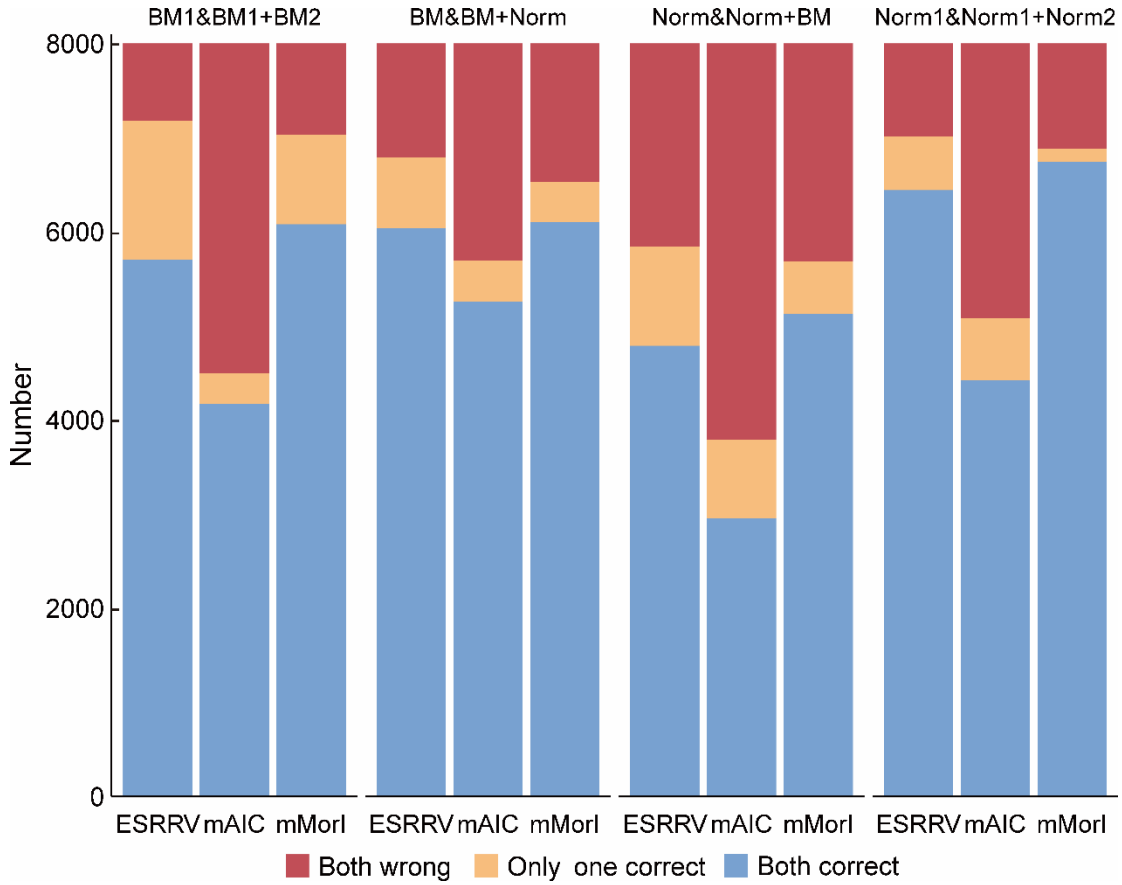

Figure 3. Distribution of correct and wrong results in the PVR models when swapping dependent and independent variables ($OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$). The figure depicts the count of correlation outcomes obtained from OLS regression models when the roles of dependent and independent variables are exchanged. Results are categorized as follows: "Both correct" denotes instances where both regression outcomes align with the gold standard. "Only one correct" indicates cases where only one of the regression results is consistent with the gold standard. "Both wrong" signifies scenarios where both regression outcomes deviate from the gold standard. Please refer to the Materials and Methods section for details regarding the three phylogenetic eigenvector selection methods (ESRRV, mAIC, and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Table 1. Comparison of the best scenario and the worst scenario of phylogenetic eigenvector selection methods in PVR accuracy[a]

	Method[b]	Correct	Incorrect	χ^2	p[c]
BM1 & BM1 + BM2	ESRRV, worst	5704	2296	391.52	$<2.2\times10^{-16}$
	mAIC, best	4500	3500		
	mMorI, worst	6081	1919	696.61	$<2.2\times10^{-16}$
	mAIC, best	4500	3500		
BM & BM + Norm	ESRRV, worst	6041	1959	37.841	7.675×10^{-10}

	mAIC, best	5696	2304		
	mMorI, worst	6102	1898	52.938	3.443×10^{-13}
	mAIC, best	5696	2304		
Norm & Norm + BM	ESRRV, worst	4790	3210	249.82	$<2.2\times10^{-16}$
	mAIC, best	3792	4208		
	mMorI, worst	5130	2870	452.91	$<2.2\times10^{-16}$
	mAIC, best	3792	4208		
Norm1 & Norm1 + Norm2	ESRRV, worst	6442	1558	570.59	$<2.2\times10^{-16}$
	mAIC, best	5085	2915		
	mMorI, worst	6743	1257	890.24	$<2.2\times10-^{16}$
	mAIC, best	5085	2915		

[a]When swapping dependent and independent variables in PVR models, three patterns emerge: consistent and both correct results, consistent and both incorrect results, and cases where only one correlation result is accurate. We define a best scenario and a worst scenario, representing the most accurate and the least accurate selection of correct results. In the best scenario, all correct results in cases where only one correlation is accurate can be accurately identified, except for those categorized as both correct. Conversely, no correct results in cases where only one correlation is accurate can be identified in the worst scenario. Please refer to the Materials and Methods section for details regarding the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

[b]ESRRV refers to the Eigenvectors Significantly Related with Response Variable method, mAIC denotes the stepwise regression with AIC minimization method, and mMorI represents the stepwise regression with residual Moran's I minimization method.

[c]By conducting a Chi-square test on the number of correct and incorrect results of the two methods, it can be determined whether there is a significant difference in performance between these two methods. The smallest significance *p*-value that can be obtained in the R environment is 2.2×10^{-16}.

Criteria for selecting dependent and independent variables in PVR analysis

Confronted with the conflicting outcomes arising from swapping dependent and independent variables, pinpointing the model that accurately reflects the real trait relationship becomes crucial, especially for empirical data, which frequently lack a gold standard. We propose three criteria to streamline this process:

1. **Phylogenetic Signal Strength (λ-Criterion)**: Prioritizing dependent variables and their corresponding models based on the relative strength of the phylogenetic signal exhibited by traits. This assessment employs Pagel's λ, which measures the

correlation between regression residuals and the underlying phylogenetic structure (Pagel 1997).

2. **R^2 Magnitude (R^2-Criterion)**: Prioritizing models based on the relative magnitude of R^2, which quantifies the proportion of variance in the dependent variable explained by the independent variable(s).
3. **AIC Magnitude (AIC-Criterion)**: Prioritizing models based on the relative magnitude of AIC, a statistical measure balancing model fit and complexity.

Results unveil that in the simulation scenarios such as BM1 & BM1 + BM2, BM & BM + Norm, and Norm & Norm + BM, models with higher R^2 significantly outperform those with lower R^2, consistenting across both ESRRV and mMorI methods (Table 2). However, the R^2-criterion fails to distinguish between models $OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$ in the Norm1 & Norm1 + Norm2 scenario. Nonetheless, this lack of discrimination is inconsequential as the evolution of X_1 and X_2 is unrelated to the phylogenetic tree, and phylogenetic autocorrelation is not critical in this context.

Table 2. The performance of the R^2-criterion for selecting the dependent variable in PVR.

	Simulation scenarios	Criteria	Correct	Incorrect	χ^2	p
ESRRV	BM1 & BM1 + BM2	R^2_{higher}	873	638	72.5	$<2.2\times10^{-16}$
		R^2_{lower}	638	873		
	BM & BM + Norm	R^2_{higher}	492	270	128.2	$<2.2\times10^{-16}$
		R^2_{lower}	270	492		
	Norm & Norm + BM	R^2_{higher}	903	177	973.4	$<2.2\times10^{-16}$
		R^2_{lower}	177	903		
	Norm1 & Norm1 + Norm2	R^2_{higher}	278	298	1.3	0.3
		R^2_{lower}	298	278		
mMorI	BM1 & BM1 + BM2	R^2_{higher}	539	428	25.0	5.6×10^{-7}
		R^2_{lower}	428	539		
	BM & BM + Norm	R^2_{higher}	260	174	33.3	7.9×10^{-9}
		R^2_{lower}	174	260		
	Norm & Norm + BM	R^2_{higher}	453	113	406.1	$<2.2\times10^{-16}$
		R^2_{lower}	113	453		
	Norm1 & Norm1 + Norm2	R^2_{higher}	71	68	0.1	0.8
		R^2_{lower}	68	71		

The instances in this table all represent situations where PVR results were conflicting after switching the dependent and independent variables. Please refer to the Materials and Methods section for details

regarding the two phylogenetic eigenvector selection methods (ESRRV and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Examining the choice of dependent and independent variables based on phylogenetic signal strength, we find significant performance disparities favoring models with stronger signals in BM & BM + Norm and Norm & Norm + BM scenarios. Yet, this distinction is absent in the other two scenarios, observed across both ESRRV and mMorI methods (Table 3).

Table 3. Performance of the λ-criterion for selecting the dependent variable in PVR.

	Simulation scenarios	Criteria	Correct	Incorrect	χ^2	p
ESRRV	BM1 & BM1 + BM2	λ_{higher}	560	546	0.3	0.6
		λ_{lower}	546	560		
	BM & BM + Norm	λ_{higher}	495	267	135.3	$<2.2\times10^{-16}$
		λ_{lower}	267	495		
	Norm & Norm + BM	λ_{higher}	903	177	973.4	$<2.2\times10^{-16}$
		λ_{lower}	177	903		
	Norm1 & Norm1 + Norm2	λ_{higher}	124	118	0.24	0.7
		λ_{lower}	118	124		
mMorI	BM1 & BM1 + BM2	λ_{higher}	343	361	0.82	0.4
		λ_{lower}	361	343		
	BM & BM + Norm	λ_{higher}	261	173	34.9	3.5×10^{-9}
		λ_{lower}	173	261		
	Norm & Norm + BM	λ_{higher}	453	113	406.1	$<2.2\times10^{-16}$
		λ_{lower}	113	453		
	Norm1 & Norm1 + Norm2	λ_{higher}	34	35	0	1
		λ_{lower}	35	34		

The instances in this table all represent situations where PVR results were conflicting after switching the dependent and independent variables. Please refer to the Materials and Methods section for details regarding the two phylogenetic eigenvector selection methods (ESRRV and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

In the BM & BM + Norm scenario, models with lower AIC values outrival, whereas in Norm & Norm + BM, the reverse holds true. However, no significant difference is noted in other instances, with this trend replicated across selection methods (Table 4).

Table 4. Performance of the AIC-Criterion for selecting the dependent variable in PVR.

	Simulation scenarios	Criteria	Correct	Incorrect	χ^2	p
ESRRV	BM1 & BM1 + BM2	AIC_{higher}	767	744	0.6	0.4
		AIC_{lower}	744	767		
	BM & BM + Norm	AIC_{higher}	267	495	135.3	$<2.2\times10^{-16}$
		AIC_{lower}	495	267		
	Norm & Norm + BM	AIC_{higher}	903	177	973.4	$<2.2\times10^{-16}$
		AIC_{lower}	177	903		
	Norm1 & Norm1 + Norm2	AIC_{higher}	301	275	2.2	0.1
		AIC_{lower}	275	301		
mMorI	BM1 & BM1 + BM2	AIC_{higher}	487	480	0.1	0.8
		AIC_{lower}	480	487		
	BM & BM + Norm	AIC_{higher}	173	261	34.9	3.5×10^{-8}
		AIC_{lower}	261	173		
	Norm & Norm + BM	AIC_{higher}	453	113	406.1	$<2.2\times10^{-16}$
		AIC_{lower}	113	453		
	Norm1 & Norm1 + Norm2	AIC_{higher}	69	70	0	1
		AIC_{lower}	70	69		

The instances in this table all represent situations where PVR results were conflicting after switching the dependent and independent variables. Please refer to the Materials and Methods section for details regarding the two phylogenetic eigenvector selection methods (ESRRV and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Comparing the three criteria's performance across diverse scenarios, prioritizing dependent and independent variables based on the model's R^2 magnitude emerges as the most reasonable approach. While the AIC criterion yields conflicting outcomes, the R^2 criterion consistently proves robust. Notably, in the BM1 & BM1 + BM2 scenario, selecting variables based on R^2 size yields superior accuracy, surpassing λ-criteria. The R^2 criterion is not universally effective. It performs well in situations where there is a phylogenetic relationship. In contexts without phylogeny, it is on par with other strategies. Moreover, when the evolution of traits is unrelated to the phylogenetic tree, the issue of phylogenetic

autocorrelation does not exist, and the necessity of phylogenetic comparative methods such as PVR is also diminished.

Furthermore, evaluating the R^2 criterion's performance across varying noise item variances (Figure 4) reveals its consistency across parameter gradients.

Given these findings, it is recommended to employ the relative magnitude of model R^2 as the criterion for selecting dependent and independent variables in PVR analysis. In cases of correlation result discrepancies, prioritize the model with the larger R^2.

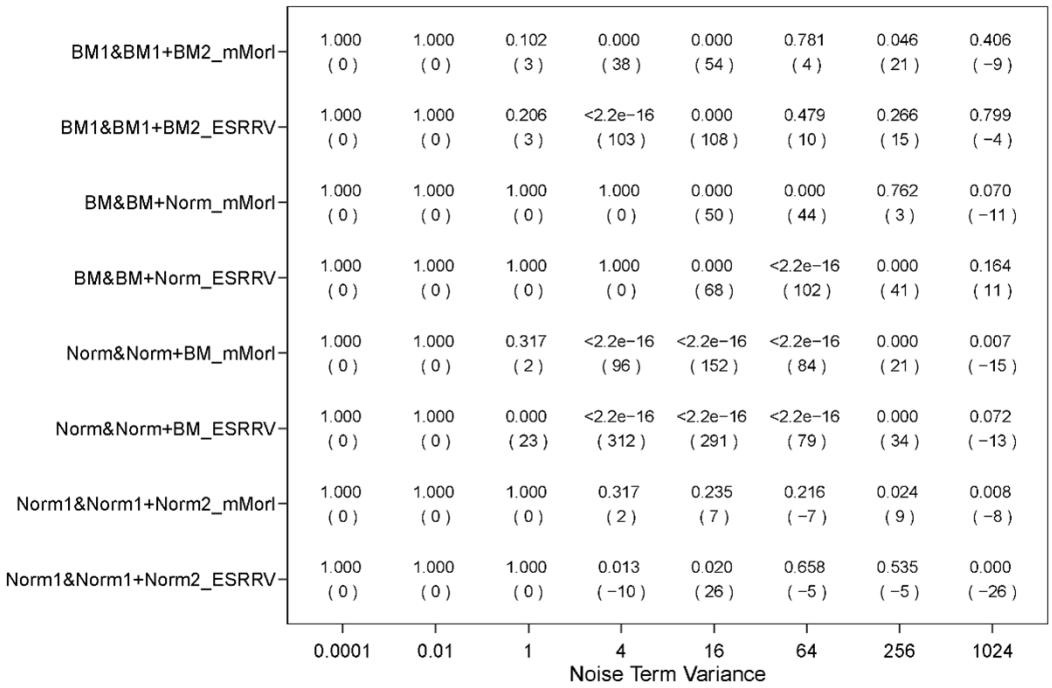

Figure 4. The performance of the R^2-criterion for dependent variable selection under different noise item variances. The numbers in parentheses represents the difference between the number of correct detections by the model with larger R^2 and the model with smaller R^2 (Number difference). The number above each cell represents the *p*-value of the Chi-square test. Please refer to the Materials and Methods section for details regarding the two phylogenetic eigenvector selection methods (ESRRV and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Comparing the performance of ESRRV and mMorI methods in PVR analysis

Utilizing R^2 as the criterion, we efficiently assigned the traits of X_1 and X_2 to dependent and independent variables. Subsequently, we compared the performance of two phylogenetic eigenvector selection methods, ESRRV and mMorI, in PVR analysis. Across most cases, no significant disparity was noted between the two methods (Figure 5). In instances where

significant differences emerged, neither method displayed a clear advantage, as each outperformed the other significantly in certain scenarios. Overall, the ESRRV method yielded slightly more accurate predictions than mMorI (25,523 vs. 25,372), with a difference of less than 1%. However, this discrepancy was not statistically significant ($\chi^2 = 2.159$, $p = 0.142$). Therefore, it can be concluded that there is no significant difference between the ESRRV and mMorI methods in detecting correlations.

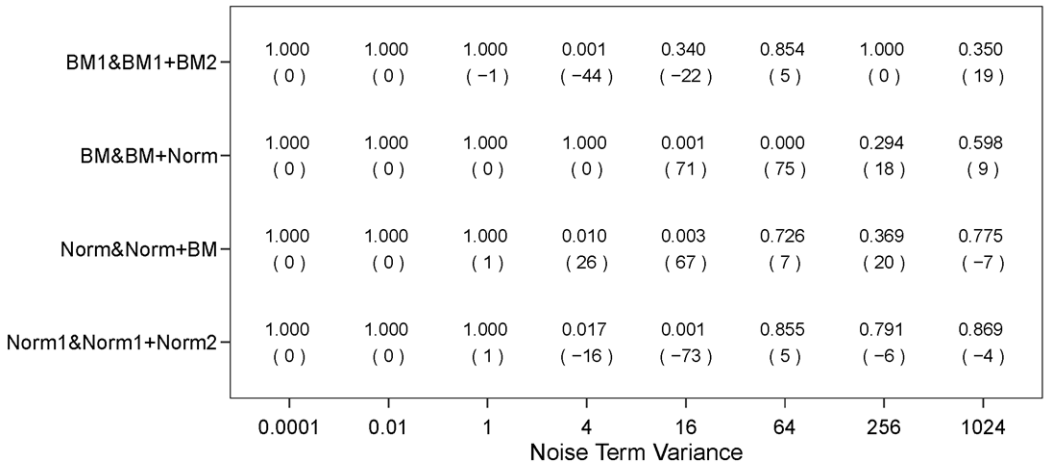

Figure 5. ESRRV and mMorI method's significance heatmap for the differences in the number of correct detections of trait correlations. Each number above each cell represents the *p*-value of the Chi-square test, and the number below represents the difference between the correct detections of the ESRRV method and the mMorI method. Please refer to the Materials and Methods section for details regarding the two phylogenetic eigenvector selection methods (ESRRV and mMorI) and the four simulation scenarios ("BM & BM + Norm," "Norm & Norm + BM," "BM1 & BM1 + BM2," and "Norm1 & Norm1 + Norm2").

Disscussion

In evolutionary biology, a fundamental inquiry often revolves around the existence of correlations between traits. By scrutinizing these correlations, evolutionary biologists can infer potential causal relationships among traits, thereby supporting or challenging hypotheses and theories within the field.

Whether employing standard linear models or phylogenetic comparative methods grounded in linear models, the overarching aim remains consistent: accurately discerning correlations between features. Essentially, researchers hypothesize the causal relationship between dependent and independent variables in linear models. In ordinary linear regression, when X_2 is the dependent variable, the *p*-value of the model's regression coefficient mirrors that of X_1 as the dependent variable, both equating to the *p*-value of the Pearson correlation test. However, not all regression models uphold consistent *p*-values when switching

dependent and independent variables (Chen et al. 2023). Consequently, this disparity results in varied *p*-values of model regression coefficients under different selections of dependent and independent variables, sometimes leading to conflicting correlation outcomes.

The *p*-value of a regression model solely reflects the presence of a significant correlation between two traits under specific selections of dependent and independent variables. Due to disparities in variable selection, even within the same method, conflicting correlation outcomes may arise. In such instances, the reliability of phylogenetic comparative methods may be questioned. We advocate for researchers to meticulously consider which trait to designate as the dependent variable and which as the independent variable based on existing information. Although conflicting correlation outcomes may arise within the same regression method, most errors can be circumvented through judicious variable selection.

This study primarily explores the impact of varied choices of dependent and independent variables on PVR analysis outcomes. Specifically, the choice of different dependent variables influences the selection of phylogenetic eigenvectors, thereby influencing PVR analysis outcomes. This influence occasionally results in conflicting correlation outcomes. For two competing models yielding conflicting results, it is observed, based on the gold standard, that the model with a relatively higher R^2 value can more accurately detect the correlation between traits compared to the model with a lower R^2 value. This observation is intuitive, as a higher R^2 value indicates a better explanatory fit of the independent variable to the dependent variable, thus enhancing the model's capacity to accurately detect correlations between traits.

Similar to phylogenetic comparative methods like PVR, many techniques in spatial statistics operate on comparable principles. Concerns arise regarding whether swapping dependent and independent variables may impact correlation analyses in commonly used spatial statistical methods such as spatial eigenvector mapping, conditional autoregressive models, simultaneous autoregressive models, generalized least squares models, and spatial generalized estimating equations (Dormann et al. 2007), potentially influencing spatial eigenvector selection or other unforeseen pathways. Further research efforts should be directed towards examining the implications of swapping dependent and independent variables on correlation analyses in spatial statistical methods.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (grant number 31671321). To enhance language and readability, this article has undergone refinement by

ChatGPT 3.5 (OpenAI 2024). Subsequently, the authors conducted a thorough review and made necessary edits, taking full responsibility for the publication's content.

Author Contributions

DKN and ZLC conceived the research hypothesis; ZLC performed the simulations and analyzed the data. ZLC drafted the manuscript and DKN revised the manuscript. Both authors agree with the final version and the submission.

References

Burns JH, Anacker BL, Strauss SY, Burke DJ (2015) Soil microbial community variation correlates most strongly with plant species identity, followed by soil chemistry, spatial location and plant genus. AoB PLANTS 7

Chen Z-L, Guo H-J, Niu D-K (2023) Dependent variable selection in phylogenetic generalized least squares regression analysis under Pagel's lambda model. bioRxiv:2023.2005.2021.541623

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution 39:1335-1351

Cottenie K (2005) Integrating environmental and spatial processes in ecological community dynamics. Ecol Lett 8:1175-1182

Dale MRT, Fortin M-J (2009) Spatial autocorrelation and statistical tests: Some solutions. Journal of Agricultural, Biological, and Environmental Statistics 14:188-206

Debastiani VJ, Bastazini VAG, Pillar VD (2021) Using phylogenetic information to impute missing functional trait values in ecological databases. Ecol Inform 63:101315

Desdevises Y, Legendre P, Azouzi L, Morand S (2003) Quantifying phylogenetically structured environmental variation. Evolution 57:2647-2652

Diniz-Filho JA, Nabout JC (2009) Modeling body size evolution in Felidae under alternative phylogenetic hypotheses. Genet Mol Biol 32:170-176

Diniz-Filho JAF, Bini LM (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. Glob Ecol Biogeogr 14:177-185

Diniz-Filho JAF, Bini LM, Rangel TF, Morales-Castilla I, Olalla-Tárraga MÁ, Rodríguez MÁ, Hawkins BA (2012) On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239-249

Diniz-Filho JAF, Cianciaruso MV, Rangel TF, Bini LM (2011) Eigenvector estimation of phylogenetic and functional diversity. Funct Ecol 25:735-744

Diniz-Filho JAF, de Sant'Ana CER, Bini LM (1998) An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247-1262

Diniz-Filho JAF, Jardim L, Mondanaro A, Raia P (2019) Multiple components of phylogenetic non-stationarity in the evolution of brain size in fossil hominins. Evol Biol 46:47-59

Diniz-Filho JAF, Rodríguez MÁ, Bini LM, Olalla-Tarraga MÁ, Cardillo M, Nabout JC, Hortal J, Hawkins BA (2009) Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. Journal of Biogeography 36:2222-2236

Divíšek J, Chytrý M, Beckage B, Gotelli NJ, Lososová Z, Pyšek P, Richardson DM, Molofsky J (2018) Similarity of introduced plant species to native ones facilitates naturalization, but differences enhance invasion success. Nat Commun 9:4631

Doreian P (1980) Linear models with spatially distributed data: spatial disturbances or spatial effects? Sociological Methods & Research 9:29-60

Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628

Dow MM, Burton ML, White DR (1982) Network autocorrelation: A simulation study of a foundational problem in regression and survey research. Social Networks 4:169-200

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

Gao X, Dai J, Tao Z, Shahzad K, Wang H (2023) Autumn phenology of tree species in China is associated more with climate than with spring phenology and phylogeny. Front Plant Sci 14:1040758

Garamszegi LZ (2014) Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Springer, Berlin

Grafen A (1989) The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci 326:119-157

Griffith DA, Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603-2613

Harvey PH, Read AF, Nee S (1995) Why ecologists need to be phylogenetically challenged. The Journal of ecology 83:535-536

He F, Zhou J, Zhu H (2003) Autologistic regression model for the distribution of vegetation. Journal of Agricultural, Biological, and Environmental Statistics 8:205-222

Keitt TH, Bjørnstad ON, Dixon PM, Citron-Pousty S (2002) Accounting for spatial pattern when modeling organism-environment interactions. Ecography 25:616-625

Lanuza JB, Rader R, Stavert J, Kendall LK, Saunders ME, Bartomeus I (2023) Covariation among reproductive traits in flowering plants shapes their interactions with pollinators. Funct Ecol 37:2072-2084

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673

LeRoy CJ, Hipp AL, Lueders K, Follstad Shah JJ, Kominoski JS, Ardón M, Dodds WK, Gessner MO, Griffiths NA, Lecerf A, Manning DWP, Sinsabaugh RL, Webster JR (2020) Plant phylogenetic history explains in-stream decomposition at a global scale. J Ecol 108:17-35

OpenAI (2024) ChatGPT (3.5) [Large language model]. https://chat.openai.com.

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zool Scr 26:331-348

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877-884
Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput Statist Data Anal 49:974-997
Peres-Neto PR, Legendre P (2010) Estimating and controlling for spatial structure in the study of ecological communities. Glob Ecol Biogeogr 19:174-184
Peres-Neto PR, Leibold MA, Dray S (2012) Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. Ecology 93:S14-S30
Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319-329
Rohlf FJ (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143-2160
Westoby M, Leishman MR, Lord JM (1995) On Misinterpreting the `Phylogenetic Correction'. J Ecol 83:531-534