Variable Dependence in Eigenvector Selection for Phylogenetic Eigenvector Regression

Zheng-Lin Chen and 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

Abstract

In evolutionary biology, accurate analysis of trait correlations is essential for understanding phylogenetic relationships and evolutionary processes. This study investigates the impact of dependent and independent variable selection on the outcomes of Phylogenetic Vector Regression (PVR) and explores the broader implications for similar statistical models. Using simulated data, we demonstrated how the choice of dependent and independent variables influences phylogenetic eigenvector selection and subsequently affects correlation results within PVR. Our findings reveal that models with higher R^2 values consistently provide more accurate detection of correlations, suggesting that R^2 is a reliable criterion for variable selection in PVR analysis. Additionally, while our primary analysis focused on PVR, preliminary investigations into other spatial statistical techniques, such as spatial eigenvector mapping, conditional autoregressive models and simultaneous autoregressive models, have also indicated similar discrepancies when swapping dependent and independent variables. These observations suggest a potentially widespread issue across various statistical models. However, to maintain focus and coherence, a detailed exploration of these findings will be the subject of future work.

Keywords

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

We constructed a phylogenetic tree with 100 species and simulated the evolution of two traits, X_1 and X_2, defined by the relationship $X_2 = X_1 + \varepsilon$, where ε represents noise. We varied the variance of ε to manipulate the traits' correlation strength, inspired by Revell (2010). We modeled four scenarios:

1. **BM & BM + Norm**: Trait X_1 evolves along the phylogenetic tree following a Brownian motion model (variance = 1), while ε follows a normal distribution with mean 0 and variance d. Values of d used are 10^{-4}, 0.01, 1, 4, 16, 64, 256, and 1024.
2. **Norm & Norm + BM**: Trait X_1 evolves under a normal distribution (mean = 0, variance =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. **BM1 & BM1 + BM2**: Trait X_1 evolves according to a Brownian motion model (variance = 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. **Norm1 & Norm1 + Norm2**: 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.

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 initiate our analysis with Cheverud's autoregressive model (Cheverud et al. 1985), defined by the equation $T = P + S$. In this model, T represents the trait being analyzed, P accounts for the influence of phylogenetic relationships, and S encompasses influences from other traits. Utilizing the PVR method, we perform principal coordinate analysis on a double-centered phylogenetic matrix ($n \times n$), resulting in $n - 1$ phylogenetic eigenvectors. We then select k eigenvectors (where typically $k < n - 1$) using the selection methods detailed later, incorporating them as covariates in the regression model. The model is expressed as $Y = V\beta + \varepsilon$, where Y is the dependent variable representing the trait, V is the matrix of selected phylogenetic eigenvectors, β is the vector of regression coefficients for these eigenvectors, and ε represents the residual component S, accounting for additional explanatory variables not included in V.

Diniz-Filho et al. (2012) proposed two principal approaches for selecting phylogenetic eigenvectors: sequential and non-sequential. In the sequential approach, the $n - 1$ phylogenetic eigenvectors derived from the matrix decomposition are sorted by their eigenvalues from largest to smallest (EV_1, EV_2, ..., EV_{n-1}). These eigenvectors are then

incrementally added to the model in descending order of importance until Moran's I for the residuals, ε, reaches zero, indicating the absence of significant phylogenetic autocorrelation. Moran's I, adapted from spatial to phylogenetic autocorrelation analysis, measures the degree of autocorrelation, ranging from –1 (total negative autocorrelation) to +1 (total positive autocorrelation), with 0 representing no autocorrelation.

The non-sequential approach includes three distinct methods. The first, evaluating significant relationships between phylogenetic eigenvectors and the response variable (ESRRV), uses regression models to identify pertinent eigenvectors (Diniz Filho and Nabout 2009, Diniz Filho et al. 2012). In this study, we first perform normality tests on the eigenvectors and the response variable. Depending on the outcome, we apply either Pearson correlation (for normally distributed data) or Spearman rank correlation (for non-normal data). The next two methods involve stepwise regression, specifically employing forward selection to sequentially add variables until no further improvement in model fit is observed. The second method optimizes model fit by minimizing the AIC (mAIC), while the third minimizes Moran's I of model residuals (mMorI), effectively encompasses the sequential approach. Therefore, this study exclusively utilizes these three non-sequential approaches, avoiding redundancy.

Once selected, these phylogenetic eigenvectors are integrated as covariates into a simple linear regression model, formulated as $y = \sum_{i=1}^{k} a_i * ev_i + b * x + \epsilon$, where ev_i represents the set of selected eigenvectors, y is the dependent variable trait, x is the independent variable trait, and ϵ is the residual term.

Gold Standard for Trait Correlations in Each Simulation

In our study, we simulated the evolution of traits X_1 and X_2 along a phylogenetic tree, recording trait values at each node. We then calculated the changes in these traits per unit

branch length, represented as $\Delta X_1/L$ and $\Delta X_2/L$, where L is the length of the branch. These changes are independent from one branch to another, which allows us to use conventional correlation tests to analyze the relationships between $\Delta X_1/L$ and $\Delta X_2/L$. We opted for correlation analysis over linear regression due to the potential non-normal distribution of the trait changes. Non-normal data does not meet the assumptions required for linear regression, particularly regarding the normality of residuals. If both $\Delta X_1/L$ and $\Delta X_2/L$ are normally distributed, Pearson correlation analysis is conducted; if not, Spearman rank correlation tests are applied.

By assessing the relationship between $\Delta X_1/L$ and $\Delta X_2/L$, we effectively capture the true evolutionary relationship between the traits. This approach enables us to evaluate the accuracy of various models by comparing their predicted trait correlations with the actual evolutionary relationships observed in the simulations.

Results

For convenience, we denote the set of phylogenetic eigenvectors selected with X_1 and X_2 as dependent variables as V_{X_1} and V_{X_2}, respectively. Ordinary least squares linear regression models were used to assess the relationship between X_1 and X_2. $OLS_{X_1 \sim X_2 + V_{X_1}}$ represents the 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.

Impact of Variable Selection on Phylogenetic Eigenvector Selection

The choice of dependent and independent variables significantly influenced the selection of phylogenetic eigenvectors. Analysis of 8,000 simulated datasets showed that the sets of selected eigenvectors often differed when dependent and independent variables were

swapped (Fig. 1). Specifically, the discrepancy rates in eigenvector selection were 81.6% for the ESRRV method, 70.6% for the mMorI method, and 93.1% for the mAIC method. These results highlight how variable selection impacts the choice of phylogenetic eigenvectors.

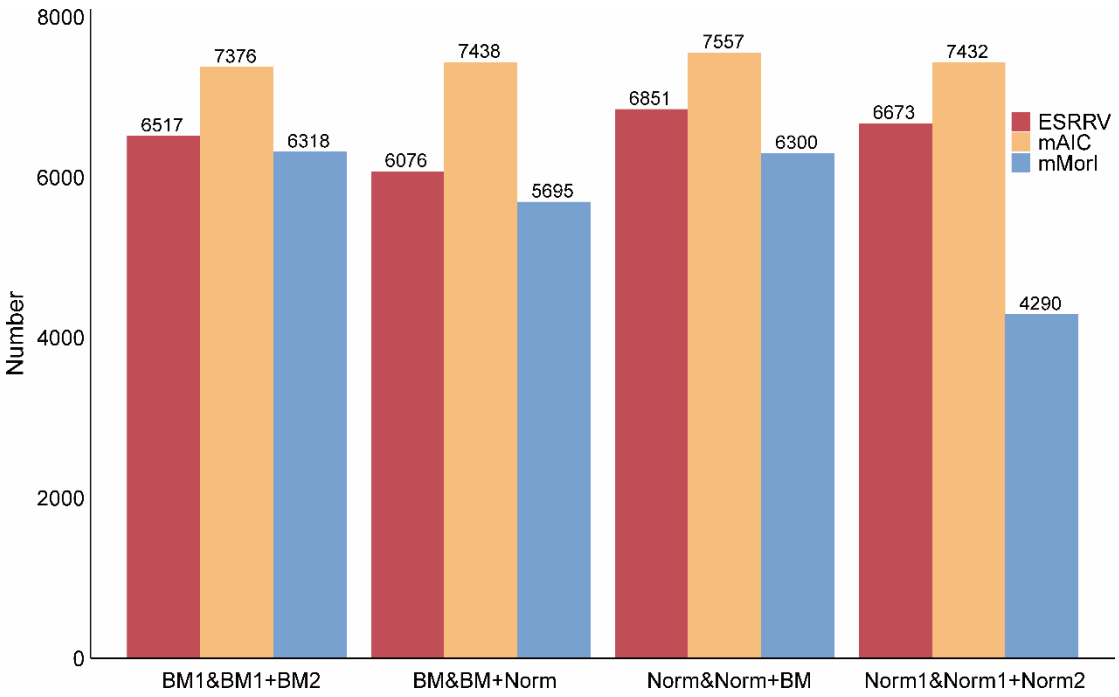

Figure 1. Dependence of phylogenetic eigenvector selection on variable exchange. This figure illustrates the count of divergent selections of phylogenetic eigenvector sets resulting from interchanging dependent and independent variables (denoted as V_{X_1} and V_{X_2}) across various phylogenetic eigenvector 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 variable selection on correlation results

The selection of dependent variables varied the sets of phylogenetic eigenvectors, leading to conflicting correlation outcomes in our simulations (Fig. 2). Out of 32,000 simulations, the ESRRV method resulted in 3,935 instances of conflicting correlation results, while the mAIC and mMorI methods showed 2,393 and 2,108 discrepancies, respectively. Approximately 10% of these results showed inconsistencies, underscoring the critical role of variable selection in correlation analysis.

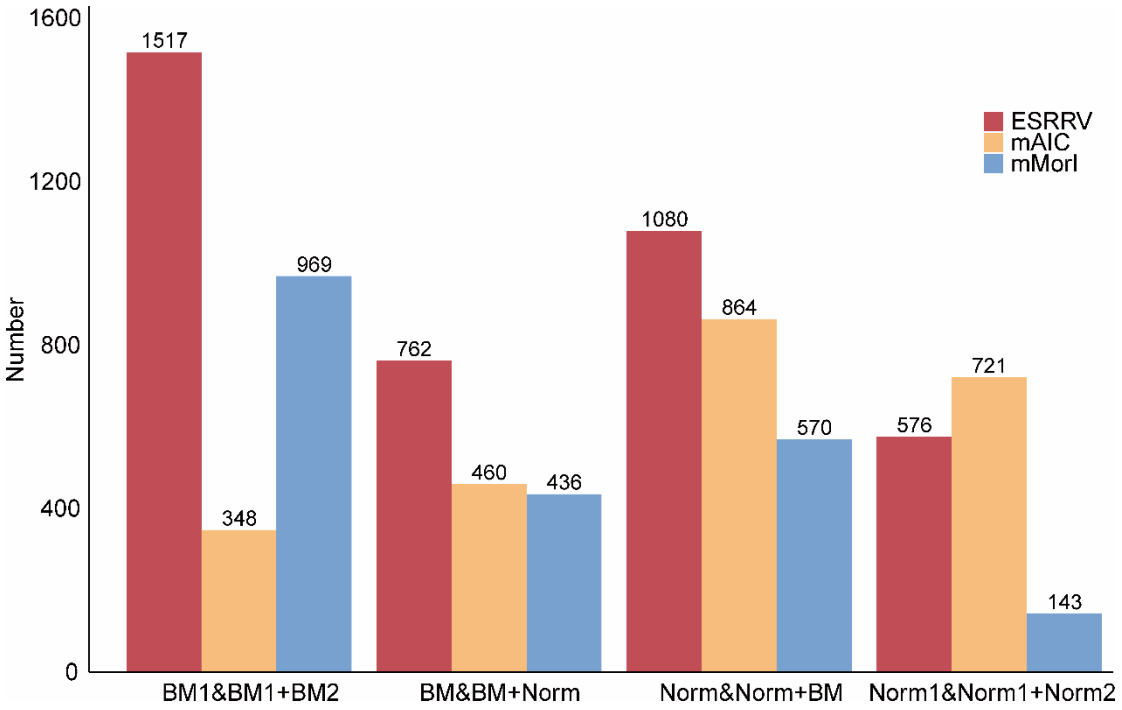

Figure 2. Impact of variable exchange on correlation results in PVR models. The figure depicts the number of conflicting correlation results resulting from switching dependent and independent variables in PVR models. This switch leads to inconsistent sets of phylogenetic eigenvectors used in PVR analysis, 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 variable selection and eigenvector selection on PVR results

To establish a baseline for model evaluation, we calculated the changes per unit length ($\Delta X_1/L$ and $\Delta X_2/L$) for each branch of the phylogenetic trees and conducted correlation tests on these changes. These test results, reflecting significant positive, significant negative, or nonsignificant correlations, serve as the gold standard for assessing the true relationship between traits X_1 and X_2. This gold standard is crucial for evaluating the performance of different phylogenetic models.

We then thoroughly analyzed the correlation results obtained from the PVR models, $OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$, especially focusing on the effects of swapping dependent and independent variables. Our analysis included evaluating the frequency of outcomes where both models yielded consistent and correct results, consistent yet incorrect results, and scenarios where only one model provided an accurate correlation result. The analysis revealed that the ESRRV and mMorI methods more consistently produced accurate correlation results across all four simulation scenarios, outperforming the mAIC method (Fig. 3).

To address the issue of conflicting outcomes from variable swapping, we developed criteria to identify the correct results in datasets lacking a gold standard, acknowledging that empirical data often do not provide such benchmarks. In preparation for introducing these criteria, we defined a 'best scenario' and a 'worst scenario'. The 'best scenario' represents the most favorable outcome where all possible methods, including our newly developed criteria, correctly identify all correct results in cases where only one correlation is accurate the true

relationships in the data. Conversely, the 'worst scenario' reflects the least favorable outcome, where no correct results are identified when only one correlation result is accurate.

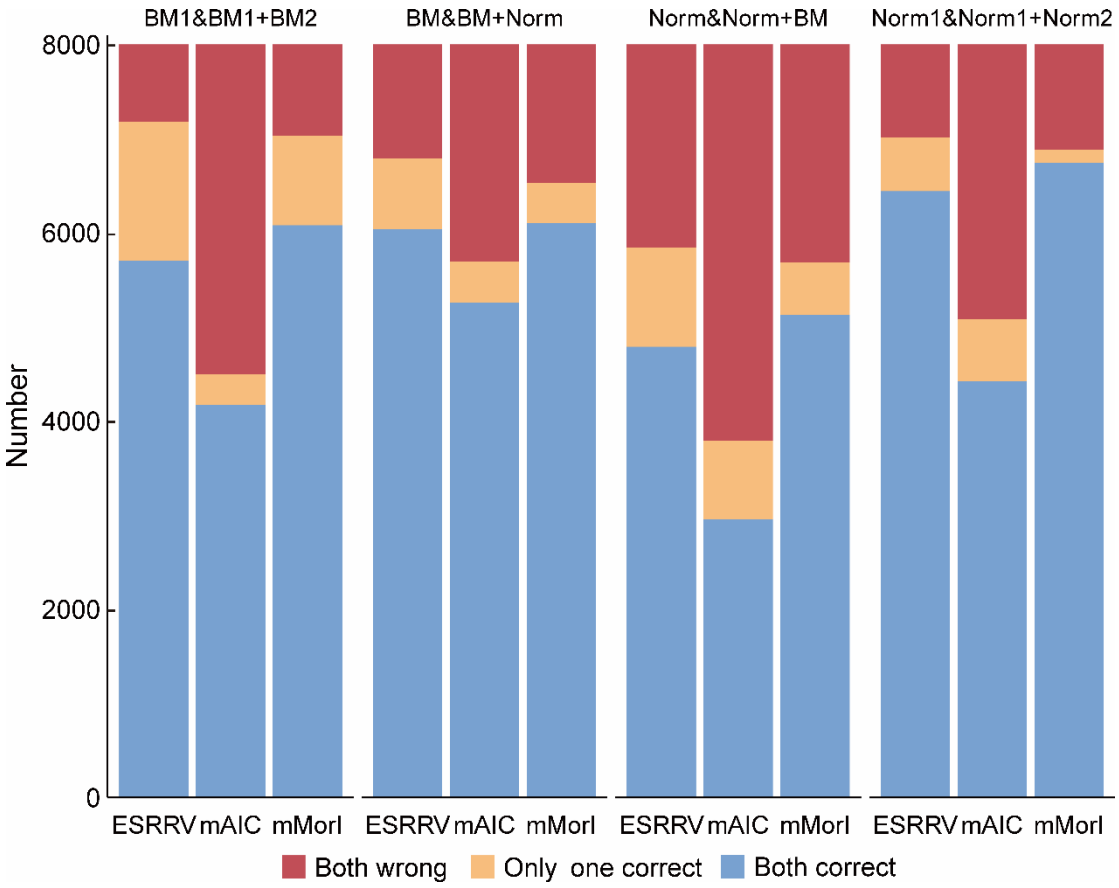

Figure 3. Distribution of correct and wrong results in the PVR models when swapping dependent and independent variables ($OLS_{X_1 \sim X_2 + V_{X_1}}$ and $OLS_{X_2 \sim X_1 + V_{X_2}}$). The figure depicts the count of correlation outcomes obtained from OLS regression models when the roles of dependent 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").

With this framework, we first evaluated the effectiveness of the three eigenvector selection methods. Comparing the accuracy of the mAIC method's predictions in the best scenario with those of the ESRRV and mMorI methods in the worst scenario highlights the superior reliability of the ESRRV and mMorI methods (Table 1). Consequently, subsequent analyses primarily concentrate on the ESRRV and mMorI methods.

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

Given the conflicting outcomes that arise from swapping dependent and independent variables, accurately identifying the model that truly reflects the relationships between traits

is critical, especially in empirical data which often lacks a gold standard. To streamline this selection process, we propose three criteria:

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.

Our results indicate that in simulation scenarios such as BM1 & BM1 + BM2, and BM & BM + Norm, models with higher R^2 consistently outperform those with lower R^2, across both ESRRV and mMorI methods (Table 2). However, the R^2-criterion does not differentiate 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, where the evolution of X_1 and X_2 is unrelated to the phylogenetic tree, rendering phylogenetic autocorrelation non-critical.

When assessing the choice of dependent and independent variables based on phylogenetic signal strength, significant performance differences emerge, favoring models with stronger signals in the BM & BM + Norm and Norm & Norm + BM scenarios. This distinction, however, is absent in the other two scenarios, as observed across both ESRRV and mMorI methods (Table 3).

In scenarios like BM & BM + Norm, models with lower AIC values excel, whereas in Norm & Norm + BM, the opposite is true. No significant difference is noted in other

scenarios, with this trend being consistent across different eigenvector selection methods (Table 4).

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").

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").

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").

Upon comparing the performance of these three criteria across various scenarios, selecting dependent and independent variables based on the model's R^2 magnitude emerges as the most effective strategy. While the AIC criterion sometimes produces conflicting outcomes, the R^2 criterion consistently demonstrates robustness. Notably, in the BM1 & BM1 + BM2 scenario, choosing variables based on R^2 size provides superior accuracy, surpassing the λ-criterion. Although the R^2-criterion is not universally effective, it excels in scenarios with a phylogenetic relationship and performs comparably to other strategies in contexts without phylogeny.

Additionally, an evaluation of the R^2-criterion's performance across varying noise item variances (Fig. 4) highlights its consistency across different parameter gradients. Based on these findings, we recommend using the relative magnitude of model R^2 as the criterion for selecting dependent and independent variables in PVR analysis. In cases where correlation results differ, the model with the larger R^2 should be prioritized.

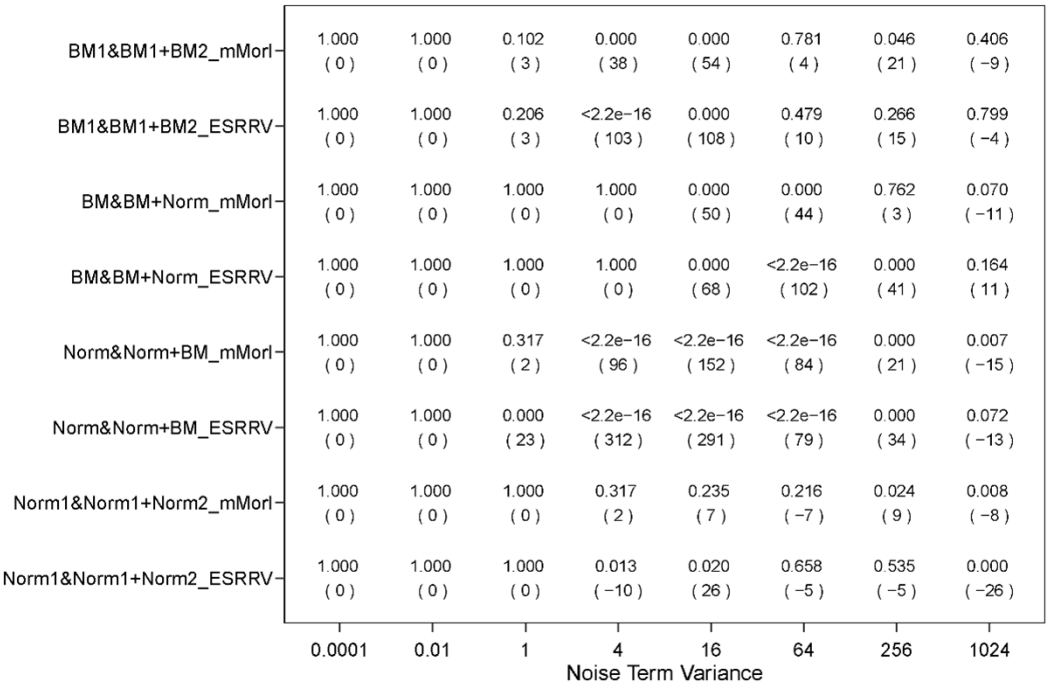

Figure 4. The performance of the R^2-criterion for dependent variable selection under different noise item variances. The numbers in parentheses represents the difference between the number of correct detections by the model with larger R^2 and the model with smaller R^2 (Number difference). The number 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").

Comparison of Eigenvector Selection Methods in PVR Analysis

Using R^2 as our selection criterion, we effectively select the dependent variable from the two traits X_1 and X_2. We then assessed the performance of two phylogenetic eigenvector selection methods, ESRRV and mMorI, within the framework of PVR analysis. Our comparative analysis revealed that in most scenarios, there was no significant difference in performance between the two methods (Figure 5). In cases where notable differences did

emerge, neither method consistently outperformed the other; each showed advantages in specific scenarios. Overall, the ESRRV method provided slightly more accurate predictions than mMorI, with respective accuracies of 25,523 versus 25,372, a difference of less than 1%. However, this discrepancy was not statistically significant ($\chi^2 = 2.159$, $p = 0.142$). Consequently, we conclude that there is no significant difference in the efficacy of the ESRRV and mMorI methods for detecting correlations in PVR analysis.

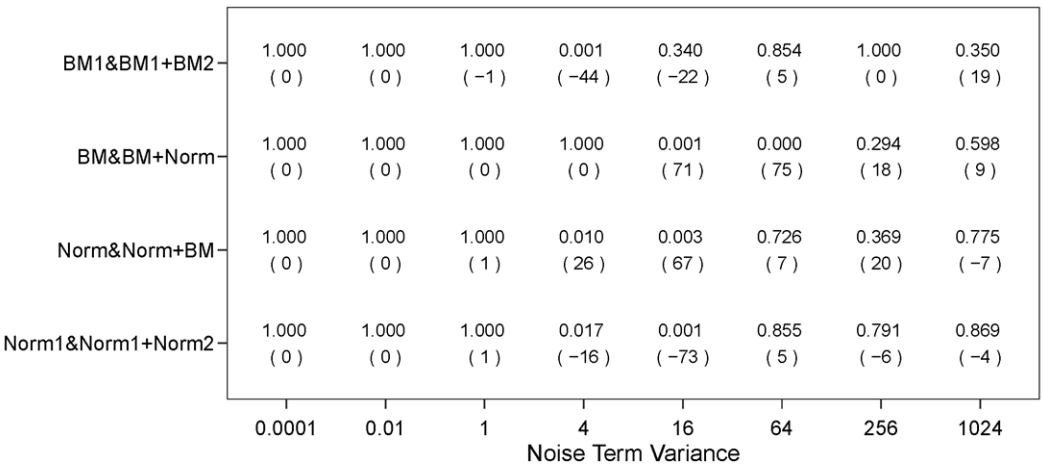

Figure 5. ESRRV and mMorI method's significance heatmap for the differences in the number of correct detections of trait correlations. Each number above each cell represents the *p*-value of the Chi-square test, and the number below represents the difference between the correct detections of the ESRRV 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, the analysis of trait correlations is fundamental, often prompting inquiries into potential causal relationships among traits. This scrutiny supports or

challenges existing hypotheses and theories within the field. Both standard linear models and phylogenetic comparative methods grounded in linear models strive to accurately discern correlations between traits. Typically, researchers hypothesize about the causal relationships between dependent and independent variables in these models.

In standard linear regression, the *p*-value of a model's regression coefficient is expected to reflect the strength and direction of the correlation between variables. However, we observe that not all regression models maintain consistent *p*-values when the roles of dependent and independent variables are reversed in phylogenetic generalized least squares, leading to potential disparities (Chen et al. 2023). Such inconsistencies can result in conflicting correlation outcomes, questioning the reliability of phylogenetic comparative methods. It is crucial for researchers to carefully select which trait is designated as dependent or independent, as inappropriate choices can lead to significant analytical errors.

This study specifically addresses how the selection of dependent and independent variables impacts the outcomes of PVR analysis. We find that these choices significantly influence the selection of phylogenetic eigenvectors, which in turn affects the correlation results obtained from the analysis. Notably, models with a higher R^2 value, indicating a better explanatory fit, are more reliable in selecting the dependent variable between traits. This underscores the importance of model selection criteria, such as the R^2 magnitude, in improving the accuracy of phylogenetic comparative analyses. Additionally, once an appropriate set of phylogenetic eigenvectors and variables is determined via criteria such as R^2, the designation of which trait functions as the dependent or independent variable becomes largely irrelevant for the regression outcomes. Mathematically, the regression is symmetric with respect to the interchange of variables, leading to equivalent regression results whether variables are swapped or not. This highlights the robustness of the selected model configuration, ensuring that the primary influence on the results is the appropriateness

of the eigenvectors and the explanatory variables selected, rather than their arrangement in the model.

Furthermore, while the principal focus of this study has been on the impacts of variable selection within PVR, our preliminary investigations into other spatial statistical techniques—such as spatial eigenvector mapping, conditional autoregressive models, simultaneous autoregressive models, and trend-surface generalized additive model—have also indicated similar discrepancies when dependent and independent variables are swapped. These findings are documented in Supplementary Table S1, which summarize the initial results and suggest that the issue of variable swapping leading to discrepancies in model outcomes may be widespread across various statistical models. Detailed analysis of these results will be pursued in future work to further elucidate the implications for a broader range of ecological and evolutionary research methodologies. This anticipates a significant expansion of our understanding of how methodological choices impact the reliability and interpretability of statistical models in these fields.

Data availability

The codes to generate the results in the main text (Code_for_maintext.zip) and that to generate the results in Supplemental Table S1 (Code_for_TableS1.zip) are included in Supplementary material.

Supplementary material

Data available from the Dryad Digital Repository.

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 J.H., Anacker B.L., Strauss S.Y., Burke D.J. 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 J.M., Dow M.M., 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 M.R.T., Fortin M.-J. 2009. Spatial autocorrelation and statistical tests: Some solutions. JABES, 14:188-206.

Debastiani V.J., Bastazini V.A.G., Pillar V.D. 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 J.A.F., Bini L.M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Glob. Ecol. Biogeogr., 14:177-185.

Diniz Filho J.A.F., Bini L.M., Rangel T.F., Morales-Castilla I., Olalla-Tárraga M.Á., Rodríguez M.Á., Hawkins B.A. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239-249.

Diniz Filho J.A.F., Cianciaruso M.V., Rangel T.F., Bini L.M. 2011. Eigenvector estimation of phylogenetic and functional diversity. Funct. Ecol., 25:735-744.

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

Diniz Filho J.A.F., 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 J.A.F., Nabout J.C. 2009. Modeling body size evolution in Felidae under alternative phylogenetic hypotheses. Genet Mol Biol, 32:170-176.

Diniz Filho J.A.F., Rodríguez M.Á., Bini L.M., Olalla-Tarraga M.Á., Cardillo M., Nabout J.C., Hortal J., Hawkins B.A. 2009. Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. J. Biogeogr., 36:2222-2236.

Divíšek J., Chytrý M., Beckage B., Gotelli N.J., Lososová Z., Pyšek P., Richardson D.M., 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? Soc. Methods Res., 9:29-60.

Dormann C.F., McPherson J.M., Araújo M.B., Bivand R., Bolliger J., Carl G., Davies R.G., Hirzel A., Jetz W., Kissling W.D., Kühn I., Ohlemüller R., Peres-Neto P.R., Reineking B., Schröder B., Schurr F.M., Wilson R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30:609-628.

Dow M.M., Burton M.L., White D.R. 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 L.Z. 2014. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin, Springer.

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

Griffith D.A., Peres-Neto P.R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology, 87:2603-2613.

Harvey P.H., Read A.F., Nee S. 1995. Why ecologists need to be phylogenetically challenged. J Ecol, 83:535-536.

He F., Zhou J., Zhu H. 2003. Autologistic regression model for the distribution of vegetation. JABES, 8:205-222.

Keitt T.H., Bjørnstad O.N., Dixon P.M., Citron-Pousty S. 2002. Accounting for spatial pattern when modeling organism-environment interactions. Ecography, 25:616-625.

Lanuza J.B., Rader R., Stavert J., Kendall L.K., Saunders M.E., 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 C.J., Hipp A.L., Lueders K., Follstad Shah J.J., Kominoski J.S., Ardón M., Dodds W.K., Gessner M.O., Griffiths N.A., Lecerf A., Manning D.W.P., Sinsabaugh R.L., Webster J.R. 2020. Plant phylogenetic history explains in-stream decomposition at a global scale. J Ecol, 108:17-35.

OpenAI. 2024. ChatGPT (4) [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 P.R., Jackson D.A., Somers K.M. 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 P.R., Legendre P. 2010. Estimating and controlling for spatial structure in the study of ecological communities. Glob. Ecol. Biogeogr., 19:174-184.

Peres-Neto P.R., Leibold M.A., Dray S. 2012. Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. Ecology, 93:S14-S30.

Revell L.J. 2010. Phylogenetic signal and linear regression on species data. Methods Ecol. Evol., 1:319-329.

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

Westoby M., Leishman M.R., Lord J.M. 1995. On Misinterpreting the `Phylogenetic Correction'. J Ecol, 83:531-534.