Genetics and Molecular Biology, 38, 3, 396-400 (2015)

DOI: http://dx.doi.org/10.1590/S1415-475738320140391

Short Communication

The best of both worlds: Phylogenetic eigenvector regression and mapping

José Alexandre Felizola Diniz Filho, Fabricio Villalobos and Luis Mauricio Bini

Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, GO, Brazil.

Abstract

Eigenfunction analyses have been widely used to model patterns of autocorrelation in time, space and phylogeny. In a phylogenetic context, Diniz-Filho *et al.* (1998) proposed what they called Phylogenetic Eigenvector Regression (PVR), in which pairwise phylogenetic distances among species are submitted to a Principal Coordinate Analysis, and eigenvectors are then used as explanatory variables in regression, correlation or ANOVAs. More recently, a new approach called Phylogenetic Eigenvector Mapping (PEM) was proposed, with the main advantage of explicitly incorporating a model-based warping in phylogenetic distance in which an Ornstein-Uhlenbeck (O-U) process is fitted to data before eigenvector extraction. Here we compared PVR and PEM in respect to estimated phylogenetic signal, correlated evolution under alternative evolutionary models and phylogenetic imputation, using simulated data. Despite similarity between the two approaches, PEM has a slightly higher prediction ability and is more general than the original PVR. Even so, in a conceptual sense, PEM may provide a technique in the best of both worlds, combining the flexibility of data-driven and empirical eigenfunction analyses and the sounding insights provided by evolutionary models well known in comparative analyses.

Keywords: evolutionary models, phylogenetic comparative methods, phylogenetic imputation, phylogenetic signal.

Received: December 26, 2014; Accepted: March 29, 2015.

Eigenfunction analyses have been widely used to model patterns of autocorrelation in time, space and phylogeny (Peres-Neto, 2006; Dray *et al.*, 2006; Kuhn *et al.*, 2009; Safi and Pettorelli, 2010; Peres-Neto and Legendre, 2010; Peres-Neto *et al.*, 2012). In general, these analyses start with a Principal Coordinate Analysis (PCoA; see Legendre and Legendre, 2012) of pairwise distance or connectivity matrices between observations (*e.g.*, species in the case of phylogenetic analysis). After, selected eigenvectors from PCoA are used to detect the magnitude of (temporal, spatial or phylogenetic) patterns in data, both in univariate and multivariate domains. In a phylogenetic context, Diniz-Filho *et al.* (1998) proposed what they called Phylogenetic Eigenvector Regression (PVR), in which eigenvectors are used as explanatory variables in a multiple regression to model trait evolution. The coefficient of determination (R^2) of PVR was interpreted as the amount of phylogenetic signal (see Diniz-Filho *et al.*, 2012a,b,c; Münkemüller *et al.*, 2012). The method was later expanded to estimate a phylogenetically corrected correlation and variance partitioning modeling (Martins *et al.*, 2002; Desdevises *et al.*, 2003).

However, PVR was criticized by Rohlf (2001), who showed that if not all phylogenetic eigenvectors are used to model the trait, then there would be a missing part of the phylogeny in the model and, as a consequence, the estimated phylogenetic signal for a trait evolving under Brownian motion would be underestimated. Also, correlation between traits, after accounting for phylogenetic relationships given by the eigenvectors, would be biased and possess and inflated Type I error rates (as shown by Martins *et al.*, 2002). More recently, Freckleton *et al.* (2011; see also Martins *et al.*, 2002; Laurin, 2010) also criticized the statistical performance of PVR and favored the Phylogenetic Generalized Least-Squares (PGLS) because this last one is formally based on evolutionary models (*e.g.* Brownian motion, Ornstein-Uhlenbeck (O-U)), whereas PVR is a purely statistical, data-driven approach.

Diniz-Filho *et al.* (2012a) recently expanded the PVR method by relating the coefficients of determination (R^2) of successive PVR models (*i.e.*, with the consecutive addition of phylogenetic eigenvectors as explanatory variables of a trait) to the cumulative eigenvalues of the eigenvectors used in the models. This approach to explore phylogenetically structured patterns of trait variation was called Phylogenetic Signal-Representation (PSR) curve. They showed that different models of trait evolution generate different patterns of relationship between the coefficients of determination (R^2) and the cumulative eigenvalues (*i.e.*, different PSR curves). For instance, under a Brownian motion model of trait evolution, the PSR curve is linear, so the R^2 estimated will depend on which eigenvectors are used [supporting Rohlf's issue (Rohlf, (2001) that missing even a

Send correspondence to José Alexandre Felizola Diniz-Filho. Departamento de Ecologia, Universidade Federal de Goiás, Caixa Postal 131, Goiânia, GO, Brazil, E-mail: diniz@ufg.br.

few eigenvectors with very small eigenvalues will cause an inflation in Type I error]. However, if evolution is not Brownian and the PSR is not linear, some of the eigenvectors can describe trait evolution and will be useful for modeling purposes.

Guénard *et al.* (2013) recently proposed a new phylogenetic eigenfunction analysis called Phylogenetic Eigenvector Mapping (PEM, akin to Moran's Eigenvector Mapping – MEM – proposed for spatial analyses – see Dray *et al.*, 2006; Griffith and Peres-Neto, 2006; Peres-Neto and Legendre, 2010). They proposed PEM particularly to model and predict species traits from phylogenetic relatedness, recently called "phylogenetic imputation" (see also Penone *et al.*, 2014; Swenson, 2014). The main advance of PEM in respect to the original PVR is the explicit incorporation of a model-based approach in which an O-U process is fitted to data and used to warp the edge lengths of the phylogeny accordingly, before extracting eigenvectors using a PCoA. Our purpose here is to empirically compare the original PVR and the new PEM approach in respect to estimated phylogenetic signal and correlated evolution under alternative evolutionary models.

We generated random phylogenies as coalescent (ultrametric) trees with the same numbers of tips (species) as those in Guénard *et al.* (2013): ranging from 50 to 400 (50, 100, 200, and 400). Subsequently, we simulated the evolution of two independent traits (Y_1 and Y_2) on these trees according to an O-U processes with four restraining forces: $\alpha = 0$ (pure diffusion or Brownian Motion), $\alpha = 0.5$ (weak selection), $\alpha = 1$ (medium selection), and $\alpha = 10$ (strong selection), creating more complex curvilinear evolutionary models (see Diniz-Filho *et al.*, 2012c), and analogous to the simulations performed in Guénard *et al.* (2013). We used 100 simulations of each combination of number of species and α values. The phylogenies generated in the first step of our analyses were back transformed to distance matrices, which in turn were used to calculate the phylogenetic eigenvectors using either PVR (as revised by Diniz-Filho *et al.*, 2012a) or PEM (as proposed by Guénard *et al.*, 2013).

We compared PVR and PEM for different sample sizes and O-U models based on four criteria: similarity of eigenvectors (estimated by Procrustes analysis), estimates of phylogenetic signal (R^2 of models), correlation between model residuals (used to estimate correlated evolution), and phylogenetic imputation ability.

Our first comparison between PVR and PEM consisted in evaluating the similarity between the eigenvectors (containing the scores of the n tips) generated by these methods. Thus, for each simulation, we used a Procrustes analysis (Legendre and Legendre, 2012) to measure the match between the configurations ("species ordinations" along the phylogenetic axes) generated by PVR and PEM considering increasingly number of eigenvectors. The m^2 values (the badness-of-fit statistic that measures the level of congruence between two ordination configurations) were transformed to Procrustes correlation (r) by calculating the square root of their complements (Oksanen *et al.*, 2013). A high value of r, say ≥ 0.75, would indicate that the configurations generated by PVR and PEM, for a given eigenvector dimensionality, are strongly concordant.

Second, we modeled traits Y_1 and Y_2, separately, as a function of the eigenvectors generated by PVR and PEM. For each method and before modeling, eigenvector selection was done with a forward stepwise procedure (Blanchet *et al.*, 2008). This step is necessary to circumvent the problem of a perfect fit when all eigenvectors are used (see Rohlf, 2001). The coefficients of determination of the regression of a trait (Y_1 or Y_2) on the selected eigenvectors, derived from PVR and PEM, were interpreted as the amounts of phylogenetic signal given by each method. We then compared these amounts (R^2_{PVR} and R^2_{PEM}) across the 100 simulations obtained under the different evolutionary models (*i.e.*, O-U with different restraining forces).

Third, we used the residuals derived from the PVR models or from the PEM models (see above) to estimate the partial correlation between the two traits after accounting for phylogenetic relatedness among species (see Martins *et al.*, 2002). A partial correlation estimated by each of these methods should then give the input correlation between the traits, defined as "the correlation of the bivariate normal distribution from which the evolutionary changes in the two traits were drawn" or a "measure of the nonhistorical correlation between the two characters, corrected for phylogenetic interdependences" (Martins, 1996). For each method, the correlation between the phylogenetically corrected traits or specific components (*i.e.*, residuals) should not be significantly different from zero. Thus, type I error rates of correlation coefficients were estimated by the ratio between the number of coefficients that differed significantly from zero and the number of simulations (Martins, 1996; Martins *et al.*, 2002).

Our last comparison between PVR and PEM was based on phylogenetic modeling and the prediction of unknown trait values for species, as proposed in Guénard *et al.* (2013). That is, we compared the ability of eigenvectors derived from PVR and PEM to predict unknown trait values for one or several species ('target species'), based on their relative phylogenetic positions, in phylogenies for which trait values were already estimated for a reduced set of species ('model species') (Guénard *et al.*, 2013). Simply put, such prediction is based on a regression model built using the loadings of the 'model species' from the selected eigenvectors derived from PVR or PEM to estimate the trait values of the 'target species'. We followed this procedure to predict trait values for each species at a time as if it were missing from the original set of species (for each combination of species numbers and restraining forces in our simulations). We removed one species ('target species') at a time from the original set of species and then calculated the scores of the remaining species ('model species') to use

them in the regression model to estimate the trait value of the missing species. We then evaluated the predictive power of PVR and PEM and calculated the prediction coefficient proposed by Guénard *et al.* (2013), which can attain values of 1 when all predictions perfectly match the observations, or below 1 indicating imperfect predictions, and values close to 0 (positive or negative) when predictions are no better than expected by chance (see also Penone *et al.*, 2014 for a recent discussion on "imputation" methods).

Phylogenies and simulations of trait evolution were, respectively, done using the functions *rcoal* and *rTraitCont* from the *ape* package (Analyses of Phylogenetics and Evolution; see Paradis, 2012). Eigenvectors from PVR and PEM were respectively extracted using the packages *PVR* (Santos *et al.*, 2013) and *MPSEM* (Modelling Phylogenetic Signals using Eigenvector Maps; Guénard and Legendre, 2013), available in the R environment for statistical computing (R Development Core Team, 2012).

Our results show that, in general, PVR and PEM are very similar according to the four criteria we devised for comparison. The Procrustes correlations between the two sets of eigenvectors tend to decrease when more eigenvectors (*i.e.*, with smallest eigenvalues) are used in the comparison, even under a Brownian motion model, where the parameter *a* of PEM (analogous to α) is set to zero (stochastic fluctuations revealing that PEM probably cause the deviations in the last eigenvectors) (Figure 1). Correlations between the first two eigenvectors derived from PVR and PEM are as high as 0.95, and decrease to no less than 0.5 when all eigenvectors are used in the Procrustes analysis.

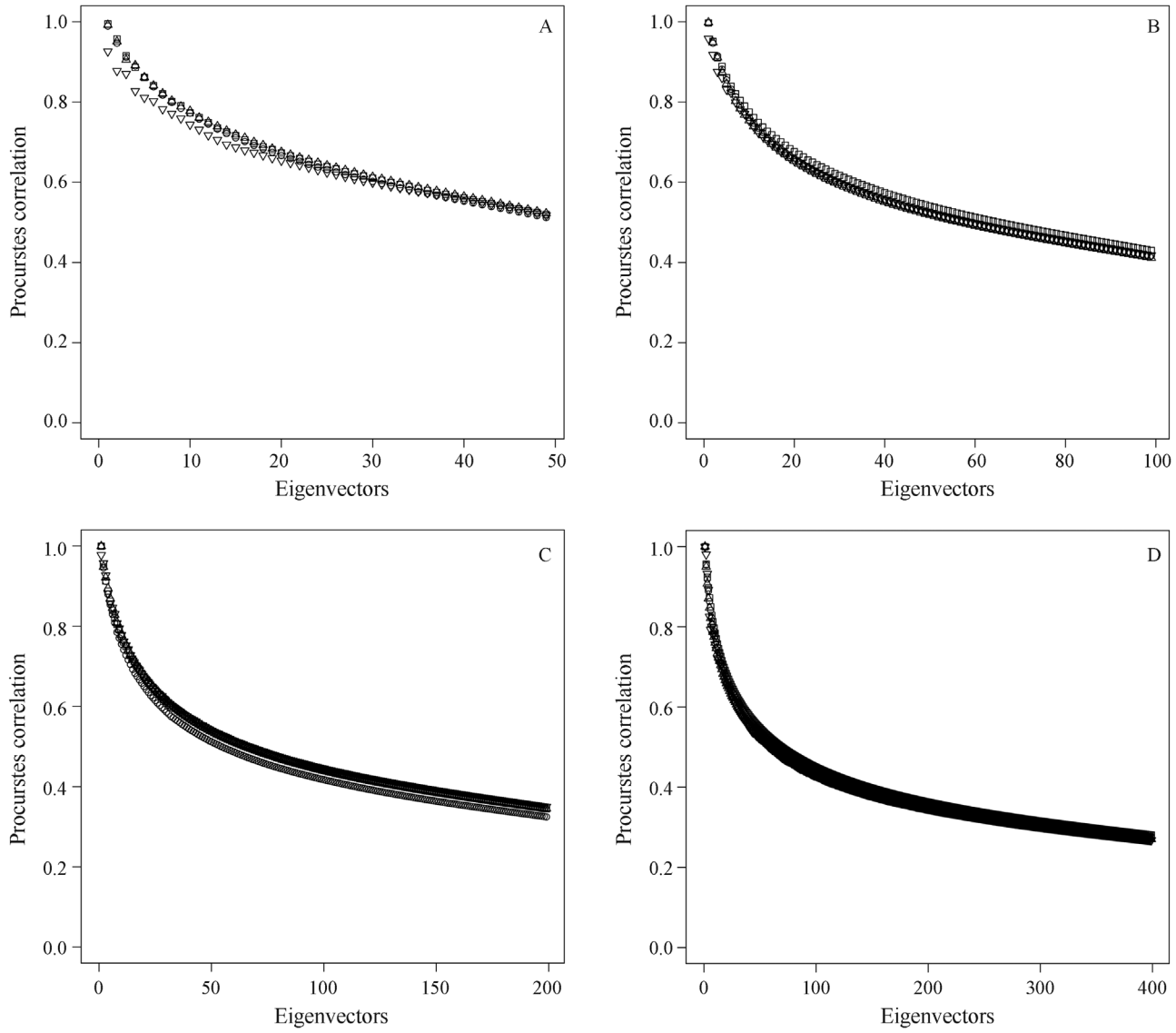

Figure 1 - Procrustes correlation between eigenvectors from PEM and PVR, for a successive set of eigenvectors extracted from phylogenies of (A) 50 species, (B) 100 species, (C) 200 species, and (D) 400 species and traits evolving under alternative O-U models (open circles: $\alpha = 0$; open squares: $\alpha = 0.5$; open triangle point-up: $\alpha = 1$; open triangle point-down: $\alpha = 10$).

This is expected as PEM warps the edge lengths to take into account patterns of evolutionary deviations (even slightly) from Brownian motion. Thus, it will tend to give more weight to edges close to the tips, producing the differences between the eigenvectors of PVR and PEM with the smallest eigenvalues (see Diniz-Filho and Nabout, 2009; Diniz-Filho *et al.*, 2012c). There is a small, albeit consistent, tendency that when PEM is fitting traits evolving under O-U processes with higher restraining forces, the correlation between the two sets of eigenvectors is slightly lower (Figure 1).

The average amounts of phylogenetic signal and Type I errors estimated by PVR (R^2_{PVR}) and PEM (R^2_{PEM}) were highly similar, independently of the sample size and evolutionary models (Table 1). Nevertheless, both methods were unable to provide a consistent type I error of 5% (average for PVR and PEM was around 9%, respectively, see Table 1) (see Rohlf, 2001; Martins *et al.*, 2002; Freckleton *et al.*, 2011), and these increase with sample sizes. On the other hand, the correlation across tips between the specific components of the simulated traits (residuals from PVR and PEM or the expected values of the traits that are independent of phylogenetic structure) was also high, but mainly when the restraining forces were low or equal to zero.

Prediction coefficients were also highly congruent between PVR and PEM, with both coefficients having values of less than 1 in all cases and varying in the same manner with the different number of species (sample size) and restraining forces (evolutionary models) used in our simulations (Table 1). PEM values are slightly larger than those from PVR, and prediction coefficients from both PVR and PEM tend to increase with sample size within a single restraining force, and to decrease altogether with higher restraining forces. At low or no restraining forces ($\alpha = 0$, $\alpha = 0.5$), prediction coefficients of both methods were similarly high, indicating the higher phylogenetic signal produced by those evolutionary models.

We understand that, despite similarity between the two approaches, PEM has a slightly higher prediction ability, especially when there is strong phylogenetic signal (low α values - see Table 1). Also, it is more general than the original PVR because it allows incorporating explicit evolutionary models. Thus, it may solve, perhaps with further improvements in the process of eigenvector selection, some of the problems raised by Freckleton *et al.* (2011) in respect to poorer (in comparison with PGLS) statistical performance of PVR. Our results show that PEM, however, does not provide entirely accurate Type I errors under Brownian motion and so does not perform better than PGLS (according to the previous analyses from the literature; *e.g.*, Freckelton *et al.* [2011]). This also reinforce the issues on phylogenetic eigenvectors theoretically pointed out by Rohlf (2001; see also Diniz-Filho *et al.*, 2012a,b,c for the same argument in the context of PSR curve). However, notice that recent papers still support the use of phylogenetic eigenvector methods, such as PEM or PVR, in the

Table 1 - Comparison of PEM and PVR for phylogenies with different sample sizes (n), simulating trait evolution with distinct restraining forces of an O-U process (α, in which $\alpha = 0$ indicates Brownian motion). The comparison includes the phylogenetic signal estimated by the two methods (R^2), the correlation between model residuals (r), the prediction coefficient used in phylogenetic imputation and the Type I errors of correlated evolution.

n	α	R^2_{PVR}	R^2_{PEM}	*r*	Prediction coefficient		Type I error	
					PVR	PEM	PVR	PEM
50	0	0.982	0.982	0.87	0.516	0.525	0.06	0.07
	0.5	0.977	0.976	0.759	0.475	0.481	0.11	0.07
	1	0.969	0.969	0.86	0.177	0.295	0.04	0.1
	10	0.889	0.889	0.53	0.116	0.196	0.09	0.08
100	0	0.983	0.983	0.933	0.591	0.586	0.07	0.09
	0.5	0.981	0.98	0.908	0.293	0.39	0.03	0.05
	1	0.979	0.979	0.894	0.29	0.436	0.08	0.11
	10	0.955	0.942	0.54	0.029	-0.103	0.04	0.06
200	0	0.977	0.977	0.954	0.83	0.853	0.06	0.07
	0.5	0.974	0.974	0.942	0.753	0.67	0.11	0.09
	1	0.972	0.972	0.896	0.681	0.77	0.1	0.1
	10	0.955	0.954	0.853	-0.057	-0.023	0.08	0.08
400	0	0.975	0.975	0.948	0.927	0.907	0.19	0.19
	0.5	0.971	0.971	0.936	0.926	0.933	0.19	0.14
	1	0.965	0.965	0.946	0.877	0.915	0.18	0.17
	10	0.933	0.93	0.935	-0.039	-0.215	0.15	0.15

context of "phylogenetic imputation" (see Guénard *et al.*, 2013; Swenson, 2014; Penone *et al.*, 2014)

Despite the similarities between PEM and PVR, from a conceptual point of view we understand that PEM may provide an alternative to the original PVR method, being more effective in taking into account phylogenetic signal in trait evolution. This is because PEM may be viewed as technique in the best of both worlds, combining the flexibility of data-driven and empirical eigenfunction analyses (see Griffith and Peres-Neto, 2006) and the sounding insights provided by evolutionary models well known in comparative analyses.

Acknowledgments

Work by JAFD-F and LMB on comparative methods and macroecology has been continuously supported by CNPq productivity fellowships and grants. Work by F.V. was supported by PDJ and BJT ("Science without Borders") grants from CNPq.

References

Blanchet FG, Legendre P and Borcard D (2008) Forward selection of explanatory variables. Ecology 89:2623-2632.

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

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

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

Diniz-Filho JAF, Rangel TF, Santos T and Bini LM (2012a) Exploring patterns of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regression. Evolution 66:1079-1090.

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

Diniz-Filho JAF, Santos T, Rangel TF and Bini LM (2012c) A comparison of metrics for estimating phylogenetic signal under alterrnative evolutionary models. Genet Mol Biol 35:673-679.

Dray S, Legendre P and Peres-Neto PR (2006) Spatial modeling: A comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecol Model 196:483-493.

Freckleton RP, Cooper N and Jetz W (2011) Comparative method as a statistical fix: the dangers of ignoring evolutionary models. Am Nat 178:E10-E17.

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

Guénard G, Legendre P and Peres-Neto PR (2013) Phylogenetic eigenvector mapping: A framework to model and predict species trait. Methods Ecol Evol 4:1120-1131.

Kuhn I, Nobis MP and Durka W (2009) Combining spatial and phylogenetic eigenvector filtering in trait analysis. Global Ecol Biogeogr 18:745-758.

Laurin M (2010) Assessment of the relative merits of a few methods to detect evolutionary trends. Syst Biol 59:689-704.

Legendre P and Legendre L (2012) Numerical Ecology. 3rd edition. Elsevier, Amsterdam, 990 pp.

Martins EP (1996) Phylogenies, spatial autoregression and the comparative method: A computer simulation test. Evolution 50:1750-1765.

Martins EP, Diniz-Filho JAF and Housworth EA (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56:1-13.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K and Thuiller W (2012) How to measure and test phylogenetic signal. Methods Ecol Evol 3:743-756.

Paradis E (2012) Analysis of Phylogenetics and Evolution with R. 2nd edition. Springer, New York, 386 pp.

Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE, Graham CH and Costa GC. (2014) Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol Evol 5:961-970.

Peres-Neto PR (2006) A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. Oecol Austral 10:105-119.

Peres-Neto PR and Legendre P (2010) Estimating and controlling for spatial autocorrelation in the study of ecological communities. Global Ecol Biogeogr 19:174-184.

Peres-Neto PR, Leibold MA and Dray S (2012) Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. Ecology 93:S14-S30.

R Development Core Team (2012) R: A language and environment for statistical computing (http://www.R-project.org). R Foundation for Statistical Computing, Vienna, Austria.

Rohlf FJ (2001) Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143-2160.

Safi K, Pettorelli N (2010) Phylogenetic, spatial and environmental components of extinction risk in carnivores. Global Ecol Biogeogr 19:352-362.

Swenson NG (2014) Phylogenetic imputation of plant functional trait databases. Ecography 37:105-110.

Internet Resources

Oksanen JFG, Blanchet R, Kindt P, Legendre PR, Minchin RB, O'Hara RB, Simpson GL, Solymos P, Stevens MHH and Wagner H (2013) Vegan: Community Ecology Package. R package version 2.0-8, http:// CRAN.R-project.org/package=vegan.

Santos T, Diniz-Filho JAF and Rangel TF (2013). PVR: Computes phylogenetic eigenvectors regression (PVR) and phylogenetic signal-representation curve (PSR) (with null and Brownian expectations). R package version 0.2.1, http://CRAN.R-project.org/package=PVR.

Associate Editor: Louis Bernard Klaczko