SUPPLEMENTAL MATERIAL FOR

Comparative analysis of principal components can be misleading

Josef C. Uyeda, Daniel S. Caetano, & Matthew W. Pennell

Department of Biological Sciences & Institute for Bioinformatics an Evolutionary Studies, University of Idaho

josef.uyeda@gmail.com

appendix: equivalency between ornstein-uhlenbeck and accelerating change models

In our paper, we investigate the scenario in which the individual traits have each evolved under a different model. To simulate the data, we drew values for the exponential rate parameter r of the accelerating/decelerating change (ACDC; Blomberg et al., 2003) model for each trait from a normal distribution with mean 0. We claim that when r is positive, the ACDC model generates traits with a structure equivalent to those produced by a single optimum Ornstein-Uhlenbeck (OU; Hansen, 1997) model. To our knowledge, this has not been previously demonstrated in the literature. Slater et al. (2012) suggested that these two models were equivalent for ultrametric trees: "Looking at extant taxa only, the outcome of [a process with accelerating rates] is very similar to an OU process, as both tend to erase phylogenetic signal" [p. 3940], though they did not provide any proof.

Conjecture. A single optimum OU process produces identical covariance matrices to those produced by the AC model when i) the tree is ultrametric and ii) the trait is assumed to be at the optimum at the root of the tree.

Proof. Consider a bifurcating tree of depth T with two terminal taxa i and j that are sampled at the present and share a common ancestor at time t_{ij} where $t_{ij} < T$. A trait Y is measured for both i and j.

Ornstein-Uhlenbeck process

First, assume that Y has evolved according to an Ornstein-Uhlenbeck (OU) process

$$dY = -\alpha(Y - \theta)dt + \sigma dW \tag{1}$$

where θ is the optimum trait value, α is the strength of the pull towards θ, and σ is the rate of the Brownian diffusion process dW (Hansen, 1997). Also assume that the process

began at the optimum, such that $Y(t=0) = \theta$. The expected value for Y_i and Y_j is equal to the root state. The expected variance for both Y_i and Y_j is given by Hansen (1997):

$$\text{Var}[Y_i] = \text{Var}[Y_j] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha T}) \tag{2}$$

The expected covariance between lineages Y_i and Y_j is given by

$$\text{Cov}[Y_i, Y_j] = \frac{\sigma^2}{2\alpha} e^{-2\alpha(T - t_{ij})}(1 - e^{-2\alpha t_{ij}}) \tag{3}$$

The correlation between Y_i and Y_j, $\rho[Y_i, Y_j]$, is defined as

$$\rho[Y_i, Y_j] = \frac{\text{Cov}[Y_i, Y_j]}{\sqrt{\text{Var}[Y_i]\text{Var}[Y_j]}}$$

Under an OU process, $\rho[Y_i, Y_j]$ is

$$\rho[Y_i, Y_j] = \frac{\frac{\sigma^2}{2\alpha} e^{-2\alpha(T - t_{ij})}(1 - e^{-2\alpha t_{ij}})}{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha T})} \tag{4}$$

With some algebra, it is straightforward to reduce Equation 4 to

$$\frac{1 - e^{2\alpha t_{ij}}}{1 - e^{2\alpha T}} \tag{5}$$

Accelerating Change Model

Next, assume that Y has evolved according to the Accelerating Change (AC) model, which describes a Brownian motion process in which the rate of diffusion σ^2 changes as function of time

$$dY(t) = \sigma(t) dW. \tag{6}$$

Specifically, we consider the functional form of $\sigma^2(t)$ to be

$$\sigma^2(t) = \sigma_0^2 e^{rt}$$

where r is constrained to be positive (Blomberg et al., 2003; Slater et al., 2012). The expected value of the AC model is also equal to the root state. The expected variance for Y_i and Y_j is given by

$$\text{Var}[Y_i] = \text{Var}[Y_j] = \int_0^T \sigma_0^2 e^{rT} dt = \sigma_0^2 \left(\frac{e^{rT} - 1}{r} \right) \tag{7}$$

(Harmon et al., 2010) and the covariance is equal to

$$\text{Cov}[Y_i, Y_j] = \int_0^{t_{ij}} \sigma_0^2 e^{rt_{ij}} dt = \sigma_0^2 \left(\frac{e^{rt_{ij}} - 1}{r} \right) \tag{8}$$

Under the AC model, $\rho[Y_i, Y_j]$ is

$$\rho[Y_i, Y_j] = \frac{\sigma_0^2 \left(\frac{e^{rt_{ij}} - 1}{r} \right)}{\sigma_0^2 \left(\frac{e^{rT} - 1}{r} \right)} \tag{9}$$

Equation 9 can be easily reduced to

$$\frac{1 - e^{rt_{ij}}}{1 - e^{rT}} \tag{10}$$

Comparing the expectations under OU and AC

Comparing equations 4 and 9, it is clear that the correlation between Y_i and Y_j under the OU model is equal to that of the AC model

$$\frac{1 - e^{2\alpha t_{ij}}}{1 - e^{2\alpha T}} = \frac{1 - e^{rt_{ij}}}{1 - e^{rT}} \tag{11}$$

when $\alpha = 0.5r$. For every value of α there is a value of r that can produce an identical correlation structure. Note that the values of σ^2 and σ_0^2 do not affect the correlation structure but do enter into the covariance structure. As $\text{Cov}[Y_i, Y_j] = \text{Var}[Y_i, Y_j]\rho[Y_i, Y_j]$ and the values of $\rho[Y_i, Y_j]$ are equivalent, we can set the variances of two models (given by Equations 2 and 7) equal to one another

$$\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha T}) = \sigma_0^2 \left(\frac{e^{rT} - 1}{r} \right) \tag{12}$$

and substitute r for 2α

$$\frac{\sigma^2}{r}(1 - e^{-rT}) = \sigma_0^2 \left(\frac{e^{rT} - 1}{r} \right)$$

Reducing algebraically, it is easy to show that

$$\sigma^2 = \sigma_0^2 e^{rT} \tag{13}$$

Therefore for any covariance matrix for Y_i and Y_j, OU and AC are completely unidentifiable and the likelihoods for the two models will be identical.

□

Notes. The two variances $\text{Var}[Y_i]$ and $\text{Var}[Y_j]$ will only be equal to one another when the tree is ultrametric. If either i or j were not sampled at the present (e.g., if one was an extinct lineage), this proof for the non-identifiability of OU and AC does not hold and one can potentially distinguish these models (Slater et al., 2012).

SUPPLEMENTARY FIGURES

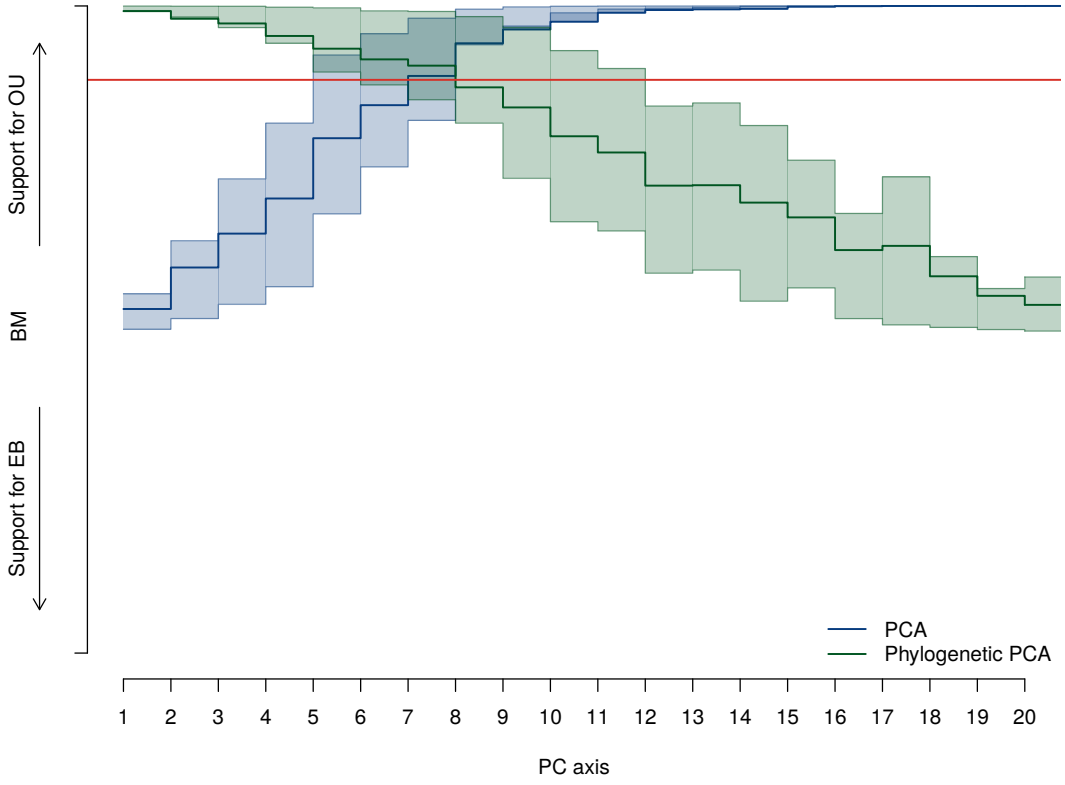

FIGURE S.1: Distribution of support for BM, OU and EB models when the generating model is a uncorrelated multivariate OU model. Support for models were transformed into a linear scale by calculating an overall model support statistic: $AICw_{OU} - AICw_{EB}$. Thus high values support OU, low values support EB, and intermediate values near 0 indicate BM-like evolution. Models were fit to each replicated dataset for each of 20 different traits which were taken either from PC scores (blue line) or phylogenetic PC scores (green line). Shaded regions indicate the 25^{th} and 75^{th} quantiles of the model support statistic for 100 replicated datasets. The red line indicates the average model support statistic averaged over all 20 original trait variables. Note that standard PCA results in Akaike weights that are skewed toward EB for the first few PCs of standard PCA, while and that later PCs subsequently favor OU models. By contrast, pPCA results in Akaike weights that are skewed toward stronger support for OU models relative to the original trait variables.

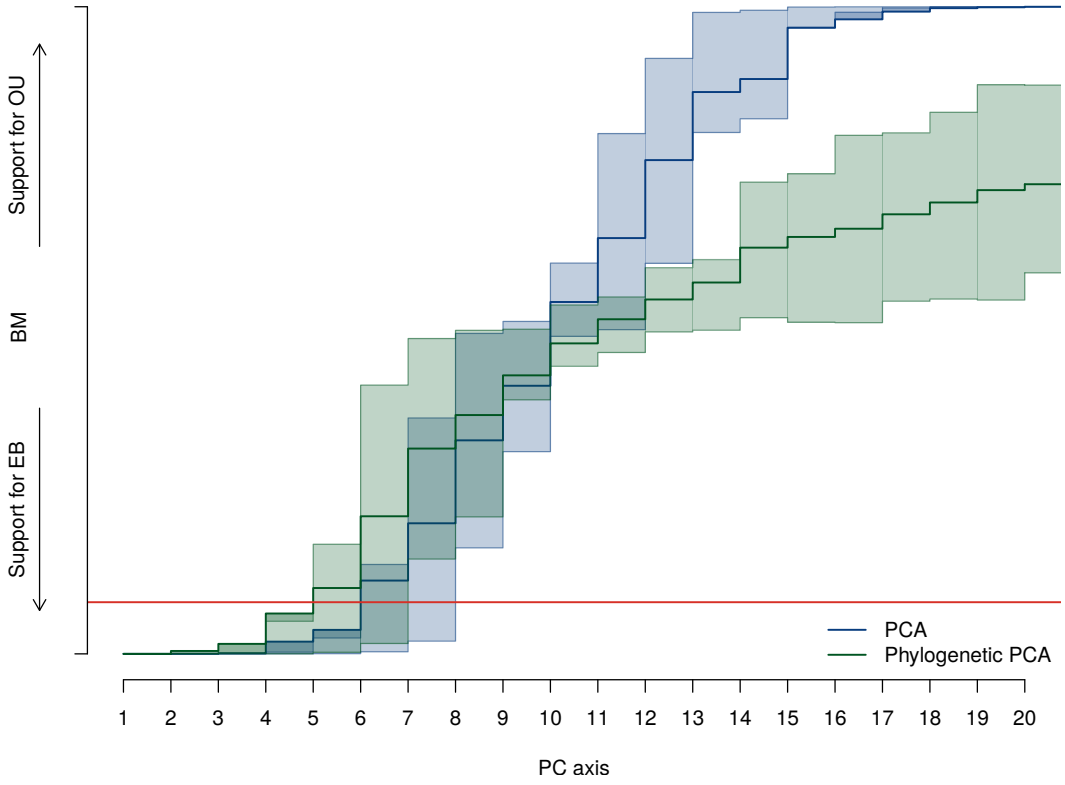

FIGURE S.2: Distribution of support for BM, OU and EB models when the generating model is an uncorrelated multivariate EB model. Support for models were transformed into a linear scale by calculating an overall model support statistic: $AICw_{OU} - AICw_{EB}$. Thus high values support OU, low values support EB, and intermediate values near 0 indicate BM-like evolution. Models were fit to each replicated dataset for each of 20 different traits which were taken either from PC scores (blue line) or phylogenetic PC scores (green line). Shaded regions indicate the 25^{th} and 75^{th} quantiles of the model–support statistic for 100 replicated datasets. The red line indicates the average model support statistic averaged over all 20 original trait variables. Note that both pPCA and PCA increase support for EB models for early PC axes.

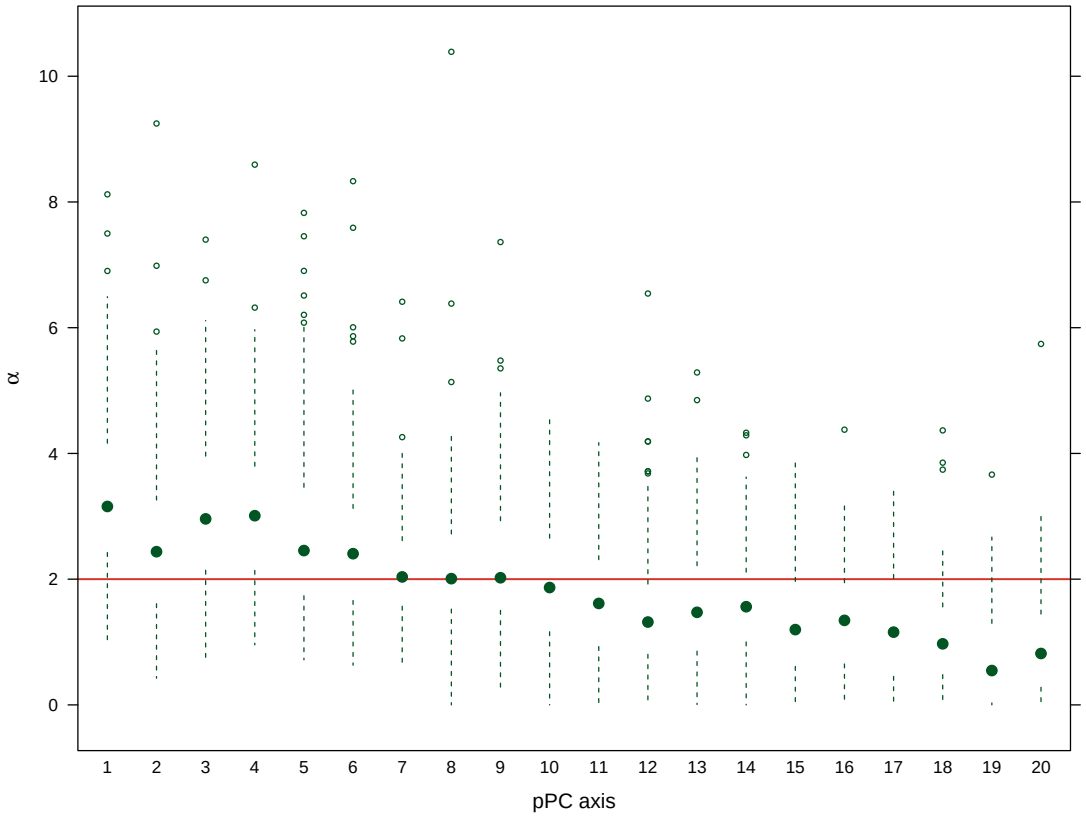

FIGURE S.3: Estimated values of the α parameter from phylogenetic PCA when data is simulated under a correlated multivariate OU model. The simulating value $\alpha = 2$ is depicted with the red line. The estimate of α is inflated in the first few pPC axes consistent with an exaggerated support for the OU model. The value of α decreases for successive pPC axes, resulting in increasing support for a BM model (Figure 1).

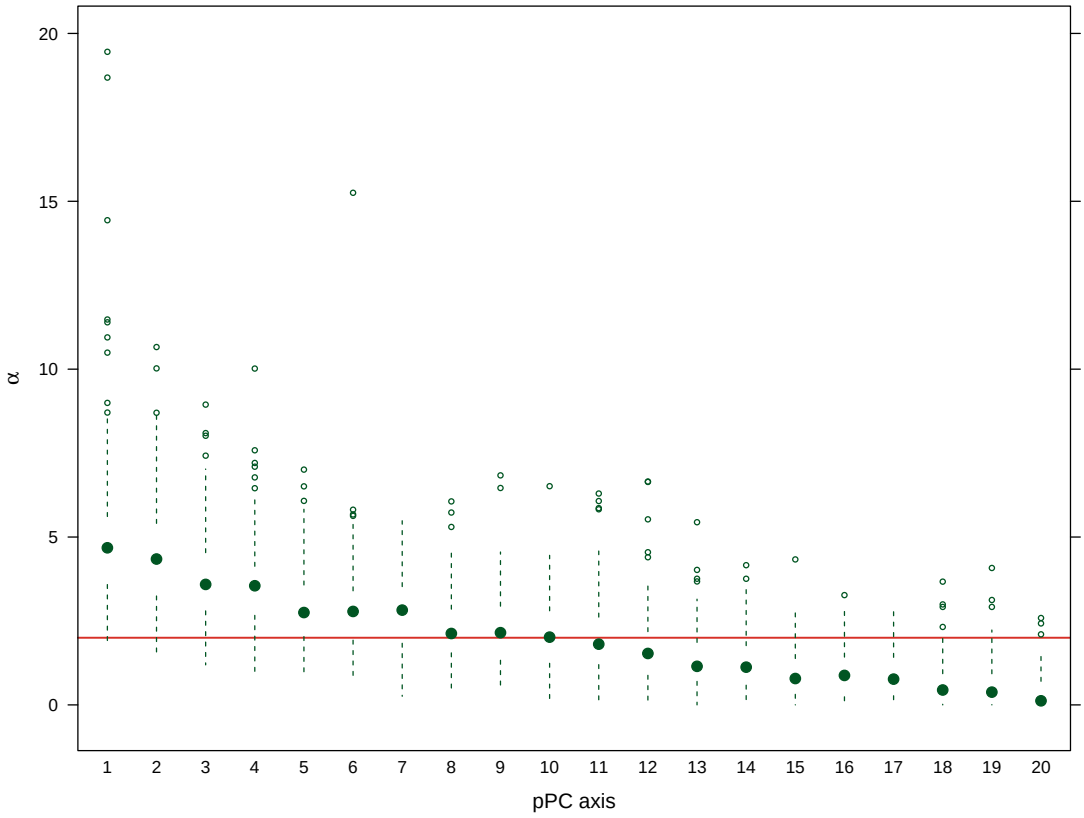

FIGURE S.4: Estimated values of the α parameter from phylogenetic PCA when data is simulated under an uncorrelated multivariate OU model. The simulating value $\alpha = 2$ is depicted with the red line. The estimate of α is inflated in the first few pPC axes consistent with an exaggerated support for the OU model. The value of α decreases for successive pPC axes, resulting in increasing support for a BM model (Figure S.1).

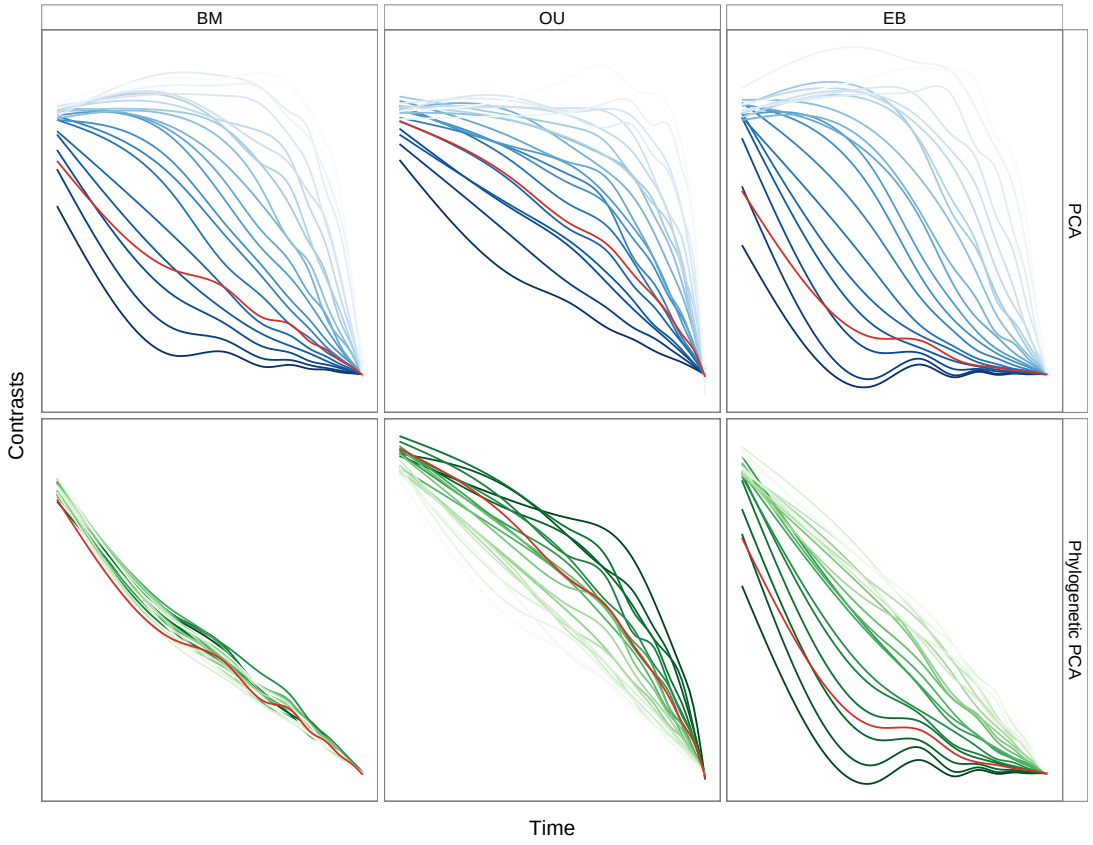

FIGURE S.5: Disparity through time plots averaged across the 100 simulated datasets. The datasets were simulated under BM (left), OU (middle) or EB (right). The analyses were then performed on PC scores (top row) and pPC scores (bottom row). The average disparity through time of all 20 original trait variables is indicated by the red line. We fit a loess curve through the relative disparities for each trait/transformation/model combination. The plots are oriented so that the left side of each panel corresponds to the root of the phylogeny, with time increasing tipward to the right. The intensity of the colors are proportional to the ranking of the PC or pPC axes, stronger lines represent the first axes. As in Fig. 3, the first few axes from the PCA show a strong pattern of high disparity early in the clades' histories with the higher components showing seemingly higher disparity towards the present. Phylogenetic PCA corrects the distortion if the generating model is multivariate BM. However, if the generating model was not BM, the first few pPC axes tend to show an exaggerated pattern of disparity relative to the original traits.

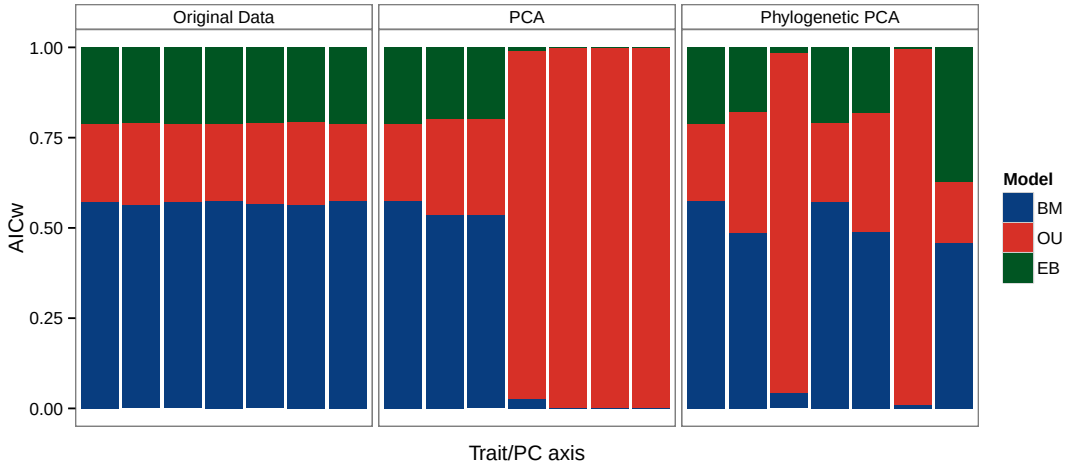

FIGURE S.6: Proportion of support for BM, OU and EB models for each of the traits/PC axes from the morphological dataset of Felidae species from Slater and Van Valkenburgh (2009) and Sakamoto et al. (2010). Traits were log transformed prior to analysis. Note that all original traits and the first axes under standard and phylogenetic PCA show strong support for a BM model.

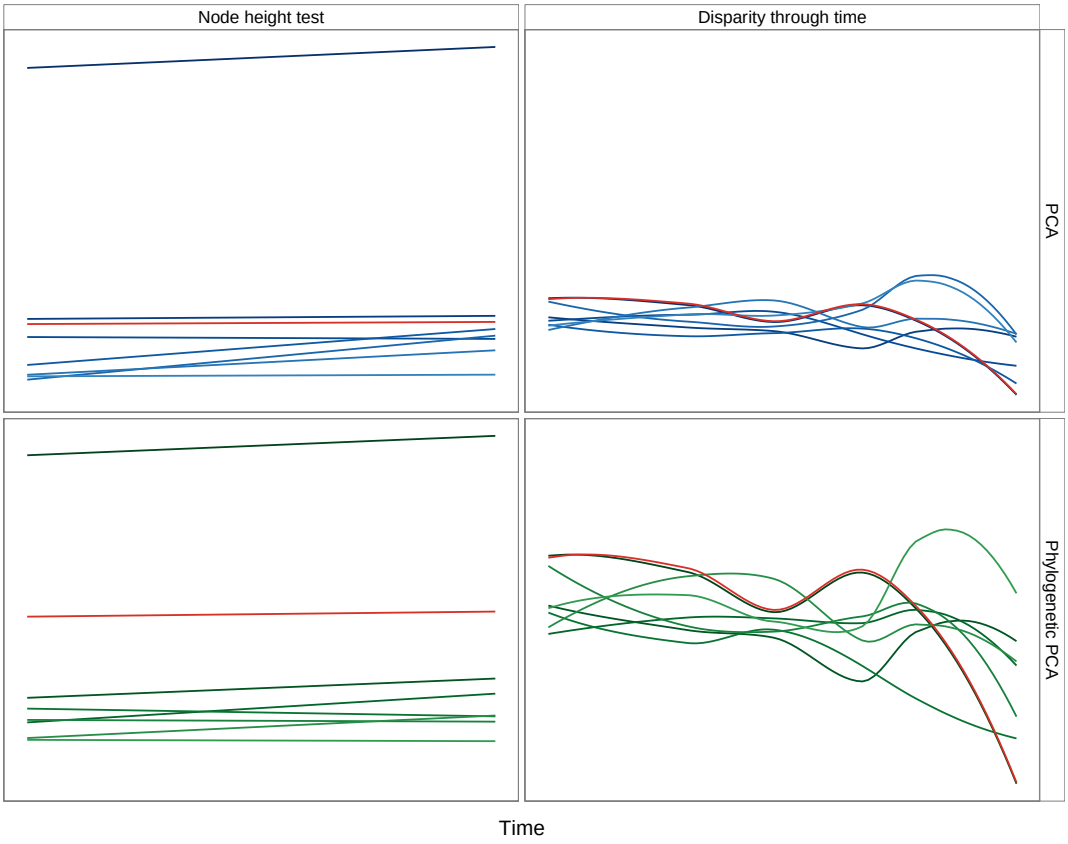

FIGURE S.7: Node height test and disparity through time plots for the morphological dataset of Felidae species. Each line represents a best–fit linear model (left) or loess curve fitted (right) to the original traits, PC or pPC scores. All traits were log transformed prior to analysis. The intensity of color is proportional to the ranking of the PC or pPC axes, stronger lines represent the first axes. Left panels show the relationship between the average phylogenetic independent contrasts and the height of the node. Red lines indicate the average value for the original trait values. Right panels show disparity through time plots. The plots are oriented so that the left side of each panel corresponds to the root of the phylogeny, with time increasing tipward to the right. Compare this highly correlated dataset with only 7 traits to the larger, less correlated dataset of *Anolis* lizards (Figure S.9).

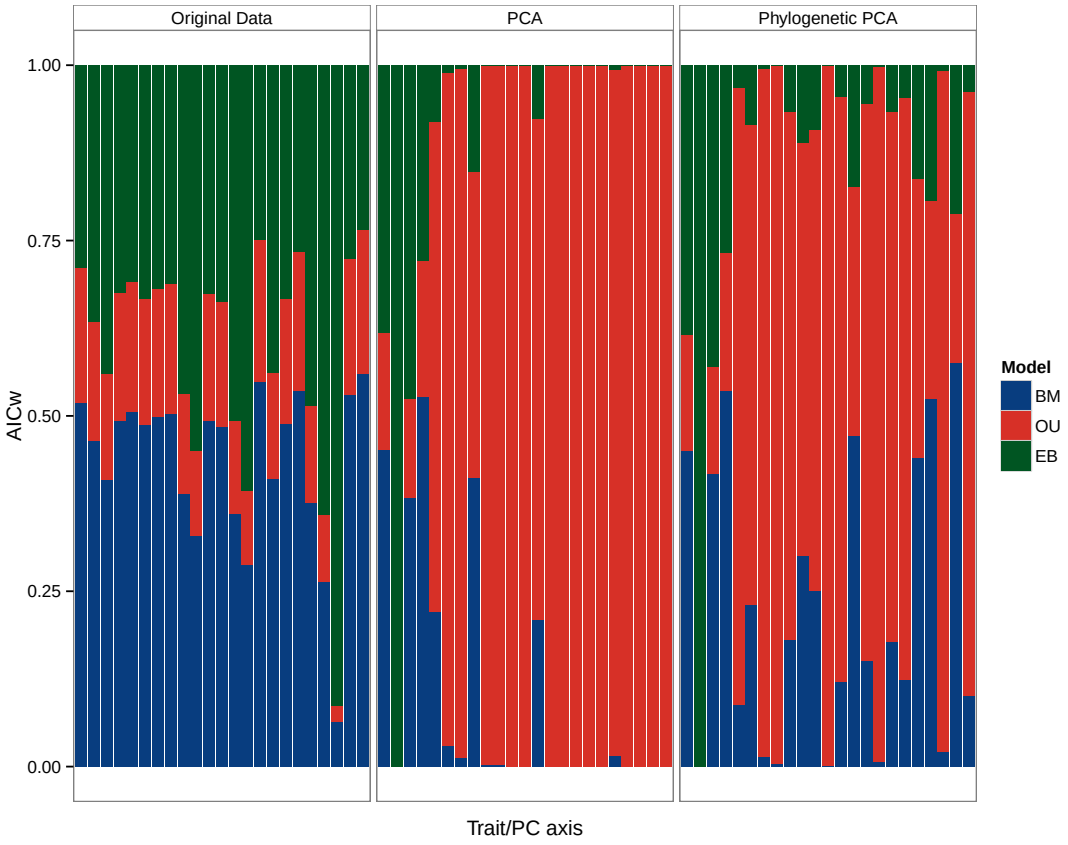

FIGURE S.8: Distribution of support for BM, OU and EB models for a 23–trait morphometric dataset taken from Mahler et al. (2010). Support is measured in Akaike weights across all original trait variables (left), as well as standard PCA (middle) and pPCA (right). For both PCA and pPCA, support for the EB model appears to be concentrated in PCs 1-4, with a suggestive pattern of decreasing support across PCs 2-4.

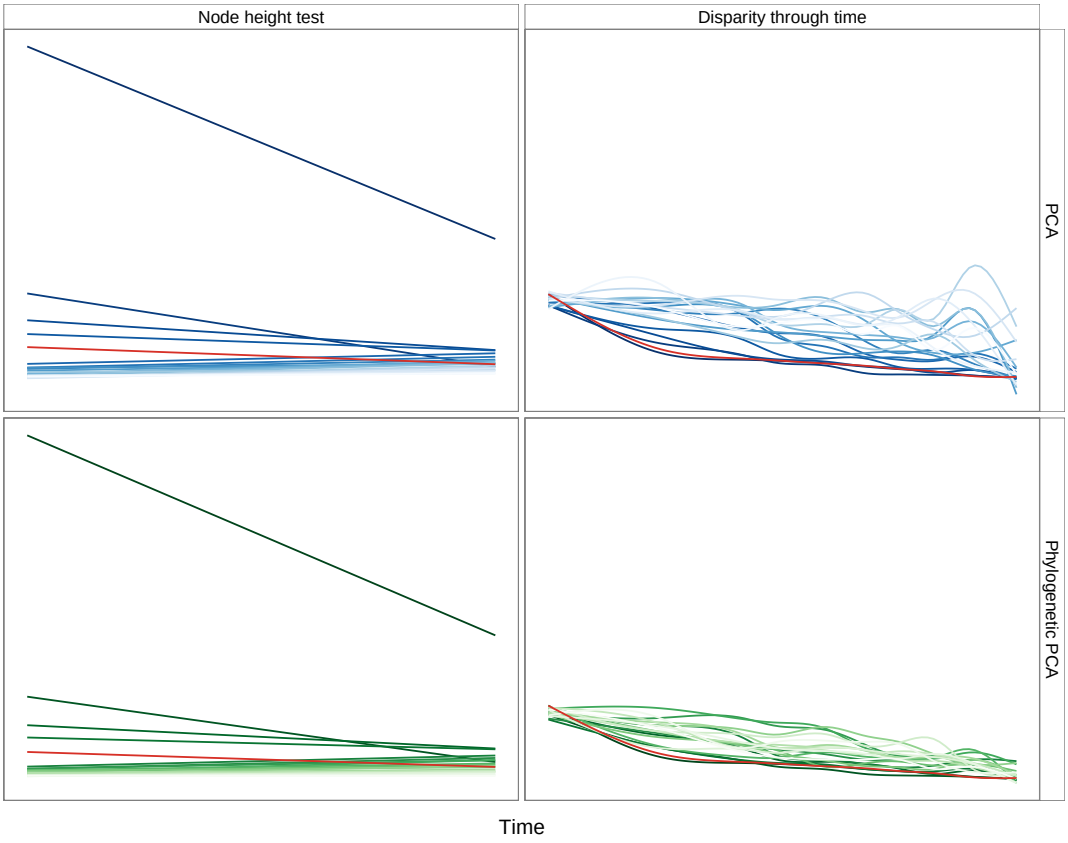

FIGURE S.9: Node height test and disparity through time plots for the morphological dataset of *Anolis* lizards. Each line represents a best–fit linear model (left) or loess curve fitted (right) to the original traits, PC or pPC scores. All traits were log transformed prior to analysis. The intensity of color is proportional to the ranking of the PC or pPC axes, stronger lines represent the first axes. Left panels show the relationship between the average phylogenetic independent contrasts and the height of the node. Red lines indicate the average value for the original trait values. Right panels show disparity through time plots. The plots are oriented so that the left side of each panel corresponds to the root of the phylogeny, with time increasing tipward to the right.

REFERENCES

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, A. Purvis, R. E. Ricklefs, D. Schluter, J. A. Schulte II, O. Seehausen, B. L. Sidlauskas, O. Torres-Carvajal, J. T. Weir, and A. Ø. Mooers. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Mahler, D. L., L. J. Revell, R. E. Glor, and J. B. Losos. 2010. Ecological opportunity and the rate of morphological evolution in the diversification of greater antillean anoles. Evolution 64:2731–2745.

Sakamoto, M., G. T. Lloyd, and M. J. Benton. 2010. Phylogenetically structured variance in felid bite force: the role of phylogeny in the evolution of biting performance. Journal of Evolutionary Biology 23:463–478.

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Slater, G. J. and B. Van Valkenburgh. 2009. Allometry and performance: the evolution of skull form and function in felids. Journal of Evolutionary Biology 22:2278–2287.