COMPARING THE STRENGTH OF MODULAR SIGNAL, AND EVALUATING ALTERNATIVE MODULAR HYPOTHESES, USING COVARIANCE RATIO EFFECT SIZES WITH MORPHOMETRIC DATA

Dean C. Adams[1,*] and Michael L. Collyer[2]

02 October, 2019

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA.

[2]Department of Science, Chatham University, Pittsburgh, Pennsylvania, USA.

*Correspondence: Dean C. Adams dcadams@iastate.edu

Keywords: Morphometrics, modularity, trait covariation, morphological evolution

Short Title: An Effect Size for Comparing Modularity

Author Contributions: DCA and MLC contributed equally to all parts of this manuscript.

Data Archiving: doi:10.5061/dryad.h18931zfs

Acknowledgments: We thank A. Kaliontzopoulou for comments on drafts of this manuscript. E. Márquez kindly provided the data for the empirical example. This work was sponsored in part by National Science Foundation Grants DEB-1556379 and DBI-1902511 (to DCA) and DEB-1737895 and DBI-1902694 (to MLC).

Introduction

Characterizing the extent to which phenotypic traits covary, and deciphering the forces that shape patterns of trait covariation, are perennial topics in evolutionary biology. Empirical research has demonstrated that patterns of covariation are unevenly dispersed across traits, with some variables exhibiting high correlations with one another while other traits are more independent (Olson and Miller 1958; Cheverud 1996; Klingenberg 2008; Goswami and Polly 2010). Indeed, studies of morphological integration and modularity are largely concerned with how these trait correlations are distributed, and the extent to which they are driven by developmental, genetic, or functional linkages among traits (Olson and Miller 1958; Wagner 1984). In particular, modularity describes patterns where trait correlations are high within subsets of variables, termed *modules*, and where trait correlations across modules are comparatively weaker (Cheverud 1982; Goswami 2006; Wagner et al. 2007). Patterns of modularity therefore result in sets of semi-autonomous traits, which have the potential to respond differentially to natural selection, and thus the capacity to promote the evolution of novelty (see Wagner and Altenberg 1996; Wagner et al. 2007; Tokita et al. 2007; Hansen and Houle 2008; Clune et al. 2013).

The past several decades have seen resurging interest in understanding the evolution of modularity, and numerous analytical approaches have been developed for quantifying patterns of modularity in phenotypic datasets (e.g., Magwene 2001; Mitteroecker and Bookstein 2007; Márquez 2008; Klingenberg 2009; Adams 2016; Goswami and Finarelli 2016). Concomitant with these advances is an increasing number of empirical studies that characterize patterns of modularity in distinct phenotypic traits, and across a wide variety of taxa (for recent examples, see: Parsons et al. 2012; Parr et al. 2016; Felice and Goswami 2018; Larouche et al. 2018; Bardua et al. 2019). Likewise, evolutionary biologists have striven to decipher whether patterns of modularity are similar among taxa and traits, and across levels of biological organization (Drake and Klingenberg 2010; Renaud et al. 2012; Sanger et al. 2012; Felice and Goswami 2018; Bardua et al. 2019; Marshall et al. 2019). Some studies have investigated whether patterns of modularity are conserved across taxa (e.g., Goswami 2006; Marshall et al. 2019), however, direct quantitative or statistical comparisons of modularity patterns are generally lacking. We assert that a critical aspect of this endeavor should be to determine whether the *strength* of modularity is similar across datasets. However, to our knowledge, no formal statistical procedure has yet been proposed to accomplish this task.

Another key challenge in the study of modularity is identifying which sets of traits represent anatomical

modules (Zelditch et al. 1990; Hallgrimsson et al. 2007). Biologically, one expects that traits subjected to common processes may exhibit high correlations with one another, resulting in modular structure (Olson and Miller 1958; Wagner 1984; Cheverud 1996; Goswami and Polly 2010). However, phenotypic traits are influenced by an array of genetic, developmental, and functional forces, not all of which act similarly. Thus, it is reasonable to expect that multiple, competing modular hypotheses may be proposed for the same set of anatomical traits (e.g., Márquez 2008; Goswami and Finarelli 2016; Felice and Goswami 2018; Bardua et al. 2019). This then raises the question: which of the alternative modular hypotheses provides the best description of the observed patterns of trait covariation? Several recent attempts to address this problem (e.g., Márquez 2008; Goswami and Finarelli 2016) have greatly sharpened our notions of modularity, and provided insights as to how phenotypic variation may be expected to evolve. For instance, one attempt (`MINT`: Márquez 2008) uses goodness of fit measures to evaluate alternative modular hypotheses, while another approach (`EMMLi`: Goswami and Finarelli 2016) uses penalized likelihood indices (AICc) to evaluate the fit of alternative modular hypotheses to the observed trait correlations.

Nonetheless, these approaches do not provide a complete analytical toolkit for evaluating alternative modular hypotheses for the full spectrum of phenotypic datasets evolutionary biologists wish to examine. One reason for this is that initial investigations into their performance were relatively limited, and suggested that future improvements or alternatives may be required. For example, simulations revealed that `MINT` tended to identify modular patterns when none existed in the data, suggesting that false positives may be a concern (Márquez 2008: Table 3). Likewise, initial investigations found that `EMMLi` tended to select highly paramaterized models [Goswami and Finarelli (2016); Goswami pers. comm.], suggesting that it may display some degree of model misspecification. However, to date, an evaluation of both approaches under a broad set of plausible scenarios has not been conducted, nor has their performance been directly compared (see below). As a consequence, it remains possible that neither approach provides a reliable means of evaluating patterns of modular structure, thereby reducing one's ability to identify evolutionary trends in trait covariation. Therefore, we suggest that a second pressing analytical need in the study of modularity is the development of robust statistical methods for comparing the strength of modular signal across alternative modular hypotheses for the same dataset.

In this article, we develop a single statistical tool that accomplishes both of these tasks. Our approach utilizes a standardized test statistic (an effect size: Z_{CR}) for measuring the degree of modularity between sets of variables, and for comparing these effect sizes statistically. Our procedure is based on the covariance ratio

(Adams 2016), and may be used for comparing the strength of modular signal across datasets, as well as for evaluating patterns of modular signal as defined by alternative modular hypotheses for the same dataset. Using computer simulations we evaluate the statistical properties of tests based on Z_{CR}, and find that they display appropriate type I error rates, high statistical power, and low levels of model misspecification. Thus, when modular signal is present, Z_{CR} is unlikely to mis-assign that signal to an incorrect modular hypothesis. We then compare these results to those found using two alternative methods for comparing modular hypotheses: `MINT` (Márquez 2008) and `EMMLi` (Goswami and Finarelli 2016). We find that both exhibit poor statistical performance (very high false positive rates, and high model misspecification rates), making it challenging to arrive at reliable biological inferences when using these approaches. We illustrate the utility of Z_{CR} with an empirical example of mandible shape in sigmodontine rodents, where alternative modular hypotheses, and the strength of modular signal across several species, is compared. Our new method is implemented in R, and is distributed through the package `geomorph` (Adams et al. 2019). Finally, the method can accommodate future analytical developments for characterizing modular signal as they are developed (see Discussion).

Methods

Analytical Development

Covariance Ratio E ect Sizes (Z_{CR}): One approach to quantifying the degree of modular signal in morphometric datasets is based on the covariance ratio (Adams 2016). This measure is preferable to alternative estimates, such as the RV coe□cient (sensu: Klingenberg 2009), because it is insensitive to the number of variables (p) and the number of specimens (n) (Adams 2016). To calculate the covariance ratio (CR), one first concatenates the variables for all modules into an $n \times p$ matrix (Y). At a minimum, the variables are represented as mean-centered data, implying that they are residuals from an intercept model. However, one could also perform additional transformations to the data, such as a phylogenetic transformation (e.g., Garland and Ives 2000; Adams 2014; Adams and Collyer 2018a) to account for non-independence among objects as a result of shared evolutionary history (see Adams and Felice 2014 for a related approach measuring morphological integration in a phylogenetic context). Using this concatenated dataset, the covariance ratio for the observed data is found as:

$$CR = \sqrt{\frac{tr(\mathbf{S}_{12}\mathbf{S}_{21})}{tr(\mathbf{S}_{11}^{*}\mathbf{S}_{11}^{*})tr(\mathbf{S}_{22}^{*}\mathbf{S}_{22}^{*})}} \quad (1)$$

where $\mathbf{S}_{11}^{*}$ & $\mathbf{S}_{22}^{*}$ represent the within-module covariance matrices for each of two modules (with zeroes replacing the diagonal elements), and $\mathbf{S}_{21} = \mathbf{S}_{12}^{T}$ represents the covariation between modules $\mathbf{Y}_1$ and $\mathbf{Y}_2$. For datasets containing more than two modules, the *CR* coefficient is calculated for each pair of modules, and the average pairwise *CR* coefficient is used (see Adams 2016). Empirically, the *CR* coefficient has a lower limit of 0, an expected value of 1 when there is no modularity, and values greater than 1 are also possible if covariances between module variables exceed covariances within modules. Thus, *CR* values that are closer to 0 describe datasets with a relatively greater degree of modularity, while *CR* values closer to 1 describe datasets with less modular signal. To evaluate *CR* statistically, a resampling procedure that randomly assigns variables to modules is used, and in each iteration a CR_{rand} is obtained under the null hypothesis of no modularity. The observed value, CR_{obs}, is then compared to the empirically-generated sampling distribution of CR_{rand} values to evaluate significance (for details see: Adams 2016).

The permutation procedure above is sufficient to statistically determine whether the degree of modular signal is greater than expected by chance for a single dataset, but evaluating modular signals across datasets requires a standardized effect size to ensure statistical comparability. To accomplish this, we make use of prior theoretical developments for effect sizes obtained from statistics from multivariate data and their empirical sampling distributions (Collyer et al. 2015; Adams and Collyer 2016). Previous studies used residual randomization in permutation procedures (RRPP) to generate sampling distributions. The important insight from Collyer et al. (2015) was that a permutation-based effect size may be obtained from any observed test statistic, found in relation to the mean and standard deviation of its empirical sampling distribution obtained from RRPP. Although the resampling procedure for generating sampling distributions of the *CR* statistic is different than RRPP, the concept of using the random outcomes to evaluate the size of the observed effect is the same. Thus, for the case of the covariance ratio, this effect size is obtained as:

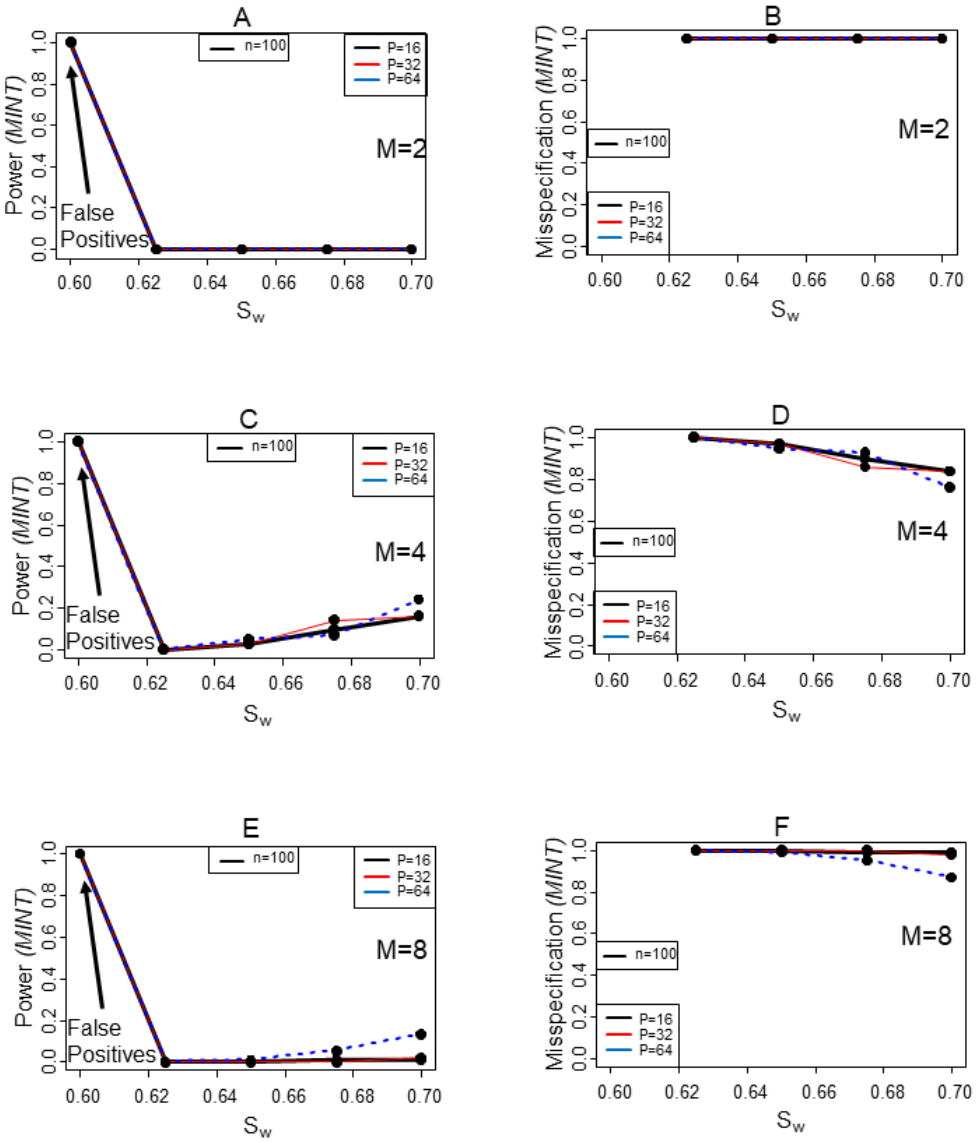
$$Z_{CR} = \frac{CR_{obs} - \hat{\mu}_r}{\hat{\sigma}_r} \quad (2)$$

where CR_{obs} is the observed covariance ratio for the dataset, $\hat{\mu}_r$ is the expected value of *CR* under the

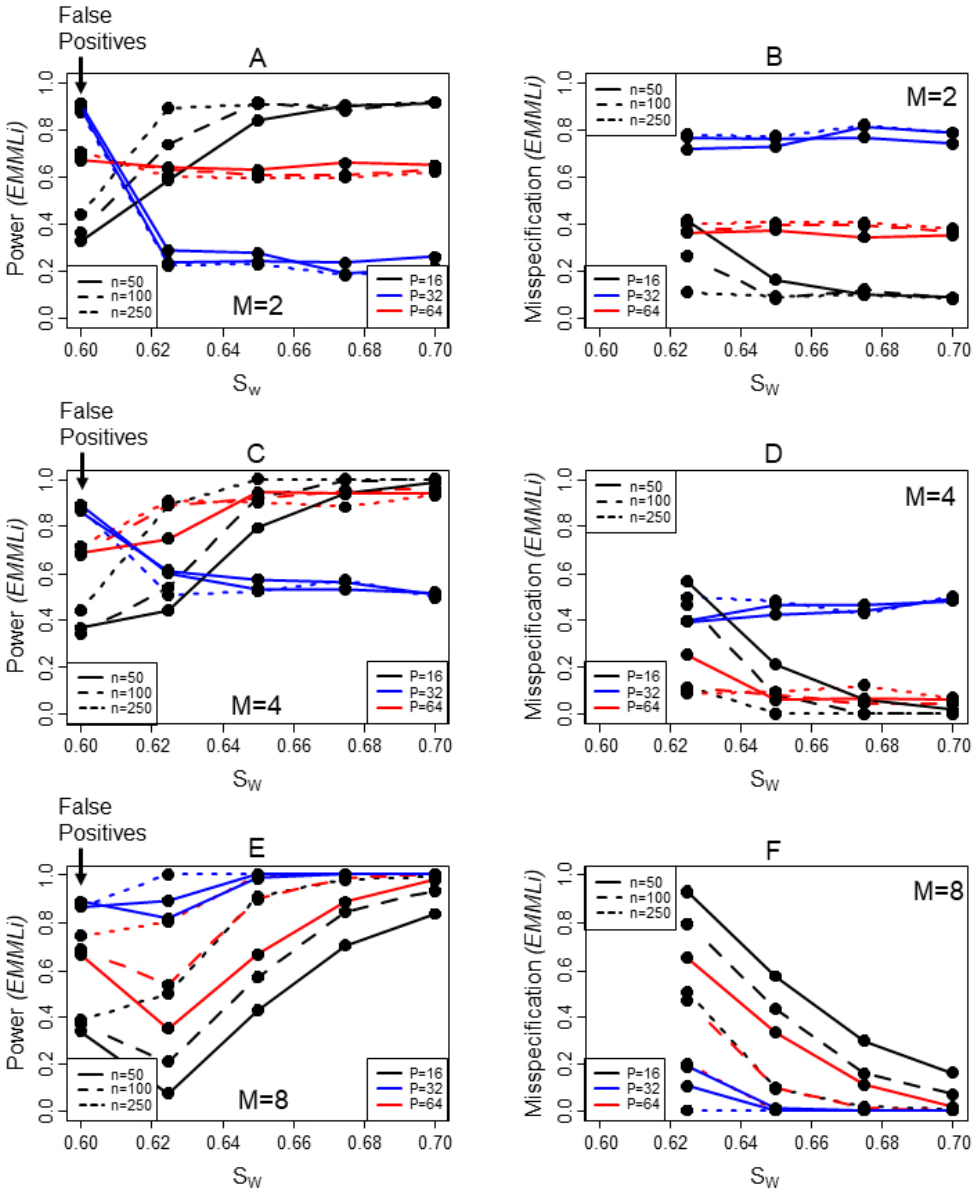
null hypothesis of no modularity (found as the mean of the empirical sampling distribution), and $\hat{\sigma}_r$ is the standard error of the mean, found as the standard deviation of the empirical sampling distribution. Note that the calculation of Z_{CR} is the same regardless of how many modules are represented, because CR_{obs} in equation 2 represents the average CR value obtained across all pairs of modules (see Adams 2016). One should recognize that more negative values of Z_{CR} represent greater modular signal, because when modular signal is present, CR_{obs} will be less than $\hat{\mu}_r$.

Importantly, it can be shown that Z_{CR} is normally distributed, and thus represents a valid standardized effect size. To demonstrate this property we performed a simulation experiment, where 500 multivariate datasets were generated at a given n and p, and with all variates drawn from a normal distribution: $N(0, 1)$. Next, variables were randomly assigned to modules, and both the CR and Z_{CR} were obtained from each dataset, with Z_{CR} obtained using 999 permutations. The mean and standard deviation of Z_{CR} across the 500 datasets were then calculated, and the sampling experiment was repeated across differing levels of n and p. As is clear from Fig. 1, under the null hypothesis of no modularity (i.e., a random association of variables to modules), the covariance ratio effect size exhibits a constant expected value near zero, and constant variance (inferred from the constant confidence interval) across the entire spectrum of sample sizes (n) and variable number (p). Furthermore, the distribution of Z_{CR} values from the 500 datasets for any set of simulation conditions was also found to be normally distributed (Results not shown). Thus, Z_{CR} represents a valid standardized effect size that characterizes the strength of modular signal in morphometric datasets.

A Two-Sample Z-Score for Comparing Modular Signals: Once the degree of modular signal has been characterized, one may wish to evaluate hypotheses that compare multiple effect sizes. For hypotheses of modularity, this is particularly useful in two biological contexts. First, it may be of interest to determine whether the *strength* of modular signal is greater in one dataset as compared to another. To date, no explicit analytical approach has been proposed to statistically compare the strength of modular signal across datasets, though some studies have evaluated whether general patterns of modularity are conserved across taxa (e.g., Goswami 2006; Sanger et al. 2012; Marshall et al. 2019). Second, if several alternative modular hypotheses have been proposed for the same structure (based on developmental, genetic, or functional grounds), it may be of interest to characterize the degree of modular signal under each of these hypotheses and determine which provides the best description of the observed patterns of morphological covariation (see Márquez 2008; Goswami and Finarelli 2016; Felice and Goswami 2018).



189x221mm (96 x 96 DPI)



188x229mm (96 x 96 DPI)

fig 6

169x203mm (96 x 96 DPI)