Consilience of methods for phylogenetic analysis of variance

Consilience of methods for phylogenetic analysis of variance

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

14 April, 2022

[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: Phylogenetic Comparative Methods, Macroevolution, Simulation/Methods, Brownian motion

Short Title: simulation-based phylogenetic anova

Author Contributions: DCA and MLC collaboratively developed the concept and contributed to all portions of this manuscript. All authors approve of the final product and are willingly accountable for any portion of the content.

Conflicts of Interests: The authors declare no conflicts of interest.

Data Archiving: R-scripts for simulation tests are available in the Supplemental Information.

Acknowledgments: This work was sponsored in part by National Science Foundation Grants DBI-1902511 (to DCA) and DBI-1902694 (to MLC).

Abstract

Simulation-based and permutation-based inferential methods are commonplace in phylogenetic comparative methods, especially as evolutionary data have become more complex and parametric methods more limited for their analysis. Both approaches simulate many random outcomes from a null model to empirically generate sampling distributions of statistics. Although simulation-based and permutation-based methods seem commensurate in purpose, results from analysis of variance (ANOVA) based on the distributions of random F-statistics produced by these methods can be quite different in practice. Differences could be from either the null model process that generates variation across many simulations or random permutations of the data, or different estimation methods for linear model coefficients and statistics. Unfortunately, because null model process and coefficient estimation are intrinsically linked in phylogenetic ANOVA methods, the precise reason for methodological differences has not been fully considered. Here we show that the null model processes of phylogenetic simulation and randomization of residuals in a permutation procedure (RRPP) are indeed commensurate, and that both also produce results consistent with parametric ANOVA, for cases where parametric ANOVA is possible. We also provide results that caution against using ordinary least-squares estimation along with phylogenetic simulation; a typical phylogenetic ANOVA implementation.

Biology in the 21^{st} century is firmly entrenched in the big data revolution. The technological advances of recent years have enabled biologists to rapidly characterize thousands of genomic (Qin et al. 2015; Papageorgiou et al. 2018), phylogenomic (Young and Gillung 2020), morphological (Goswami et al. 2019), physiological (Orphanidou 2019), climatic (Stockwell 2006), and behavioral attributes (Kabra et al. 2013), from hundreds to thousands of observations representing individuals, populations, species, and communities. These large biological datasets, interrogated with mechanistic and phenomenological models (Otto and Day 2007; Maruvka et al. 2013; Connolly et al. 2017; Mitov et al. 2019; Otto and Rosales 2020), have extended the scope of inquiry in ecology and evolution, leading to major insights. Although computation-intensive analysis in not new (Crowley 1992), significant challenges still remain in how to extract biological signal from large (Li and Chen 2014) – and even traditionally-sized – data sets. In particular, sparse data tables, ill-conditioned covariance matrices, convergence issues with likelihood estimation, and models that lack a known probability density distribution, all lead to computational and statistical complexities for evaluating trends in biological data.

Monte Carlo (simulation) and resampling methods (Manly 2007) have a rich history in evolutionary biology research (Martins and Garland 1991; Crowley 1992; Garland et al. 1993, 2005). As computers have become more powerful, biologists are increasingly turning to computation-intensive approaches, especially for analyses that lack parametric solutions or for data that violate parametric assumptions. Computation-intensive approaches can include simulation methods (Garland et al. 2005; Rangel et al. 2018; Cornell et al. 2019) and resampling procedures (Manly 2007; Collyer and Adams 2018). With respect to the former, a family of procedures known as *simulation-based inference* methods are gaining prominence (Diggle and Gratton 1984; Gourieroux and Monfort 1993; Cranmer et al. 2020; Brehmer et al. 2020). These approaches, which include approximate Bayesian computation (ABC), likelihood and probability estimation, neural networks and neural learning (among other approaches), are particularly useful for characterizing models that describe complex biological dynamics, even when the probability distribution of the model is intractable (Beaumont 2019; Cranmer et al. 2020; Brehmer et al. 2020). Likewise, resampling procedures are increasingly used, especially for highly multivariate data that can preclude multivariate (M) analysis of variance (ANOVA). Resampling procedures are frequently used for multivariate statistics in MANOVA (Clavel and Morlon 2020), ANOVA based on dissimilarity matrices (Anderson 2001), or ANOVA using univariate-like F-statistics calculated from traces of covariance matrices (Collyer et al. 2015). These methods use random permutations of data or linear model residuals. Bootstrap resampling can also be performed for ANOVA (see, e.g., Figueiredo 2017), if resampling data with replacement is preferred.

Conceptually, simulation-based and resampling-based inference procedures, when used for hypothesis testing, are straightforward. First, random samples of observations are drawn (simulated) from a specified generating process that describes the biological null model under investigation. Next, test statistics are obtained for each simulated dataset, which compose null sampling distributions for the statistics. The chief difference is whether data are newly simulated, perhaps drawing a sample from, e.g., a normal distribution, or redistributed in random permutations of the existing data or linear model residuals. Performed many times, either approach allows a sampling distribution proxy (of a real but perhaps intractable sampling distribution) of a test-statistic to be empirically generated, and inferences about the observed data are made based on the location of the observed statistic in the sampling distribution. It has been shown that empirical sampling distributions obtained from simulation-based inference approaches can accurately approximate both likelihood profiles (Diggle and Gratton 1984; Gourieroux and Monfort 1993), and theoretical sampling distributions of summary statistics (Kac 1949; O'Hara 2019), for models that could equally use parametric probability distributions as proxies for sampling distributions. Likewise, various tests based on resampling data or residuals of linear models have been shown to have good statistical properties in terms of type I error, statistical power, and asymptotic convergence on exact tests (Anderson and Robinson 2001; Manly 2007).

One challenge for comparative data is that the observations under scrutiny (e.g., species) are correlated with one another due to shared phylogenetic history (Felsenstein 1985). Failure to account for phylogenetic history in analysis of data can result in spurious conclusions (Garland et al. 2005; Rezende and Diniz-Filho 2011). The non-independence of observations because of phylogenetic relatedness can be described by an object covariance matrix, $\mathbf{C}$, which describes the expected correlation among species due to common ancestry from a Brownian Motion (BM) model of evolutionary divergence (Grafen 1989; Martins and Hansen 1997; Rohlf 2001; O'Meara et al. 2006). This matrix has had multiple uses in phylogenetic comparative methods (PCM). It has been used to simulate data for species with expected phylogenetic relatedness as a null model process for distributions of F-statistics in ANOVA and analysis of covariance (ANCOVA) (Garland et al. 1993). For such analyses, many thousands of data sets are generated and one or more F-statistics are calculated for linear model effects, using ordinary least squares (OLS) estimation of linear model coefficients. A distribution of F-statistics from the random data sets is used as reference to calculate the percentile of the observed F-statistic (from the real data) as a P-value for a hypothesis test. This simulation-based method assures that the null model process accounts for phylogenetic correlations in the data. $\mathbf{C}$ can also be used to condition the estimation of linear model coefficients, such that residuals are independent, via generalized least squares (GLS, Martins and Hansen 1997). Letting $\mathbf{\Omega} = f(\mathbf{C})$, a transformation of the $\mathbf{C}$ matrix, ANOVA

can be performed on linear model effects, with the violation of the assumption of independent observations comfortably abated by GLS estimation. This approach can yield the same linear model coefficients as using phylogenetically independent contrasts (PIC) (Felsenstein 1985) for single-factor linear models and if $\mathbf{\Omega} = \mathbf{C}$. Therefore, parametric ANOVA results for the effects based on PIC or GLS coefficients will be the same if $\mathbf{C}$ is not transformed in any way (Blomberg et al. 2012).

Using simulated data sets is a non-parametric PCM for testing linear model effects – henceforth, abbreviated here as sim-pANOVA – and phylogenetic (P) GLS estimation is a PCM that allows for parametric ANOVA under certain circumstances. However, there are cases (e.g., multivariate data) that might be better approached with a non-parametric method. Several permutation tests using PGLS or PICs have been recently introduced. Klingenberg and Marugán-Lobón (2013) introduced a method of obtaining PICs, variable by variable, which can be concatenated in a matrix whose rows can be shuffled in a permutation procedure. Adams and Collyer (2015) demonstrated this method had inferior type I error rates, compared to randomizing residuals in a permutation procedure (RRPP) (Adams 2014; Collyer et al. 2015) and performing ANOVA by calculating univariate-like F-statistics based on PGLS coefficients in random permutations to generate distributions of F-statistics, much like the purpose of sim-pANOVA. The RRPP-ANOVA method was refined by Adams and Collyer (2018b) to use phylogenetically-transformed residuals, which had better and appropriate type I error rates under more conditions. In comparison to sim-pANOVA, RRPP-ANOVA had greater statistical power, although both methods had appropriate type I error rates. (See also, Revell 2013). Furthermore, for circumstances that parametric ANOVA following PGLS was appropriate (assumptions met), the random F distributions tracked parametric F-distribution, making the methods commensurate (Adams and Collyer 2018b).

The need for non-parametric PCMs apply inasmuch as the need to analyze complex data exists, but under conditions that traditional, parametric ANOVA is possible, one would expect consistent results between parametric and simulation-based or permutation-based (resampling) methods. However, previous comparison of sim-pANOVA and RRPP-ANOVA did not take into account that both methods have two components – estimation and a null model process – that could be considered independently to ascertain why one method performs better than another with regard to type I error rate or statistical power. Estimation simply refers to whether OLS or GLS is used to estimate coefficients, which thus impacts the calculation of F-statistics in ANOVA. A null model process is the process that generates random outcomes over many permutations from a null model. Simulation of residuals or RRPP might be sufficient null model processes, provided the fixed effects of a null model are preserved (Collyer et al. 2015). Alternatively, simulating or randomizing data

(rather than residuals) for hypothesis tests on multiple effects of a linear model would be less appropriate, as such a strategy would lack approximate exchangeability (Commenges 2003); the data would not have the same expectation (mean) as the error (zero). Recent statistical research for RRPP-ANOVA has demonstrated that phylogenetically-transformed residuals from null-models that use generalized least-squares (GLS) estimation of coefficients have appropriate exchangeability (*sensu* Commenges 2003), meaning that random pseudo-data created by this resampling procedure have approximately the same null-model residual variance for single traits (covariances and variances for multiple traits), across permutations. The simulation of data using the same $\mathbf{C}$ matrix in simulation runs should also produce data sets that have similar variance. We are, however, unaware of any previous research that has directly compared the consistency of simulation and permutation of transformed residuals as null model processes.

As described, sim-pANOVA combines simulation of data (from a BM model of evolutionary divergence) as a null model process, with estimation by OLS. By contrast, RRPP-ANOVA combines resampling (randomization) of transformed residuals as a null model process, with estimation by GLS. (The typical application of sim-pANOVA is to simulate data, rather than residuals, but for a single-factor model, this is not an issue as the null model contains only an intercept to estimate the mean.) It might be clear that OLS versus GLS estimation is one instance where greater statistical power should be expected with PGLS, as in RRPP-ANOVA. However, while GLS generally exhibits higher statistical power when compared with methods that ignore the correlations among observations (Revell 2010), if $\mathbf{\Omega}$ is not estimated properly, parameter estimates can be biased, and model evaluation procedures can be compromised (Gourieroux et al. 1984; Zeger et al. 1988; Koreisha and Fang 2001; Chavance and Escolano 2016; for discussion in a phylogenetic context see: Revell 2010; Blomberg et al. 2012).

It is possible to separately evaluate estimation and the null-model processes used in sim-pANOVA and RRPP-ANOVA. In this study we use a $3 \times 3 \times 2$ design of data type (three different levels of phylogenetic signal), F-statistic calculation (OLS or two different forms of GLS – see below), and null-model process (simulation or RRPP), to better ascertain whether it is the null-model process or estimation, or both, that leads to differences in performance (statistical power and other attributes) among methods. We compare empirically-generated sampling distributions of F-statistics to parametric F-distributions, and we evaluate the consistency of F-statistics, P-values, and effect sizes estimated from the various methods. Finally, we discuss under which conditions a particular methodological approach would be best.

Methods

Throughout our methods and results, we distinguish between "data generating models" and "analytics". The former focuses on how data were simulated with known properties (such as with evolutionary correlations) and the latter refers to how data were analyzed, as if obtained without knowledge of the properties that were simulated. Analytics refers to the choice a researcher might make with data, like choosing between simulation- or permutation-based approaches, and between OLS and GLS solutions for estimating coefficients, and thus, ANOVA statistics.

Simulation strategy. To discern how the choice of analytics affect the approximation of empirical sampling distributions and statistical power, we performed a series of stochastic sampling experiments. The sampling experiments were based on a linear model, using a single-factor ANOVA design. We simulated data with an expected variance, $\mathbf{\Sigma} = \mathbf{\Omega}\sigma^2$, with $\sigma^2 = 1$ (standard normal distribution) and $\mathbf{\Omega}$ varied for three distinct data-generating models: phylogenetic independence, phylogenetic correlation based on a BM model of evolutionary divergence, and an intermediate amount of phylogenetic correlation; i.e., three levels of phylogenetic signal in the data. For any simulated data set, three different coefficient estimation methods were used. Coefficients were estimated as, $\hat{\beta} = \left(\mathbf{X}^T\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$, with three different versions of $\hat{\mathbf{\Omega}}$, corresponding to OLS, GLS based on a $\hat{\mathbf{\Omega}} = \mathbf{C}$ (the covariance matrix representing a BM model of evolutionary divergence), and GLS using a $\hat{\mathbf{\Omega}}$ matrix, based on a maximum likelihood fit of data to the phylogenetic tree, relative to phylogenetic signal. Two distinct null model processes were also used to generate empirical sampling distributions of F-statistics: simulation of residuals using $\hat{\mathbf{\Sigma}} = \mathbf{C}\hat{\sigma}^2$, where $\hat{\sigma}^2$ was calculated following coefficient estimation, and RRPP. This $3 \times 3 \times 2$ design made it possible to isolate the impact of estimation and null model process on ANOVA results, whether data were independent with respect to phylogeny.

For any simulation run, we generated 100 or 200 'observed' datasets, each containing $n = 250$ independent observations drawn from a generating model. The data sets corresponded to 100 or 200 pure-birth phylogenies containing $n = 250$ species, each. Datasets (y) were simulated as: $y = \mathbf{X}\beta + \epsilon_{\mathcal{N}(0,\Sigma)}$; where ϵ was a vector of independent residuals drawn from a normal distribution, $\mathcal{N}(0, \Sigma)$; the model design matrix, $\mathbf{X}$, was a matrix comprising 0s and 1s as dummy variables to indicate group association (with no *a priori* association to y) for 10 groups, each with 25 species; and β was a vector of coefficients to add group effects (difference in means between groups). (Initial trials that varied group number and the number of species per group indicated that these variables were not consequential for the results, but using 10 groups produced sampling distributions

that were easy to compare among methods.) For these simulations, no group effect could be included (i.e., all $\beta = 0$), which was tantamount to generating random response data ($y = \epsilon_{\mathcal{N}(0,\Sigma)}$) and randomly assigning those observations to groups.

Comparison of sampling distributions. For our first set of simulations, we varied $\mathbf{\Omega}$, only, and set $\beta = 0$; i.e., no expected differences among groups. One version of $\mathbf{\Omega}$ was an unscaled matrix based on a BM model of evolutionary divergence (Felsenstein 1985; Grafen 1989; Rohlf 2001; Huey et al. 2019), i.e., $\mathbf{\Omega} = \mathbf{C}$. The other two changed the amount of covariance (phylogenetic signal in the data) by the scaling parameter, Pagel's λ, which scales the internal branch lengths of the tree, optimizing the fit of the data to the tree (Pagel 1997). Letting, $\mathbf{D} = diag(\mathbf{C})$ – a diagonal matrix of only the phylogenetic variances of $\mathbf{C}$ – a rescaled form of $\mathbf{C}$ is $\mathbf{\Omega} = \lambda(\mathbf{C} - \mathbf{D}) + \mathbf{D}$ (Collyer et al. 2022). We used $\lambda = 0, 0.5$, and 1 (*sensu* Clavel and Morlon 2020) to scale covariance matrices from random trees, yielding data sets with phylogenetic independence, intermediate phylogenetic signal (correlation), and phylogenetic signal as expected with BM (unscaled tree), respectively. These simulations allowed us to consider the correspondence between null sampling distributions using different combinations of estimation and null model process, in the absence of group effects.

Because data were simulated such that assumptions for parametric ANOVA should be met, we compared results to parametric ANOVA several ways. First, we mapped empirical F-distributions on a parametric F-distribution. Doing so revealed how well the combinations of null model process and coefficient estimation matched theoretical expectation. Second, we estimated P-values as the percentiles of observed cases in their corresponding distributions, which could be compared to P-values estimated via integration of the probability density function of the F-distribution, based on degrees of freedom. Third, we estimated effect sizes as Z-scores, the standard deviate of observed F-statistics in their normalized distributions (*sensu*, Adams and Collyer 2018b). Finally, "pairs" plots of P-values and Z-scores were used to evaluate the consistency of statistics among the different methods, including parametric ANOVA for P-values.

Comparison of statistical power. For our second set of simulations, we repeated the design of the first set of simulations, but varied the first β parameter, from 0 to 8, in increments of 2. (This approach increased the mean of the first group from the other groups by 0, 2, 4, 6, and 8, to create an effect.) These simulations allowed us to consider differences in type I error rate ($\beta = 0$) and statistical power ($\beta > 0$) among different combinations of estimation and null model process. Initial trials suggested 200 simulated trees was sufficient to obtain a reliable estimate of null hypothesis rejection rate at an expected significance level of $\alpha = 0.05$.

In all simulation runs, coefficient estimation was performed on simulated data three different ways. OLS estimation – as was used by Garland et al. (1993) – does not attempt to account for phylogenetic correlation in the data, even though data might be simulated to have phylogenetic correlations; i.e., $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, where ^{T} represents vector transposition, $^{-1}$ represents matrix inversion, and $\mathbf{y}$ is a vector of data. GLS estimation uses $\mathbf{\Omega}$ in estimation of coefficients; i.e., $\hat{\beta} = \left(\mathbf{X}^T\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$. $\hat{\mathbf{\Omega}}$ for coefficient estimation can be determined one of two different ways. First, it can be assumed to correspond to either a BM model of evolutionary divergence or the covariances can be scaled by an *a priori* notion of what the covariance should be. For example, scaling the covariances by $\lambda = 0$ produces coefficients that are no different than OLS estimation; i.e., OLS estimation is the same as assuming phylogenetic independence in GLS estimation. Second, a maximum-likelihood estimate of $\mathbf{\Omega}$ can be obtained by finding the value of λ that maximizes the likelihood of the data, given the tree; i.e., $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\lambda})$ (see, Collyer et al. 2022). We performed coefficient estimation for $\lambda = 0$ (OLS), $\lambda = 1$ (typical with PGLS analysis), and the maximum likelihood estimate of λ, $\hat{\lambda}$ for every data set produced in every simulation run.

F-statistics were calculated for every model, from the coefficients estimated, as $F = (k-1)(n-k)^{-1}(\mathbf{r} - \mathbf{r}_0)^T\hat{\mathbf{\Omega}}^{-1}(\mathbf{r} - \mathbf{r}_0)(\mathbf{r}^T\hat{\mathbf{\Omega}}^{-1}\mathbf{r})^{-1}$, where $\mathbf{r}$ is a vector of residuals found as $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\beta}$ and k is the number of model parameters. The residuals, $\mathbf{r}_0$, were likelise calculated from a model with only an intercept, $\mathbf{X}_0$ and its estimated coefficient, $\hat{\beta}_0$. The two null model processes were applied to each case, using 999 random permutations (which along with observed statistics generated distributions of 1,000 random F-statistics). For simulation as a null model process, $\epsilon_{\mathcal{N}(0,\Sigma)}$ were newly obtained; for RRPP, $\epsilon_{\mathcal{N}(0,\Sigma)}$ were estimated by randomizing the transformed residuals, of null model containing only an intercept (mean). One important caveat is that only if new data are simulated are the null model process and coefficient estimation truly independent, but resampling residuals means they are intrinsically linked.

This can be appreciated by the formula for residual variance for univariate data and a model with k parameters, $\hat{\sigma}^2 = (n-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})^T\hat{\mathbf{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$, which can be equivalently written with Cholesky decomposition of $\hat{\mathbf{\Omega}}$ as, $\hat{\sigma}^2 = (n-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$. Thus, the equation can be updated as, $\hat{\sigma}^2 = (n-k)^{-1}\left(\mathbf{\Psi}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})\right)^T\left(\mathbf{\Psi}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})\right)$, where $\left(\mathbf{\Psi}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})\right)$ are the phylogenetically transformed residuals and exchangeable units under the null hypothesis (Adams and Collyer 2018b). Because transformed residuals require $\mathbf{\Omega}$, both in the calculation of $\hat{\beta}$ and the transformation of the residuals, the null model process (randomization of residuals) is not independent of coefficient estimation.

In comparing the different combinations of null model process and estimation, the answers for the following four questions were sought. Do the distributions of random F-statistics comport as expected, compared to theoretical (parametric) distributions? From the distributions of random F-statistics, do the different combinations produce consistent results with regard to null hypothesis tests (correlation of P-values across simulation runs)? From the distributions of random F-statistics, do the different combinations produce consistent effect sizes (correlation of Z-scores across simulation runs)? Do the different combinations have similar statistical power over a range of simulated effects? All simulations were performed in R 4.1.2. (R Core Team 2021). The functions, `phytools::pbtree` (Revell 2012) and `geiger::rescale.phylo` (Harmon et al. 2008) were used to randomly generate and rescale phylogenetic trees, respectively. Support functions from `RRPP` (Collyer and Adams 2018, 2021b) and `geomorph` (Adams et al. 2021; Baken et al. 2021) were used along with new functions written by the authors for simulations and analysis. All `R` scripts used are provided as supporting information.

Results

Comparison of sampling distributions. Our results indicated that when the null model process and estimation were commensurate – estimation matched the type of $\mathbf{\Omega}$ matrix used to generate random outcomes – random F-statistics formed distributions that were consistent with theoretical expectation, irrespective of whether the null model process used simulation or RRPP (Fig. 1). Most notably, when GLS was performed with optimization of λ, consistent empirical F-distributions were produced, irrespective of null model process, and matched well to the parametric F-distribution. Other cases in which there was a good match appeared to be incidental. For example, simulating data with a BM process and using GLS estimation with $\lambda = 1$ produced consistent distributions, regardless of whether data were simulated with $\lambda = 0, 0.5$, or 1. The most notable departures from this trend occurred with simulation with a BM process as a null model process, but OLS estimation, or GLS estimation with optimized λ, when the data were obtained from a model other than $\lambda = 1$ (Fig. 1). There were also noticeable inconsistencies among distributions when RRPP was performed with $\lambda = 1$, for data not simulated from a model with $\lambda = 1$ (BM). However, the mismatches between theses distributions and the parametric F-distribution were far less severe than those between optimized GLS estimation for simulation of BM data as null model process.

Most notably, the combination of BM simulation as a null model process and OLS estimation – the combination used in sim-pANOVA – did not produce empirical F-distributions that resemble parametric F-distributions. These results were basically replicated with optimized λ in GLS estimation, presumably because the optimized value would be near or equal to 0. Collectively, the results indicate that (1) λ should be optimized and (2) if this is done, and if simulation is used, $\hat{\mathbf{\Omega}}$ should be a rescaled form of $\mathbf{C}$ (the internal branches of the tree are rescaled) for the null model process. This combination of simulation and GLS estimation using a rescaled $\mathbf{C}$ matrix and RRPP using the rescaled $\mathbf{C}$ matrix based on $\hat{\lambda}$ yielded unequivocally consistent distributions, regardless of the value of λ.

The consistency of empirical sampling distributions with parametric distributions might not be a cause for concern for hypothesis tests (more on this below), provided there is consistency in null hypothesis test outcomes. In general, the correspondence between P-values was noteworthy (Pearson $r > 0.99$) for comparable methods whether using OLS or either form of GLS, and for data generated with no phylogenetic signal (Fig. 2), data with intermediate phylogenetic signal (Fig. 3), or data with phylogenetic signal as expected from a BM model of evolutionary divergence (Fig 4.). In all cases, when the null model process and estimation matched (the $\mathbf{\Omega}$ matrix was the same in the null model process and estimation), a 1:1 relationship between P-values from parametric ANOVA and the non-parametric alternatives was evident. However, if either simulation or RRPP was used as a null model process along with GLS and $\lambda = 1$, the correlation was not as strong for data simulated with $\lambda \neq 1$, as it was if λ was optimized. It was also quite apparent that if simulation was used as a null model process, it was important to rescale $\mathbf{C}$ in order to retain a linear relationship between P-values (from parametric F-statistics).

There were no obvious relationships among different estimation methods unless incidentally because data were simulated to have no phylogenetic signal ($\lambda = 0$) or phylogenetic signal expected with a BM model of evolutionary divergence ($\lambda = 1$), in which case the P-values from GLS with optimized λ and either OLS, or GLS with $\lambda = 1$, respectively, were highly correlated. These results only reinforce that λ optimization is an important step that yields consistent results with OLS and traditional PGLS estimation, at the extremes. Additionally, using OLS estimation when data had phylogenetic signal appeared to produce P-values that were consistently near 0, regardless of null model process.

These patterns were generally the same among the non-parametric tests for effect sizes (Z-scores; see Supporting Information); though here there is no basis for comparison to parametric tests, which do not have an

obvious transformation to obtain a Z-score. Collectively, these results confirm that choice of null model process is not as important as estimation, provided λ optimization is performed at all stages. Ignoring optimization, under no circumstances does simulation from a BM model of evolutionary divergence as a null model process coupled with OLS estimation make sense. Rather, estimation and null model process should be seen as an analytical pairing that requires matching $\mathbf{C}$ matrices (inherent in RRPP), which performs best if $\mathbf{C}$ is scaled by λ.

Comparison of statistical power. Some of the inconsistencies noted in the first set of simulations were more obvious as pathologies in the second set of simulations, meant to address type I error and statistical power. Estimation with OLS or GLS without λ optimization tended to result in higher type I error rates (Fig. 5), unless the λ assumed for estimation happened to match the λ used to simulate data. For example, data simulated with $\lambda = 0$ had appropriate type I error rates (at $\beta = 0$) for OLS estimation and RRPP, but OLS estimation used on data with any phylogenetic signal had exceptionally large type I error rates; data simulated to have phylogenetic signal consistent with a BM model of evolutionary divergence had appropriate type I error rates if GLS was performed with $\lambda = 1$ but had elevated type I error rates for data simulated with $\lambda < 1$, with the rate increased for data simulated with $\lambda = 0$ compared to $\lambda = 0.5$ (further departure from $\lambda = 1$). The only obvious difference between null model processes was that OLS estimation with RRPP (intrinsic relationship between residuals and estimation) and OLS estimation with BM data simulation had strikingly different results. Type I error rates were appropriate, irrespective of data type, if simulation was used. If simulation and RRPP assumed the same $\mathbf{C}$ matrix, type I error rate and power curves were indistinguishable (Fig. 5).

For statistical power, the method of estimation appeared to be more important than the null model process. This can be appreciated by the consistency of statistical power curves among the different data types, irrespective of null model process, juxtaposed with the disparity among power curves if GLS was not based on optimization of λ. Regarding the latter, statistical power was considerably higher if the value of λ used for GLS estimation incidentally matched the λ used to simulate data. A greater departure from this ($\lambda = 0$) resulted in a greater reduction in statistical power. The simulated effect (β) also contributed to differences in statistical power. The only disparity between statistical power curves occurred at small β (2 or 4), with data simulated with stronger phylogenetic signal having greater statistical power.

The result of previous research (Adams and Collyer 2018b) was also confirmed; BM simulation as a null model process with OLS estimation (sim-pANOVA) has less statistical power than RRPP with GLS, but only

if data were simulated with a BM model of evolutionary divergence. The enhanced statistical power of RRPP with GLS appears to be an amelioration afforded by simulating only data with a BM model of evolutionary divergence. Simulating data with weaker phylogenetic signal renders low statistical power using GLS, assuming $\lambda = 1$. However, optimizing λ provides the most statistical power, regardless of null model process or the apparent amount of phylogenetic correlation in the data. These results are consistent with those of previous work (Collyer et al. 2022), in which statistical power for detecting phylogenetic signal was increased by optimizing λ.

Discussion

This research revealed several important points. First, whether simulation of data (residuals, more precisely) or RRPP is used as a null model process is inconsequential. Both methods produce reliable results. Second, the method of estimation is exceedingly important. Performing ANOVA with a method of estimation that does not appropriately estimate the phylogenetic covariances of the data can be detrimental. Statistical research varying the strength of phylogenetic signal has been performed for evaluating methods that test phylogenetic signal strength (e.g., Münkemüller et al. 2012; Collyer et al. 2022) but might be comparatively rarer for research that evaluates the proficiency of methods to test hypotheses with linear models. One exception is the work of Clavel and Morlon (2020), who varied λ as we did in this study, for the comparison of various multivariate methods. In their comparisons, two methods mimicked methods used in this study. They used BM simulation with OLS estimation of residual covariance matrices and RRPP with GLS estimation of residual covariance matrices, assuming $\lambda = 1$. The former used MANOVA statistics and the latter used univariate-like F-statistics (Collyer et al. 2015) for test statistics, but for a single variable, these statistics would be comparable to the BM simulation plus OLS and RRPP plus GLS with $\lambda = 1$ cases we considered here. Clavel and Morlon (2020) noted that using RRPP plus GLS with $\lambda = 1$ resulted in drastically increased statistical power but also high type I error rates, when considered for only the first principal component if data were simulated with a BM model of evolutionary divergence, but statistical power was otherwise considerably lower than with multivariate likelihood or penalized likelihood statistics, which used $\hat{\mathbf{\Omega}}$ instead of $\mathbf{C}$. Our results are remarkably consistent with theirs for the univariate consideration in this study, and suggest that the inherent optimization step in their statistical approach is the important component of the cross-validated statistics they introduced. The null model process they used was similar to RRPP, except that residuals were not transformed prior to permutation but were transformed after permutation, unlike the recommendation of Commenges (2003) to maintain approximate second-moment exchangeability, and $\mathbf{C}$ was used rather than $\hat{\mathbf{\Omega}}$ for transformation. The issues we found with

BM simulation plus OLS were also consistent with the issues they found (limited statistical power). Not only did we find consilience of ANOVA methods for our comparisons of parametric ANOVA, simulation-based ANOVA, and RRPP-based ANOVA, when $\hat{\mathbf{\Omega}}$ is based on an optimized value of λ, but we found some consistency with the increased statistical power of other multivariate methods that use $\hat{\mathbf{\Omega}}$ rather than $\mathbf{C}$.

Performing statistical research to evaluate methods that should have more applicability for multivariate data by simulating univariate data might seem unintuitive. However, there were two important reasons for doing this. First, simulating univariate data meant we could map random F-statistics on the density plot of parametric F-distribution to consider the behavior of the different combinations of null model process and estimation. We could have alternatively considered examples with conditions that multivariate statistics have exact F-distributions but this would have also meant generalizing sim-pANOVA – a commonly used method – for MANOVA. Second, and more importantly, generalizing the maximum-likelihood estimation of λ for multivariate data is not a trivial exercise. As noted by Collyer et al. (2022), λ optimization can be quite complex, starting with the consideration of whether λ is free to vary across multivariate variables. Finding the determinant of a residual covariance matrix, $|\mathbf{\Sigma}|$, is essential for estimating model likelihood. $\mathbf{\Sigma}$ is found for multivariate data as, $\mathbf{R} \otimes \hat{\mathbf{\Omega}}$, where $\otimes$ indicates a Kronecker product, and $\mathbf{R}$ is the residual covariance matrix generalization of $\hat{\sigma}^2$, found from an $n \times p$ matrix of data, $\mathbf{Y}$, as $\mathbf{R} = (n-k)^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})^T\hat{\mathbf{\Omega}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$. However, this definition implies single $\mathbf{R}$ and $\hat{\mathbf{\Omega}}$ matrices, which means a scalar λ must be used in estimation. Allowing for p different $\mathbf{R}$ and $\hat{\mathbf{\Omega}}$ matrices is possible with an algorithm to solve $\mathbf{\Sigma}$ for likelihood estimation, with maximum-liklihood estimates of λ first found p times for each variable (Collyer et al. 2022), a likely computationally intense procedure.

Clavel and Morlon (2020) offered that a maximum likelihood estimate of $\mathbf{\Omega}$ could be made with either a scalar or vector of λ, but it is not clear how either would be obtained. Whether λ should be a scalar or vector must consider the type of multivariate data (see Adams and Collyer 2018a, 2019). For example, shape data found through generalized Procrustes analysis (Rohlf and Slice 1990) uses multiple variables to describe one organism attribute, shape. Thus, assuming that natural selection acts independently on sub-components of a set of shape variables – as is explicitly the case with optimizing separate λ for each trait dimension – would not make much sense. Thus, for such multidimensional traits, use of a common scalar of λ would be more appropriate. Yet this recommendation is in contrast to that for, say, life history data, where a multivariate dataset comprises multiple traits that could (in theory) be independent. Here, separate λ for each trait may be envisioned (for a similar approach with evolutionary rates see: Adams 2013). Although

a scalar λ offers a simpler calculation of $\mathbf{\Sigma}$, and thus, a simpler estimation of model likelihood, allowing independent λ optimization offers a potentially simpler solution for optimization, as λ is optimized for each variable with a univariate optimization strategy. It is the multivariate generalization of a single λ parameter that is potentially more complex.

Although the optimization of λ is straightforward – the value of λ that maximizes the likelihood estimator – calculating the likelihood is fraught with estimation issues as $\mathbf{\Sigma}$ becomes singular (when the number of variables approaches or exceeds the number of observations). In this study, using univariate examples posed no issue, but generalizing the maximum likelihood estimate for λ for multivariate data would require further investigation, especially for high-dimensional data, in which case $\mathbf{\Sigma}$ is certainly singular but tractable test statistics based on traces of residual covariance matrices could be used. Clavel and Morlon (2020) introduced penalized likelihood, which offers an ability to find $\hat{\lambda}$, even for high-dimensional data, but this approach might only be useful for likelihood-based statistics as test statistics.

We envision four possible scenarios for generalizing λ optimization to find a scalar that maximizes the likelihood of a model for high-dimensional data (for the general formula for multivariate likelihood estimation, see, Revell and Harmon 2008), and for test statistics that do not necessarily rely on calculating model likelihood. The simplest generalization – if data dimensionality is not an issue – finds alternative multivariate likelihoods based on residual covariance matrix estimates, spanning λ from 0 to 1. However, this approach could fail to consider strong latent phylogenetic signal, restricted to only a portion of the data space (see Collyer and Adams 2021a). For example, for shape data where strong phylogenetic signal is localized to a portion of a more comprehensive anatomical configuration, a solution that converges toward $\lambda = 0$ might be found. (In this case, allowing λ to vary might be warranted.) A second approach, which would mitigate the potential issues of the first approach, is to find optimized values of λ variable by variable, and average them. This solution, however, would bias lambda optimization toward 0.5 unless all variables either have or lack phylogenetic signal. Based on our statistical power results (Fig. 5), this might not be such a worrisome outcome, as the most egregious issues occurred for cases where λ used in analysis had a large departure from its optimized value. (A tendency toward an intermediate value would preclude larger disparity that could exist between two λ values.)

A third alternative would be to determine the data dimensions that have most phylogenetic signal and rotate the data space with respect to these dimensions, an analysis called phylogenetically aligned component analysis (PACA, Collyer and Adams 2021a). Multivariate optimization of λ could thus be confined to the

dimensions where phylogenetic signal is present. This procedure would likely bias λ in a positive direction, which could yield traditional PGLS solutions (assuming $\lambda = 1$) even if phylogenetic signal is constrained to a small portion of the variables. Finally, the likely best solution is also the most computationally exhaustive. RRPP or simulation could be performed for a reliable number of permutations for, e.g., $\lambda = 0, 0.1, 0.2, ..., 1$, and a spline function could find the optimal λ that maximizes a Z-score for characterizing phylogenetic signal, *sensu* Collyer et al. (2022). This approach has used sampling distributions of log-likelihood statistics measuring phylogenetic signal to obtain, Z, which has been shown to have a linear association with λ. Thus, measuring likelihood was less important than its role in measuring the phylogenetic signal effect size. The same approach could be used on Z-scores obtained from distributions of random univariate-like F-statistics based on traces of residual covariance matrices. Whether such an approach yields comparable or better statistical power than penalized likelihood approaches would also be a useful future research endeavor.

If λ should be considered free to vary, a multivariate generalization may not be needed, as p univariate solutions would be acquired for p variables. However, this approach also assumes each variable could be considered independent, which is perhaps a risky assumption. An alternative solution is to use the independent solutions as a starting point for an iterative procedure that finds the optimal combination of values that maximizes likelihood, *sensu* Adams (2013). Such an approach might have intrigue as an analysis that considers the modularity of anatomical sub-configurations for landmark shape data (*sensu* Klingenberg 2009; Adams 2016; Zelditch and Goswami 2021), as suites of contrasting λ values for groups of variables might be evidence for natural selection acting differently on anatomical components.

This discussion highlights the current tension and needed research direction. Although research to evaluate the most appropriate methods to estimate $\mathbf{\Omega}$ for multivariate data, especially with respect to statistics used for hypothesis tests, will require a thorough investigation, what can researchers do currently to assuage concerns about inappropriately estimated $\mathbf{\Omega}$ matrices? As a heuristic, at least for a scalar form of λ, simply performing a range of analyses first with, e.g., $\lambda = 0, 0.1, 0.2, ..., 1$ for a linear model that contains only an intercept, and choosing a value of λ that yields the largest effect size, Z, based on a distribution of random log-likelihoods (Collyer et al. 2022) will be a valuable analytical strategy. This is consistent with one of the proposed optimization strategies we outlined above. Furthermore, if it is found that alternative optimization strategies work well, the advantage would be saved computation time but likely not improved statistical power, over this approach. It can be expected that a solution that maximizes Z also maximizes likelihood, using RRPP (Collyer et al. 2022). If one wishes to allow λ to

vary among different variables, then finding λ variable by variable might not maximize model likelihood compared to an alternative solution, but might have a solution that maximizes model likelihood better than a scalar. As a minimum, these are two approaches that should work well – especially compared to assuming $\lambda = 1$ in PGLS or using OLS in sim-pANOVA – and could possibly be improved with further research.

Given that estimation is important, the question turns to whether residuals should be resampled or simulated, as a null process? We have shown there is no real analytical concern, as results will be consistent with simulated data, but under which conditions would this question be swayed to a particular answer? Simulation assumes a parametric distribution from which data are sampled. RRPP uses the residuals that are calculated. An advantage to simulation is it can be assured that the null process correctly asserts an appropriate distribution. An advantage to RRPP is that an assumption about the distribution of residuals is not required. It will require further research to evaluate if the two methods have contrasting results with residuals that are not normally distributed or are heteroscedastic. However, if the two methods can be relied on to produce consistent results, shuffling transformed residuals is probably computationally faster than simulating residuals from a null distribution that must be modified to have phylogenetic correlation in each permutation.

We feel that a few potential updates to software packages that offer either sim-pANOVA or RRPP-ANOVA should be strongly considered. First, estimation of $\mathbf{\Omega}$ based on an optimized λ should be made available. If data are simulated, rescaling the phylogenetic tree used for simulation by $\hat{\lambda}$ should be an essential step. Simulation should also simulate residuals that are added to null model fitted values, which are estimated with PGLS, with $\hat{\mathbf{\Omega}}$ based on $\hat{\lambda}$, rather than the simulation of new data in every permutation. This is especially true for linear models with multiple effects. Currently, software packages that offer sim-pANOVA do so only for single factor models, in which case simulating new data is no different than simulating residuals. However, for multiple linear model effects, multiple null models and therefore, multiple null model processes are required. The simulation of new data implicitly considers the same model with only an estimated mean as a null model, which would probably not make much sense for ANOVA based on type I, type II, or type III sums of squares and cross-products.

More broadly, our work exposed the fact that when evaluating macroevolutionary trends across a phylogeny, it is the appropriate conditioning of the data on the phylogeny during the analysis, and not phylogenetic simulation alone, that was responsible for obtaining adequate sampling distributions (which should be a goal for making correct biological inferences). That is, the use of OLS estimation – as is common in some

implementations of sim-pANOVA – yields incorrect sampling distributions, and could thus lead to incorrect statistical inferences regarding patterns in cross-species data. Although previous research has illustrated that GLS estimation is an obvious improvement, we have shown that PGLS assuming $\lambda = 1$ can also be fraught with type I error rate and statistical power issues. Realizing that $\hat{\lambda} = 0$ and $\hat{\lambda} = 1$ are possible optimization outcomes, PGLS using $\hat{\lambda}$ should be viewed as a universal solution. Our research emphasized that statistical inference via phylogenetic comparative methods requires statistical methods that condition data on the phylogeny; not merely data that are simulated from a phylogenetic process alone. From this it follows that simulation-based approaches to macroevolutionary inference must account for phylogenetic non-independence at two stages: the generation of random samples via simulation (e.g., Martins and Garland 1991; Garland et al. 1993), and in the analytics that are used to obtain statistics (e.g., Martins and Hansen 1997). When both of these conditions are met, macroevolutionary inferences derived form simulation-based approaches are appropriate and reliable.

References

Adams, D. C. 2014. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution 68:2675–2688.

Adams, D. C. 2013. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. Systematic Biology 62:181–192.

Adams, D. C. 2016. Evaluating modularity in morphometric data: challenges with the RV coefficient and a new test measure. Methods in Ecology and Evolution 7:565–572.

Adams, D. C., and M. L. Collyer. 2018a. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. Systematic biology 67:14–31.

Adams, D. C., and M. L. Collyer. 2015. Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: What you shuffle matters. Evolution 69:823–829.

Adams, D. C., and M. L. Collyer. 2018b. Phylogenetic ANOVA: Group-clade aggregation, biological challenges, and a refined permutation procedure. Evolution 72:1204–1215.

Adams, D. C., and M. L. Collyer. 2019. Phylogenetic comparative methods and the evolution of multivariate phenotypes. Annual Review of Ecology, Evolution, and Systematics 50:405–425.

Adams, D. C., M. L. Collyer, A. Kaliontzopoulou, and E. K. Baken. 2021. Geomorph: Software for geometric morphometric analyses. R package version 4.0. R Foundation for Statistical Computing, Vienna, Austria.

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46.

Anderson, M. J., and J. Robinson. 2001. Permutation tests for linear models. Australian & New Zealand Journal of Statistics 43:75–88.

Baken, E. K., M. L. Collyer, A. Kaliontzopoulou, and D. C. Adams. 2021. Geomorph v4. 0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. Methods in Ecology and Evolution 12:2355–2363.

Beaumont, M. A. 2019. Approximate bayesian computation. Annual Review of Statistics and its Application 6:379–403.

Blomberg, S. P., J. G. Lefevre, J. A. Wells, and M. Waterhouse. 2012. Independent contrasts and PGLS regression estimators are equivalent. Systematic Biology 61:382–391.

Brehmer, J., G. Louppe, J. Pavez, and K. Cranmer. 2020. Mining gold from implicit models to improve likelihood-free inference. Proceedings of the National Academy of Sciences, U.S.A. 117:5242–5249.

Chavance, M., and S. Escolano. 2016. Misspecification of the covariance structure in generalized linear

mixed models. Statistical Methods in Medical Research 25:630–643.

Clavel, J., and H. Morlon. 2020. Reliable phylogenetic regressions for multivariate comparative data: Illustration with the MANOVA and application to the effect of diet on mandible morphology in phyllostomid bats. Systematic Biology 69:927–943.

Collyer, M. L., and D. C. Adams. 2021a. Phylogenetically aligned component analysis. Methods in Ecology and Evolution 12:359–372.

Collyer, M. L., and D. C. Adams. 2021b. R: RRPP: Linear model evaluation with randomized residuals in a permutation procedure. R Foundation for Statistical Computing, Vienna, Austria.

Collyer, M. L., and D. C. Adams. 2018. RRPP: An r package for fitting linear models to high-dimensional data using residual randomization. Methods in Ecology and Evolution 9:1772–1779.

Collyer, M. L., E. K. Baken, and D. C. Adams. 2022. A standardized effect size for evaluating and comparing the strength of phylogenetic signal. Methods in Ecology and Evolution, doi: 10.1111/2041-210X.13749.

Collyer, M. L., D. J. Sekora, and D. C. Adams. 2015. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. Heredity 115:357–365.

Commenges, D. 2003. Transformations which preserve exchangeability and application to permutation tests. Journal of Nonparametric Statistics 15:171–185.

Connolly, S. R., S. A. Keith, R. K. Colwell, and C. Rahbek. 2017. Process, mechanism, and modeling in macroecology. Trends in Ecology and Evolution 11:835–844.

Cornell, S. J., Y. F. Suprunenko, D. Finkelshtein, P. Somervuo, and O. Ovaskainen. 2019. A unified framework for analysis of individual-based models in ecology and evolution. Nature Communications 10:4716.

Cranmer, K., J. Brehmer, and G. Louppe. 2020. The frontier of simulation-based inference. Proceedings of the National Academy of Sciences, U.S.A. 117:30055–30062.

Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. Annual Review of Ecology and Systematics 23:405–447.

Diggle, P. J., and R. J. Gratton. 1984. Monte carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society, B. 46:193–227.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1–15.

Figueiredo, A. 2017. Bootstrap and permutation tests in ANOVA for directional data. Computational Statistics 32:1213–1240.

Garland, T., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. Journal of experimental Biology 208:3015–3035.

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Systematic Biology 42:265–292.

Goswami, A., A. Watanabe, R. N. Felice, C. Bardua, A.-C. Fabre, and P. D. Polly. 2019. High-density morphometric analysis of shape and integration: The good, the bad, and the not-really-a-problem. Integrative and comparative biology 59:669–683.

Gourieroux, C., and A. Monfort. 1993. Simulation-based inference: A survey with special reference to panel data models. Journal of Econometrics 52:681–700.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo-maximum likelihood methods: theory. Econometrica 59:5–33.

Grafen, A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London B, Biological Sciences 326:119–157.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: Investigating evolutionary radiations. Bioinformatics 24:129–131.

Huey, R. B., T. Garland Jr, and M. Turelli. 2019. Revisiting a key innovation in evolutionary biology: Felsenstein's "phylogenies and the comparative method." The American Naturalist 193:755–772.

Kabra, M., A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. 2013. JAABA: Interactive machine learning for automatic annotation of animal behavior. Nature Methods 10:64–67.

Kac, M. 1949. On deviations between theoretical and empirical distributions. Proceedings of the National Academy of Sciences, U.S.A. 35:252–257.

Klingenberg, C. P. 2009. Morphometric integration and modularity in configurations of landmarks: Tools for evaluating a priori hypotheses. Evolution & Development 11:405–421.

Klingenberg, C. P., and J. Marugán-Lobón. 2013. Evolutionary covariation in geometric morphometric data: Analyzing integration, modularity, and allometry in a phylogenetic context. Systematic biology 62:591–610.

Koreisha, S. G., and Y. Fang. 2001. Generalized least squares with misspecified serial correlation structure. Journal of the Royal Statistical Society, B 63:515–531.

Li, Y., and L. Chen. 2014. Big biological data: Challenges and opportunities. Genomics, Proteomics, and Bioinformatics 12:187–189.

Manly, B. F. J. 2007. Randomization, bootstrap, and monte carlo methods in biology. 3rd ed. Taylor & Francis.

Martins, E. P., and T. Garland. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. Evolution 45:534–557.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: A general approach

to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist 149:646–667.

Maruvka, Y. E., N. M. Shnerb, D. A. Kessler, and R. E. Ricklefs. 2013. Model for macroevolutionary dynamics. Proceedings of the National Academy of Sciences, U.S.A. 110:E2460–E2469.

Mitov, V., K. Bartoszek, and T. Stadler. 2019. Automatic generation of evolutionary hypotheses using mixed gaussian phylogenetic models. Proceedings of the National Academy of Sciences, U.S.A. 116:16921–16926.

Münkemüller, T., S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffers, and W. Thuiller. 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3:743–756.

O'Hara, M. 2019. Teaching hypothesis testing with simulated distributions. International Review of Economics Education 30:100138.

O'Meara, B. C., C. Ane, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Orphanidou, C. 2019. A review of big data applications of physiological signal data. Biophysical Reviews 11:83–87.

Otto, S. P., and T. Day. 2007. A biologist's guide to mathematical modeling in ecology and evolution. Princeton, NJ: Princeton University Press.

Otto, S. P., and A. Rosales. 2020. Theory in service of narratives in evolution and ecology. American Naturalist 195:290–299.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. Zoologica Scripta 26:331–348.

Papageorgiou, L., P. Eleni, S. Raftopoulou, M. Mantaiou, V. Megalooikonomou, and D. Vlachakis. 2018. Genomic big data hitting the storage bottleneck. EMBnet J 24:e910.

Qin, Y., H. K. Yalamanchili, J. Qin, B. Yan, and J. Wang. 2015. The current status and challenges in computational analysis of genomic big data. Big Data Research 2:12—18.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rangel, T., N. R. Edwards, P. B. Holden, and J. A. F. Diniz-Filho. 2018. Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. Science 361:eaar5452.

Revell, L. J. 2013. Http://blog.phytools.org/2013/02/type-i-error-and-power-of-phylogenetic.html.

Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1:319–329.

Revell, L. J. 2012. Phytools: An r package for phylogenetic comparative biology (and other things): Phytools: R package. Methods in Ecology and Evolution 3:217–223.

661 Revell, L. J., and L. J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary
662 rate matrix for continuous characters. Evolutionary Ecology Research 10:311–331.
663 Rezende, E. L., and J. A. F. Diniz-Filho. 2011. Phylogenetic analyses: Comparing species to infer
664 adaptations and physiological mechanisms. Comprehensive Physiology 2:639–674.
665 Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations.
666 Evolution 55:2143–2160.
667 Rohlf, F. J., and D. E. Slice. 1990. Extensions of the procrustes method for the optimal superimposition
668 of landmarks. Systematic Zoology 39:40–59.
669 Stockwell, D. R. B. 2006. Improving ecological niche models by data mining large environmental datasets
670 for surrogate models. Ecological Modeling 192:188–196.
671 Young, A. D., and J. P. Gillung. 2020. Phylogenomics — principles, opportunities and pitfalls of big-data
672 phylogenetics. Systematic Entomology 45:225–247.
673 Zeger, S. L., K.-Y. Liang, and P. S. Albert. 1988. Models for longitudinal data: A generalized estimating
674 equation approach. Biometrics 44:1049–1060.
675 Zelditch, M. L., and A. Goswami. 2021. What does modularity mean? Evolution & Development
676 23:377–403.

Figures

Figure 1. Density plots for random F-statistics, from different combinations of estimation and null model process, illustrating sampling distribution behavior. 100 density curves for every level of simulated λ (colored differently) is overlayed in every frame. Parametric F-distributions are shown as black curves.

Figure 2. Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 0$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

Figure 3. Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 0.5$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

Figure 4. Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 1$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

Figure 5. Statistical power curves for all combinations of estimation and null model process, and for the three data types based on λ. Null hypothesis rejection rates (each point) were based on 200 simulations.

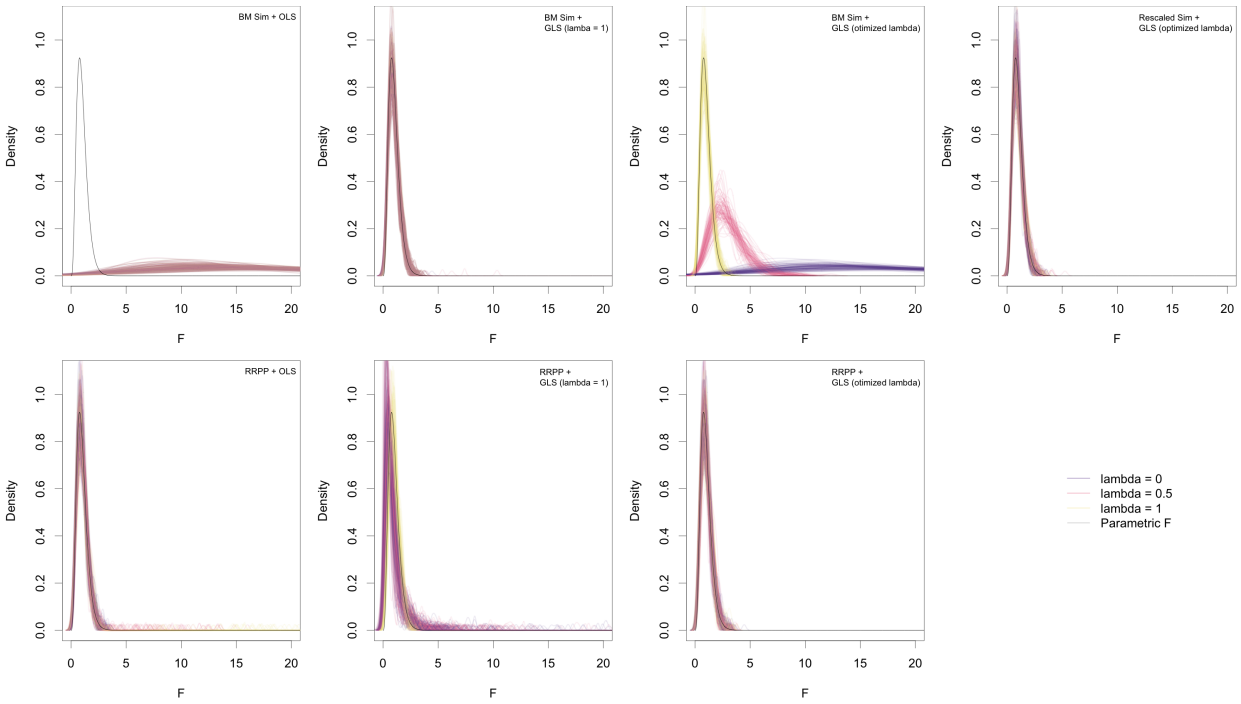

Figure 1: Density plots for random F-statistics, from different combinations of estimation and null model process, illustrating sampling distribution behavior. 100 density curves for every level of simulated λ (colored differently) is overlayed in every frame. Parametric F-distributions are shown as black curves.

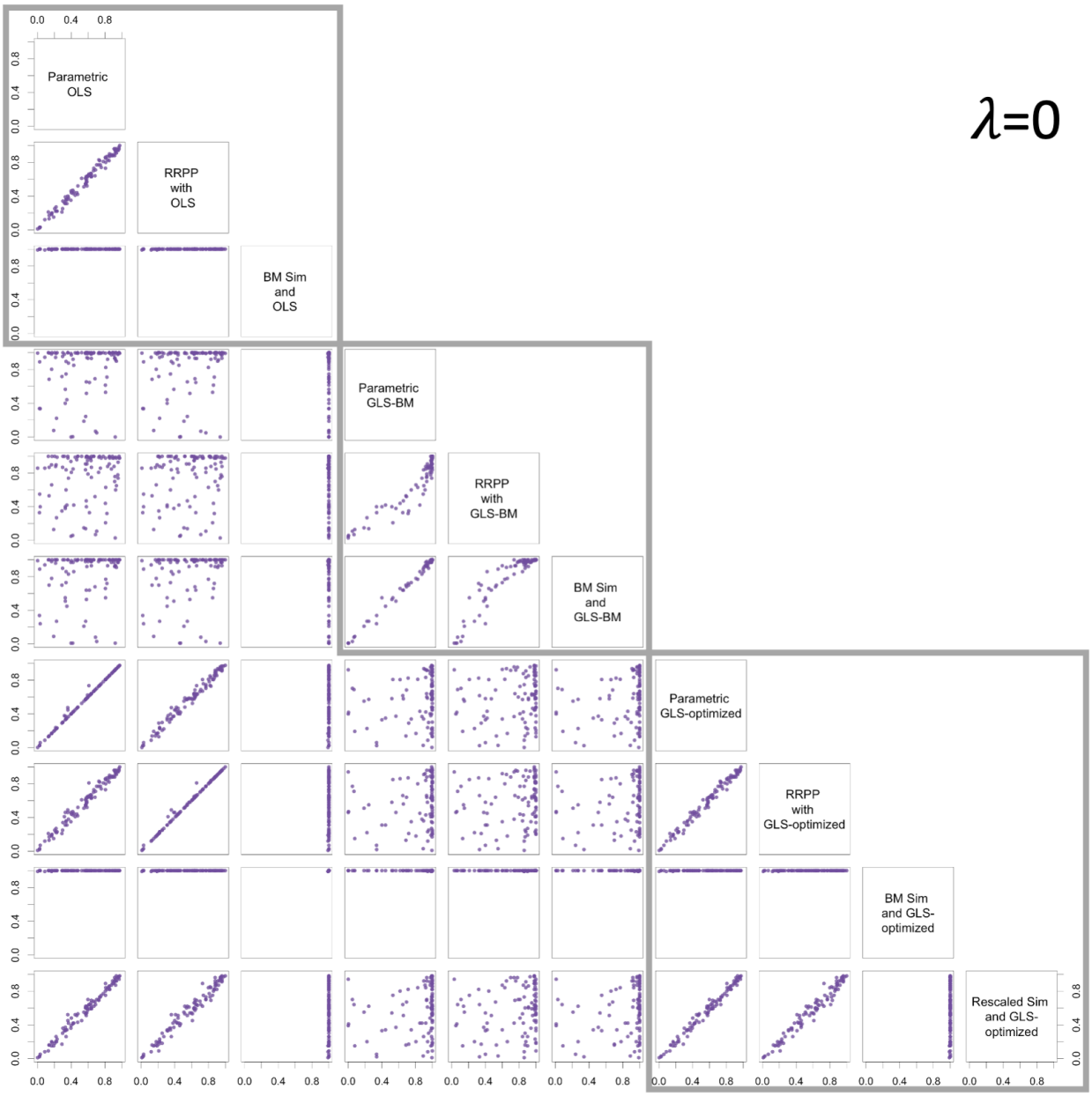

Figure 2: Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 0$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

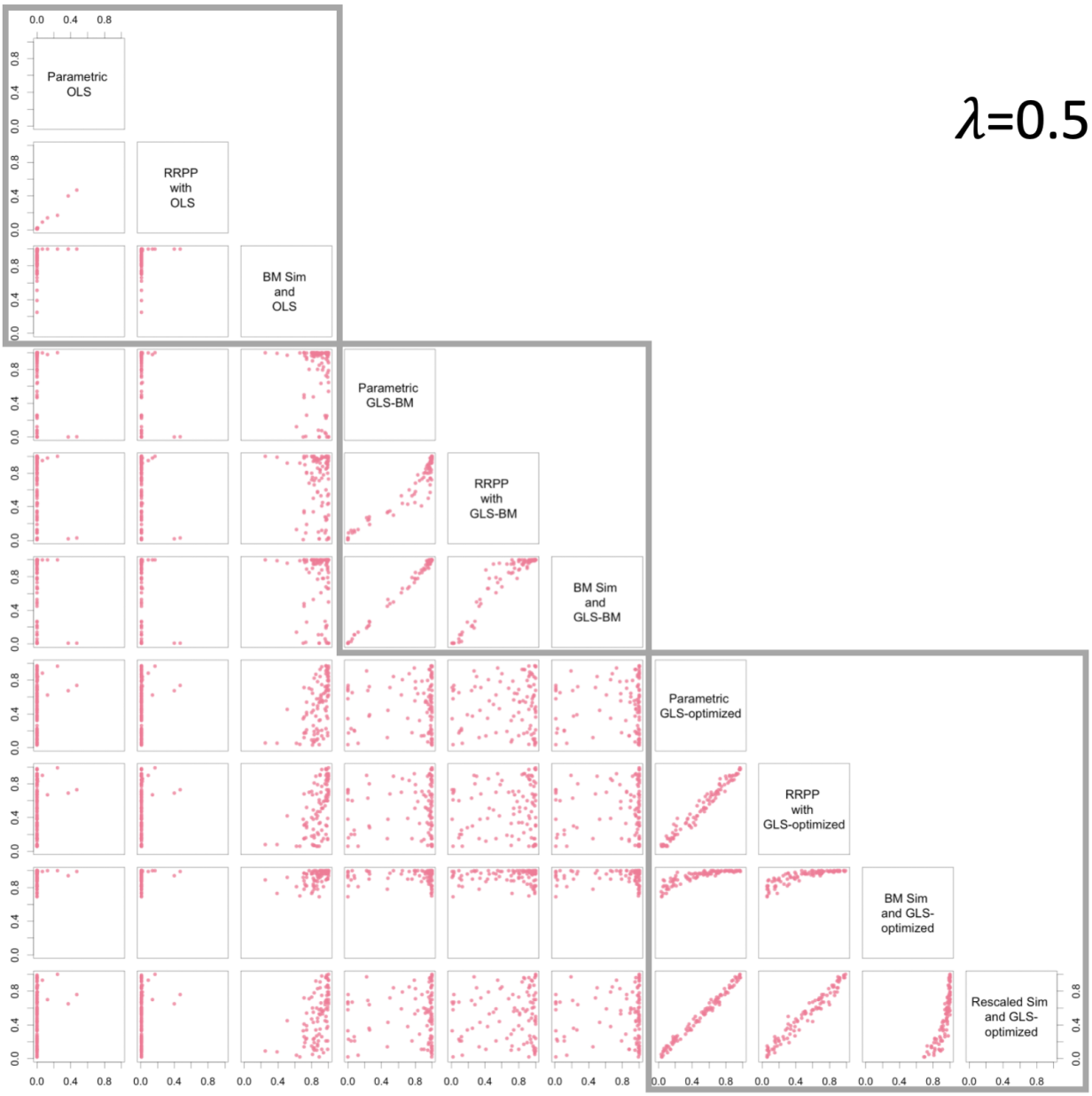

Figure 3: Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 0.5$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

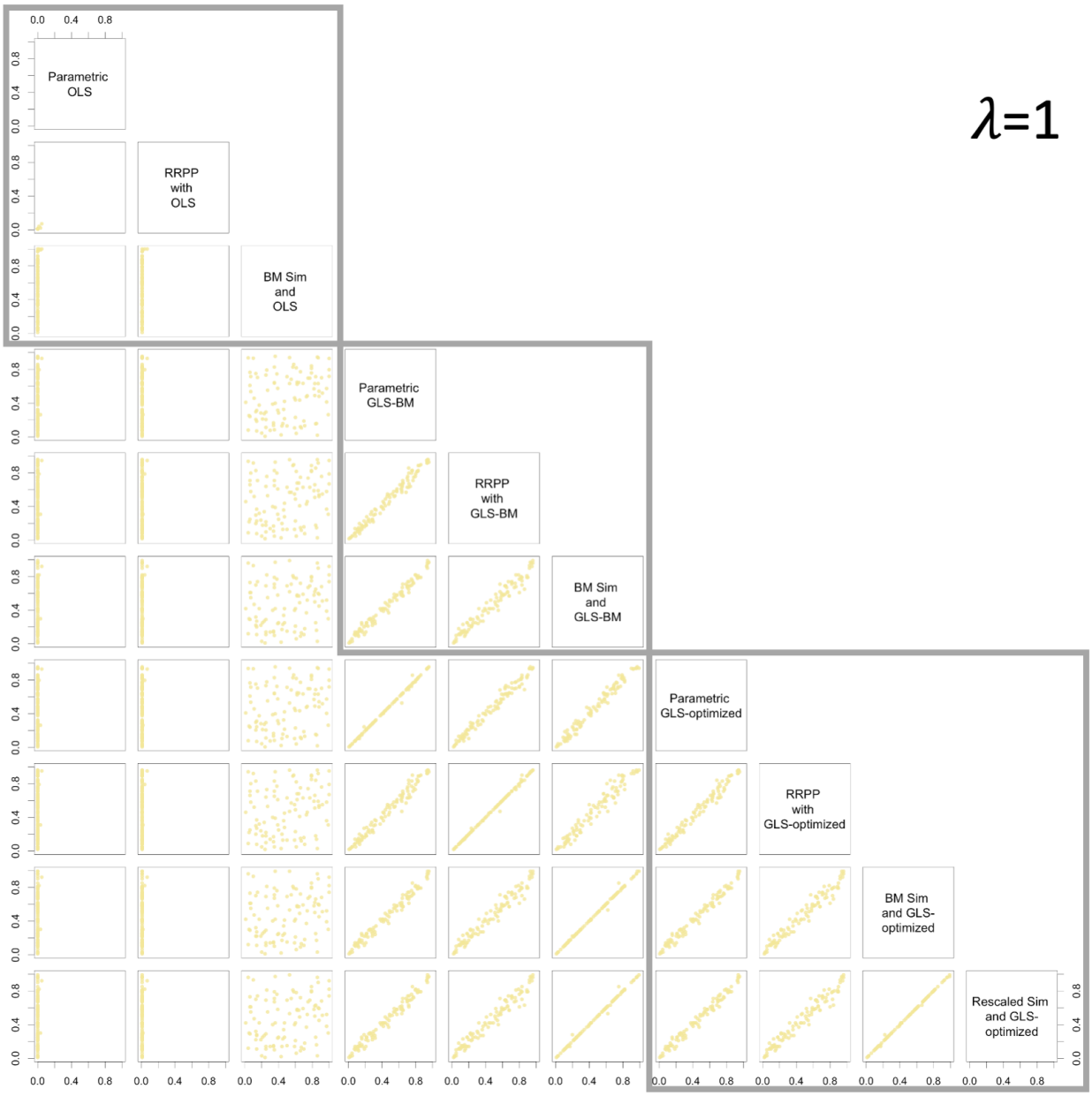

Figure 4: Pairs plots for P-values from different combinations of estimation and null model process, for data generated with $\lambda = 1$, illustrating the consistency of different combinations of null model process and estimation. Gray boxes surround plots with the same estimation method.

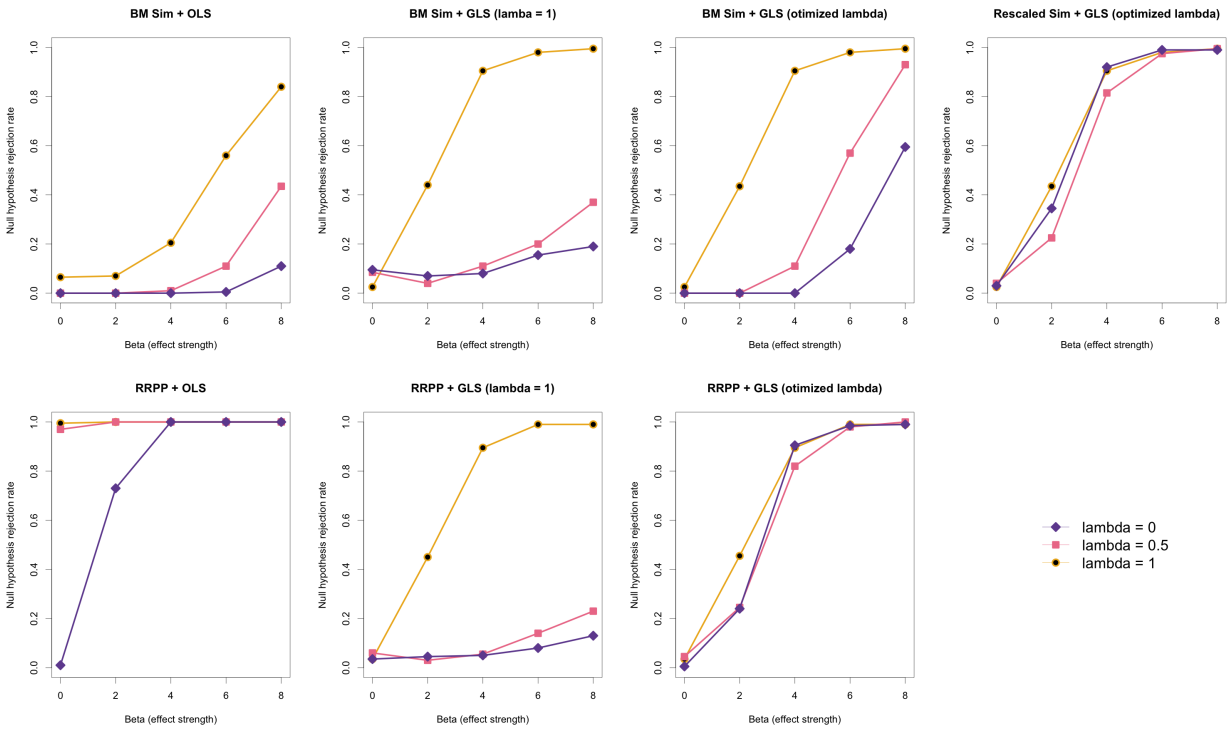

Figure 5: Statistical power curves for all combinations of estimation and null model process, and for the three data types based on λ Null hypothesis rejection rates (each point) were based on 200 simulations.