A Penalized Likelihood Framework for High-Dimensional Phylogenetic Comparative Methods and an Application to New-World Monkeys Brain Evolution

Julien Clavel, Leandro Aristide, Hélène Morlon

TITLE: A Penalized Likelihood Framework For High-Dimensional Phylogenetic Comparative Methods And An Application To New-World Monkeys Brain Evolution

RUNNING HEAD: PENALIZED MULTIVARIATE COMPARATIVE METHODS

CLAVEL Julien[1]

ARISTIDE Leandro[1]

MORLON Hélène[1]

[1]*École Normale Supérieure, Paris Sciences et Lettres (PSL) Research University, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.* *clavel@biologie.ens.fr* *morlon@biologie.ens.fr*

ABSTRACT

Working with high-dimensional phylogenetic comparative datasets is challenging because likelihood-based multivariate methods suffer from low statistical performances as the number of traits *p* approaches the number of species *n* and because some computational complications occur when *p* exceeds *n*. Alternative phylogenetic comparative methods have recently been proposed to deal with the large *p* small *n* scenario but their use and performances are limited. Here we develop a penalized likelihood framework to deal with high-dimensional comparative datasets. We propose various penalizations and methods for selecting the intensity of the penalties. We apply this general framework to the estimation of parameters (the evolutionary trait covariance matrix and parameters of the evolutionary model) and model comparison for the high-dimensional multivariate Brownian (BM), Early-burst (EB), Ornstein-Uhlenbeck (OU) and Pagel's lambda models. We show using simulations that our

penalized likelihood approach dramatically improves the estimation of evolutionary trait covariance matrices and model parameters when *p* approaches *n*, and allows for their accurate estimation when *p* equals or exceeds *n*. In addition, we show that penalized likelihood models can be efficiently compared using Generalized Information Criterion (GIC). We implement these methods, as well as the related estimation of ancestral states and the computation of phylogenetic PCA in the R package *RPANDA and mvMORPH.* Finally, we illustrate the utility of the new proposed framework by evaluating evolutionary models fit, analyzing integration patterns, and reconstructing evolutionary trajectories for a high-dimensional 3-D dataset of brain shape in the New World monkeys. We find a clear support for an Early-burst model suggesting an early diversification of brain morphology during the ecological radiation of the clade. Penalized likelihood offers an efficient way to deal with high-dimensional multivariate comparative data.

KEYWORDS

Evolutionary covariances, Phylogenetic signal, Phylogenetic PCA, Regularization, Ridge, LASSO, large *p* small *n*, Shrinkage, approximate LOOCV, Generalized Information Criterion, New World monkeys, brain shape.

Methods for modeling the evolution of species traits on phylogenetic trees, also known as phylogenetic comparative methods, have exploded since the seminal paper of Felsenstein (1985). In particular, multivariate models, in which several traits coevolve, have allowed the explicit investigation of several aspects of phenotypic evolution that cannot be addressed by separate univariate analyses (Butler and King 2009; Revell and Collar 2009; Bartoszek et al. 2012; Clavel et al. 2015; Caetano and Harmon 2017; Goolsby et al. 2017; Manceau et al.

2017; Bastide et al. 2018). These multivariate phylogenetic models have for example allowed us to analyze the evolutionary integration of traits – i.e. how the evolution of one trait is associated to the evolution of others (Revell and Harmon 2008; Revell and Collar 2009; Bartoszek et al. 2012; Clavel et al. 2015). Estimating the correlated evolution of traits is also a necessary step in several multivariate statistical methods accounting for the evolutionary history of species such as canonical correlation Analysis (CCA) (Revell and Harrison 2008), principal component analysis (PCA) (Revell 2009), two-block partial least-squares (PLS) (Adams and Felice 2014), and tests associated with traditional versions of phylogenetic multivariate analysis of variance (MANOVA) (Garland 1992, 1993), and multivariate generalized least squares regressions (GLS) (Grafen 1989; Martins and Hansen 1997). A first step common to these multivariate phylogenetic comparative methods involves computing the maximum likelihood estimate of traits evolutionary variance covariance matrices - often referred to as the rate matrix **R** (e.g., Revell and Harmon 2008) - under specific models of phenotypic evolution (often, but not always, the multivariate Brownian model).

In many respects working with multivariate methods is a challenging task as the computational time, memory requirement and the number of parameters to estimate for fitting these models grows up non-linearly with increasing dimensions (Ho and Ané 2014; Clavel et al. 2015; Goolsby et al. 2017). The challenge is even greater when working with high-dimensional datasets - where the number of traits (p) is large compared to the number of species (n). In this case, complications arise to compute and/or invert the maximum likelihood estimate (MLE) of traits evolutionary variance covariance matrices. When p approaches n, the MLE of the covariance matrix is no longer a good approximation of the true (population) covariance matrix (James and Stein 1961; Schäfer and Strimmer 2005; van Wieringen 2017). For instance, the leading eigenvalues are systematically overestimated while the last eigenvalues are underestimated (Ledoit and Wolf 2004; Schäfer and Strimmer 2005; Ledoit

and Wolf 2012). As a consequence, multivariate phylogenetic comparative methods based on maximum likelihood are no longer reliable and lead to increased model misspecification or loss of statistical power (Dunn et al. 2013; Adams 2014a, 2014b; Goolsby 2016; Adams and Collyer 2018). Things get even worse when *p* equals or exceeds *n.* In this case, several eigenvalues equal zero and result in a non-invertible MLE of the covariance matrix, which precludes the use of most phylogenetic comparative methods (Adams 2014a, 2014c; Goolsby 2016; Adams and Collyer 2018). In such a situation, a common practice is to reduce the dimensionality of the dataset through PCA and then treat the first axes as independent traits in downstream comparative analysis. But doing so can lead to erroneous conclusions about the underlying evolutionary model (Revell 2009; Uyeda et al. 2015; see also Bookstein 2012 for a similar problem with non independent observations in time-series). Alternative dimension reduction methods have recently been proposed (Tolkoff et al. 2017), but they are currently limited to the Brownian motion model.

As large *p* small *n* datasets become more and more widespread, as for example in geomorphometric and gene expression studies (Dunn et al. 2013; Goolsby 2016; Cross 2017), there have been some recent efforts to tackle the issue. First, Adams (2014a, b, c) developed an approach, implemented in the geomorph R package, to compare evolutionary rates across clades, measure phylogenetic signal and perform regression (PGLS) analyses for high-dimensional data by transforming the data using analytical results known for the Brownian motion model. This approach is useful, yet restricted to Brownian motion processes (in fact to simple cases of BM that do not require computing a likelihood). More recently, Goolsby (2016) showed that Adams' method suffers from low statistical performances, and proposed an alternative approach based on pairwise composite likelihoods, implemented in the phylocurve R package. The approach considers pairs of traits at a time and computes the corresponding pair-level likelihoods under any evolutionary model (such as Brownian

motion, Ornstein-Uhlenbeck and Early burst); a pseudo-likelihood is then computed by summing the pair-level likelihoods. This is again a helpful approach, yet using a pseudo-likelihood prevents applying classical statistical techniques; for example the parametric evaluation of uncertainty around parameters estimates and model selection procedures based on Information Criterion cannot be performed, and we have to resort to simulations (although analogues to conventional Information criteria have been derived for composite-likelihoods; Varin et al. 2011). In his paper, Goolsby (2016) focused on model comparison, not estimates of the covariance matrix; such estimates can be obtained, but as discussed earlier we expect the obtained covariance to be a poor estimate of the true covariance when the ratio p/n approaches 1 or to be singular when $p>n$. Moreover, Goolsby's approach has also recently been criticized for being sensitive to potential data transformation such as rigid rotations, which limits its use, for example in the case of geomorphometric data (Adams and Collyer 2018). Hence, while the two developments above advanced the field by providing some new tools for high-dimensional datasets, they did not directly address the problem of estimating accurate covariance matrices, and thus do not allow extending the full arsenal of available phylogenetic comparative methods to small n large p datasets.

In this article we develop a general penalized likelihood framework for estimating and inverting trait covariance matrices that allows multivariate phylogenetic comparative methods to be extended to high-dimensional data. Penalized likelihood (Huang et al. 2006; Levina et al. 2008; Warton 2008; Friedman et al. 2008; van Wieringen and Peeters 2016) is a modern formulation of the regularization/shrinkage approaches commonly used in the statistical literature (Hoerl and Kennard 1970a; Dempster 1972; Friedman 1989; Tibshirani 1996; Ledoit and Wolf 2004; Schäfer and Strimmer 2005). The goal of regularization approaches is to constrain the estimates of the parameters, such as the MLE, in order to make them more accurate and to solve ill-posed problems. For instance, the ridge regularization was one of the

first methods used in large multiple regression problems to circumvent singularity issues due to over parameterization and to obtain estimates with reduced mean squared errors (Hoerl and Kennard 1970a, 1970b). Regularization techniques have been successfully applied in phylogenetics for reconstructing trees (Kim and Sanderson 2008), estimating molecular rates and divergence times (Sanderson 2002; Smith and O'Meara 2012), and more recently for estimating lineage-specific evolutionary rates and ancestral states under univariate Brownian models of trait evolution (Kratsch and McHardy 2014) or detecting shifts in the optimum of Ornstein-Uhlenbeck models (Khabbazian et al. 2016). In all these studies, regularization has proven to be an efficient alternative to Bayesian techniques, generally used to integrate over parameter rich spaces. While the idea of using regularization approaches to address the large *p* small *n* problem has been suggested before in another context (Dunn et al. 2013), it has not been developed in practice. In the case of covariance matrix estimation, the goal of the penalized likelihood approach is to constrain the MLE to a symmetric positive definite and thus invertible matrix with a simpler structure and improved statistical properties (e.g., correcting for the over dispersion of the eigenvalues or reducing the quadratic loss). This is achieved by adding a penalty term to the likelihood that favors solutions with the required structure and by finding the optimal solution that maximizes the penalized likelihood.

We consider four commonly used penalizations that fall in two broad types: the ridge penalty, also called L_2 regularization (Huang et al. 2006; Warton 2008; van Wieringen and Peeters 2016), and the LASSO penalty (Least Absolute Shrinkage and Selection Operator), also called L_1 regularization (Tibshirani 1996; Levina et al. 2008; Friedman et al. 2008; Bien and Tibshirani 2011). We propose a simple method for jointly estimating model parameters and the level of penalization. In addition, we show how to perform model selection using Information Criteria (Akaike 1974; Konishi and Kitagawa 1996).

We apply our penalized framework to the estimation of evolutionary trait covariance matrices, model fit and model comparison for the high-dimensional multivariate Brownian (BM), Early-burst (EB), Ornstein-Uhlenbeck (OU) and Pagel's lambda models. We perform a series of simulations to test the performance of our penalized likelihood approach in terms of estimation of the variance covariance matrix and of evolutionary model parameters, as well as model comparison. When alternative approaches were available (i.e. ML when $p<n$ for both parameter estimation and model comparison based on information criteria, and Goolsby's pairwise likelihood when $p>n$ for parameter estimation), we compared the performance of our approach to these alternatives. We also implemented the estimation of ancestral states (Martins and Hansen 1997; Cunningham et al. 1998) and phylogenetic PCA (Revell 2009) under the above-mentioned processes.

We illustrate potential applications of our approach by analyzing an extremely high-dimensional 3-D geometric morphometric dataset describing brain shape variation for 48 species of New World monkeys. Previous model-fitting analyses on a subset of PCs axes derived from this dataset, together with disparity-through-time plots, have indicated that brain shape has quickly diversified during the early history of the clade, according to the occupation of different ecological niches (i.e., an adaptive radiation; Aristide et al. 2016). However, it has been suggested that, under certain conditions, comparative analyses conducted only over the first PC axes of multivariate datasets may tend to artificially favor "early burst" like processes and should instead be ideally performed using fully multivariate approaches that avoid dimensionality reduction steps (Uyeda et al. 2015). Here we apply our penalized likelihood framework to this high-dimensional empirical dataset, estimating the fit of alternative multivariate evolutionary models (BM, EB, and OU) and comparing the results to those obtained previously. Additionally, we utilize the best fitting model parameters to explore the global patterns of covariation among shape variables (i.e. evolutionary integration) derived

from a phylogenetic PCA, as well as to obtain an ancestral reconstruction of the New World monkey's brain shape and evolutionary trajectories of its change through time.

Finally, we discuss our results and the broad potential of penalized-likelihood approaches for extending the multivariate comparative toolbox to high-dimensional data.

MATERIALS AND METHODS

Below we detail how to fit a class of multivariate trait evolution models to high-dimensional comparative data using penalized likelihood. These models include the multivariate Brownian model, and all other multivariate models (e.g. EB, OU, and Pagel's λ) for which a single (between traits) evolutionary variance covariance matrix $\boldsymbol{R}$ applies across the phylogeny ($\boldsymbol{R}$ is often noted $\boldsymbol{\Sigma}$, the multivariate counterpart of σ^2, the diffusion parameter of the univariate Brownian model), and all traits share a common phylogenetic variance-covariance matrix $\boldsymbol{C}$ representing the expected scaled variance for each species (a function of the height s_{ii} above the root of the *ith* taxa and the parameters of the model) and between species scaled covariances (a function of the shared evolutionary history s_{ij} between taxa *i* and *j* and the parameters of the model; Felsenstein 1985; Rohlf 2006). For these "simple" models, we can efficiently compute the corresponding log-likelihood using the multivariate normal density in Equation (1). For more complex models with for instance distinct $\boldsymbol{C}$ matrices for each trait or with distinct $\boldsymbol{R}$ matrices for different clades (e.g., the "BMM" model in mvMORPH; Clavel et al. 2015), different algorithms are required for computing the corresponding log-likelihood efficiently (e.g., Freckleton 2012; Clavel et al. 2015; Caetano & Harmon 2017; Goolsby et al. 2017). The general penalized-likelihood framework presented here can be applied to such models, provided these algorithms are available.

The Likelihood

The log-likelihood of traits is given by the multivariate normal density:

$$\mathcal{L} = -\frac{1}{2}\{np\log(2\pi) + p\log|\boldsymbol{C}| + n\log|\boldsymbol{R}| + \mathrm{tr}[\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)^T\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)]\} \quad (1a)$$

Where n is the number of species, p is the number of traits, $\boldsymbol{Y}$ is the n by p matrix of trait values, $\boldsymbol{X}$ the design matrix mapping ancestral states to species for each trait (a n dimensional column vector of one in what follows, but this matrix could have different column dimensions in general; e.g. for regression models such as PGLS), β a vector of size p representing the ancestral states for each trait, $\boldsymbol{C}$ is the phylogenetic covariance matrix of size n by n (see above), and $\boldsymbol{R}$ the traits covariance matrix of size p by p (see details in Felsenstein 2004; Revell and Harmon 2008; Freckleton 2012; Clavel et al. 2015). For a given matrix **A**, tr($\boldsymbol{A}$) stands for the trace (the sum of the diagonal elements), $|\boldsymbol{A}|$ for the determinant, $\boldsymbol{A}^{-1}$ for the inverse, $\boldsymbol{A}^{T}$ for the transpose and $\boldsymbol{A}^{-T}$ for the transpose of the inverse. The likelihood expression written in Equation (1a) can also be written as a function of a Kronecker product of $\boldsymbol{R}$ and $\boldsymbol{C}$, and also as a function of the inverse of the traits covariance matrix $\boldsymbol{R}$, which is commonly referred to as the precision matrix $\boldsymbol{\Omega}$. We report these expressions in Appendix 1, as they are commonly used in the multivariate statistics literature; the expression with $\boldsymbol{\Omega}$ is also the one we use to obtain some of our results.

In addition to this classical log-likelihood, we consider the restricted (RE) log-likelihood, given by (Felsenstein 1973, 2004; Freckleton 2012):

$$\mathcal{L} = -\frac{1}{2}\{(n-1)p\log(2\pi) + p\log|\boldsymbol{C}| + (n-1)\log|\boldsymbol{R}| + \mathrm{tr}[\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)^T\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)] + p\log|\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}|\} \quad (1b)$$

Estimates based on the unrestricted likelihood are known to be biased (Harville 1977), and the restricted likelihood is thus more appropriate for estimating the residuals and variance-covariance parameters of the models (Felsenstein 1973, 2004; Freckleton 2012). Here we consider both types of likelihoods because there are straightforward extensions of the approach where restricted likelihood won't be appropriate for model comparison (e.g. cases when the number of ancestral states – or fixed-effects – coded in the design matrix $\boldsymbol{X}$ differ across models Freckleton 2012; but see Gurka 2006).

The maximum likelihood estimate (MLE) of the evolutionary covariance matrix $\boldsymbol{R}$ is given by:

$$\widehat{\boldsymbol{R}} = \frac{(\boldsymbol{Y}-\boldsymbol{X}\beta)^T \boldsymbol{C}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)}{n} \quad (2)$$

or by using *n-1* in the denominator of Equation (2) for the restricted log-likelihood. Unfortunately when the number of variables *p* approaches the number of species *n*, $\widehat{\boldsymbol{R}}$ (the sample estimate) does not necessarily provide a good estimation of the true $\boldsymbol{R}$, and when *p* exceeds *n*, $\widehat{\boldsymbol{R}}$ is singular, meaning that we cannot compute its inverse nor the log of its determinant, and thus lose the ability to compute the associated log-likelihood.

The Penalized Likelihood

In order to solve the large *p* small *n* problem above, we use the penalized likelihood framework, which aims to provide estimators for the covariance matrix $\boldsymbol{R}$ that are well conditioned, positive definite, and thus invertible. We consider four penalties (Table 1): a linear ridge penalty hereafter referred to as the "archetypal ridge penalty", two types of quadratic ridge penalties, and the LASSO penalty. These penalties are motivated by the bias-

variance tradeoff and aim to find a compromise between the high-variance and low-bias $\widehat{\boldsymbol{R}}$ and a well-behaved low-variance (but higher bias) target matrix $\boldsymbol{T}$. Typical $\boldsymbol{T}$ matrices used in the literature are the null matrix (a matrix full of zeros), the identity matrix (or a multiple of the identity matrix), and the diagonal matrix composed of the estimated variances for each trait, that is the diagonal elements of $\widehat{\boldsymbol{R}}$ (Hoffbeck and Landgrebe 1996; Schäfer and Strimmer 2005; van Wieringen and Peeters 2016). The latter is often considered as the best choice as it has less bias than the other matrices (Schäfer and Strimmer 2005; van Wieringen and Peeters 2016, but see Lancewicki and Aladjem 2014).

The archetypal ridge estimator is given by (e.g., Schäfer and Strimmer 2005; Warton 2008; van Wieringen and Peeters 2016):

$$\boldsymbol{R}(\gamma)_{Archetypal\ ridge} = (1-\gamma)\widehat{\boldsymbol{R}} + \gamma\boldsymbol{T} \quad (3)$$

where γ is a regularization parameter (often also called tuning parameter) bounded between 0 and 1 that controls the intensity of the penalty, and $\boldsymbol{T}$ is a target matrix. We took $\boldsymbol{T}$ to be the diagonal matrix composed of the estimated variances for each trait (Hoffbeck and Landgrebe 1996; Schäfer and Strimmer 2005; van Wieringen and Peeters 2016). This estimator reduces to the MLE when $\gamma = 0$ and to $\boldsymbol{T}$ when $\gamma = 1$ (a situation in which traits are estimated to be independent when $\boldsymbol{T}$ is diagonal); intermediate γ values provide a good tradeoff between bias and variance reduction. We explain later how the value of the regularization parameter is selected. The archetypal ridge estimator results from maximizing the following penalized-likelihood (van Wieringen and Peeters (2016), see Appendix 2):

$$\mathcal{L}_P = -\frac{1}{2}\{np\log(2\pi) + p\log|\boldsymbol{C}| + n\log|\boldsymbol{R}| + (1-\gamma)\mathrm{tr}[\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)^T\boldsymbol{C}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)] + n\gamma\mathrm{tr}[\boldsymbol{R}^{-1}\boldsymbol{T}]\} \quad (4)$$

More generally, different types of regularizations are obtained by amending a penalty term to the log-likelihood (Eq. 1). The first type of quadratic ridge penalized-likelihood that we consider (hereafter referred to as ridge quad. (var)) is given by (van Wieringen and Peeters 2016):

$$\mathcal{L}_P = \mathcal{L} - \frac{n\gamma}{4}\mathrm{tr}[(\boldsymbol{R}^{-1}-\boldsymbol{T}^{-1})^T(\boldsymbol{R}^{-1}-\boldsymbol{T}^{-1})] \quad (5)$$

where again we take the target matrix $\boldsymbol{T}$ to be the diagonal elements of $\widehat{\boldsymbol{R}}$, and the regularization parameter γ can take any value in $[0,\infty[$.

Finally, we consider two penalized-likelihoods given by:

$$\mathcal{L}_P = \mathcal{L} - \frac{n\gamma}{4}\|\boldsymbol{R}^{-1}\|^q \quad (6)$$

where the penalty $\|\cdot\|^q = \sum_{i=1}^{p^2}|\cdot|^q$ is the matrix norm and γ can take any value in $[0,\infty[$. With *q=2* (Huang et al. 2006; Witten and Tibshirani 2009; van Wieringen and Peeters 2016), this penalty writes $\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\boldsymbol{R}_{ij}^{-1}\right)^2 = \mathrm{tr}[\boldsymbol{R}^{-T}\boldsymbol{R}^{-1}]$ and is equivalent to the quadratic penalty given by Equation (5), but with $\boldsymbol{T}$ chosen to be the null matrix. We refer to this penalty as ridge quad. (null). With *q=1*, we obtain the penalty well known as the LASSO (Tibshirani 1996; Friedman et al. 2008; Witten and Tibshirani 2009). The LASSO also shrinks $\widehat{\boldsymbol{R}}$ towards the null matrix, but does so by progressively assigning the least significant covariance terms to zero instead of homogeneously reducing all terms as in the ridge quad. (null).

These penalties have been successfully applied to the estimation of large covariance matrices in other contexts (e.g., Ledoit and Wolf 2004; Schäfer and Strimmer 2005; Huang et al. 2006; Friedman et al. 2008; Witten and Tibshirani 2009; van Wieringen and Peeters 2016). The archetypal ridge estimator was historically considered because efficient methods can be used to compute and solve it (Friedman 1989; Hoffbeck and Landgrebe 1996; Ledoit and Wolf 2004; Schäfer and Strimmer 2005; Warton 2008; Theiler 2012). The quadratic ridge estimator resembles more to the original ridge penalty used in regression analysis and shows improved statistical properties (van Wieringen and Peeters 2016; van Wieringen 2017), but its analytical solution is more computationally demanding. The LASSO exhibits the same interesting shrinkage properties as the ridge penalizations, and in addition provides sparse matrix estimates, which provides an efficient way to perform covariance selection, but it requires even more intensive iterative algorithms (Friedman et al. 2008; Witten et al. 2011; Bien and Tibshirani 2011). An important difference between the various penalizations is that some of them (e.g. the ridge quad null) are rotation-invariant, meaning that they are robust to data orientation, while others (e.g. the ridge quad. var. and the LASSO) are not (Table 1, Ledoit and Wolf 2004; Warton 2008; Ledoit and Wolf 2012, 2015). When working on untransformed data, methods that are not rotation-invariant can be used. However in other cases, for example for variables generated from shape spaces in geomorphometrics which usually depend on the arbitrary choice of the orientation of a baseline shape (Rohlf 1999), only methods insensitive to these rotations should be used (Rohlf 1999; Adams and Collyer 2018). The merits of the various types of penalizations are discussed at greater length in our Discussion and summarized in Table 1.

In what follows, we need to compute first and second order derivatives of the penalized-likelihood for the different types of penalizations. These derivations are detailed in Appendix 2.

<TABLE_1>

Solving the Penalized Likelihood

Unlike the archetypal ridge estimator (Eq. 3), which formulation preceded the corresponding penalized-likelihood (Warton 2008; van Wieringen and Peeters 2016; Appendix 2), the three other estimators are found by computing the *argmax* of the various penalized-likelihoods.

The two quadratic ridge estimators (with $\boldsymbol{T}$ diagonal or $\boldsymbol{T}$ null) can be obtained analytically by finding the **R** matrix for which the first derivative of Equation (5) (and Eq. 6 with *q*=2) equals zero (Witten and Tibshirani 2009; van Wieringen and Peeters 2016). For a fixed value of the regularization parameter γ in $[0, \infty[$, the quadratic ridge estimator is given by:

$$\boldsymbol{R}(\gamma)_{Quadratic\ ridge} = \left[\gamma \boldsymbol{I_p} + \frac{1}{4}\left(\widehat{\boldsymbol{R}} - \gamma \boldsymbol{T}^{-1}\right)^2\right]^{1/2} + \frac{1}{2}\left(\widehat{\boldsymbol{R}} - \gamma \boldsymbol{T}^{-1}\right) \quad (7)$$

where $\boldsymbol{I_p}$ is the identity matrix of dimension *p*.

The LASSO estimator, obtained by solving the penalized likelihood given by Eq. 6 with *q*=1, cannot be computed analytically. Several algorithms have been proposed to compute it iteratively (Tibshirani 1996; Friedman et al. 2008), such as the *"graphical lasso"* algorithm implementation in the R package "glasso*"* (Friedman et al. 2008; Witten et al. 2011). This algorithm provides a sparse estimate of $\boldsymbol{R}^{-1}$ rather than $\boldsymbol{R}$. Other algorithms that

directly introduce zeros in $\boldsymbol{R}$ (Bien and Tibshirani 2011) are more computationally demanding and have not been as well tested. As we encountered non-termination convergence issues with the "glasso*"* package, we instead implemented the *GLASSOFAST* algorithm proposed by Sustik and Calderhead (2012), which is also generally faster than the one proposed by Friedman et al. (2008). We made our implementation of this general-purpose algorithm publicly available on github (https://github.com/JClavel/glassoFast) and in an R package called "glassoFast*"* on the CRAN repository (R Development Core Team 2016).

Estimating the regularization and model parameters

We used cross-validation to obtain the *regularization parameter* γ (Hastie et al. 2009). This parameter cannot be computed jointly with the model parameters by directly maximizing the penalized log-likelihood (Eqs. 4-6), because it would then always be estimated to 0. This is because the true covariance $\boldsymbol{R}$ in Equations (4-6) is generally not available and is instead replaced by a covariance $\widehat{\boldsymbol{R}}$ estimated from the data by taking the MLE. Hence, the same data is used both to evaluate the likelihood of each observation and to estimate the covariance matrix. One way to solve this issue is to approach the likelihood by an expression that uses independent data to evaluate the likelihood and to estimate the covariance matrix. Here we use a leave-one-out cross-validation (LOOCV) of the penalized likelihood for which the general idea is to average the likelihoods over each observation (testing samples) with a covariance matrix $\widehat{\boldsymbol{R}}$ estimated without the observation (training samples) (Hoffbeck and Landgrebe 1996; Hastie et al. 2009; Theiler 2012). We develop a leave-one-out cross-validation scheme that consists first in directly removing the phylogenetic correlation between

species in the data; this avoids defining arbitrary blocks of independent species based on phylogenetic distance (e.g., Roberts et al. 2017). We note

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{Y}, \widetilde{\boldsymbol{X}} = \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{X}$$

the transformed, decorrelated data, where $\boldsymbol{C}^{-\frac{1}{2}}$ is the square root of $\boldsymbol{C}^{-1}$. β, the generalized least squares (GLS) estimate of the ancestral states ($\beta = (\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{Y}$, e.g. Rohlf 2001), is then directly given by $\beta = \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{Y}}$. The leave-one-out cross-validated penalized log-likelihood is expressed as:

$$\mathcal{L}_{CV} = -\frac{1}{2}\Big[np\log(2\pi) + p\log|\boldsymbol{C}| + \sum_{i=1}^{n}\Big(log\left|\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}\right| + tr\left[\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}^{-1}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\tilde{Y}_i - \tilde{X}_i\beta\right)^T\right]\Big)\Big], \quad (8)$$

Where $\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}}$ and $\widetilde{\boldsymbol{X}}_{\boldsymbol{i}}$ are the p by 1 column vectors made of the i^{th} row of the corresponding matrix (i.e. the row related to taxon i; $\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)^T$ is thus a p by p matrix) and $\widetilde{\boldsymbol{R}}(\gamma)_{(-\boldsymbol{i})}$ is the p by p penalized likelihood estimator computed as described in Eq. 3 for the archetypal ridge, Eq. 7 for the quadratic ridge, and the solution of Eq. 6 (with q=1) for the LASSO, with $\widehat{\boldsymbol{R}}$ replaced by:

$$\widehat{\boldsymbol{R}}_{(-\boldsymbol{i})} = \frac{\left(\widetilde{\boldsymbol{Y}}_{(-\boldsymbol{i})} - \widetilde{\boldsymbol{X}}_{(-\boldsymbol{i})}\beta_{(-\boldsymbol{i})}\right)^T\left(\widetilde{\boldsymbol{Y}}_{(-\boldsymbol{i})} - \widetilde{\boldsymbol{X}}_{(-\boldsymbol{i})}\beta_{(-\boldsymbol{i})}\right)}{n-1}$$

where the subscript $(-i)$ refers to matrices computed without the row corresponding to taxon i.

We can jointly estimate the regularization parameter and the model parameters by maximizing the LOOCV (Equation 8). Calculating the LOOCV can be very computationally intensive, and so we consider faster implementations (for the archetypal ridge penalty, Appendix 3) and approximations (for the quadratic ridge and LASSO penalties, Appendix 3). Also, for the simple models considered here with $\boldsymbol{X}$ a n dimensional column vector of one (corresponding to the case when only one ancestral trait value is reconstructed for each trait, vs. models with specific subclades that can have distinct ancestral trait values for a given trait, such as when there are sudden variations in trait values during the evolutionary history of clades – O'Meara et al. 2006), we can compute the LOOCV directly on the phylogenetic independent contrasts scores that are, in this case, closely related to the residuals $\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta$ in Equation (8) (Appendix 3; see also Stone 2011; Blomberg et al. 2012). We estimate the model and regularization parameters by maximizing the LOOCV, considering both the likelihood and the restricted likelihood, using the L-BGFS-B routine implemented in the *optim* function in R.

We also show in Appendix 2 how to compute standard errors around parameter estimates and provide some examples in the Supplementary Material (at http://datadryad.org; doi:10.5061/dryad.rf7317t).

Model selection

Penalized likelihoods cannot be compared using traditional information criteria such as the Akaike Information Criterion (AIC, Akaike 1974; Burnham and Anderson 2002), but they can be compared using the Generalized Information Criterion (GIC), which is an extension of the AIC to Maximum likelihood-type estimators (or M-estimators, Konishi and Kitagawa 1996, 2008). Consider a series of independent data $x_1, x_2, \ldots, x_n$ and a fitted model $f(x|\theta)$ with

$\theta = \theta_1, \theta_2, \ldots, \theta_k$ the parameters of the model; the M-estimator $\hat{\theta}$ is given by the solution of an equation of the type $\sum_{i=1}^{n} \psi(x_i, \theta) = 0$. Then, the GIC is defined by:

$$GIC = -2 \sum_{i=1}^{n} \log f\left(x_i \middle| \hat{\theta}\right) + 2\mathrm{tr}(\boldsymbol{J}^{-1}\boldsymbol{I}) \quad (9)$$

where

$$\boldsymbol{J} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \psi(x_i,\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} \text{ and } \boldsymbol{I} = \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i, \hat{\theta}\right) \frac{\partial f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} \quad (10)$$

Applied to our case, *f* is the likelihood function and the GIC is obtained by taking $x_i = \widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta$ (i.e., the phylogenetically transformed residuals). In the case of ML estimation, $\psi(x, \theta)$ is simply the derivative of the log-likelihood with respect to θ. In the case of penalized likelihood estimation, it is the derivative of the penalized log-likelihood. The bias term $\mathrm{tr}(\boldsymbol{J}^{-1}\boldsymbol{I})$ in Equation (9) can be seen as an effective number of degrees of freedom – a monotonic function of the regularization parameter (e.g., Rondeau et al. 2003; Abbruzzo et al. 2014; Vinciotti et al. 2016). It can be computed as $\sum_{\tau=1}^{k} \mathrm{tr}(\boldsymbol{J}_{\boldsymbol{\tau}}^{-1}\boldsymbol{I}_{\boldsymbol{\tau}})$, with $\boldsymbol{J}_{\boldsymbol{\tau}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \psi(x_i,\theta)}{\partial \theta_\tau}\Big|_{\theta=\hat{\theta}}$ and $\boldsymbol{I}_{\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i, \hat{\theta}\right) \frac{\partial f(x_i|\theta)}{\partial \theta_\tau}\Big|_{\theta=\hat{\theta}}$ (Konishi and Kitagawa 2008). In our case we take the parameters of the model to be $\theta_1 = \boldsymbol{\Omega}$ (to simplify the computation – Appendix 1), $\theta_2 = \beta$, and θ_3 the parameter of the trait evolution model (e.g. α, *r* or λ). In the case of ML estimation, we compute the derivatives with respect to $\boldsymbol{\Omega}$ and β using the well-known closed form derivatives of the log-likelihood of the multivariate normal density (e.g., McCulloch 1982; Anderson and Olkin 1985; Appendix 2). In the case of the penalized likelihood estimation, we derive the derivatives and provide efficient solutions to compute them in Appendix 2. Finally, in order to speed up the computation of the GIC, we approximate the third term in the sum (i.e. the term corresponding to *k*=3) by the number of

parameters of the trait evolution model, which corresponds to the number of degrees of freedom. Model selection with information criteria under REML estimation is often achieved with different formulations for the bias term (see discussions in Gurka, 2006). Here we follow Gurka (2006) and use the restricted log-likelihood with the bias term estimated as above, that is, by accounting for the estimation of all the parameters in the model.

Implementation

We implemented our penalized-likelihood fit of multivariate evolution for the BM, EB, OU and Pagel's lambda model. The implementation is publicly available on github (https://github.com/hmorlon/PANDA/tree/PenalizedMLPhylo) and in the R packages RPANDA (Morlon et al. 2016, function *fit_t_pl*) and mvMORPH (Clavel et al. 2015, function *mvgls*). We allowed (as an option in the function) for the possibility to account for intra-specific variation and measurement errors (they may otherwise bias the model selection, Silvestro et al. 2015) by directly estimating them from the data in the fitting process (Housworth et al. 2004; Ives et al. 2007; Hansen and Bartoszek 2012, see also the Supplementary Material). Three additional functions compute the corresponding GIC scores (function *GIC*), perform the reconstruction of ancestral states (Martins and Hansen 1997; Cunningham et al. 1998, function *ancestral*), and output the phylogenetic PC axes based on the estimated variance-covariance matrix (Revell 2009, function *phyl.pca_pl*).

Testing the performance of the PL approach with simulations

We use simulations to assess the ability of our penalized likelihood (PL) approach to recover parameter estimates and to select the proper model, for various multivariate evolutionary

models. We compare the performance of our approach to that of maximum Likelihood (ML) and pairwise composite likelihood (PCL) (Goolsby 2016) approaches. We focus on four simple and widely used models of phenotypic evolution: the Brownian Motion (BM), the Early-Burst (EB), the Ornstein-Uhlenbeck (OU) and Pagel's λ models (a measure of phylogenetic signal). We follow Goolsby (2016) and Bastide et al. (2018) and consider the pull matrix of the multivariate OU process (sometimes called α or $\boldsymbol{A}$; Bartoszek et al. 2012, Clavel et al. 2015) to be diagonal – that is, the coevolution between traits is entirely modeled in $\boldsymbol{R}$ (i.e. in the stochastic, not deterministic part of the process); in addition, the same pull parameter α applies to all the traits. For the OU and EB models we use parameter values (α for OU and r for EB) corresponding to 0 (BM) and 0.5, 1, 3, 5 and 10 half-lives elapsing over the tree height ($t_{1/2} = \log(2)/(r \; or \; \alpha)$), held identical across traits. For Pagel's λ transformation we use values of: 0, 0.2, 0.4, 0.6, 0.8 and 1 (BM), held identical across traits.

For each model and each parameter set, we simulate 1,000 trait datasets. We fix the size of the phylogeny to n=32 tips and vary the number of traits p from 2 to 50 (p=2, 5, 10, 25, 31, 50). Each simulation consists in the following steps: i) simulate a pure-birth tree with n=32 tips scaled to unit height using the *pbtree* function in "phytools" (Revell 2012), ii) transform the branch-lengths of the tree according to the phenotypic evolution model considered, using the stretching approach (Pagel 1999; O'Meara 2012; this step is skipped for BM that does not require any transformation), iii) simulate a random $\boldsymbol{R}$ matrix using the approach of Uyeda et al. (2015), which intends to reflect the structure of typical evolutionary rate matrices (e.g., Mezey and Houle 2005); this matrix is obtained by sampling the eigenvalues from an exponential distribution with rate $\lambda = 1/100$, assuring a certain degree of correlation between traits, and finally iv) simulate the traits on the transformed tree under a multivariate BM with evolutionary rate matrix $\boldsymbol{R}$ using the *mvSIM* function in "mvMORPH"

(Clavel et al. 2015). All simulations were performed on a Linux platform with R version 3.2.3 (R Development Core Team 2016).

Next, to each simulated dataset, we apply our penalized likelihood approach with the archetypal ridge, the two quadratic ridge, and the LASSO penalties, as well as the PCL approach implemented in the *evo.model* function from the R package "phylocurve" (Goolsby 2016). Under conditions when it is applicable (that is, when $p<n$), we also apply the simple ML using the *evo.model* function. In addition, we apply our approximation of the LOOCV to the LASSO and ridge quad. (var) penalties; the approximation of ridge quad. (null) is expected to perform similarly as that of ridge quad. (var). We found that, while the approximations were very good for large n compared to p, they were not when p approached or exceeded n. Thus, we apply the approximations only for $p<n$; they then provide substantial computational advantage compared to the regular LOOCV (Equation 8). Finally, for each of these analyses, we use both the likelihood and the restricted likelihood formulations. In total, we thus compare the performance of a series of 8 approaches with the likelihood (and the same 8 with the restricted likelihood) when $p<n$, and a series of 5 approaches with the likelihood (and the same 5 with the restricted likelihood) when $p>n$. For each simulated dataset and each approach we obtain estimates of the parameters of the phenotypic model. In our optimization, we constrained r from the EB model to be negative, α from the OU process to be positive, and Pagel's λ to be between 0 and 1. We also obtain estimates of the traits evolutionary covariance matrix $\widehat{\boldsymbol{R}}$. For the ML and PCL approaches, the *evo.model* function directly provides estimates of the parameters of the phenotypic models. These can be used to compute $\boldsymbol{C}$, which is then plugged into Equation 2 to obtain $\widehat{\boldsymbol{R}}$, which is not returned by default by the *evo.model* function. We compare $\widehat{\boldsymbol{R}}$ to the true (i.e. simulated) matrix $\boldsymbol{R}$ using the quadratic loss function (e.g., van Wieringen and Peeters 2016):

$$\Delta_Q\left(\widehat{\mathbf{R}}, \mathbf{R}\right) = \left\|\widehat{\mathbf{R}}^{-1}\mathbf{R} - I_p\right\|^2 \quad (11)$$

This loss measure is zero when the matrices are perfectly similar. When p is equal or larger than n, non-regularized estimates of the $\boldsymbol{R}$ matrix (e.g., Eq. 2) are likely to be singular so we compute the estimate of $\widehat{\boldsymbol{R}}^{-1}$ in the quadratic loss function using the pseudoinverse routine implemented in the R package "corpcor" (Schäfer et al. 2013). We also used instead the nearest positive definite matrix using the *nearPD* function in the "Matrix" R package (Bates and Maechler 2017) following the approach undertaken by Denton and Adams (2015) as well as Goolsby (2016); this resulted in even larger quadratic loss for the non-regularized estimates (results not shown). Finally, we measure our ability to recover the proper model by estimating the proportion of time the generating model was selected by the GIC criterion. This was done for all approaches except the PCL approach for which we cannot compute the GIC. For data simulated under BM, EB and OU, these 3 models are compared. For data simulated under Pagel's λ model, we compared Pagel's λ and BM (the typical test performed to assess phylogenetic signal).

An empirical example: the evolution of brain shape in New World monkeys

We analyzed the fit of three different evolutionary models (BM, OU and EB) to a comparative dataset of brain shape for 48 New World monkey species in order to better understand the diversification process of brain morphology in the clade. We used brain shape data from Aristide et al. (2016) which consists of a total of 26 anatomical landmark and 373 semi-landmark 3D coordinates (totaling 1197 variables) on digitized endocasts obtained by X-ray computed tomography and micro-computed tomography. Because we are considering 3D geometric morphometric landmark coordinates obtained after Generalized Procrustes

Analysis (see details in Aristide et al. 2016), model fit must be performed with a rotation-invariant procedure (Rohlf 1999; Adams and Collyer 2018). We applied the *fit_t_pl* function to the data, using the rotation-invariant "ridge quadratic null" penalty and accounting for intra-specific variation and measurement errors in the model fit by setting the option *SE* to TRUE. The relative fit of the three models was assessed using the GIC criterion with the *GIC* function. We considered phylogenetic uncertainty by performing the analyses over a sample of 100 fossil-calibrated trees from the Bayesian posterior distribution of the phylogenetic analysis described in Aristide et al. (2015) (pruned to the 48 species considered here).

Additionally, we studied the patterns of evolutionary integration and modularity (i.e., the correlated/independent change of traits through evolution; Klingenberg and Marugán-Lobón 2013) in brain shape by analyzing the estimated evolutionary variance-covariance matrix derived from the model fitting analysis. We performed a principal component analysis of this matrix using the *phyl.pca_pl* function, which allowed us to extract evolutionary independent axes of correlated change among variables that broadly represent evolutionary integration patterns across the structure (Klingenberg 2008). Finally, with the *ancestral* function, we utilized the parameters derived from the evolutionary model that best explained our data to obtain brain shape reconstructions through time that illustrate the hypothetical evolutionary changes leading to the two most different species in the dataset. To generate 3D model visualizations of the reconstructed shapes, a 3D surface model of the sample's mean shape was warped to each target ancestral shape by aligning the reconstructed and mean shape landmark data and using a TPS interpolation (Wiley et al. 2005), as implemented in the "Morpho" package for R (Schlager 2017). Color maps depicting shape changes were generated by computing the radial distance between each pair of corresponding vertices in two given 3D surfaces.

RESULTS

As expected, for all the methods we considered, we found a better performance of the restricted likelihood than of the unrestricted one. All the results shown in the main text are based on the restricted likelihood, while results based on the (unrestricted) likelihood are presented in the Supplementary Material.

Parameter estimation

With the restricted likelihood, parameters of the EB, OU and Pagel's λ models are estimated accurately by all methods for low dimensional traits (p=2-5) (Fig. 1 and S1, S3 and S5). Estimators based on the unrestricted likelihood are on the other hand systematically biased, regardless of the method used (Fig. S2, S4, and S6), as is expected from unrestricted estimators (Harville 1977): the parameter of the OU process (α) is overestimated, while the parameters of the EB (r) and Pagel (λ) models are underestimated. This bias is higher for the pairwise composite likelihood, and even more so for the maximum likelihood, than for the penalized likelihood approaches.

With both the restricted and unrestricted likelihoods, as the trait dimension p increases, parameters are estimated with increased precision when using the various non-approximated penalized approaches or the PCL approach, while the classical maximum likelihood estimate becomes completely inaccurate (Fig. S1-S6), and it cannot be computed at all when $p>n$.

The various non-approximated LOOCV approaches perform similarly, with no significant difference between the chosen penalties (Fig. 1 and S1-S6). The approximated LOOCV approaches perform as well as the non-approximated ones for small p values (Fig. 1); they lose precision as p increases (in particular in the case of the LASSO and when p approaches the number of species, Fig. S1-S6), but are still as efficient as the ML or pairwise

composite ML when n is still larger than p (i.e., the situations where the use of the LOOCV is generally a major computational bottleneck).

<FIGURE_1>

Estimation of the traits variance-covariance matrix R

The error in the estimate of $\boldsymbol{R}$, measured with the Q-loss function, is always smaller with the penalized likelihood approaches than with other approaches (Fig. 2 and Fig. S7-14). The Q-loss is also slightly smaller when using the restricted *versus* unrestricted likelihood (Fig. S7-14). Even with relatively small dimensions (p=2 or p=5), the penalized estimates outperform the PCL and ML estimates (Fig. S7-14). The difference in performance between the PL estimate and the other estimates increases as the dimension p increases (Fig. S7-14).

Among the penalized estimates obtained with the non-approximated LOOCV, the quadratic ridge penalty with a null target is the most efficient; next, the quadratic ridge with the diagonal target matrix composed of the traits' variances and LASSO penalties perform similarly. Finally, the archetypal ridge penalty performs slightly less well, consistent with observations from previous studies (Ledoit and Wolf 2012; van Wieringen and Peeters 2016). The penalized estimates obtained with the approximated LOOCV are not as good as those obtained with the non-approximated ones, but they still always outperform the PCL and ML approaches (Fig. 2 and Fig. S7-14).

<FIGURE_2>

Model comparison

Consistent with the other results, model comparison performs better with the restricted than with the unrestricted likelihood (Fig. S15 and S16); in particular, there is a high false-discovery rate for the Ornstein-Uhlenbeck model with both the ML and the PL unrestricted approaches, which could be linked to the systematic overestimation of α obtained in this case (Fig. S4).

For low trait dimensions (p=2-10), the ability to select the generating model is roughly similar with the penalized and maximum likelihood approaches (Fig. 3 and Figs. S15-17). However when p>10 the penalized likelihood approaches clearly outperform the maximum likelihood (Fig. 3 and Figs. S15-17): the power to detect the true model increases with increasing trait dimensions for the penalized likelihoods while maximum likelihood fails to detect the generating model when p is close to n (*i.e.,* p=31; Fig. 3 and Fig. S15-S17). The different types of penalizations perform similarly well (Fig. 3 and Fig. S15-17). When models are fitted using the approximated LOOCV, model misspecification increases when p is close to n, but not as badly as when models are fitted using maximum likelihood (Figs. S17).

<FIGURE_3>

Evolution of brain morphology in New World monkeys

Applying our penalized likelihood framework to the New World monkey brain shape dataset (Fig. 4a, b), we found substantial support for an Early-burst model and these conclusions were robust to the uncertainty in tree topology and dating (Table 2). The BM model was not supported in any of the 100 trees; the OU model was best supported for only 3 trees, and in these cases the difference in GIC with the EB model was very small (Table 2). The average parameter estimate for the Early-burst model ("r", Table 2) describes an evolutionary rate decay with a half-life – the time it takes for the rate to decay half of its initial value – of ~19.4

Ma for an average root height of ~25.5 Ma, indicative of a mild early-burst pattern of brain shape evolution. Considering the first two principal components of the trait evolutionary variance-covariance matrix obtained under the best fitting model (EB), we were able to extract the major patterns of evolutionary integration in brain shape across the clade. Variation represented by the first PC axis (PC1) mainly reveals a pattern of correlated evolution among the stem, the base, and the parietal regions of the brain (Fig. 4c). Noticeably, variation along PC2 broadly indicates that the temporal, occipital and frontal regions evolve in a concerted fashion but independently from those aspects represented by PC1 (Fig. 4c).

Finally, we generated a hypothesis about the ancestral New World monkey brain shape, and the potential shape evolutionary trajectories leading to two species at opposite extremes in the brain morphospace (*Alouatta macconelli* and *Saimiri boliviensis*; Fig. 4b, d). The ancestral brain shape presents a relatively reduced occipital lobe compared to *S. boliviensis*, but a relatively expanded neocortex region and a more flexed base compared to *A. macconnelli* (Fig. 4d). Moreover, in agreement with the recovered EB model, the reconstructions reveal that for both considered lineages most changes occur at the intergeneric level rather than within each genus (Fig. 4b, d). This suggests that the brain shape of extant species was attained relatively early during the diversification process of the clade.

<FIGURE_4>

<TABLE_2>

DISCUSSION

Phenotypic evolution is multivariate by nature given that traits covary due to pleiotropic effects, genetic linkages, developmental and functional constraints or because of correlated

selection on multiple traits (Felsenstein 1988; Armbruster and Schwaegerle 1996; Walsh 2007; Walsh and Blows 2009; Armbruster et al. 2014). In this article we developed a penalized likelihood approach for phylogenetic multivariate comparative methods. We demonstrated through simulations that this approach performs well in estimating parameters of trait evolution models, and performs better than maximum likelihood (when $p<n$) and current alternatives (when $p>n$) in estimating model parameters. Even a rough penalization using approximations of the cross-validated log-likelihood to select the *regularization*/*tuning* parameter outperforms standard estimates when $p<n$. The approach allows the use of generalized information criteria (GIC) for model selection, therefore avoiding computationally intensive simulation-based model comparison techniques.

Parameter estimation

Our penalized likelihood approach provides an efficient way to estimate parameters of various multivariate phenotypic evolution models as well as to estimate corresponding evolutionary covariance matrices in high-dimensional datasets (with p large compared to n). Maximum likelihood estimates are highly biased when p approaches n, and they can no longer be computed when p exceeds n. Penalized estimates generally perform better than maximum likelihood for estimating covariance matrices or their inverses (Ledoit and Wolf 2004, 2012, 2015; van Wieringen 2017), and we have illustrated here that it is also the case for the evolutionary covariance in phylogenetic comparative methods. To date, the only alternative comparative method is the pairwise composite likelihood (PCL) approach proposed by Goolsby (2016). This method performs as well as the penalized likelihood for estimating the parameters of various classical phenotypic evolution models; however, the penalized likelihood is much more accurate for estimating the evolutionary covariance matrix. Thus, the penalized likelihood approach developed here is particularly useful for studying the

correlated evolution of traits in high-dimensional settings. It should also be useful for extending comparative methods, most of which rely on accurate estimates of the traits evolutionary variance-covariance matrix, to such high-dimensional situations.

Model selection

The penalized likelihood approach also provides an efficient way to compare the relative fit of various multivariate phenotypic evolution models to high-dimensional trait datasets using information criteria. Model selection based on maximum-likelihood is inaccurate when p approaches n, and it is non applicable when p exceeds n. Previous attempts to deal with such situations have used parametric bootstrap to compare pairs of models (Goolsby 2016). Parametric bootstrap approaches have appropriate statistical performances (e.g., type I and type II errors; Good 2005) but are computationally intensive and may be impractical when comparing several alternative models. In contrast, within the penalized likelihood framework, multiple model comparison as well as model averaging can be easily and efficiently performed using the Generalized Information Criterion (GIC). Information theoretic approaches (e.g., AIC, BIC, GIC) are commonly used in comparative studies for model selection and/or model averaging and we showed in our simulations that the GIC combined with penalized likelihoods clearly outcompetes information criteria based on ML when p approaches n and still performs well when p exceeds n.

Performing model selection using GIC requires computing a bias correction term for the log-likelihood. We have shown here how to derive this term analytically. However, this might be hard to achieve for more complex models (e.g. models with distinct trait covariation structures and/or evolutionary regimes in distinct subclades). In addition, we have derived a first-order correction term, and higher-order corrections could still improve the statistical performance of the model selection scheme (Konishi and Kitagawa 2003). Future

developments could use numerical approximations to compute higher level bias correction terms even under complex models (e.g., Ueki and Fueda 2010). Alternatively, parametric bootstrap approaches such as the Bootstrap Information Criterion (or EIC, Efron Information Criterion; Ishiguro et al. 1997; Konishi and Kitagawa 2008) could be used in combination with our penalized likelihood framework to do multiple model comparison. This would be more computationally demanding but would automatically achieve higher-order bias correction, and efficient bootstrap strategies have already been developed (Ishiguro et al. 1997; Konishi and Kitagawa 2008; Kitagawa and Konishi 2010).

Properties of the various penalizations

There are several regularization/shrinkage estimators for covariance matrices (reviewed in Engel et al. 2017), none of which performs best in all situations; the choice must be guided by practical use (Table 1). The first choice concerns the type of penalization (linear versus quadratic). The archetypal ridge estimator shrinks the eigenvalues of the covariance matrix linearly, and is thus expected to perform well when the eigenvalues are similar in magnitude (Ledoit and Wolf 2004). The quadratic ridge estimators shrink the eigenvalues non-linearly, and are thus expected to perform better when the difference in magnitude between the leading and other eigenvalues is high (i.e. when the eigenvalues are dispersed, Ledoit and Wolf 2012). As the dispersion of eigenvalues partly reflects the numbers of clusters of integrated traits, their respective numbers of traits, and the average correlation between the traits within clusters (diffuse correlation when eigenvalues are of same magnitude and strong correlation within few clusters when eigenvalues are dispersed), we expect linear estimators to perform well when traits tend to evolve independently and their covariance matrix is not tightly integrated; the LASSO is also expected to perform well when many pairs of traits are not correlated. Quadratic estimators, on the other hand, should

perform better when traits are highly correlated within few clusters, as is often observed in empirical data (Wagner 1984). Unsurprisingly, in our simulations where eigenvalues are sampled from an exponential distribution, the quadratic ridge estimators clearly outperform the other penalties. In practice, when studying the evolution of traits that are thought to coevolve, quadratic estimators are probably more reliable; however, they are more computationally demanding. There is thus a tradeoff between reliability and computational load (see Table 1). In very high dimensions where extreme values of the regularization parameter may be necessary, archetypal and quadratic ridge estimators are expected to perform equally as they both tend towards the target matrix (van Wieringen and Peeters 2016).

The second choice to make when choosing a penalized likelihood estimator concerns the target matrix. In our study we analyzed the performance of the Quadratic ridge estimator when using either the null or the diagonal matrix of traits' variances. Here we obtained better $\boldsymbol{R}$ estimates in terms of quadratic loss when using the null target matrix, consistently with previous studies (van Wieringen and Peeters 2016); however, these results could vary with the choice of the metric for measuring statistical loss. In addition, each target matrix has a different variance-bias tradeoff, and so other targets may be a better choice depending on the goal of the study (Schäfer and Strimmer 2005; Lancewicki and Aladjem 2014).

Maybe a more important thing to consider is whether a rotation-invariant approach is needed. Approaches affecting the eigenvectors of the sample covariance matrix (e.g. Equation 2) are not rotation-invariant, while those that leave the eigenvectors unchanged are (Ledoit and Wolf 2004; Warton 2008; Ledoit and Wolf 2012, 2015). As discussed above, rotation-invariant approaches should be used when there is no natural orientation of the data, i.e. when an arbitrary choice of orientation is made such as in geomorphometrics (Rohlf 1999). In most cases however, there is a natural orientation of the data. In this case it is not necessary to

apply rotation-invariant methods, and this can even yield better estimates (Schäfer and Strimmer 2005; Friedman et al. 2008; Warton 2008; Lancewicki and Aladjem 2014; van Wieringen and Peeters 2016); indeed, shrinking the eigenvectors can improve the estimation of covariances (Pourahmadi 2011), which can for example be particularly interesting for ordination analyses (Engel et al. 2017).

Although computationally intensive, the LASSO penalty has the advantage over alternative penalizations that it directly identifies independencies between traits. Indeed, in the estimates of $\boldsymbol{R}$ or $\boldsymbol{R}^{-1}$, the least significant entries are forced to zero. Such zeros in the covariance or its inverse correspond, respectively, to marginal independency between traits (pairs of traits are not correlated even when considering potential indirect factors) or conditional independency (a weaker form of independence where the pairs of traits are not directly related only when considering one or several other putative traits) (Dempster 1972; Friedman et al. 2008; Bien and Tibshirani 2011). Identifying such independencies is of major interest, for instance, in the study of patterns of phenotypic integration and modularity (Magwene 2001, 2008; Goswami and Polly 2010). We considered here the graphical-LASSO penalization where the target of the optimization is the precision matrix $\boldsymbol{R}^{-1}$ rather than the covariance $\boldsymbol{R}$ (Friedman et al. 2008); an alternative algorithm will need to be used to target the covariance matrix (Bien and Tibshirani 2011) in order to analyze marginal independencies.

Computational considerations

A limit to the applicability of the penalized likelihood framework can be the computational cost associated with the estimation of the regularized covariance matrix and its inverse when p is large. The least to most computationally intensive of the estimators we have considered here are the archetypal ridge, the quadratic ridge with null target, the quadratic

ridge with diagonal matrix of variances, and finally the LASSO estimator. In our simulations, the computational time needed to cross-validate the log-likelihood with the LASSO penalty was order of magnitude higher than with ridge penalties. These computational considerations can guide the choice of the penalization strategy if there is no specific reason to choose one estimator *versus* another; all the penalized estimates improve upon the standard estimates. There are several strategies to gain in computational efficiency. We considered approximations of the LOOCV score and have shown that even this rough regularization generally outcompetes maximum-likelihood estimation, but only when $p<n$. Parallel computing and/or alternative cross validation strategies such as k-fold CV (Hastie et al. 2009) could potentially be used. Approaches avoiding cross-validation for estimating the regularization parameter could also be developed. For instance, the regularization parameter could be chosen such that the deviation in likelihood from the maximum is insignificant (Meyer 2011) or by using *a priori* criteria depending on n, p, and quantities related to the model such as the number of free parameters (e.g., Foygel and Drton 2010).

New World monkeys brain shape evolution

We demonstrated the applicability of our framework by analyzing a highly-dimensional empirical morphometric dataset describing brain shape variation in New World monkeys. Our analyses recovered as the best fit an Early-Burst model of evolution, a model which is generally considered to better represent the expectations of adaptive radiations than the BM or OU models (Harmon et al. 2010). Aristide et al. (2016) had already found evidence in favor of a model that supports an adaptive radiation scenario in the brain shape data, but using only the first PC axes. Given the potential issues arising from comparative analyses of PC variables (Uyeda et al. 2015), this result was questionable. The fully multivariate approach used here on the complete morphometric dataset thus provides a stronger support to this

result. Brain shape evolution in the clade shows a pattern of early diversification consistent with the hypothesis of an adaptive radiation (Aristide et al. 2016). This result is noteworthy because an early-burst pattern is generally hard to detect even on univariate data (Slater and Pennell 2014) and highlights both the singularity of the New World monkey radiation (Aristide et al. 2015) and the strength of the multivariate framework we developed.

Additionally, we were able to extract the main patterns of evolutionary integration in brain shape, which indicated that most correlated changes are concentrated along well defined regions of the brain (e.g. frontal, occipital, parietal, etc.) and that some of these regions vary independently from the others. The functional and evolutionary implications of this remains to be assessed but ours constitutes the first attempt to analyze evolutionary integration patterns by using models of evolution in an extremely high-dimensional geometric morphometrics dataset. As stated previously, further developments that allow extracting these patterns from evolutionary covariance matrices of landmark data will help to provide a more detailed picture of brain evolution in the clade.

Finally, using the model parameters estimated for the fully multivariate dataset, we generated ancestral reconstructions for brain shape that can serve as hypotheses regarding various aspects of the evolution of the clade. For example, as brain shape is associated to social group size and other ecological traits in New World monkeys (e.g. Aristide et al. 2016), a careful consideration of reconstructed ancestral shapes may help to understand the evolution of traits that are in general not directly preserved in the fossil record.

Future directions

We see several directions for future developments. The models we have considered here are simple models with a common $\boldsymbol{R}$ matrix shared across the entire clade and a common phylogenetic covariance matrix $\boldsymbol{C}$ shared across traits. More complex models with multiple

regime/clade specific covariance matrices could be developed. For such models, joint optimization of multiple penalties (each associated to a given $\boldsymbol{R}$) could be envisioned (e.g., Guo et al. 2011; Danaher et al. 2014). In even more complex multivariate models described by parameter rich stochastic processes with deterministic parts that are themselves non-independent (e.g., in multivariate OU processes, if the pull matrix $\mathbf{A}$ is not diagonal, Bartoszek et al. 2012; Reitan et al. 2012; Clavel et al. 2015), penalties associated to parameters other than $\boldsymbol{R}$ (e.g. $\mathbf{A}$) could be incorporated to reduce the variance in parameter estimates. Penalized approaches for such models have been shown to outperform alternative approaches in terms of model selection and parameter estimation in a non-phylogenetic context (Wang et al. 2016). In a phylogenetic context, the models are already available in a maximum likelihood framework, but they are not applicable with high-dimensional data and are generally estimated with large uncertainty (Bartoszek et al. 2012; Clavel et al. 2015).

Second, the regularized estimates provided by the penalized likelihood framework in high-dimensional settings can be used to improve and extend the multivariate comparative toolbox to high-dimensional datasets, beyond fitting multivariate trait evolution models. Regularized versions of traditional (non-phylogenetic) multivariate statistics (such as the Wilks lambda or Pillai trace used in multivariate regressions and MANOVA) or multivariate classification methods (e.g., discriminant analysis, CCA, etc.) have already been shown to perform well and with increased power when the sample size is small compared to the number of variables (Vinod 1976; Friedman 1989; Warton 2008; Ullah and Jones 2015; Engel et al. 2015). Regularized estimates of evolutionary covariance matrices should likewise allow developing adequate phylogenetic equivalents of these multivariate statistics in high dimension. Similarly, penalized likelihood approaches for high-dimensional (non-phylogenetic) data have proven to be extremely efficient in estimating missing data (Allen and Tibshirani 2010); missing data is also a common feature of high-dimensional phenotypic

datasets (Clavel et al. 2014, 2015; Goolsby et al. 2017), and phylogenetic equivalents of these regularized missing values imputation techniques should solve this challenging task.

Finally, the regularization/shrinkage approaches we describe in this paper can be efficiently extended to Bayesian inferences (e.g., Caetano and Harmon 2017) where shrinkage priors are routinely used as penalties analogues – penalized likelihood estimates are often interpreted as maximum *a posteriori* (MAP) estimates (Green 1990) – to reduce statistical risks or to obtain well-conditioned symmetric positive definite matrices (Daniels and Kass 2001; Lu and Ades 2009; Wang 2012; Khondker et al. 2013). Some of the computational tricks we developed here for computing the LOOCV should be useful for developing efficient proposals and sampling strategies in Bayesian MCMC approaches in high-dimensional parameter space.

CONCLUSION

Phylogenetic comparative methods for high-dimensional datasets are challenging, yet sorely needed. Regularization techniques are a powerful avenue to address this challenge. By providing the tools for properly estimating the evolutionary covariance matrix using penalized likelihood, we open the door to the extension of current multivariate phylogenetic comparative methods to high-dimensional datasets. This should allow addressing long-standing questions related to the correlated evolution of traits.

SUPPLEMENTARY MATERIAL

Our Supplementary Material as well as example codes for running analyses in the paper are available from the Dryad Digital Repository at http://datadryad.org, doi:10.5061/dryad.rf7317t.

FUNDING

This work was supported by European Research Council grant ERC 616419-PANDA (to HM).

ACKNOWLEDGEMENTS

The authors wish to thank Ivan Vujačić and Wessel N. van Wieringen for precisions on the estimation of the regularization parameter in their papers. We thank Eric W. Goolsby for assistance with the R package "phylocurve" and Mátyás A. Sustik for supplying source codes for the graphical LASSO. Finally, we thank Renske Gudde, Eric Lewitus, Odile Maliet, Marc Manceau, Olivier Missa, Benoît Perez, Guilhem Sommeria-Klein, Thomas Near, Luke Harmon, Michael Landis and two anonymous reviewers for helpful comments on the manuscript.

REFERENCES

Abbruzzo A., Vujačić I., Wit E., Mineo A.M. 2014. Generalized information criterion for model selection in penalized graphical models. arXiv.:1–29.

Adams D.C. 2014a. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. Syst. Biol. 63:166–177.

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

Adams D.C. 2014c. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. Syst. Biol. 63:685–697.

Adams D.C., Collyer M.L. 2018. Multivariate Phylogenetic comparative methods: evaluations, comparisons, and recommendations. Syst. Biol. 67:14–31.

Adams D.C., Felice R.N. 2014. Assessing trait covariation and morphological integration on

phylogenies using evolutionary covariance matrices. Plos One. 9:1–8.

Akaike H. 1974. A new look at the statistical model identification. Autom. Control IEEE Trans. On. 19:716–723.

Allen G.I., Tibshirani R. 2010. Transposable regularized covariance models with an application to missing data imputation. Ann Appl Stat.:764–790.

Anderson T.W., Olkin I. 1985. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. Linear Algebra Its Appl. 70:147–171.

Aristide L., dos Reis S.F., Machado A.C., Lima I., Lopes R.T., Perez S.I. 2016. Brain shape convergence in the adaptive radiation of New World monkeys. Proc. Natl. Acad. Sci. 113:2158–2163.

Aristide L., Rosenberger A.L., Tejedor M.F., Perez S.I. 2015. Modeling lineage and phenotypic diversification in the New World monkey (Platyrrhini, Primates) radiation. Mol. Phylogenet. Evol. 82, Part B:375–385.

Armbruster W.S., Pélabon C., Bolstad G.H., Hansen T.F. 2014. Integrated phenotypes: understanding trait covariation in plants and animals. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369.

Armbruster W.S., Schwaegerle K.E. 1996. Causes of covariation of phenotypic traits among populations. J. Evol. Biol. 9:261–276.

Bartoszek K., Pienaar J., Mostad P., Andersson S., Hansen T.F. 2012. A phylogenetic comparative method for studying multivariate adaptation. J. Theor. Biol. 314:204–215.

Bastide P., Ané C., Robin S., Mariadassou M. 2018. Inference of Adaptive Shifts for Multivariate Correlated Traits. Syst. Biol.:syy005-syy005.

Bates D.M., Maechler M. 2017. Matrix: sparse and dense matrix classes and methods. R package version 1.2-14.

Bien J., Tibshirani R.J. 2011. Sparse estimation of a covariance matrix. Biometrika. 98:807–

820.

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

Bookstein F.L. 2012. Random walk as a null model for high-dimensional morphometrics of fossil series: geometrical considerations. Paleobiology. 39:52–74.

Burnham K.P., Anderson D.R. 2002. Model selection and multi-model inference: a practical information-theoric approach. New York: Springer-Verlag.

Butler M.A., King A.A. 2009. Multivariate comparative analysis using OUCH. Integr. Comp. Biol. e24.

Caetano D.S., Harmon L.J. 2017. ratematrix: an R package for studying evolutionary integration among several traits on phylogenetic trees. Methods Ecol. Evol.

Clavel J., Escarguel G., Merceron G. 2015. mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311–1319.

Clavel J., Merceron G., Escarguel G. 2014. Missing Data Estimation in Morphometrics: How Much is Too Much? Syst. Biol. 63:203–218.

Cross R. 2017. The inside story of 20,000 vertebrates. Science. 357:742–743.

Cunningham C.W., Omland K.E., Oakley T.H. 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. 13:361–366.

Danaher P., Wang P., Witten D.M. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. J. R. Stat. Soc. Ser. B Stat. Methodol. 76:373–397.

Daniels M.J., Kass R.E. 2001. Shrinkage estimators for covariance matrices. Biometrics. 57:1173–1184.

Dempster A.P. 1972. Covariance Selection. Biometrics. 28:157–175.

Denton J.J., Adams D.C. 2015. A new phylogenetic test for comparing multiple high-dimensional evolutionary rates suggests interplay of evolutionary rates and modularity in

lanternfishes (Myctophiformes; Myctophidae). Evolution. 69:2425–2440.

Dunn C.W., Luo X., Wu Z. 2013. Phylogenetic Analysis of Gene Expression. Integr. Comp. Biol. 53:847–856.

Dwyer P. 1967. Some applications of matrix derivatives in multivariate analysis. J. Am. Stat. Assoc. 62:607–625.

Engel J., Blanchet L., Bloemen B., van den Heuvel L.P., Engelke U.H.F., Wevers R.A., Buydens L.M.C. 2015. Regularized MANOVA (rMANOVA) in untargeted metabolomics. Anal. Chim. Acta. 899:1–12.

Engel J., Buydens L., Blanchet L. 2017. An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics. J. Chemom. 31:e2880.

Fan J., Li R. 2001. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. J. Am. Stat. Assoc. 96:1348–1360.

Felsenstein. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Evol. Syst. 19:445–471.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, Massachusetts, USA: Sinauer Associates.

Foygel R., Drton M. 2010. Extended Bayesian Information Criteria for Gaussian graphical models. Adv. Neural Inf. Process. Syst. 23.:604–612.

Freckleton R.P. 2012. Fast likelihood calculations for comparative analyses. Methods Ecol. Evol. 3:940–947.

Friedman J., Hastie T., Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 9:432–441.

Friedman J.H. 1989. Regularized Discriminant Analysis. J. Am. Stat. Assoc. 84:165–175.

Good P. I. 2005. Permutation, Parametric, and Bootstrap tests of hypotheses. New York: Springer-Verlag.

Goolsby E.W. 2016. Likelihood-Based Parameter Estimation for High-Dimensional Phylogenetic Comparative Models: Overcoming the Limitations of "Distance-Based" Methods. Syst. Biol. 65:852–870.

Goolsby E.W., Bruggemann J., Ané C. 2017. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. Methods Ecol. Evol. 8:22–27.

Goswami A., Polly P.D. 2010. Methods for studying morphological integration and modularity. Quantitative Methods in Paleobiology. John Alroy & Gene Hunt. p. 213–243.

Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. B. 326:119–157.

Green P.J. 1990. On the use of the EM for penalized likelihood estimation. J. R. Stat. Soc. Ser. B Methodol. 52:443–452.

Guo J., Levina E., Michailidis G., Zhu J. 2011. Joint estimation of multiple graphical models. Biometrika. 98:1–15.

Gurka M.J. 2006. Selecting the best linear mixed model under REML. Am. Stat. 60:19–26.

Hansen T.F., Bartoszek K. 2012. Interpreting the evolutionary regression : the interplay between observational and biological errors in phylogenetic comparative studies. Syst. Biol. 61:413–425.

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

Harville D.A. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. J. Am. Stat. Assoc. 72:320–338.

Hastie T., Tibshirani R., Friedman J.H. 2009. The elements of statistical learning. Berlin: Springer.

Henderson H.V., Searle S.R. 1979. Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics. Can. J. Stat. Rev. Can. Stat. 7:65–81.

Ho L.S.T., Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol. 63:397–408.

Hoerl A.E., Kennard R.W. 1970a. Ridge regression: applications to nonorthogonal problems. Technometrics. 12:69–82.

Hoerl A.E., Kennard R.W. 1970b. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 12:55–67.

Hoffbeck J.P., Landgrebe D.A. 1996. Covariance matrix estimation and classification with limited training data. IEEE Trans. Pattern Anal. Mach. Intell. 18:763–767.

Housworth E.A., Martins E.P., Lynch M. 2004. The phylogenetic mixed model. Am. Nat. 163:84–96.

Huang J.Z., Liu N., Pourahmadi M., Liu L. 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika. 93:85–98.

Ishiguro M., Sakamoto Y., Kitagawa G. 1997. Bootstrapping log likelihood and EIC, an extension of AIC. Ann. Inst. Stat. Math. 49:411–434.

Ives A.R., Midford P.E., Garland T.J. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol. 56:252–270.

James W., Stein C. 1961. Estimation with Quadratic Loss. :361–379.

Khabbazian M., Kriebel R., Rohe K., Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol. 7:811–824.

Khondker Z.S., Zhu H., Chu H., Lin W., Ibrahim J.G. 2013. The Bayesian Covariance Lasso. Stat. Interface. 6:243–259.

Kim J., Sanderson M.J. 2008. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. Syst. Biol. 57:665–674.

Kitagawa G., Konishi S. 2010. Bias and variance reduction techniques for bootstrap information criteria. Ann. Inst. Stat. Math. 62:209–234.

Klingenberg C.P. 2008. Morphological integration and developmental modularity. Annu. Rev. Ecol. Evol. Syst. 39:115–132.

Klingenberg C.P., Marugán-Lobón J. 2013. Evolutionary Covariation in Geometric Morphometric Data: Analyzing Integration, Modularity, and Allometry in a Phylogenetic Context. Syst. Biol. 62:591–610.

Konishi S., Kitagawa G. 1996. Generalised Information Criteria in Model Selection. Biometrika. 83:875–890.

Konishi S., Kitagawa G. 2003. Asymptotic theory for information criteria in model selection - functional approach. J. Stat. Plan. Inference. 114:45–61.

Konishi S., Kitagawa G. 2008. Information Criteria and Statistical Modeling. Springer-Verlag New York.

Kratsch C., McHardy A.C. 2014. RidgeRace: ridge regression for continuous ancestral character estimation on phylogenetic trees. Bioinformatics. 30:i527–i533.

Lancewicki T., Aladjem M. 2014. Multi-Target Shrinkage Estimation for Covariance Matrices. IEEE Trans. Signal Process. 62:6380–6390.

Ledoit O., Wolf M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88:365–411.

Ledoit O., Wolf M. 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann. Stat. 40:1024–1060.

Ledoit O., Wolf M. 2015. Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. J. Multivar. Anal. 139:360–384.

Levina E., Rothman A., Zhu J. 2008. Sparse estimation of large covariance matrices via a nested lasso penalty. Ann. Appl. Stat. 2:245–263.

Lian H. 2011. Shrinkage tuning parameter selection in precision matrices estimation. J. Stat. Plan. Inference. 141:2839–2848.

Lu G., Ades A.E. 2009. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics. 10:792–805.

Magnus J.R., Neudecker H. 1985. Matrix differential calculus with applications to simple, hadamard, and kronecker products. J. Math. Psychol. 29:474–492.

Magnus J.R., Neudecker H. 2007. Matrix differential calculus with applications in statistics and econometrics. Chichester: John Wiley & Sons.

Magwene P.M. 2001. New tools for studying integration and modularity. Evolution. 55:1734–1745.

Magwene P.M. 2008. Using Correlation Proximity Graphs to Study Phenotypic Integration. Evol. Biol. 35:191–198.

Manceau M., Lambert A., Morlon H. 2017. A Unifying Comparative Phylogenetic Framework Including Traits Coevolving Across Interacting Lineages. Syst. Biol. 66:551–568.

Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.

McCulloch C.E. 1982. Symmetric Matrix Derivatives with Applications. J. Am. Stat. Assoc. 77:679–682.

Meyer K. 2011. Performance of penalized maximum likelihood in estimation of genetic

covariances matrices. Genet. Sel. Evol. 43:39.

Mezey J.G., Houle D. 2005. The dimensionality of genetic variation for wing shape in Drosophila melanogaster. Evolution. 59:1027–1038.

Moneta G.B. 1991. Implicit construction of McCulloch's G matrix for the numerical evaluation of Fisher information matrixes. Comput. Stat. Data Anal. 11:333–344.

Morlon H., Lewitus E., Condamine F.L., Manceau M., Clavel J., Drury J. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7:589–597.

O'Meara B.C. 2012. Evolutionary inferences from phylogenies: a review of methods. Annu. Rev. Ecol. Evol. Syst. 43:267–285.

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

Pagel M.D. 1999. Inferring the historical patterns of biological evolution. Nature. 401:877–884.

Pourahmadi M. 2011. Covariance estimation: the GLM and regularization perspectives. Stat. Sci. 26:369–387.

R Development Core Team. 2016. R: A language and environment for statistical computing. Vienna, Austria. URL http://www.R-project.org .

Reitan T., Schweder T., Henderiks J. 2012. Phenotypic evolution studied by layered stochastic differential equations. Ann. Appl. Stat. 6:1531–1551.

Revell liam J. 2009. Size-correction and principal components for interspecific comparative studies. Evolution. 63:3258–3268.

Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Revell L.J., Collar D.C. 2009. Phylogenetic analysis of the evolutionary correlation using

likelihood. Evolution. 63:1090–1100.

Revell L.J., Harmon L.J. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res. 10:311–331.

Revell L.J., Harrison A.S. 2008. PCCA: a program for phylogenetic canonical correlation analysis. Bioinformatics. 24:1018–1020.

Roberts D.R., Bahn V., Ciuti S., Boyce M.S., Elith J., Guillera-Arroita G., Hauenstein S., Lahoz-Monfort J.J., Schröder B., Thuiller W., Warton D.I., Wintle B.A., Hartig F., Dormann C.F. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 40:913–929.

Rohlf F.J. 1999. Shape statistics: Procrustes superimpositions and tangent spaces. J. Classif. 16:197–223.

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

Rondeau V., Commenges D., Joly P. 2003. Maximum Penalized Likelihood Estimation in a Gamma-Frailty Model. Lifetime Data Anal. 9:139–153.

Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Schäfer J., Opgen-Rhein R., Zuber V., Ahdesmäki M., Silva P.D., Strimmer K. 2013. Corpcor: Efficient estimation of covariance and (partial) correlation. R package version 1.6.9.

Schäfer J., Strimmer K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4:1–29.

Schlager S. 2017b. Morpho and Rvcg -- Shape Analysis in {R}. Statistical Shape and Deformation Analysis. Guoyan Zheng, Shuo Li, Gabor Szekely. p. 217–256.

Silvestro D., Kostikova A., Litsios G., Pearman P.B., Salamin N. 2015. Measurement errors should always be incorporated in phylogenetic comparative analysis. Methods Ecol. Evol. 6:340–346.

Slater G.J., Pennell M.W. 2014. Robust Regression and Posterior Predictive Simulation Increase Power to Detect Early Bursts of Trait Evolution. Syst. Biol. 63:293–308.

Smith S.A., O'Meara B.C. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 28:2689–2690.

Stegle O., Lippert C., Mooij J.M., Lawrence N.D., Borgwardt K.M. 2011. Efficient inference in matrix-variate Gaussian models with iid observation noise. Adv. Neural Inf. Process. Syst. 24 NIPS 2011.:630–638.

Stone E.A. 2011. Why the phylogenetic regression appears robust to tree misspecification. Syst. Biol. 60:245–260.

Sustik M.A., Calderhead B. 2012. GLASSOFAST: An efficient GLASSO implementation. UTCS Technical Report TR-12-29 :1–3.

Theiler J. 2012. The incredible shrinking covariance estimator. *In*: Automatic Target Recognition XXII. International Society for Optics and Photonics, p. 83910P.

Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 58:267–288.

Tolkoff M.R., Alfaro M.E., Baele G., Lemey P., Suchard M.A. 2018. Phylogenetic Factor Analysis. Syst. Biol. 67: 384–399

Ueki M., Fueda K. 2010. Optimal tuning parameter estimation in maximum penalized likelihood method. Ann. Inst. Stat. Math. 62:413–438.

Ullah I., Jones B. 2015. Regularised Manova for High-Dimensional Data. Aust. N. Z. J. Stat. 57:377–389.

Uyeda J.C., Caetano D.S., Pennell M.W. 2015. Comparative Analysis of Principal

Components Can be Misleading. Syst. Biol. 64:677–689.

Varin C., Reid N., Firth D. 2011. An overview of composite likelihood methods. Stat. Sin. 21:5–42.

Vinciotti V., Augugliaro L., Abbruzzo A., Wit C.E. 2016. Model selection for factorial Gaussian graphical models with an application to dynamic regulatory networks. Stat. Appl. Genet. Mol. Biol. 15:193–212.

Vinod H.D. 1976. Canonical ridge and econometrics of joint production. J. Econom. 4:147–166.

Vujačić I., Abbruzzo A., Wit E. 2015. A computationally fast alternative to cross-validation in penalized Gaussian graphical models. J. Stat. Comput. Simul. 85:3628–3640.

Wagner G.P. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: Evidence for a nonrandom organization of quantitative character variation. J. Math. Biol. 21:77–95.

Walsh B. 2007. Escape from flatland. J. Evol. Biol. 20:36–38.

Walsh B., Blows M.W. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. Annu. Rev. Ecol. Evol. Syst. 40:41–59.

Wang H. 2012. Bayesian Graphical Lasso models and efficient posterior computation. Bayesian Anal. 7:867–886.

Wang Y., Tang Y., Zhang X. 2016. CGMM LASSO-type estimator for the process of Ornstein-Uhlenbeck type. J. Korean Stat. Soc. 45:114–122.

Warton D.I. 2008. Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices. J. Am. Stat. Assoc. 103:340–349.

van Wieringen W.N. 2017. On the mean squared error of the ridge estimator of the covariance and precision matrix. Stat. Probab. Lett. 123:88–92.

van Wieringen W.N., Peeters C.F.W. 2016. Ridge estimation of inverse covariance matrices from high-dimensional data. Comput. Stat. Data Anal. 103:284–303.

Wiley D.F., Amenta N., Alcantra A., Ghosh D., Kil Y.J., Delson E., Harcourt-Smith W., Rohlf F.J., John K.S., Hamann B. 2005. Evolutionary morphing. Proc. IEEE Vis. 2005.:431–438.

Witten D.M., Friedman J.H., Simon N. 2011. New Insights and Faster Computations for the Graphical Lasso. J. Comput. Graph. Stat. 20:892–900.

Witten D.M., Tibshirani R. 2009. Covariance-Regularized Regression and Classification for High Dimensional Problems. J. R. Stat. Soc. Ser. B Stat. Methodol. 71:615–636.

APPENDIX 1. THE LIKELIHOOD WRITTEN WITH THE KRONECKER PRODUCT AND AS A FUNCTION OF THE PRECISION MATRIX

Expression with the Kronecker product

The log-likelihood formulation often used in studies of multivariate trait evolution is (e.g., Clavel et al. 2015; Goolsby 2016):

$$\mathcal{L} = -\frac{1}{2}\{np\log(2\pi) + \log|\boldsymbol{R}\otimes\boldsymbol{C}| + vec(\boldsymbol{Y}-\boldsymbol{X}\beta)^T(\boldsymbol{R}\otimes\boldsymbol{C})^{-1}vec(\boldsymbol{Y}-\boldsymbol{X}\beta)\}$$

Where $(\otimes)$ is the Kronecker product, and the *vec* operator stacks the columns of a matrix into a column vector (Henderson and Searle 1979). This formulation is intuitive, but it is computationally prohibitive. The formulation we use in the paper (Eq. 1a) is equivalent to this formulation, using the following Kronecker product identities: $|\boldsymbol{A}\otimes\boldsymbol{B}| = |\boldsymbol{A}|^n|\boldsymbol{B}|^p$ where $\boldsymbol{A}$ is a p by p matrix and $\boldsymbol{B}$ is a n by n matrix, $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$ and $tr(\boldsymbol{A}\boldsymbol{C}^T\boldsymbol{B}\boldsymbol{C}) = vec(\boldsymbol{C})^T(\boldsymbol{A}\otimes\boldsymbol{B})vec(\boldsymbol{C})$ (see Magnus and Neudecker 2007, p. 32-35).

Expression with the precision matrix

We can re-express the log-likelihood in Equation (1a) as a function of the precision matrix $\boldsymbol{\Omega}$, defined as the inverse of the covariance matrix $\boldsymbol{R}$ (Anderson and Olkin 1985) as follows:

$$\mathcal{L} = -\frac{1}{2}\left\{np\log(2\pi) + p\log|\mathbf{C}| - n\log|\boldsymbol{\Omega}| + n\mathrm{tr}\left(\widehat{\boldsymbol{R}}\boldsymbol{\Omega}\right)\right\} \text{ (A1)}$$

where $\widehat{\boldsymbol{R}}$ is the MLE estimate of $\boldsymbol{R}$ given by $\widehat{\boldsymbol{R}} = \frac{(\boldsymbol{Y}-\boldsymbol{X}\beta)^T\boldsymbol{C}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)}{n}$ (Equation 2). When the inverse of $\widehat{\boldsymbol{R}}$ is used as an estimate for the precision matrix $\boldsymbol{\Omega}$, the last term in A1 is equal to $n \times p$ (see also Eq. 23.31 in Felsenstein 2004). Following Anderson and Olkin (1985), for

computational reasons we use A1 (instead of Equation 1) in Appendices 2 and 3. The penalized estimate of $\boldsymbol{\Omega}$, noted $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ in what follows, is obtained as $\boldsymbol{\Omega}_{\boldsymbol{\gamma}} = \boldsymbol{R}(\gamma)^{-1}$ (e.g., van Wieringen and Peeters 2016). Likewise, $\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} = \left[\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}\right]^{-1}$ in the equation of the LOOCV (Eq. 8), hence the LOOCV scores obtained maximizing with respect to $\boldsymbol{\Omega}$ are equal to LOOCV scores obtained maximizing with respect to $\boldsymbol{R}$.

APPENDIX 2. COMPUTATION OF THE FIRST AND SECOND ORDER DERIVATIVES OF THE PENALIZED LIKELIHOOD AND APPLICATION TO THE COMPUTATION OF THE GIC

Model comparison using the generalized information criterion (Eq. 9) needs the computation of the first and second order derivatives of the penalized log-likelihood with respect to the parameters estimates (Eq. 10). This is also the case for the LOOCV approximation we describe in Appendix 3. The first and second order derivatives are called the gradient and the Hessian, respectively.

We recall first some matrix notations and differentiation rules (e.g., Dwyer 1967; Magnus and Neudecker 1985, 2007). If ϕ is a differentiable scalar function of an $n \times 1$ vector $\boldsymbol{x}$, then the $1 \times n$ vector made of the partial derivatives of ϕ with respect to each element of $\boldsymbol{x}$, noted $D\phi(x) = \frac{\partial \phi(x)}{\partial x^T}$ is called the derivative of ϕ. If f is an $m \times 1$ differentiable vector function of $\boldsymbol{x}$, the derivative of f is the $m \times n$ matrix noted $Df(x) = \frac{\partial f(x)}{\partial x^T}$ where the i$^{\text{th}}$ row of the matrix is given by $Df_i(x)$. Finally, if $\boldsymbol{F}$ is a differentiable $m \times p$ function of an $n \times q$ matrix $\boldsymbol{X}$, the derivative of $\boldsymbol{F}$ at $\boldsymbol{X}$ (noted $\frac{\mathrm{d}\boldsymbol{F}(\boldsymbol{X})}{\mathrm{d}\boldsymbol{X}}$) is the $mp \times nq$ matrix $D\boldsymbol{F}(\boldsymbol{X}) = \frac{\partial vec\{\boldsymbol{F}(\boldsymbol{X})\}}{\partial \{vec(\boldsymbol{X})\}^T}$.
We also use the following matrix differentiation rules (e.g., Dwyer 1967; Magnus and Neudecker 2007):

$$\frac{\mathrm{d}\log|\mathbf{X}|}{\mathrm{d}\mathbf{X}} = vec(\boldsymbol{X}^{-1})^T; \frac{\mathrm{d}\mathrm{tr}(\boldsymbol{AX})}{\mathrm{d}\mathbf{X}} = vec(\boldsymbol{A})^T; \frac{\mathrm{d}\mathrm{tr}(\boldsymbol{XX}^T)}{\mathrm{d}\mathbf{X}} = \frac{\mathrm{d}\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})}{\mathrm{d}\mathbf{X}} = vec(2\boldsymbol{X})^T; \frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}\mathbf{X}} = \boldsymbol{I}_{p^2}; \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}\mathbf{X}} =$$

$$\mathbf{0}_{p^2}; \frac{\mathrm{d}\boldsymbol{X}^{-1}}{\mathrm{d}\mathbf{X}} = -(\boldsymbol{X}^T)^{-1} \otimes \boldsymbol{X}^{-1}$$

where $\boldsymbol{X}$ and $\boldsymbol{A}$ are matrices of size p by p with $\boldsymbol{A}$ a constant matrix, $\boldsymbol{I}_{p^2}$ and $\mathbf{0}_{p^2}$ are respectively an identity matrix and a matrix full of zero both of size p^2 by p^2.

Estimation of the first and second order derivatives with respect to $\boldsymbol{\Omega}$ for the quadratic penalty

The penalized log-likelihood with quadratic ridge penalization is given by Equation 5 (Witten and Tibshirani 2009; van Wieringen and Peeters 2016). Combined with Equation A1 we have:

$$\mathcal{L}_P(\boldsymbol{\Omega}; \widehat{\boldsymbol{R}}) = -\frac{1}{2}\Big\{np\log(2\pi) + p\log|\mathbf{C}| - n\log|\boldsymbol{\Omega}| + n\mathrm{tr}(\widehat{\boldsymbol{R}}\boldsymbol{\Omega}) + \frac{n\gamma}{2}\mathrm{tr}[(\boldsymbol{\Omega} - \boldsymbol{T})^T(\boldsymbol{\Omega} - \boldsymbol{T})]\Big\} \quad \text{(B1)}$$

where $\boldsymbol{T}$ is the diagonal matrix corresponding to the diagonal elements of $\widehat{\boldsymbol{R}}^{-1}$. The first derivative of the penalized likelihood w.r.t. the precision matrix is:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathcal{L}_P(\boldsymbol{\Omega}; \widehat{\boldsymbol{R}}) = \frac{n}{2}\left[vec\{\boldsymbol{\Omega}^{-1}\}^T - vec\{\widehat{\boldsymbol{R}}\}^T - \frac{\gamma}{2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathrm{tr}[(\boldsymbol{\Omega} - \boldsymbol{T})^T(\boldsymbol{\Omega} - \boldsymbol{T})]\right] \quad \text{(B2)}$$

We can expand the derivative of the penalty term:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathrm{tr}[(\boldsymbol{\Omega} - \boldsymbol{T})^T(\boldsymbol{\Omega} - \boldsymbol{T})] = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}[\mathrm{tr}(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \mathrm{tr}(\boldsymbol{\Omega}^T\boldsymbol{T}) - \mathrm{tr}(\boldsymbol{T}^T\boldsymbol{\Omega}) + \mathrm{tr}(\boldsymbol{T}^T\boldsymbol{T})]$$

$$= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathrm{tr}(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathrm{tr}(\boldsymbol{\Omega}^T\boldsymbol{T}) - \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathrm{tr}(\boldsymbol{T}^T\boldsymbol{\Omega})$$

$$= vec\{2\boldsymbol{\Omega}\}^T - 2vec\{\mathbf{T}\}^T$$

$$= 2[vec\{\boldsymbol{\Omega}\}^T - vec\{\mathbf{T}\}^T]$$

Which leads to:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = \frac{n}{2}\left[vec\{\boldsymbol{\Omega}^{-1}\}^T - vec\{\widehat{\boldsymbol{R}}\}^T - \gamma(vec\{\boldsymbol{\Omega}\}^T - vec\{\mathbf{T}\}^T)\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = \frac{n}{2}vec\{\boldsymbol{\Omega}^{-1} - \widehat{\boldsymbol{R}} - \gamma\boldsymbol{\Omega} + \gamma\boldsymbol{T}\}^T \text{ (B3)}$$

The second derivative of the penalized likelihood is:

$$\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\Omega}^2}\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = -\frac{n}{2}\{\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1} + \gamma\boldsymbol{I}_{\boldsymbol{p}^2}\} \text{ (B4)}$$

Estimation of the first and second order derivatives with respect to $\boldsymbol{\Omega}$ *for the archetypal ridge penalty*

The penalized log-likelihood with archetypal ridge penalization is given by Equation 4 (Warton 2008, van Wieringen and Peeters 2016); expressed as a function of $\boldsymbol{\Omega}$, we have:

$$\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = -\frac{1}{2}\{np\log(2\pi) + \mathrm{plog}|\boldsymbol{C}| - n\log|\boldsymbol{\Omega}| + n(1-\gamma)\mathrm{tr}(\widehat{\boldsymbol{R}}\boldsymbol{\Omega}) + n\gamma\mathrm{tr}[\boldsymbol{T}\boldsymbol{\Omega}]\}$$

(B5)

The first and second order derivatives with respect to $\boldsymbol{\Omega}$ are:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = \frac{n}{2}vec\{\boldsymbol{\Omega}^{-1} - (1-\gamma)\widehat{\boldsymbol{R}} - \gamma\boldsymbol{T}\}^T \text{ (B6)}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\Omega}^2}\mathcal{L}_P(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}) = \frac{n}{2}\{-\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\} \text{ (B7)}$$

Note that setting B6 to zero, we find that the solution of the penalized-likelihood for the archetypal ridge estimator (Equation 4) is indeed $(1-\gamma)\widehat{\boldsymbol{R}}+\gamma\boldsymbol{T}$, as given in Equation 3.

Efficient inversion of the Hessian matrix and computation of the GIC

Computing the GIC criterion (Eq. 9) and the approximate LOOCV scores (Appendix 3) involves computing the inverse of the second order matrix derivative of the likelihood (or penalized likelihood) with respect to the covariance or the precision matrix (the Hessian, Eq. B4 and B7). The Hessian matrix is of dimensions $p^2 \times p^2$ and its direct inversion and storage is computationally infeasible even for a dataset of moderate dimension. For instance with the New-World Monkeys brain dataset studied here the Hessian is a matrix with 2×10^{12} entries, which would require more than 15 TB of memory to store (for double precision numerical data). Instead of directly inverting the Hessian, we use identities that reduce this operation to inversions and multiplications of matrices of size *p* by *p*.

In the case of maximum-likelihood, the negative Hessian ($\boldsymbol{J}$ in equation 9 and 10) is given by $\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\Omega}^2}\mathcal{L}\left(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}\right) = \frac{n}{2}\{\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\}$ and the first derivative is given by $\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Omega}}\mathcal{L}\left(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}\right) = \frac{n}{2}vec\{\boldsymbol{\Omega}^{-1}-\widehat{\boldsymbol{R}}\}^T$ (e.g., Anderson and Olkin 1985). $\boldsymbol{J}^{-1}$ is easy to compute using the kronecker identity $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1}=\boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$. In addition, an efficient computation of the GIC can be done by using the computational trick of Abbruzzo et al. (2014; Eq. 20) – see also Lian (2011) and Vujačić et al. (2015):

$$\mathrm{tr}(\boldsymbol{J}^{-1}\boldsymbol{I}) = \frac{1}{2n}\sum_{i=1}^{n} vec\left(\widehat{\boldsymbol{R}}_i\right)^T vec\{\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{R}}_i\widehat{\boldsymbol{\Omega}}\} - \frac{1}{2}vec\left(\widehat{\boldsymbol{R}}\right)^T vec\{\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{R}}\widehat{\boldsymbol{\Omega}}\} \quad \text{(B8)}$$

where $\widehat{\boldsymbol{R}}_i = \left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)^T$ and $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{R}}^{-1}$.

For the archetypal ridge, the Hessian is also of type $\boldsymbol{A} \otimes \boldsymbol{B}$ (Eq. B7), and its inverse is easy to compute using the kronecker identity $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The GIC can be computed using:

$$\mathrm{tr}[\boldsymbol{J}^{-1}\boldsymbol{I}] = \mathrm{tr}\left[\boldsymbol{J}^{-1}(\boldsymbol{G_P}^T\boldsymbol{G}/n)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}\left[\boldsymbol{J}^{-1}vec(\boldsymbol{G}_{\boldsymbol{P}_i})vec(\boldsymbol{G}_i)^T\right] =$$

$$\frac{1}{n}\sum_{i=1}^{n} vec(\boldsymbol{G}_{\boldsymbol{P}_i})^T\boldsymbol{J}^{-1}vec(\boldsymbol{G}_i) \text{ (B9)}$$

where $\boldsymbol{G_P}$ is the matrix of gradient scores, i.e. the n by p^2 matrix which i,j element is given by $\psi(x_i, \left[\boldsymbol{\Omega}_\gamma\right]_j)$ (Eq. 10). With $\boldsymbol{J}$ of the form $\boldsymbol{A} \otimes \boldsymbol{B}$ and using the identity $vec(\boldsymbol{ACB}) = (\boldsymbol{B}^T \otimes \boldsymbol{A})vec(\boldsymbol{C})$ (Henderson and Searle 1979; Magnus and Neudecker 2007), each element in the sum above can be efficiently computed as:

$$\mathrm{tr}((\boldsymbol{A} \otimes \boldsymbol{B})^{-1}vec(\boldsymbol{C})vec(\boldsymbol{C})^T) = vec(\boldsymbol{C})^T(\boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1})vec(\boldsymbol{C})$$

$$= vec(\boldsymbol{C})^T\, vec(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{B}^{-1}) = \sum[\boldsymbol{C} \odot\ \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{B}^{-1}] \text{ (B10)}$$

Where $\odot$ is the Hadamard (element wise) product.

For the quadratic ridge, the Hessian is of type $\left(\boldsymbol{A} \otimes \boldsymbol{B} + \gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right)$ (Eq. B4) and its inversion can be done by using the eigen-decomposition of Kronecker products identity (e.g., Magnus and Neudecker 2007 p. 33; see also Stegle et al. 2011):

$$\left(\boldsymbol{A} \otimes \boldsymbol{B} + \gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right) = (\boldsymbol{U}_A \otimes \boldsymbol{U}_B)\left(\boldsymbol{S}_A \otimes \boldsymbol{S}_B + \gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right)(\boldsymbol{U}_A^T \otimes \boldsymbol{U}_B^T)$$

where $\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{S}_A\boldsymbol{U}_A^T$ is the eigenvalue decomposition of $\boldsymbol{A}$, and $\boldsymbol{B} = \boldsymbol{U}_B\boldsymbol{S}_B\boldsymbol{U}_B^T$ the eigenvalue decomposition of $\boldsymbol{B}$ ($\boldsymbol{U}$ is the matrix of eigenvectors and $\boldsymbol{S}$ is the diagonal matrix of eigenvalues $\boldsymbol{d}$). It follows that:

$$\left(\boldsymbol{A}\otimes\boldsymbol{B}+\gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right)^{-1}=(\boldsymbol{U}_A\otimes\boldsymbol{U}_B)diag(1./(\boldsymbol{d}_A\otimes\boldsymbol{d}_B+\gamma))(\boldsymbol{U}_A^T\otimes\boldsymbol{U}_B^T) \text{ (B11)}$$

Where $diag(v)$ is the diagonal matrix with diagonal elements given by vector *v* and ./ is the element wise inverse operator. In order to efficiently compute the GIC we make use of the various Kronecker product identities given above, and obtain each element in the last sum of Equation (B9) as:

$$\mathrm{tr}\left(\left(\boldsymbol{A}\otimes\boldsymbol{B}+\gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right)^{-1}vec(\boldsymbol{C})vec(\boldsymbol{C})^T\right)=vec(\boldsymbol{C})^T\left(\boldsymbol{A}\otimes\boldsymbol{B}+\gamma\boldsymbol{I}_{\boldsymbol{p}^2}\right)^{-1}vec(\boldsymbol{C})=$$

$$\sum[vec(\boldsymbol{U}_B^T\boldsymbol{C}\boldsymbol{U}_A)\odot vec[1./(\boldsymbol{d}_A\otimes\boldsymbol{d}_B+\gamma)]\odot vec(\boldsymbol{U}_B^T\boldsymbol{C}\boldsymbol{U}_A)] \text{ (B12)}$$

Note that in our case (Eqs. B4 and B7), the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are identical and further simplify the computations.

Computation of the GIC for the penalized likelihood with LASSO

The LASSO penalty is not easily differentiable (Lian 2011; Abbruzzo et al. 2014; Vujačić et al. 2015). Instead of computing the derivatives of the LASSO penalized likelihood with respect to the covariance matrix, we follow Abbruzzo et al. (2014) and Vujačić et al. (2015) and use instead the derivatives of the log-likelihood with a sparse estimate assumption as an approximation. This allows us to estimate the bias term in Equation (9) efficiently using the following formula:

$$\mathrm{tr}(\boldsymbol{J}^{-1}\boldsymbol{I})\approx\frac{1}{2n}\sum_{i=1}^{n}vec\left(\widehat{\boldsymbol{R}}_i\odot\boldsymbol{D}_\gamma\right)^T vec\left\{\boldsymbol{\Omega}_\gamma\left(\widehat{\boldsymbol{R}}_i\odot\boldsymbol{D}_\gamma\right)\boldsymbol{\Omega}_\gamma\right\}-\frac{1}{2}vec\left(\widehat{\boldsymbol{R}}\odot\boldsymbol{D}_\gamma\right)^T vec\left\{\boldsymbol{\Omega}_\gamma\left(\widehat{\boldsymbol{R}}\odot\boldsymbol{D}_\gamma\right)\boldsymbol{\Omega}_\gamma\right\} \text{ (B13)}$$

Where $\boldsymbol{\Omega}_{\gamma}$ is the regularized estimate that maximize the LASSO penalized likelihood for a given γ, $\boldsymbol{D}_{\gamma}$ is an indicator matrix where the entries are 1 when the corresponding entries are non-zero in $\boldsymbol{\Omega}_{\gamma}$ and zero otherwise (see the details for the derivation of this approximation in Abbruzzo et al. (2014)), and $\widehat{\boldsymbol{R}}_i$ is as above $\widehat{\boldsymbol{R}}_i = \left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)^T$. Note that when every element of the indicator matrix $\boldsymbol{D}_{\gamma}$ is equal to one we obtain the formula described above (Eq. B8) for the MLE (i.e., when $\gamma = 0$).

Estimation of the first and second order derivatives with respect to β

The gradient and Hessian with respect to the ancestral states β are the same whether we consider the log-likelihood or the penalized log-likelihood (except in the case of the archetypal ridge, where they are equal up to $(1-\gamma)$, which does not affect model comparison using GIC), and they can be obtained analytically. From matrix calculus and differential rules (*e.g.*, Magnus and Neudecker 2007, p. 201), we have, if $\boldsymbol{A}$ is a symmetric matrix:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} tr[\boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{X}\beta)^T(\boldsymbol{Y}-\boldsymbol{X}\beta)]$$

$$= \frac{\mathrm{d}}{\mathrm{d}\beta}[\mathrm{tr}(\boldsymbol{A}\boldsymbol{Y}^T\boldsymbol{Y}) - \mathrm{tr}(\boldsymbol{A}\boldsymbol{Y}^T\boldsymbol{X}\beta) - \mathrm{tr}(\beta^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{A}) + \mathrm{tr}(\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta\boldsymbol{A})]$$

$$= -\frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{tr}(\boldsymbol{A}\boldsymbol{Y}^T\boldsymbol{X}\beta) - \frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{tr}(\beta^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{A}) + \frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{tr}(\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta\boldsymbol{A})$$

$$= -vec\{\boldsymbol{A}\boldsymbol{Y}^T\boldsymbol{X}\}^T - vec\{\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{A}\}^T + vec\{\boldsymbol{X}^T\boldsymbol{X}\beta\boldsymbol{A} + [\beta^T\boldsymbol{X}^T\boldsymbol{X}]^T\boldsymbol{A})\}^T$$

$$= vec\{-2\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{A} + 2\boldsymbol{X}^T\boldsymbol{X}\beta\boldsymbol{A}\}^T$$

Therefore, we have:

$$\frac{\mathrm{d}\mathcal{L}_P(\beta;\boldsymbol{\Omega})}{\mathrm{d}\beta} = \frac{\mathrm{d}\mathcal{L}(\beta;\boldsymbol{\Omega})}{\mathrm{d}\beta} = -\frac{1}{2}vec\left\{-2\,\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{Y}}\boldsymbol{\Omega} + 2\,\widetilde{\boldsymbol{X}}^T\,\widetilde{\boldsymbol{X}}\beta\boldsymbol{\Omega}\right\}^T = vec\left\{\,\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{Y}}\boldsymbol{\Omega} - \,\widetilde{\boldsymbol{X}}^T\,\widetilde{\boldsymbol{X}}\beta\boldsymbol{\Omega}\right\}^T \quad \text{(B14)}$$

From equation (B14) and results from Magnus and Neudecker (2007, p. 198, 360), the second order derivatives are given by:

$$\frac{\mathrm{d}^2\mathcal{L}_P(\beta;\boldsymbol{\Omega})}{\mathrm{d}\beta^2} = \frac{\mathrm{d}^2\mathcal{L}(\beta;\boldsymbol{\Omega})}{\mathrm{d}\beta^2} = -\left\{\boldsymbol{\Omega} \otimes \widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}\right\} \text{ (B15)}$$

The inverse of the negative Hessian $-\left[\frac{\mathrm{d}^2\mathcal{L}(\beta;\boldsymbol{\Omega})}{\mathrm{d}\beta^2}\right]^{-1}$ and the GIC criterion (Eqs. 9-10) can then be efficiently computed using the tricks based on kronecker product identities described above for the precision matrix.

Standard error of parameter estimates

Taking $\mathcal{L}_p(\theta) = \mathcal{L}(\theta) - P(\theta)$ to be the penalized likelihood with respect to parameters θ with penalty term *P*, the standard error (SE) of the parameter estimates $\hat{\theta}$ can be computed using either the observed Fisher information matrix – the negative Hessian – or the so-called "sandwich" formula (Fan and Li 2001; Rondeau et al. 2003). A first simple approximation of the asymptotic variance-covariance matrix for the parameter estimates $\hat{\theta}$ is given by:

$$\widehat{cov}\left(\hat{\theta}\right)_1 = \hat{H}^{-1}\left(\hat{\theta}\right) \text{ (B16)}$$

where $\hat{H}\left(\hat{\theta}\right) = -\frac{\mathrm{d}^2\mathcal{L}_p(\hat{\theta})}{\mathrm{d}\theta^2}$ is the negative Hessian of the penalized log-likelihood. Another robust estimator of the covariance matrix for the parameter estimates $\hat{\theta}$ is given by the "sandwich" formula (Fan and Li 2001; Rondeau et al. 2003):

$$\widehat{cov}\left(\hat{\theta}\right)_2 = \hat{H}^{-1}\left(\hat{\theta}\right)\hat{I}\left(\hat{\theta}\right)\hat{H}^{-1}\left(\hat{\theta}\right) \text{ (B17)}$$

where $\hat{H}\left(\hat{\theta}\right)$ is as previously defined and $\hat{I}\left(\hat{\theta}\right) = -\frac{\mathrm{d}^2\mathcal{L}(\hat{\theta})}{\mathrm{d}\theta^2}$ or $\hat{I}\left(\hat{\theta}\right) = cov\left(\frac{\mathrm{d}\mathcal{L}(\hat{\theta})}{\mathrm{d}\theta}\right)$. The standard errors for the parameter estimates $\hat{\theta}$ are then obtained by taking the square root of the diagonal elements of either $\widehat{cov}\left(\hat{\theta}\right)_1$ or $\widehat{cov}\left(\hat{\theta}\right)_2$.

When computing standard errors around the precision matrix $\boldsymbol{\Omega}$ (or the covariance $\boldsymbol{R}$), there are $p(p+1)/2$ parameter estimates (not p^2), because $\boldsymbol{\Omega}$ (or $\boldsymbol{R}$) is symmetric. Therefore, instead of computing the derivatives with respect to $\boldsymbol{\Omega}$, we compute the derivatives with respect to $vech(\boldsymbol{\Omega})$, where *vech* is an operator that stacks the columns of a square matrix starting at its diagonal elements into a column vector (Henderson and Searle 1979). We have: $\frac{\mathrm{d}^2}{\mathrm{d}vech(\boldsymbol{\Omega})^2}\mathcal{L}_P\left(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}\right) = \boldsymbol{G}^T\frac{\mathrm{d}^2\mathcal{L}_P\left(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}\right)}{\mathrm{d}\boldsymbol{\Omega}^2}\boldsymbol{G}$ where $\boldsymbol{G}$, known as the "duplication matrix", is a p^2 by $p(p+1)/2$ matrix with indicator entries 0 or 1 that satisfies $vec(\boldsymbol{A}) = \boldsymbol{G}vech(\boldsymbol{A})$ (Henderson and Searle 1979; Moneta 1991). For the archetypal ridge for example, from B7 we deduce (e.g., McCulloch 1982):

$$\frac{\mathrm{d}^2}{\mathrm{d}vech(\boldsymbol{\Omega})^2}\mathcal{L}_P\left(\boldsymbol{\Omega};\widehat{\boldsymbol{R}}\right) = \frac{n}{2}\{\boldsymbol{G}^T(-\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\boldsymbol{G}\} \quad \text{(B18)}$$

Codes for this calculation and the associated computation of standard error for the precision matrix are provided as an illustrative example in our Supplementary Material.

APPENDIX 3. SPEEDING-UP THE COMPUTATIONS OF THE LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV)

The ordinary approach for computing the LOOCV score (Eq. 8) requires the inversion of *n* matrices of size *p* by *p* for a given value of the regularization parameter, which is computationally intensive, especially for large *n*. Here we develop approaches for reducing computation time. In the case of the archetypal ridge penalty, we use an analytical trick and illustrate the use of independent contrasts scores instead of the $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{X}}$ matrices in Equation (8). In the case of the quadratic ridge and LASSO penalties, we derive an approximation of the LOOCV score.

Exact and efficient computation of the LOOCV with the archetypal ridge penalty

The LOOCV of the penalized log-likelihood (Eq. 8) with the archetypal ridge estimator, target matrix $\boldsymbol{T}$, and regularization parameter γ can be efficiently computed using the rank-one update trick proposed by Hoffbeck and Landgrebe (1996; section 3.3-3.4). These authors showed that we could obtain $\widehat{\boldsymbol{R}}_{(-\boldsymbol{i})}$ required to compute $\widetilde{\boldsymbol{R}}(\gamma)_{(-\boldsymbol{i})}$ in Equation (8) using a rank-one update of the covariance matrix $\widehat{\boldsymbol{R}}$:

$$\widehat{\boldsymbol{R}}_{(-\boldsymbol{i})} = \frac{n^*}{n^* - 1}\widehat{\boldsymbol{R}} - \frac{n}{(n-1)(n^*-1)}\widehat{\boldsymbol{R}}_{\boldsymbol{i}}$$

with $n^* = n$ for the maximum likelihood estimator and $n^* = n - 1$ for the restricted maximum likelihood (see also Theiler 2012; eq. 50). Hence, we can re-express the archetypal ridge estimate as:

$$\widetilde{\boldsymbol{R}}(\gamma)_{(-\boldsymbol{i})} = (1-\gamma)\widehat{\boldsymbol{R}}_{(-\boldsymbol{i})} + \gamma\boldsymbol{T} = \boldsymbol{G}_\gamma - \phi\widehat{\boldsymbol{R}}_{\boldsymbol{i}}$$

Where $\boldsymbol{G}_\gamma = \frac{(1-\gamma)n^*}{(n^*-1)}\widehat{\boldsymbol{R}} + \gamma\boldsymbol{T}$, and $\phi = \frac{(1-\gamma)n}{(n-1)(n^*-1)}$.

Using this new expression, it is possible to rewrite $\mathrm{tr}\left[\widetilde{\boldsymbol{R}}(\gamma)^{-1}_{(-i)}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)^T\right]$ in the LOOCV (Eq. 8). If we note $v = \left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)$, we have $\widehat{\boldsymbol{R}}_{\boldsymbol{i}} = vv^T$ and the Sherman-Morrison-Woodbury formula yields:

$$\left(\boldsymbol{G}_\gamma - \phi\widehat{\boldsymbol{R}}_{\boldsymbol{i}}\right)^{-1} = \left(\boldsymbol{G}_\gamma - \phi vv^T\right)^{-1} = \boldsymbol{G}_\gamma^{-1} + \phi\frac{\boldsymbol{G}_\gamma^{-1}vv^T\boldsymbol{G}_\gamma^{-1}}{1 - \phi v^T\boldsymbol{G}_\gamma^{-1}v}$$

We deduce:

$$\mathrm{tr}\left[\left(\boldsymbol{G}_\gamma - \phi\widehat{\boldsymbol{R}}_{\boldsymbol{i}}\right)^{-1}vv^T\right] = v^T\left(\boldsymbol{G}_\gamma - \phi\widehat{\boldsymbol{R}}_{\boldsymbol{i}}\right)^{-1}v$$

$$= v^T\left[\boldsymbol{G}_\gamma^{-1} + \phi\frac{\boldsymbol{G}_\gamma^{-1}vv^T\boldsymbol{G}_\gamma^{-1}}{1 - \phi v^T\boldsymbol{G}_\gamma^{-1}v}\right]v = v^T\boldsymbol{G}_\gamma^{-1}v + \phi\frac{v^T\boldsymbol{G}_\gamma^{-1}vv^T\boldsymbol{G}_\gamma^{-1}v}{1 - \phi v^T\boldsymbol{G}_\gamma^{-1}v}$$

$$= r_i + \phi \frac{{r_i}^2}{1 - \phi r_i} = \frac{r_i}{1 - \phi r_i}$$

With $r_i = v^T \boldsymbol{G}_\gamma^{-1} v$ (see also Theiler 2012, eq. 16-18).

In addition, we can compute $\left|\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}\right|$ in Equation (8) using Sylvester's determinant theorem (see section 3.3 in Hoffbeck and Landgrebe 1996):

$$\left|\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}\right| = \left|\boldsymbol{G}_\gamma - \phi v v^T\right| = \left|\boldsymbol{G}_\gamma\right|(1 - \phi r_i)$$

Using these expressions, the LOOCV (Eq. 8) of the log-likelihood for the archetypal ridge is given by:

$$\mathcal{L}_{CV} = -\frac{1}{2}\left[n^* p \log(2\pi) + p\log|\boldsymbol{C}| + n^* \log\left|\boldsymbol{G}_\gamma\right| + \sum_{i=1}^{n}\left(log(1 - \phi r_i) + \frac{r_i}{1 - \phi r_i}\right)\right] \text{ (C1)}$$

This expression is an exact equivalent to the LOOCV score in Equation (8), yet the computational cost is greatly reduced (from $O(np^3)$ to $O(p^3)$+$O(np^2)$ operations) because the computation of the determinant and inverse of $\boldsymbol{G}_\gamma$ is done only once (Hoffbeck and Landgrebe 1996; Theiler 2012). Moreover, the determinant and the inverse can be cheaply obtained from the factorization of $\boldsymbol{G}_\gamma$ (see Clavel et al. 2015; Appendix S1). We note that in their original paper, Hoffbeck and Landgrebe (1996; section 3.4) considered a slight variant of the LOOCV with $\text{tr}\left[\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}^{-1}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta_{(-i)}\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta_{(-i)}\right)^T\right]$ instead of $\text{tr}\left[\widetilde{\boldsymbol{R}}(\gamma)_{(-i)}^{-1}\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} - \widetilde{\boldsymbol{X}}_{\boldsymbol{i}}\beta\right)^T\right]$ where $\beta_{(-i)} = \left(\widetilde{\boldsymbol{X}}_{(-i)}^T\widetilde{\boldsymbol{X}}_{(-i)}\right)^{-1}\widetilde{\boldsymbol{X}}_{(-i)}^T\widetilde{\boldsymbol{Y}}_{(-i)}$; in this case, an efficient computation is done by replacing the term $\frac{r_i}{1-\phi r_i}$ in Eq. C1 by $\left(\frac{n}{n-1}\right)^2\frac{r_i}{1-\phi r_i}$. However, in our case, we found Equation C1 to provide better estimates than those obtained with the formula used by Hoffbeck and Landgrebe (1996).

To further speed up the computations, we use the phylogenetic independent contrasts scores (Felsenstein 1973, 1985, 2004) instead of the $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{X}}$ matrices in Equation (8). We note $\boldsymbol{U}$ the $(n-1)$ by p matrix of phylogenetic independent contrasts scores. We further note V the column vector of size n containing the variance associated to the estimation of all contrasts and the root state (Felsenstein 1973; Freckleton 2012). Applying the same calculations as above, we can write the LOOCV as:

$$\log(\mathcal{L})_{CV} = -\frac{1}{2}\left[n^* p \log(2\pi) + p\sum_{i=1}^{n}\log(V_i) + n^*\log\left|\boldsymbol{G}_\gamma\right| + \sum_{i=1}^{n-1}\left(\log(1-\phi r_i) + \frac{r_i}{1-\phi r_i}\right)\right] \quad \text{(C2)}$$

where $r_i = u_i^T \boldsymbol{G}_\gamma^{-1} u_i$ with $\boldsymbol{G}_\gamma = n^*\phi\widehat{\boldsymbol{R}} + \gamma\boldsymbol{T}$ and u_i the column vector made of the i^{th} row of matrix $\boldsymbol{U}$ and $\phi = \frac{1-\gamma}{n^*-1}$. Here $\widehat{\boldsymbol{R}}$ is simply computed as $\frac{\boldsymbol{U}^T\boldsymbol{U}}{n^*}$ (Felsenstein 2004). Note that because the contrasts scores have a zero mean, $\boldsymbol{G}_\gamma$ and ϕ are different from those used for Equation (C1) above (Theiler 2012). For REML estimation, $n^* = n - 1$ and the variance term associated to the root state in V is omitted.

Approximation of the LOOCV with the quadratic ridge penalty

Previous authors have shown that it is possible to efficiently approximate the LOOCV score for the multivariate normal log-likelihood and used these results to approximate the LOOCV score for the graphical LASSO (Lian 2011; Vujačić et al. 2015; see below *Efficient approximation of the LOOCV with the LASSO penalty*). Building on these previous developments we give here the derivation of the approximate LOOCV score for the multivariate normal penalized likelihood with the quadratic ridge penalty.

From Lian (2011; Appendix A), and Vujačić et al. (2015; section 4.1), the LOOCV score for the penalized likelihood can be approximated by:

$$\mathcal{L}_{CV}\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}};\widehat{\boldsymbol{R}}\right) \approx -\frac{1}{n}\mathcal{L}\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}};\widehat{\boldsymbol{R}}\right) - \frac{1}{2n}\sum_{i=1}^{n}\left\{\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}}_{i};\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}} vec\left(\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} - \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)\right\} \text{ (C3)}$$

where g is the (scaled) penalized log-likelihood, which, for the quadratic ridge penalty, is defined as:

$$g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}\right) = \log|\boldsymbol{\Omega}| - \mathrm{tr}\left(\widehat{\boldsymbol{R}}\boldsymbol{\Omega}\right) - \frac{\gamma}{2}\mathrm{tr}[(\boldsymbol{\Omega}-\boldsymbol{T})^{T}(\boldsymbol{\Omega}-\boldsymbol{T})]$$

We can approximate the term $vec\left(\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} - \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)$ in Equation (C3) using the first order Taylor expansion of the function $\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}};\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}}$ around $\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)$, evaluated at $\left(\widehat{\boldsymbol{R}}_{(-i)},\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}}\right)$ (Lian 2011; Vujačić et al. 2015, 4.1):

$$\begin{aligned}\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}}_{(-i)},\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}}\right)}{\mathrm{d}\boldsymbol{\Omega}} &\\ &\simeq \left\{\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}}\right\} + \frac{\mathrm{d}^{2}g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}^{2}} vec\left(\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} - \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)\\ &+ \frac{\mathrm{d}^{2}g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}\mathrm{d}\widehat{\boldsymbol{R}}} vec\left(\widehat{\boldsymbol{R}}_{(-i)} - \widehat{\boldsymbol{R}}\right)\end{aligned}$$

Since $\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}} = 0$ (it is the derivative – Eq. B3 – of the penalized-likelihood g evaluated at the maximum penalized-likelihood estimates $\widehat{\boldsymbol{R}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$) and equivalently $\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}}_{(-i)},\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}}\right)}{\mathrm{d}\boldsymbol{\Omega}} = 0$, when $\widehat{\boldsymbol{R}}_{(-i)}$ and $\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}}$ are sufficiently close from $\widehat{\boldsymbol{R}}$ and $\boldsymbol{\Omega}(\gamma)$ it follows that:

$$vec\left(\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} - \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right) \simeq -\left(\frac{\mathrm{d}^2 g(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}})}{\mathrm{d}\boldsymbol{\Omega}^2}\right)^{-1} \frac{\mathrm{d}^2 g(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}})}{\mathrm{d}\boldsymbol{\Omega}\mathrm{d}\widehat{\boldsymbol{R}}} vec\left(\widehat{\boldsymbol{R}}_{(-i)} - \widehat{\boldsymbol{R}}\right) \text{ (C4)}$$

This approximation is accurate when $\widehat{\boldsymbol{R}}_{(-i)} - \widehat{\boldsymbol{R}}$ is sufficiently small.

Therefore, the term $\frac{\mathrm{d}g(\widehat{\boldsymbol{R}}_i,\boldsymbol{\Omega}_{\boldsymbol{\gamma}})}{\mathrm{d}\boldsymbol{\Omega}} vec\left(\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{\gamma}_{(-i)}} - \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)$ on the right hand side of Equation (C3) can be approximated by:

$$\frac{\mathrm{d}g\left(\widehat{\boldsymbol{R}}_i,\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}}\left[-\left(\frac{\mathrm{d}^2 g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}^2}\right)^{-1} \frac{\mathrm{d}^2 g\left(\widehat{\boldsymbol{R}},\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)}{\mathrm{d}\boldsymbol{\Omega}\mathrm{d}\mathbf{R}} vec\left(\widehat{\boldsymbol{R}}_{(-i)} - \widehat{\boldsymbol{R}}\right)\right]$$

From Appendix 2 we have $\frac{\mathrm{d}^2 g(\widehat{\boldsymbol{R}},\boldsymbol{\Omega})}{\mathrm{d}\boldsymbol{\Omega}^2} = -\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1} - \gamma \boldsymbol{I}_{p^2}$, $\frac{\mathrm{d}g(\widehat{\boldsymbol{R}},\boldsymbol{\Omega})}{\mathrm{d}\boldsymbol{\Omega}} = vec\left(\boldsymbol{\Omega}^{-1} - (\widehat{\boldsymbol{R}} - \gamma \boldsymbol{T}) - \lambda\boldsymbol{\Omega}\right)^T$, $\frac{\mathrm{d}\widehat{\boldsymbol{R}}}{\mathrm{d}\widehat{\boldsymbol{R}}} = \boldsymbol{I}_{p^2}$ and thus $\frac{\mathrm{d}^2 g(\widehat{\boldsymbol{R}},\boldsymbol{\Omega})}{\mathrm{d}\boldsymbol{\Omega}\mathrm{d}\mathbf{R}} = \frac{\mathrm{d}\left(\boldsymbol{\Omega}^{-1} - (\widehat{\boldsymbol{R}} - \gamma \boldsymbol{T}) - \lambda\boldsymbol{\Omega}\right)}{\mathrm{d}\mathbf{R}} = -\boldsymbol{I}_{p^2}$.

Therefore the term above is equal to:

$$vec\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} - \left(\widehat{\boldsymbol{R}}_i - \gamma \boldsymbol{T}\right) - \lambda\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)^T \left[\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} \otimes \boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} + \gamma \boldsymbol{I}_{p^2}\right)^{-1} vec\left(-\boldsymbol{I}_p\left(\widehat{\boldsymbol{R}}_{(-i)} - \widehat{\boldsymbol{R}}\right)\right)\right]$$

Using the eigen-decomposition of $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ and similar calculus as above (see Appendix 2), we find that this term is equal to:

$$\sum_{p^2}\left[vec\left(\boldsymbol{U}^T\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} - \left(\widehat{\boldsymbol{R}}_i - \gamma \boldsymbol{T}\right) - \lambda\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)\boldsymbol{U}\right) \odot \left(vec[1./(d \otimes d + \gamma)] \odot vec\left(\boldsymbol{U}^T\left(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{R}}_{(-i)}\right)\boldsymbol{U}\right)\right)\right] \text{ (C5)}$$

Finally, after simplification we obtain the first order Taylor approximation of the LOOCV score for the penalized log-likelihood with the quadratic ridge penalty (see also Eq. 6 in Vujačić et al. 2015):

$$\mathcal{L}_{CV} \approx -\frac{1}{n}\mathcal{L}\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}; \widehat{\boldsymbol{R}}\right) + 1/2n(n-1)\sum_{i=1}^{n}\left\{\sum_{p^2}\left[vec\left(\boldsymbol{U}^T\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} - \left(\widehat{\boldsymbol{R}}_i - \gamma\boldsymbol{T}\right) - \gamma\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right)\boldsymbol{U}\right) \odot \left(vec[1./(d \otimes d + \gamma)] \odot vec\left(\boldsymbol{U}^T\left(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{R}}_i\right)\boldsymbol{U}\right)\right)\right]\right\} \text{ (C6)}$$

Further, we note that once $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1}$ $(= \boldsymbol{R}(\gamma))$ has been computed (using Eq. 7), $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ can be easily obtained from the eigen-decomposition of $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1}$ or by noticing that $\boldsymbol{\Omega}_{\boldsymbol{\gamma}} = \frac{1}{\gamma}\left[\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} - (\widehat{\boldsymbol{R}} - \gamma\boldsymbol{T})\right]$ (van Wieringen and Peeters 2016; section 3.1).

Efficient approximation of the LOOCV with the LASSO penalty

For the LASSO penalty we use the LOOCV approximation of the multivariate normal log-likelihood with the assumption of a sparse estimate for $\boldsymbol{\Omega}$ (i.e., $\boldsymbol{\Omega} = \boldsymbol{\Omega}_{\boldsymbol{\gamma}_{LASSO}}$) given by Vujačić et al. (2015 - Eq. 6 and 9):

$$\mathcal{L}_{CV} \approx -\frac{1}{n}\mathcal{L}\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}; \widehat{\boldsymbol{R}}\right) + 1/2n(n-1)\sum_{i=1}^{n} T_i$$

with

$$T_i = \sum\left\{\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{-1} - \widehat{\boldsymbol{R}}_i\right) \odot \boldsymbol{D}_{\gamma} \odot \boldsymbol{\Omega}_{\boldsymbol{\gamma}}\left\{\left(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{R}}_i\right) \odot \boldsymbol{D}_{\gamma}\right\}\boldsymbol{\Omega}_{\boldsymbol{\gamma}}\right\} \text{ (C7)}$$

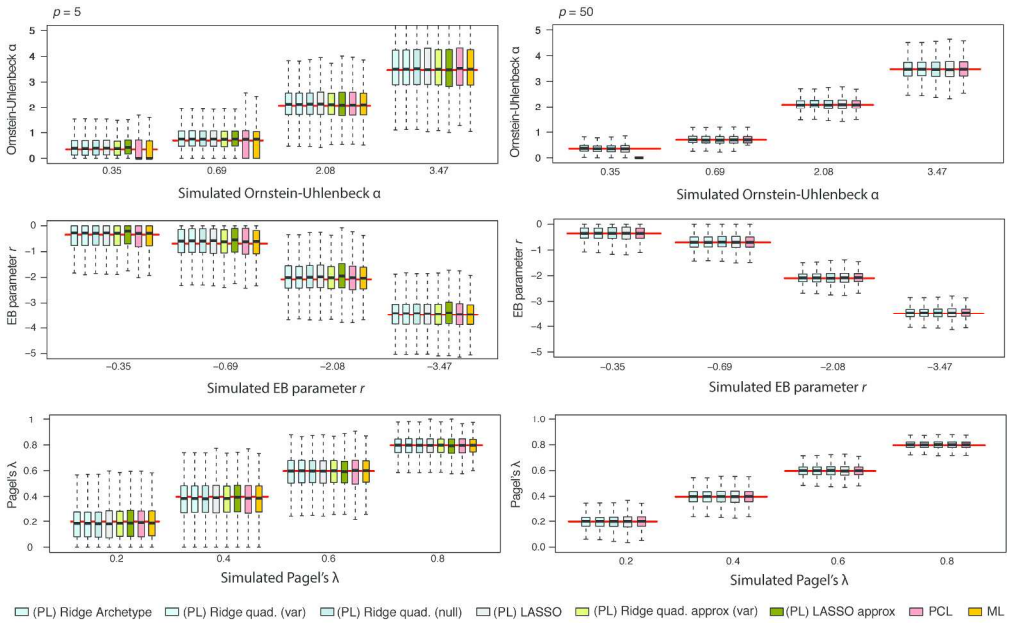

Figure 1. Performance of penalized likelihood for the estimation of model parameters and comparison with other approaches. Estimates of the parameters of the OU, EB and Pagel's λ models obtained with the various (restricted) penalized likelihood approaches, the (restricted) pairwise composite likelihood approach, and the (restricted) maximum likelihood. Boxplots represent the median, first and 3rd quartile and range of estimates obtained over 1000 simulated datasets. The red line represents simulated parameter values. Left panels (p=5) illustrate results with p<n (n=32), and right panels (p=50) illustrate results with p>n. Results with other p values and unrestricted likelihoods are presented in Figure S1-S6. For p>n, the approximated LOOCV approaches are not reliable (and thus not shown), and the maximum likelihood approach is not applicable. For each trait model, we report results for only 4 of the 6 simulated parameters for presentation purposes.

283x175mm (300 x 300 DPI)

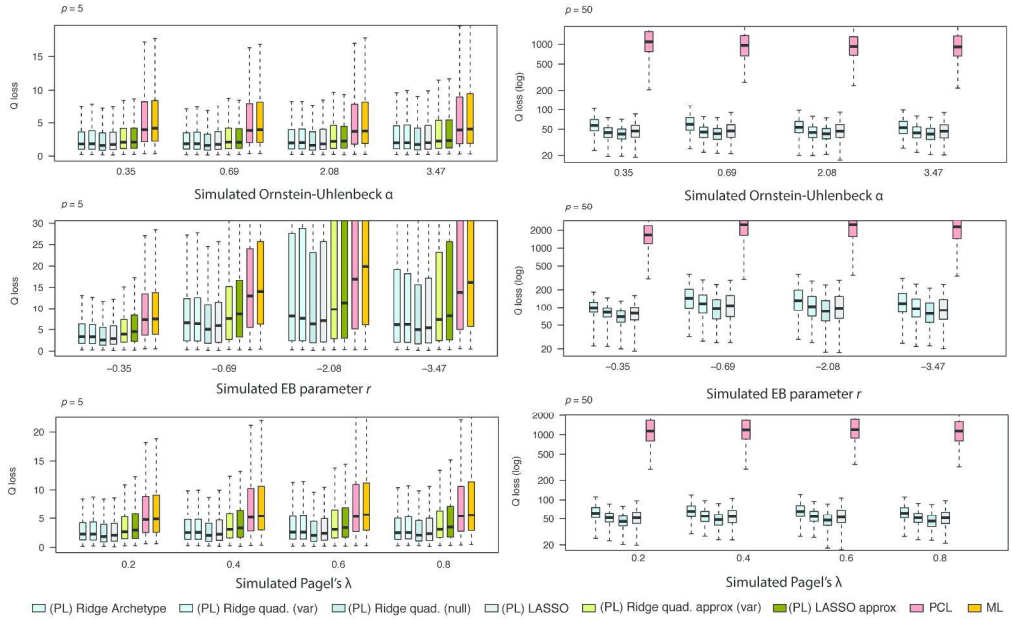

Figure 2. Performance of penalized likelihood for the estimation of the variance covariance matrix and comparison with other approaches. Quadratic loss between the simulated and estimated traits evolutionary variance-covariance matrix R obtained with the various (restricted) penalized likelihood approaches, the (restricted) pairwise composite likelihood approach, and the (restricted) maximum likelihood. Boxplots represent the median, first and 3rd quartile and range of Q loss values obtained over 1000 simulated datasets. Left panels (p=5) illustrate results with p<n (n=32), and right panels (p=50) illustrate results with p>n. Results with other p values and unrestricted likelihoods are presented in Figures S7-S10.

289x177mm (300 x 300 DPI)

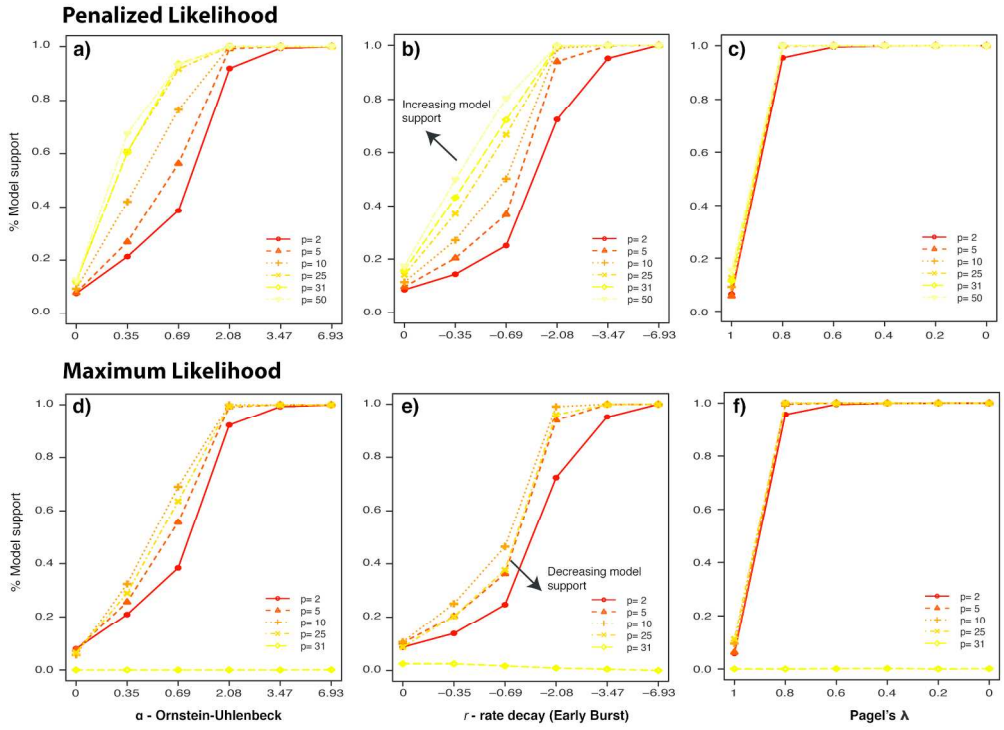

Figure 3. Performance of penalized likelihood for model selection and comparison with maximum likelihood. Model support, measured as the proportion of simulated datasets (over 1000) for which the generating model has the lowest GIC score, represented as a function of the model parameters and for p ranging from smaller to larger than n (n=32) (p = 2, 5, 10, 25, 31, 50), with: (a-c) the restricted likelihood with archetypal ridge penalization, and (d-f) the restricted maximum likelihood approach. With penalized likelihood, the ability to recover the generating model increases with increasing dimensions, while with maximum likelihood it decreases as p approaches n. Results with the unrestricted likelihood and other types of penalizations are presented in Figures S15-S17. Note that the OU with α=0, EB with r=0 and Pagel's model with λ=1 are equivalent to BM, such that the % model support for these parameter values represent the false recovery rate.

212x153mm (300 x 300 DPI)

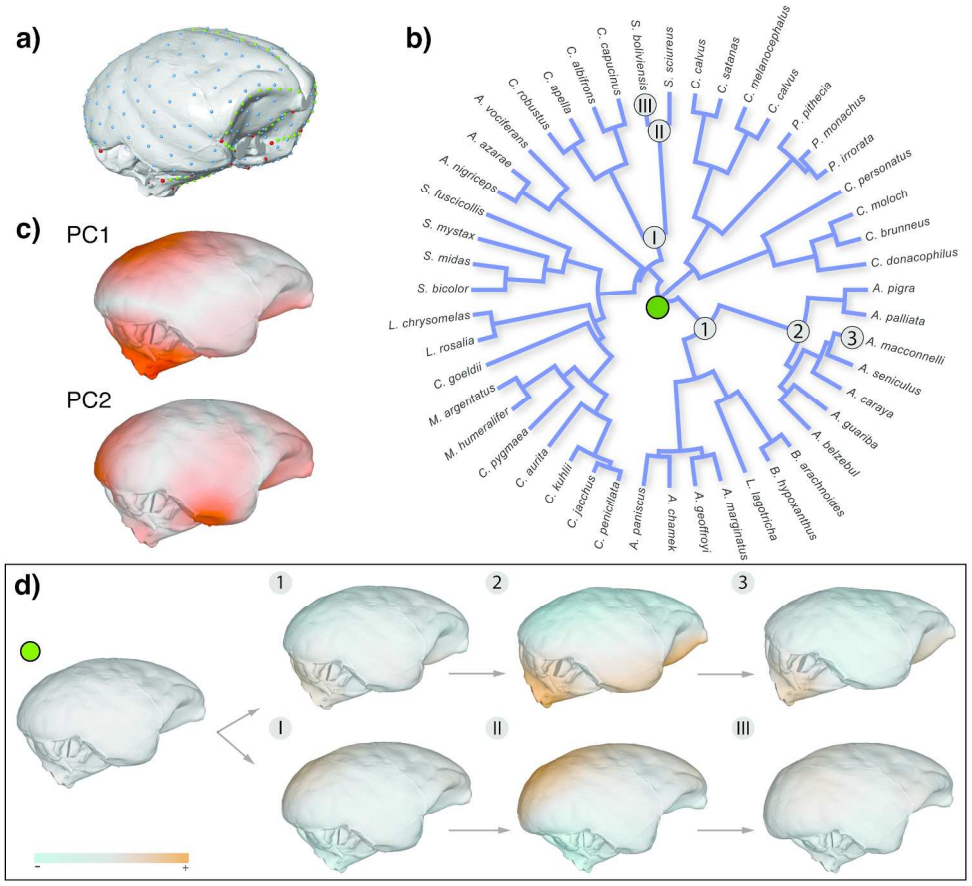

Figure 4. New-World Monkeys Brain Evolution. a) Brain endocast (of Pithecia irrorata) showing the position of the anatomical landmarks (red), and surface (blue) and curve (green) semi-landmarks that were taken on each species (see Aristide et al. 2016). b) Maximum clade credibility tree showing the phylogenetic relationships for the 48 platyrrhine species considered in this study (see Aristide et al. 2015). Number on nodes indicates the location of reconstructions in d). c) Patterns of evolutionary integration and modularity in brain shape as represented by the first two Principal Component (PC) axes of the estimated evolutionary covariance matrix. Red colored areas represent regions changing concertedly along each PC axis. d) New world monkey ancestral brain shape reconstruction, and reconstructed evolutionary trajectories for two selected species. Colors depict the amount of change between a given state and the previous one, in terms of expansion (orange) or contraction (cyan) with respect to the center of the surface.

210x196mm (300 x 300 DPI)

Table 1. *Summary of various penalized likelihood approaches, associated target matrices, and their properties. The four penalties evaluated in the paper are highlighted in bold. I is the identity or multiple of the identity matrix. V is a diagonal matrix with distinct diagonal values, such as the estimated variances for each trait. N is the null matrix (a matrix full of zero).*

Penalty	Target matrix	Rotation-invariance	Computational complexity*	Flexibility**
ridge archetypal	I	YES	+	+
	V	**NO**	+	++
ridge quadratic	**N**	**YES**	++	++
	I	YES	+++	++
	V	**NO**	+++	+++
LASSO	**N**[***]	**NO**	++++	++++

* Computational complexity refers to both the computation of the solution of the penalized likelihood and the LOOCV.** Flexibility refers to the robustness of the penalization procedure with respect to structure in the data ; more confidence in the estimates (and thus also in model selection) will be provided by more flexible penalizations, at the expense of computational efficiency. *** In some LASSO implementations, the diagonal elements of $\boldsymbol{R}$ (or $\boldsymbol{R}^{-1}$) are not penalized and $\boldsymbol{R}$ (or $\boldsymbol{R}^{-1}$) is shrunk towards $diag(\boldsymbol{R})$ (or $diag(\boldsymbol{R}^{-1})$) which can be then seen as the target matrix (e.g., Bien and Tibshirani 2011).

Table 2. *Support for BM, OU and EB models for the evolution of New World monkeys' brain shape over 100 trees from the Bayesian posterior distribution. The EB model is preferred (lowest GIC value) in 97% of the trees.*

Model	GIC (mean±2sd)	ΔGIC (2.5%-97.5% range)	% trees preferred	Parameters* (mean±2sd)
BM	-726505±410	[4.95 - 258]	0	–
OU	-726504±409	[1.03 - 260.4]	3	α = 9.34e-5±4.8e-4
EB	-726618±379	[0 - 0.30]	97	r = -0.91±0.69

*****The Brownian parameters are given in the multidimensional **R** matrix (see also Figure 4).*