

Chapter 12

Multimodel-Inference in Comparative Analyses

László Zsolt Garamszegi and Roger Mundry

Abstract Multimodel inference refers to the task of making a generalization from several statistical models that correspond to different biological hypotheses and that vary in the degree of how well they fit the data at hand. Several approaches have been developed for such purpose, and these are widely used, mostly for intraspecific data, i.e., in a non-phylogenetic framework, to draw inference from models that consider different predictor variables in different combinations. Adding the phylogenetic component, in theory, calls for a more extended exploitation of these techniques as several hypotheses about the phylogenetic history of species and about the mode of evolution should also be considered, all of which can be flexibly incorporated and combined with different statistical models. Here, we highlight some biological problems that inherently imply multimodel approaches and show how these problems can be tackled in the phylogenetic generalized least squares (PGLS) modeling framework based on information-theoretic approaches (e.g., by using Akaike's information criterion, AIC) or maximum likelihood. We present a conceptual framework of model selection for phylogenetic comparative analyses, where the goal is to generalize across models that involve different combinations of predictors, phylogenetic hypotheses, parameters describing the mode of evolution, and error structures. Although this overview suggests that a model selection strategy may be useful in several situations, we note that the performance of the approach in the phylogenetic context awaits further evaluation in simulation studies.

L. Z. Garamszegi (✉)

Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC,
Seville, Spain

e-mail: laszlo.garamszegi@ebd.csic.es

R. Mundry

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

e-mail: roger_mundry@eva.mpg.de

12.1 Introduction

The world is so complex that researchers are often confronted with the challenge of assessing a large number of biological explanations for a given phenomenon (Chamberlin 1890). Making drawing inference from multiple hypotheses traditionally involves the evaluation of the appropriateness of different statistical models that describe the relationship among the considered variables. This task can be seen as a model selection problem, and there are three general approaches that allow such inference based on statistical analysis. The approach that dominated applied statistics for decades is that of null-hypothesis significance testing (NHST) (Cohen 1994). Applying NHST, one typically states a null-hypothesis of no influence or no difference, which is then rejected or not based on a significance threshold (conventionally, $P = 0.05$ that specifies the probability that one would obtain the observed data given the null hypothesis were true). In this framework, nested multiple models can be examined in a stepwise fashion, in which terms can be eliminated or added based on their significance following a backward or forward process (but see, e.g., Mundry and Nunn 2008; Whittingham et al. 2006; Hegyi and Garamszegi 2011 for problems with stepwise model selection). The second approach is Bayesian inference where one considers a range of ‘hypotheses’ (e.g., model parameters) and incorporates some prior knowledge about the probability of the particular model parameter values to update one’s ‘belief’ in what are more and less likely model parameters (Congdon 2003; Gamerman and Lopes 2006). Bayesian inference has a long history, but only recent increases in computer power made its application feasible for a wide range of problems (for relevance for comparative studies, see Chaps. 10 and 11). The third, relatively recent approach to statistical inference is based on information theory (IT) (Burnham and Anderson 2002; Johnson and Omland 2004; Stephens et al. 2005). Here, a set of candidate models, which represent different hypotheses, is compared with regard to how well they fit the data. A key component of the IT approach is that the measure of model fit is penalized for model complexity (i.e., the number of estimated parameters), and, as such, IT-based inference aims at identifying models that represent a good compromise between model fit and model complexity. Most frequently, IT-based inference goes beyond simply choosing the best model (out of the set of candidate models) and allows accounting for model selection uncertainty (i.e., the possibility that several models receive similar levels of support from the data).

Although model selection is classically viewed as a solution to the problem caused by the large number of potential combinations of predictors that may affect the response variable, here we propose that the comparative phylogenetic framework involves a range of questions that require multimodel inference and approaches based on IT. In particular, we emphasize that in addition to the variables included in the candidate models, the models can also differ in terms of other parameters that describe the mode of evolution, or account for phylogenetic uncertainty and heterogeneities in sampling effort. In this chapter, we present general strategies for drawing inference from multiple evolutionary models in the framework of

phylogenetic generalized least squares (PGLS). We formulate our suggestions merely on a conceptual basis with the hope that these will stimulate further research that will assess the performance of the methods based on simulations. We envisage that such simulation studies are crucial steps before implementing model selection routines into the practice of phylogenetic modeling. Our discussion is accompanied with an Online Practical Material (hereafter OPM) available at <http://www.mpcm-evolution.com>, which demonstrates how our methodology can be applied to real data in the *R* statistical environment (R Development Core Team 2013).

12.2 The Fundamentals of IT-based Multimodel Inference

Given that a considerable number of primary and secondary resources discuss the details of the IT-based approach (Burnham and Anderson 2002; Claeskens and Hjort 2008; Garamszegi 2011; Konishi and Kitagawa 2008; Massart 2007), we avoid giving an exhausting description here. However, in order to make our subsequent arguments comprehensible for the general readership, we first provide a brief overview on the most important aspects of the approach.

12.2.1 Model Fit

The central idea of an IT-based analysis is to compare the fit of different models in the candidate model set (see below). However, it is trivial that more complex models show better fits (e.g., larger R^2 or smaller deviance). Hence, an IT-based analysis aims at identifying those models (in the set of candidate models; see below) that represent a good compromise between model complexity and model fit, in other words, parsimonious models. Practically, this is achieved by penalizing the fit of the models by their complexity. One way of doing this is to use Akaike's information criterion (AIC), namely

$$\text{AIC} = -2 \ln \mathcal{L}_{(\text{model}|\text{data})} + 2k, \quad (12.1)$$

where

$\mathcal{L}_{(\text{model}|\text{data})}$ maximum likelihood of the model given the data and the parameter estimates,

k the number of parameters in the model ($-2 \ln \mathcal{L}_{(\text{model}|\text{data})}$ is known as “deviance”).

Two models explaining the data equally well will have the same likelihood, but they might differ in the number of parameters estimated. Then, the model with the smaller number of parameters will reveal the smaller AIC (and the difference in the AIC values of the more complex and the simpler model will be twice the

difference of the numbers of parameters they estimate). Hence, in an IT-based analysis, the model with the smaller AIC is considered to be ‘better’ because it represents a more parsimonious explanation of the response investigated.¹ Noteworthy, some argue that AIC-based inference can select for overly complex models and suggest alternative information criteria (Link and Barker 2006). Here, we continue focusing on AIC with the notion that the framework can be easily tailored for other metrics.

The core result of an IT-based analysis is a set of AIC values associated with a set of candidate models. However, unlike P values, AIC values do not have an interpretation in themselves but receive meaning only by comparison with AIC values of other models, fitted to the exact same response. The model with the smallest AIC is the ‘best’ (i.e., best compromise between model parsimony and complexity) in the set of models. However, in contrast to an NHST analysis, it would be misleading to simply select the best model and discard the others. This is because the best model according to AIC (i.e., the one with the smallest AIC) might not be the model that explicitly describes the truth (in fact, it is unlikely to ever be). Such discrepancies can happen for various reasons, including stochasticity in the sampling process (i.e., a sample is used to draw inference about a population), measurement error in the predictors and/or the response, or unknown predictors not being in the model, to mention just a few. An analysis in the framework of a phylogenetic comparative analysis expands this list considerably to include, for instance, imperfect knowledge about the phylogenetic history or the underlying model of evolution (e.g., Brownian motion or Ornstein-Uhlenbeck). An IT-based analysis allows dealing elegantly with such model selection uncertainty by explicitly taking it into account (see below).

12.2.2 Candidate Model Set

A key component in an IT-based analysis is the candidate set of models to be investigated, which classically includes models with different combinations of predictors. The validity of the analysis is conditional on this set, and if the candidate model set is not a reasonable one, the results will be deceiving (Burnham and Anderson 2002; Burnham et al. 2011). Hence, the development of the candidate set needs much care and is a crucial and potentially challenging step of an IT-based analysis. First of all, different models might represent different research hypotheses. For instance, one might hypothesize that brain size might have co-evolved with social complexity (e.g., group size), ecological complexity (e.g., seasonality in food availability), or both. However, in biology, it is frequently not easy to come up with such a clearly defined set of potentially competing models,

¹ When drawing inference in an IT framework, it is essential to not mix it up with the NHST framework. Most crucially, it does not make sense to select the best model based on AIC and then test its significance or the significance of the predictors it includes.

and hence one frequently sees candidate model sets that encompass all possible models that can be built out of a set of predictors. Furthermore, in the context of phylogenetic comparative analysis, different models in the candidate set might represent different evolutionary models (e.g., Brownian motion or an Ornstein-Uhlenbeck process) or different phylogenies. It is important to emphasize that in a phylogenetic comparative analysis both these aspects (and also other ones) can be reflected in a single candidate set of models; that is, the candidate set might comprise models that represent combinations of hypotheses about the coevolution of traits, the model of evolution, and the phylogenetic history.

12.2.3 Accounting for Model Uncertainty

There are several ways of dealing with model selection uncertainty (i.e., with the fact that not only one model is unanimously selected as best). One way is to consider Akaike weights. Akaike weights are calculated for each model in the set and can be thought of as the probability of the actual model to be the best in the set of models (although there are warnings against such interpretations, e.g., see Bolker 2007). From Akaike weights, one can also derive the evidence ratio of two models, which is the quotient of their Akaike weights and tells how much more likely is one of the two models (i.e., the one in the numerator of the evidence ratio) to be the best model. Akaike weights can also be used to infer about the importance of individual predictors by summing Akaike weights for all models that contain a given predictor. The summed Akaike weight for a given predictor then can be considered analogous to the probability of it being in the best model of the set (see also Burnham et al. 2011; Symonds and Moussalli 2011).

12.2.3.1 Model Averaging

One can also use Akaike weights for model averaging of the estimated coefficients associated with the different predictors. Here, the estimated coefficients (e.g., regression slopes) are averaged across all models (or across a confidence set of best models²) weighted by the Akaike weights of the corresponding models (see also Burnham et al. 2011; Symonds and Moussalli 2011). Hence, an estimate of a coefficient from a model having a large Akaike weight contributes more to the

² Another way of dealing with model selection uncertainty is to consider the best model confidence set, which contains the models that can be considered as best with some certainty. Different criteria do exist to identify the best model confidence set among which the most popular are to include those models that differ in AIC from the best model by at most some threshold (e.g., 2 or 10) or, alternatively, to include those models for which their summed cumulative Akaike weights (from largest to smallest) just exceed 0.95. In this chapter, we do not consider such subjective thresholds further, and throughout the remaining discussion we refer to model averaging in a sense that it is made across the full model set.

averaged value of the coefficient. When using model averaging of the estimated coefficients, there are two ways of treating models in which a given predictor is not present: one is to simply ignore them (the ‘natural’ method), and one is to set the estimated coefficient to zero for models in which the given predictor is not included (the ‘zero’ method; Burnham and Anderson 2002; Nakagawa and Hauber 2011). Using the latter penalizes the estimated coefficient when it is mainly included in models with low Akaike weights, and to us, this seems to be the better method.

12.3 Model Selection Problems in Phylogenetic Comparative Analyses

There can be several biological questions involving phylogenies, which necessitate inference from more than one model that are equally plausible hypothetically. Most readers might have encountered such a challenge when judging the importance of different combinations of predictor variables. However, in addition to parameters that estimate the effects of different predictors, in a phylogenetic model, there are several other parameters that deal with the role of phylogenetic history or with another error term (e.g., within-species variance). The statistical modeling of these additional parameters often requires multiple models that differently combine them, even at the same set of predictors. Below, we demonstrate that in most of these situations the observer is left with the classical problem of model selection, when s/he needs to draw inferences from a pool of models based on their fit to the data. Accordingly, the same general framework can be applied: Competing biological questions are first translated into statistical models, and then, multimodel inference is used for generalization.

12.3.1 *Selecting Among Evolutionary Models with Different Combinations of Predictors*

The classical problem of finding the most plausible combination of predictors to explain interspecific variation in the response variable while accounting for the phylogeny of species is well exemplified in the comparative literature. Starting from a pioneering study by Legendre et al. (1994), a good number of studies exist that evaluate multiple competing models to assess their relative explanatory value and to draw inferences about the effects of particular predictors. Below, as an appetizer, we provide summaries of two of these studies to demonstrate the diversity of questions that can be addressed by using the model selection framework. In the OPM, we give the *R* code that can be easily tailored to any biological problem requiring an AIC-based information-theoretic approach.

Terribile et al. (2009) investigated the role of four environmental hypotheses mediating interspecific variation in body size in two snake clades. These hypotheses

emphasized the role of heat balance as given by the surface area-to-volume ratio, which in ectothermic vertebrates may influence heat conservation (e.g., small-bodied animals may benefit from rapid heating in cooler climates), habitat availability (habitat zonation across mountains limits habitat areas that ultimately select for smaller species), primary productivity (low food availability can reduce growth rate and delay sexual maturity, which would in turn result in small-bodied species in areas with low productivity), and seasonality (large-bodied species may be more efficient in adapting to seasonally fluctuating resources that often include periods of starvation). To test among these hypotheses, the authors estimated the extent to which the patterns of body size are driven by current environmental conditions as reflected by mean annual temperature, annual precipitation, primary productivity, and range in elevation. They challenged a large number of models with data and chose the best model that offered the highest fit relative to model complexity to draw inference about the relative importance of different hypotheses. This best model included all main predictors, but the amount of variation explained differed between Viperidae and Elapidae, the two snake clades investigated. Moreover, the relative importance for each predictor also varied, as indicated by the summed Akaike weights. Consequently, none of the proposed hypotheses was overwhelmingly supported or could be rejected, and the mechanisms constraining body size in snakes can even vary from one taxonomic group to another.

A recent phylogenetic comparative analysis of mammals focused on the determinants of dispersal distance, a variable of major importance for many ecological and evolutionary processes (Whitmee and Orme 2013). Dispersal distance can be hypothesized as a trait being influenced by several constraints arising from life history, a situation that necessitates multipredictor approaches. For example, larger body size can allow longer dispersal distances because locomotion is energetically less demanding for larger-bodied animals. Second, home range size may be important, as dispersing individuals of species using larger home ranges may need to move longer distances to find empty territories. Furthermore, trophic level, reflecting the distribution of resources, may mediate dispersal distance with carnivores requiring more dispersed resources than herbivores or omnivores. Intraspecific competition may also affect dispersal: species maintaining higher local densities may also show higher frequencies of distantly dispersing individuals which thereby encounter less competition. Finally, investment in parental activities can be predicted to negatively influence dispersal, as species that wean late and mature slowly will create less competitive conditions for their offspring than species with fast reproduction. To simultaneously evaluate the plausibility of these predictors, Whitmee and Orme (2013) applied a model selection strategy based on the evaluation of a large number of models composed of the different combinations (including their quadratic terms) of the considered predictors. Even the best-supported multipredictor models had low Akaike weights, indicating no overwhelming support for any particular model. Therefore, they applied model averaging to determine the explanatory role of particular variables, which indicated that home range size, geographic range size, and body mass are the most important terms across models.

12.3.2 Dealing with Phylogenetic Uncertainty: Inference Across Models Considering Different Phylogenetic Hypotheses

While phylogenetic comparative studies necessarily require a phylogenetic tree, the true phylogeny is never known and must be estimated from morphological or, more recently, from genetic data; thus, phylogenies always contain some uncertainty (see detailed discussion in Chap. 2). In several cases, more than one phylogenetic hypothesis (i.e., tree) can be envisaged for a given set of species, and it might be desirable to test whether the results found for a given phylogenetic tree are also apparent for other, similarly likely trees.

With GenBank data and nucleotide sequences for phylogenetic inference, the above problem is not restricted anymore to the comparison of a handful of alternative trees corresponding to different markers. Nonetheless, the reconstruction of phylogenies from the same molecular data still raises uncertainty issues at several levels. Different substitution models and multiple mechanisms can be considered for sequence evolution, each leading to different sets of phylogenies that can be considered (note that this is also a model selection problem). Moreover, even the same substitution model can lead to various phylogenetic hypotheses with similar likelihoods. As a result, in the recent day's routine, several hundreds or even thousands of phylogenetic trees are often available for the same list of species used in a comparative study. The most common way to deal with such a large sample of trees is the use of a single, consensus tree in the phylogenetic analysis. However, although this approximation is convenient from a practical perspective, using an 'average' tree does not capture the essence of uncertainty, which lies in the variation across the trees. The whole sample of similarly likely trees defines a confidence range around the phylogenetic hypothesis (de Villemereuil et al. 2012; Pagel et al. 2004).

For the appropriate treatment of phylogenetic uncertainty, one needs to incorporate an error component that is embedded in the pool of trees that can be envisaged for the species at hand. Martins and Hansen (1997) proposed that most questions in relation to the evolution of phenotypic traits across species can be translated into the same general linear model:

$$\mathbf{y} = \beta\mathbf{X} + \boldsymbol{\varepsilon}, \quad (12.2)$$

where

- y** is a vector of characters or functions of character states for extant or ancestral taxa,
- X** is a matrix of states of other characters, environmental variables, phylogenetic distances, or a combination of these,
- β** is a vector of regression slopes,
- ε** is a vector of error terms with an assumed structure.

ε is composed of at least three types of errors that can be assembled in a complex way: ε_S , the error due to common ancestry; ε_M , the error due to within-species variance or measurement error; and ε_P , the error due phylogenetic uncertainty. The regression technique based on PGLS when combined with maximum likelihood (ML) model fitting offers a flexible way to handle and combine the errors ε_S and ε_M (for example, they can be treated additively if they are independent, see Chaps. 5 and 7). However, simultaneously handling the third error, the one that is caused by phylogenetic uncertainty, ε_P , is more challenging, because it is not an independent and additive term (Martins 1996). Approaches based on Bayesian sampling that are discussed in Chap. 10 offer a potential solution. They allow the use of a large number of similarly likely phylogenetic trees by effectively weighting parameter estimates across their probability distribution and can also incorporate errors due to within-species variance (de Villemereuil et al. 2012). However, widely available Bayesian methods can be sensitive to prior settings and are not yet implemented in the commonly used statistical packages.

We propose a simpler solution and suggest that when combined with multimodel inference, approaches based on PGLS can be used to deal with uncertainties in the phylogenetic hypothesis. The underlying philosophy of this approach is that when a list of trees is available, each of them can be used to fit the same model describing the relationship between traits using ML. Subsequently, parameter estimates (e.g., intercepts and slopes) can be obtained from the resulting models, which can then be averaged with a weight that is proportional to the relative fit of the corresponding model to the data. The output will not only provide a single average effect (as is the case when using a single model fitted to a consensus tree) but will also include a confidence or error range as obtained from the variance of model parameters across models associated with different trees. This interval can be interpreted as a consequence of the uncertainty in the phylogenetic hypothesis, that is, the mean estimate (model-averaged slope, or the slope that is based on the consensus tree) with the associated uncertainty component (variance among particular slopes) will form the results together. The logic of analyzing the interspecific data on each possible phylogeny to obtain a sample of estimates and then to calculate summary statistics from this distribution was already proposed by Martins (1996). Our favored method differs with regard to that it applies a model-averaging technique to derive the mean and confidence interval from the frequency distribution of parameters. This can be important, because if the pool of the trees across which the models are fitted reflects the likelihood of particular trees explaining the evolution of taxa, the resulting model-averaged parameter estimates will also reflect this variation.

Although apparently different trees are used in each model, drawing inference across them does not violate the fundamentals of information theory that assumes that each model is fitted to the very same data. Different trees can be regarded as different hypotheses that arise from identical nucleotide sequence information. They are actually just different statistical translations of the same biological information and act like scaling parameters on the tree. The approach may be particularly useful when a large number of alternative trees are at hand (e.g., in the

form of a Bayesian sample originating from the same sequence data). When only a handful of phylogenies is available (e.g., from other published papers), model-averaged means and variances can also be calculated, but conclusions would be conditional on the phylogenies considered (i.e., some alternative phylogenies may have not been evaluated). Furthermore, fitting models to trees that correspond to different marker genes calls for philosophical issues about the underlying assumption concerning the use of the same data.

In Fig. 12.1, we illustrate how our proposed model averaging works in practice (the underlying computer codes are available in the OPM). In this example, we tested for the evolutionary relationship between brain size and body size in primates by using PGLS regression methods with ML estimation of parameters. We considered a sample of reasonable phylogenetic hypotheses in the form of 1,000 trees as obtained from the *10KTrees Project* (Arnold et al. 2010). When using the consensus tree from this tree sample, we can estimate that the phylogenetically corrected allometric slope is 0.287 ($SE = 0.039$, solid line in Fig. 12.1). However, using different trees from *10KTrees* pool in the model provides slightly different results for the phylogenetic relationship between traits, as the obtained slopes vary (gray lines in the left panel of Fig. 12.1). The model-averaged regression slope yields 0.292 (model-averaged $SE = 0.041$, dashed line in the left panel of Fig. 12.1). This mean estimate is quite close to what one can obtain based on the consensus tree, but the variation between the particular slopes corresponding to different trees in the sample delineates some uncertainty around the averaged allometric coefficient. Few models in the ML sample provide extreme estimates (note that, model fitting with one particular tree even results in a negative slope, left panel of Fig. 12.1). However, these models were characterized by a very poor model fit; thus, their potential influence is downweighted in the model-averaged mean estimate.

The benefit of using the AIC-based method to account for phylogenetic uncertainty over Bayesian approaches is that the former does not require prior information on model parameters that would affect the posterior distribution of parameters, an issue that is often challenging in the Bayesian framework (Congdon 2006) and that is also demonstrated in Fig. 12.1. In the right panel, we applied Markov chain Monte Carlo (MCMC) procedure to estimate the posterior distribution of parameter values from the same PGLS equation by using (Pagel et al. 2004; Pagel and Meade 2006) *BayesTraits* with the same interspecific data and pool of trees (see also Chap. 10). Supposing that we have no information to make an expectation about the range where parameter estimates should fall, we are constrained to use flat and uniform prior distributions (e.g., spanning from -100 to 100).³ When we used MCMC to

³ It may not be necessarily applied to the current biological example, because allometric regressions are intensively studied (e.g., Bennett and Harvey 1985; Hutcheon et al. 2002; Iwaniuk et al. 2004; Garamszegi et al. 2002). Therefore, results from a large number of studies on other vertebrate taxa may be used to define a narrower and more informative prior. However, in this example simulated on the general situation when no preceding information on the expected relationship is available. Note that technically *BayesTraits* only allows uniform priors for continuous data.

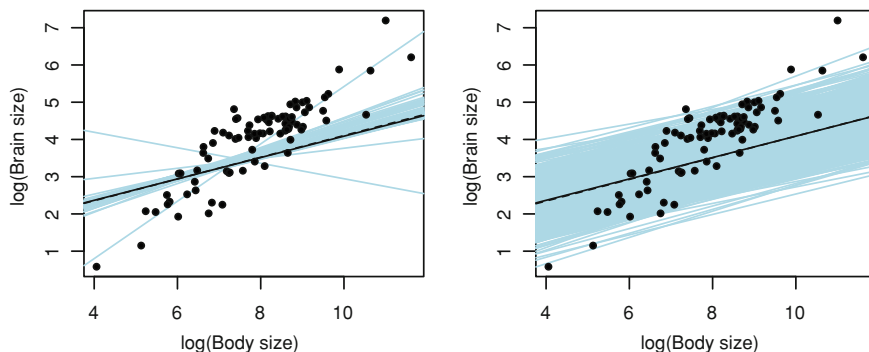


Fig. 12.1 Estimated regression lines for the correlated evolution of two traits (body size and brain size in primates) when different hypotheses for the phylogenetic relationships of species are considered and when ML (*left panel*) or MCMC (*right panel*) estimation methods are used in the AIC-based or Bayesian framework, respectively. *Gray lines* show the regression slopes that can be obtained for alternative phylogenetic trees (*left panel* 1,000 ML models fitted to different trees, *right panel* 1,000 models that the MCMC visited in the Bayesian framework). The alternative trees originate from a sample of 1,000 similarly likely trees that can be proposed for the same nucleotide sequence data (Arnold et al. 2010). The *dashed bold line* represents the slope estimate that can be derived by model averaging over the particular ML estimates (*left panel*) or by taking the mean of the posterior distribution from the MCMC sample of 1,000 models (*right panel*). Both methods provide a mean estimate over the entire pool of trees by incorporating the uncertainty in the underlying phylogenetic hypothesis. The *solid bold line* shows the regression line that can be fitted when the single consensus tree is used. The model-averaged slope, the mean of the posterior distribution, and the one that corresponds to the consensus highly overlap in this example (which may not necessarily be the case). However, the precision by which the mean can be estimated is different between ML and MCMC approaches, as the latter introduces a larger variance in the slopes in the posterior sample

sample from a large number of models with different parameters and trees and took 1,000 estimates from the posterior distribution of slopes, we detected that the estimate is accompanied by a considerable uncertainty (Fig. 12.1, right panel). For comparison, the 95 % confidence interval of the allometric coefficients obtained from the ML sample is 0.278–0.312, while it is 0.211–0.373 for the MCMC sample (i.e. the confidence interval obtained from the Bayesian framework is almost five times wider than that from the AIC-based inference). Consequently, the Bayesian approach introduces an unnecessary uncertainty due to the dominance of the prior distribution on the posterior distribution.

Another benefit of using ML model fitting over a range of phylogenetic hypotheses in conjunction with model averaging is that by doing so we can exploit the flexibility of the PGLS framework. For example, as we discussed above, one can evaluate different sets of predictor variables when defining models, or as we explain below, one can also take into account additional error structures (e.g., due to within-species variation) or different models of trait evolution (e.g., Brownian motion or an Ornstein-Uhlenbeck process). These different scenarios can be simultaneously considered during model definition, but can also be combined with

alternative phylogenetic trees (some examples are given in the OPM). This will result in a large number of candidate models representing different evolutionary hypotheses, over which model averaging may offer interpretable inference.

Box 12.1 A simulation strategy for testing the performance of multimodel inference

The behaviour of the AIC-based framework to account for phylogenetic uncertainty requires simulation studies that consist of the following steps. First, one needs to simulate a tree for a considered number of species and under some scenario for the underlying model (e.g., time-dependent birth–death model or just a random tree). The next step is then to simulate species-specific trait data along the branches of the generated phylogeny. To obtain simulated tip values, we also need to consider a model to describe the evolutionary mechanism in effect (e.g., Brownian motion or an OU process). We might also consider other constraints for trait evolution, for example, by defining a correlation structure (a zero or a nonzero covariance) for two coevolving traits. These parameters will serve as generating values, and the underlying tree and the considered covariance structure will reflect the truth that we want to recover in the simulation. If the interest is to examine the performance of the model-averaging strategy to account for phylogenetic uncertainty, we need to generate a sample of trees that integrates a given amount of variance (e.g., both the topology and branch lengths are allowed to vary to some pre-defined degree). For each simulation, we can then fit a model estimating the association between the two traits by controlling for phylogenetic effects. The phylogeny used in this model to define the expected variance–covariance structure on the one hand can be the consensus tree calculated for the whole sample of trees. On the other hand, we can also fit the model to each tree in the sample and then do a model averaging to obtain an overall estimate for the parameter of interest (e.g., slope or correlation as calculated from the model). By simulating new trait data (and optionally new pools of trees), we can repeat the whole process a large number of times (i.e., 1,000 or 10,000 times). At each iteration, we will, hence, obtain estimates (either over the consensus tree or over the entire sample of trees through model averaging) for the parameter of interest. Finally, we can compare the distribution of these parameters over simulations with the generating parameter state. The difference between the mean of the distribution and the generating value will inform about bias of the approach, while the width of the distribution informs about precision (the uncertainty in parameter estimation).

As an important cautionary note, we emphasize that the performance of the AIC-based method based on model averaging still requires further assessment with both simulated and empirical data. In Box 12.1, we describe the philosophy of an

appropriate simulation study that can efficiently test the performance of averaging parameters over a large number of models corresponding to different hypotheses about phylogenies or other evolutionary patterns.

12.3.3 Variation Within Species

One of the advantages of the PGLS approach is that it allows accounting for within-species variation, which broadly includes true individual-to-individual or population-to-population variation, and also other sources of variation in the estimates of taxon trait values such as measurement error (see Ives et al. 2007; Hansen and Bartoszek 2012; and Chap. 7). Given that these different sources of error can be translated into different models, selecting among these may also be performed by model selection. Does a model that considers within-species variation perform better than a model that neglects such variation? Such simple questions can be developed further as by applying the general Eq. 12.2, in which different error structures (e.g., phylogenetic errors and measurement errors, or measurement error on one trait may correlate with measurement error on another trait) can be combined in different ways.

For example, when considering intraspecific variation in an interspecific context, we can evaluate at least four models and compare them based on their relative fit (here, we are only focusing on the main logic; for details on how to take into account intraspecific variation, see Chap. 7). First, as a null model, we can fit a model that is defined as an ordinary least squares regression (i.e., with a covariance matrix for the residuals based on a star phylogeny and measurement errors being zero). Then, we can investigate a model that does account for phylogeny but not for the uncertainty in the species-specific trait values (conditioned on the true phylogeny, while measurement errors are assumed to be equal to zero), and also a model that considers measurement error but ignores the phylogenetic structure (unequal and nonzero values along the diagonal of the measurement error matrix, and a phylogenetic covariance matrix representing a star phylogeny). Finally, we wish to test a model that includes both error structures (the joint variance–covariance matrix reflecting the phylogeny and the known measurement errors). To obtain parameter estimates and to make appropriate evolutionary conclusions, the observer can rely on the model that offers the best fit to the data as indicated by the corresponding AIC (but only if one model is unanimously supported over the others). Such a simple model selection strategy can be followed in the OPM of Chap. 7. Note that for the appropriate calculation of AIC according to Eq. 12.1 (and thus for the meaningful comparison of models), it is required that the number of estimated parameters is determined, which may be difficult when parameters in both the mean and variance components are estimated. This problem can be avoided by a smart definition of models (e.g., by defining analog models that estimate the same number of parameters even if these are known to be zero). In any case, the approach requires further validation by simulation studies.

Methods that account for within-species variation can also deal with a situation, in which different sample sizes (n) are available for different species, implying that data quality might be heterogeneous (i.e., larger errors in taxa with lower sampling effort; see Chap. 7 for more details). For example, if within-species variances or standard errors are unknown, one can fit a measurement error model by using $1/n$ as an approximation of within-species variance.

Another way to incorporate heterogeneous sampling effort across species into the comparative analyses is to apply statistical weights in the model. A particular issue arising in this case is that weighting can result in a large number of models (with potentially different results). For example, by using the number of individuals sampled per taxon as statistical weights in the analysis, we enforce weights differing a lot between species that are already sampled with sufficient intensity (e.g., the underlying sample size is 20 at least) but still differ in the background research effort (e.g., 100 individuals are available for one species, while 1,000 for another). However, if we log- or square-root-transform within-species sample sizes and use these as weights, more emphasis will be given on differences between lightly sampled species than on differences between heavily sampled species. Continuing this logic, and applying the appropriate transformation, we can create a full gradient that scales differences in within-species sample sizes along a continuum spanning from no differences to large differences between species with different within-species sample sizes.

For illustrative purposes (Fig. 12.2, left panel), we have created such a gradient of statistical weights by the combination of the original species-specific sample sizes (n) and an emphasis parameter (the ‘weight of weights’) that we will label ω ; ω is simply an elevation factor that ranges from 1 to $1/100$ and defines the exponent of n . If ω is 1, the original sample sizes are used as weights in the analysis. If ω is $1/2 = 0.5$, the square-root-transformed values serve as weights, and differences between small sample sizes become more emphasized than differences between species with larger sample sizes. $\omega = 1/\infty = 0$ represents the scenario in which all species are considered with equal weight ($n^0 = 1$), so the model actually represents a model that does not take into account heterogeneity in sampling effort. Other transformations on sample sizes based on different scaling factors that create a gradient can also be envisaged.

Using the parameter ω , we provide an example for the study of brain size evolution based on the allometric relationship with body size (Fig. 12.2 right panel, the associated R code is provided in the OPM). We have created a set of phylogenetic models that also included statistical weights in the form of the ω exponent of within-species sample sizes. The scaling factor ω varied from 0 to 1. We challenged these models with exactly the same data using ML; thus, model fit statistics (e.g., AIC) are comparable. We found that when accounting for phylogenetic relationships, the $\omega = 0$ scenario provides by far the best fit, implying that weighting species based on sample size is not important. This finding is not surprising, given that both traits, brain size and body size, have very high repeatability ($R > 0.8$). Thus, relatively few individuals provide reliable information on the species-specific trait values. Giving different weights to different

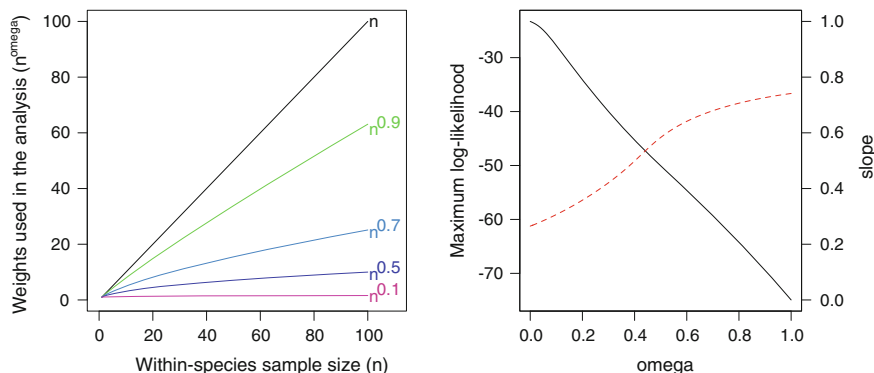


Fig. 12.2 The effect of using different transformations of the number of individuals as statistical weights. The *left panel* shows how differences between species are scaled when the underlying within-species sample sizes are transformed by exponentiating them with the exponent ω . ω varied between 1 (untransformed sample sizes maximally emphasizing differences in data quality between species) and 0 (all species have the same weight; thus, data quality is considered to be homogeneous). The *right panel* shows the maximized log-likelihood (*black solid line*) and the estimated slope parameters of models (*red dashed line*) for the brain size/body size evolution that implement weights that are differently scaled by ω

species based on the underlying sample sizes would actually be misleading; the use of different ω values leads to qualitatively different parameter estimates for the slope of interest (Fig. 12.2, right panel). This indicates that the results and conclusions are highly sensitive to how differences in sampling effort are treated in the analysis. Note that the above exercise only makes sense if (1) there is a considerable variation in within-species sample size and (2) if there is no phylogenetic signal in sample sizes. These assumptions require some diagnostics prior to the core phylogenetic analysis (see an example in Garamszegi and Møller 2012).

Garamszegi and Møller (2007) relied on a similar approach in a study of the ecological determinants of the prevalence of the low pathogenic subtypes of avian influenza in a phylogenetic comparative context. It was evident that there was a vast variance in sampling effort across species, as within-species sample size varied between 107 and 15,657. Therefore, when assessing the importance of the considered predictors, it seemed unavoidable to simultaneously account for common ancestry and heterogeneity in data quality. The application of the strategy of scaling the weight factor yielded that, contrary to the above example, the highest ML was achieved by a certain combination of the weight and phylogenetic scaling parameters. That finding was probably driven by the relatively modest repeatability of the focal trait (prevalence of avian influenza), suggesting that, due to different sample sizes, data quality truly differed among species.

We advocate that the importance of a correction for sample size differences between species is an empirical issue that can vary from data to data, which could (and should) be evaluated. We provided a strategy by which the optimal scaling of weight factors can be determined. In these examples, an unambiguous support could

be obtained for a single parameter combination. However, we can imagine situations, in which more than one model offers relatively good fit to the data, in which case inference would be better made based on model averaging (corresponding codes are given in the OPM) instead of focusing on a single parameter combination. Furthermore, the evaluation of the sample size scaling factor (as well as the assessment of within-species variance) can be combined with the evaluation of alternative phylogenetic hypotheses, as the IT-based framework offers a potential for the exploration of a multidimensional parameter space. Accordingly, each scaling factor can be incorporated into various models considering different phylogenies (or each phylogenetic tree can be evaluated along a range of scaling factors), and the model selection or model-averaging routines may be used for drawing inference from the resulting large number of models. Again, the performance of these methods necessitates further investigations by simulation approaches.

12.3.4 Dealing with Models of Evolution

12.3.4.1 Comparison of Models for Different Evolutionary Processes

Several phylogenetic comparative methods (e.g., phylogenetic autocorrelation, independent contrasts, and PGLS) assume that the model of trait evolution can be described by a Brownian motion (BM) random-walk process. However, this assumption might be violated in certain cases, and other models might need to be considered. For example, a model based on the Ornstein-Uhlenbeck (OU) process is another choice that takes into account stabilizing selection toward a single or multiple adaptive optima (Butler and King 2004; Hansen 1997; see also discussion in Chaps. 14 and 15). Other model variants of the BM or OU models, such as the model for accelerating/decelerating evolution (AC/DC, Blomberg et al. 2003) or the model for a single stationary peak (SSP, Harmon et al. 2010), can also be envisaged.

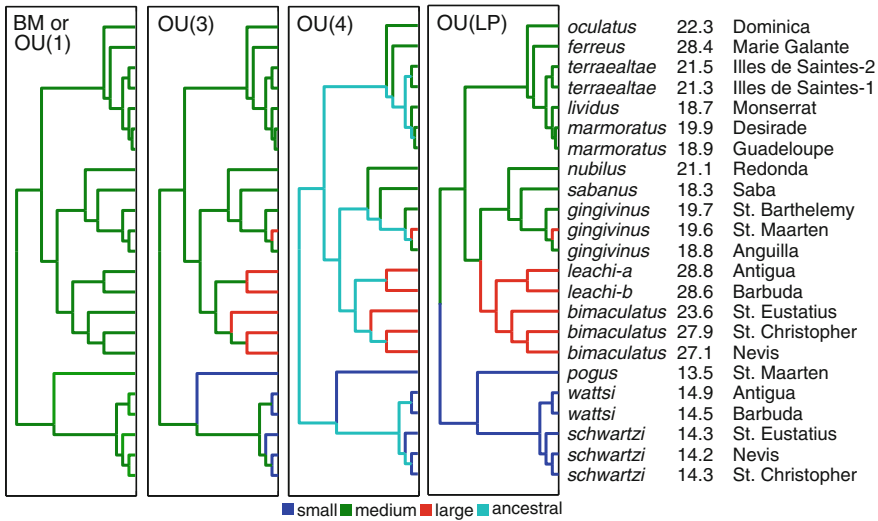
Given that we usually do not have prior information about the ‘true’ model of evolution, alternative hypotheses about how traits evolved could be considered in statistical modeling. If the considered evolutionary models are mathematically tractable (there are cases when they are not! see Kutsukake and Innan 2013), they can be translated into statistical models suitable for a model selection framework. Accordingly, each model can be fitted to the data, and once finding the one that offers the highest explanatory power, it can be used for making evolutionary inferences. This does not only control for phylogenetic relationships, but knowing which is the most likely evolutionary model can give insight about the strength, direction, and history of evolution acting on different taxa. Importantly, when using a model selection strategy in this context, the observer aims at identifying the single best model that accounts for the mode of evolution; thus, model averaging may not make sense. Therefore, for making robust conclusions, we need to obtain results in which models are well separated based on their AIC in a way that

one model reveals overwhelming support as compared to the others. Alternatively, one could use model averaging to estimate regression parameters (and also to estimate the parameters of the evolutionary model if parameters of different models are analogous), thus accounting for the uncertainty in the assessment of the underlying evolutionary process.

To demonstrate the use of model selection to choose among different evolutionary models, we provide an example from Butler and King (2004), but other illustrative analyses are also available in the literature (Collar et al. 2009, 2011; Harmon et al. 2010; Hunt 2006; Lajeunesse 2009; Scales et al. 2009). Butler and King (2004) re-examined character displacement in *Anolis* lizards on the Lesser Antilles, where lizards live either in sympatry or in allopatry. Where two species coexist, these differ substantially in size, while on islands that are inhabited by only one species, lizards are of intermediate size. Therefore, one can hypothesize that body size differences on sympatric islands result from character displacement (i.e., when two intermediate-sized species came into contact with one another when colonizing an island, they subsequently diverged into a different direction). This hypothesis can be evaluated using alternative models of body size evolution that differ in the degree of how they incorporate processes due to directional selection and character displacement. The authors, therefore, evaluated five different models: (1) BM; (2) an OU process with a single optimum; (3) an OU process with three optima corresponding to large, intermediate, and small body size; (4) another OU model that includes an additional parameter to the three-optima model to deal with the adaptive regimes occurring on the internal branches as an estimable ancestral state; and (5) a model implementing a linear parsimony reconstruction of the colonization events (arrival history of species on the islands). Only the last model assumes character displacement. These models were compared by different methods including AIC, a Bayesian (Schwarz's) information criterion (SIC), and likelihood ratio tests that unanimously revealed that the best-fitting model was the OU model with the reconstructed colonization events (Fig. 12.3). Altogether, the results support the hypothesis that character displacement had an effect on the evolution of body size in *Anolis* lizards that colonized the Lesser Antilles.

12.3.4.2 Parameterization of Models

Another way to cope with the mode of evolution and to improve the fit of any model can be achieved by the appropriate setting of parameters that describe the fine details of the evolutionary process. For example, BM models can be adjusted using the parameters κ , δ , or λ that apply different branch-length transformations on the phylogeny (e.g., κ stretches or compresses phylogenetic branch lengths and thus can be used to model trait evolution along a gradient from punctuational to gradual evolution, while δ scales overall path lengths in the phylogeny and thus can be used to characterize the tempo of evolution) or that assess the contribution of the phylogeny (λ weakens or strengthens the phylogenetic signal in the data)



	BM	OU(1)	OU(3)	OU(4)	OU(LP)
-2lnL	-34.66	-34.66	-40.21	-47.22	-49.69
AIC	-30.66	-26.66	-28.21	-33.22	-37.69
LR		0	5.55	12.56	15.03
P		1	0.24	0.028	0.0046

Fig. 12.3 Graphical representation of five evolutionary models considered for the evolution of body size in Anolis lizards inhabiting the islands of Lesser Antilles. *BM* Brownian model; *OU(1)* Ornstein-Uhlenbeck process with a single optimum; *OU(3)*, *OU(4)* Ornstein-Uhlenbeck process with three or four optima, respectively i.g., Ornstein-Uhlenbeck process with four optima, one of which is an ancestral state; *OU(LP)* Ornstein-Uhlenbeck process with implementing a linear parsimony reconstruction of the colonization events, which thus considers character displacement. The table shows the model fit statistics of different models: deviance, Akaike’s information criterion (see Eq. 12.1), and likelihood ratio test comparing the given model with the *BM* model (LR and the associated *P* values). Modified from Butler and King (2004) with the permission of the authors and University of Chicago Press

(Pagel 1999). Furthermore, the importance of the rates of evolutionary change in character states can also be assessed via estimation of the corresponding parameter (Collar et al. 2009; O’Meara et al. 2006; Thomas et al. 2006). Finally, *OU* models also operate with particular parameters (such as α for the strength of selection and θ for the optimum) that can take different values (Butler and King 2004; Hansen 1997, see also discussion in Chaps. 14 and 15).

The parameterization of models is a task that requires the investigator to choose among alternative models with different parameter settings, which is typically a model selection problem. This task is usually addressed with likelihood ratio tests, in which a null model (e.g., with a parameter set to be zero) is contrasted with an

alternative model (e.g., with a parameter set to a nonzero value). If the test turns out significant, the alternative model is accepted and used for further analyses (e.g., tests for correlations between traits) and for making evolutionary implications. Another strategy is to evaluate the ML surface of the parameter space and then set the parameter to the value where it reveals the maximum likelihood (i.e., the strategy that most PGLS methods apply). Furthermore, AIC-based information-theoretic approaches can be used to obtain the parameter combinations that offer the best fit to the data.⁴

However, such a best model approach is not always straightforward. Parameter states can span a continuous scale, and it is possible that a broad range of parameter values are similarly likely. For example, the optimal phylogenetic scaling parameter λ is usually estimated using maximum likelihood. This estimation might be robust if the peak of the likelihood surface is well defined (i.e., few parameter states in a narrow range have a very high likelihood, while the remaining spectrum falls into a small likelihood region, Fig. 12.4, upper panels). Our experience, however, is that the likelihood surfaces are rather flat and vary considerably if single species are added or removed from the analysis (especially at modest interspecific sample sizes, Freckleton et al. 2002). This means that a broad range of parameter values describe the data similarly well (Fig. 12.4, lower panels), thus arbitrarily choosing a single parameter value on a flattish surface for further analysis may be deceiving.

We suggest that such uncertainty in parameter estimation can easily be incorporated using model averaging. Applying the philosophy that we followed for dealing with multiple trees or scenarios for the correction for heterogeneous data quality, we can also estimate the parameters of interest (e.g., ancestral state, slope, or correlation between two traits) at a wide range of the settings of the evolutionary parameters. Given that IT-based approaches typically compare sets of discrete models, we need to create a large number of categories for the continuous parameter (e.g., by defining a finite number, such as 100 or 1,000, bins for λ in increasing order between the interval of 0–1) that can be used to condition different models. Then, inference across this large number of models based on their relative fit to the data can be made, and given that intermediate states between the large number of categories are meaningful, interpretations can be extended to a continuous scale. Therefore, evolutionary conclusions can be formulated based on the parameter estimates that are averaged across models receiving different levels of support instead of obtaining them from a single model. In theory, λ can be model averaged as well, but when the maximum likelihood surface is flat (meaning that many models with different λ s will have similar AIC), deriving a single mean estimate may be misleading. In such a situation, only estimates together with their model-averaged standard errors (or confidence intervals) make sense.

In the OPM, we show for λ how this model averaging works in practice. We also provide examples for the case when the exercise for model parameters is

⁴ As long as the number of parameters is equal, AIC and ML reveal the same.

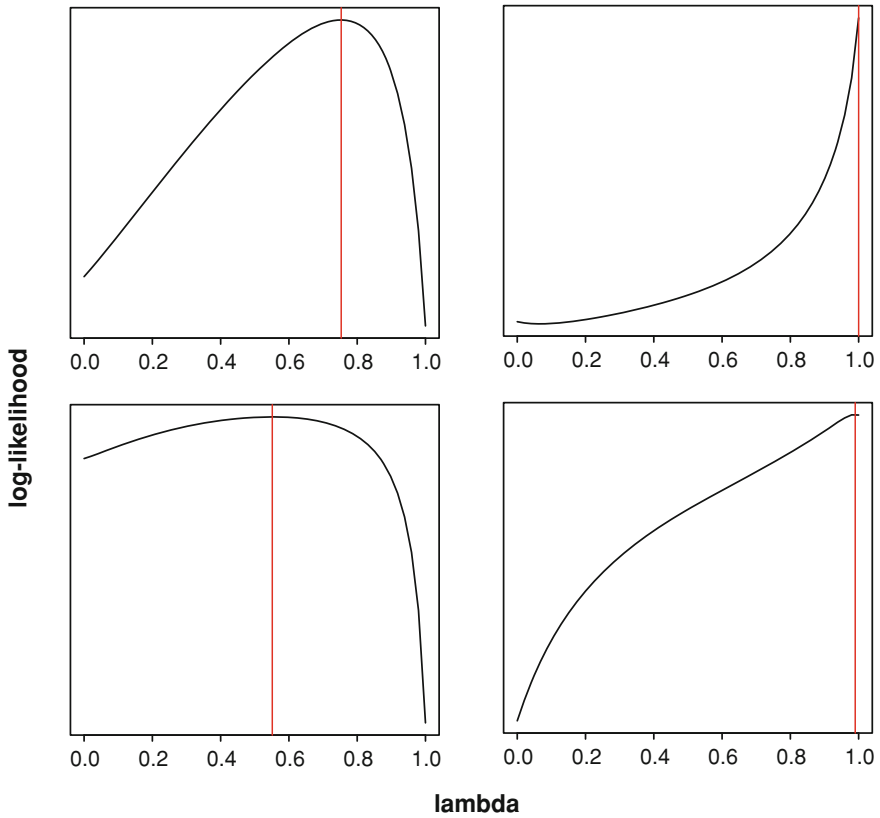


Fig. 12.4 Typical shapes of maximum likelihood surfaces of the phylogenetic scaling factor lambda (λ). The *upper figures* show two examples, in which the surface has a distinct peak and only a narrow range of parameter values are likely. In contrast, the *bottom graphs* depict two cases in which the likelihood surface is rather flat, thus incurring a considerable uncertainty when choosing a single value. *Vertical red lines* give the values at the maximum likelihood. For the illustrative purposes, it is assumed that y-axes have the same scale

combined with multimodel inference for statistical weights (Fig. 12.5). We keep on emphasizing that our suggestions merely stand on theoretical grounds; the performance of model averaging in dealing with the uncertainty of model parameterizations awaits future tests (based on both empirical and simulated data).

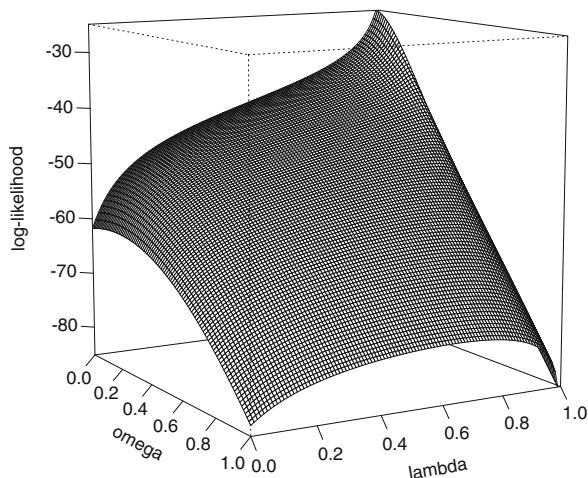


Fig. 12.5 Likelihood surface when the phylogenetic signal (lambda, λ) and the data heterogeneity (omega, ω) parameters are estimated in a set of models using different parameter combinations for the brain size/body size evolution in primates (data are shown in Fig. 12.1). The surface shows the log-likelihoods of a large number of fitted models that differ in their λ and ω parameters. These parameters are allowed to vary between 0 and 1 (with steps of 0.01) in all possible combinations. For a definition of ω , see Fig. 12.2

12.3.5 The Performance of Different Phylogenetic Comparative Methods

The logic of model selection can also be applied to assess whether any particular comparative method is more appropriate than others. For example, in a meta-analysis, Jhwueng (2013) estimated the goodness of fit of four phylogenetic comparative approaches. He collected more than a hundred comparative datasets from the published literature, to which he applied the following methods to estimate the phylogenetic correlation between two traits: the non-phylogenetic model (i.e., treating the raw species data as being independent), the independent contrasts method (Felsenstein 1985), the autocorrelation method (Cheverud et al. 1985), the PGLS method incorporating the Ornstein-Uhlenbeck process (Martins and Hansen 1997), and the phylogenetic mixed model (Hadfield and Nakagawa 2010; Lynch 1991). Model fits obtained for different approaches were compared based on AIC, which revealed that the non-phylogenetic model and the independent contrasts model offered the best fit. However, the parameter estimates for the phylogenetic correlation were quite similar across models, indicating that the studied comparative methods were generally robust to describe evolutionary patterns present in interspecific data.

12.4 Further Applications

So far, we mostly focused on the potential that the IT framework provides in association with the PGLS framework, when models are fitted with ML. However, multimodel inference also makes sense in a broader context, and related issues are known to exist in a range of other phylogenetic situations. We provide some examples below (without the intention of being exhaustive), but further applications can also be envisaged. This short list may illustrate that the benefits of multimodel inference can be efficiently exploited in relation to interspecific data.

A typical model selection problem is present in phylogenetics, when the interest is to find the best model that describes patterns of evolution for a given nucleotide or amino acid sequence. As briefly discussed in Chap. 2 (but see in-depth discussion in Alfaro and Huelsenbeck 2006; Arima and Tardella 2012; Posada and Buckley 2004; Ripplinger and Sullivan 2008), several models have been developed to deal with different substitution rates and base frequencies that ultimately influence the evolutionary outcome. The reliance on different models for phylogenetic reconstructions can result in phylogenetic trees that vary in their branching pattern and the underlying stochastic processes of nucleotide sequence changes that generate branch lengths. Given that a priori information about the appropriateness of different evolutionary models is generally lacking, those who wish to establish a phylogenetic hypothesis from molecular sequences are often confronted with a model selection problem. Accordingly, several evolutionary models need to be fitted to the sequence data, and the one that offers the best fit (e.g., as revealed by likelihood ratio test or an AIC-based comparison or Bayesian methods) should be used for further inferences about the phylogenetic relationships.

An intriguing example for the application of IT approaches in the phylogenetic context is the use of likelihood methods to detect temporal shifts in diversification rates. By fitting a set of rate-constant and rate-variable birth–death models to simulated phylogenetic data, Rabosky (2006) investigated which rate parameter combination (e.g., rate constant or rate varying over time) results in the model with the lowest AIC. The results suggested that selecting the best model in this way causes inflated Type I error, but when correcting for such error rates, the birth–death likelihood approach performed convincingly.

Eklöf et al. (2012) applied IT methods to understand the role of evolutionary history for shaping ecological interaction networks. The authors approached the effect of phylogeny by partitioning species into taxonomic units (e.g., from kingdom to genus) and then by investigating which partitioning best explained the species' interactions. This comparison was based on likelihood functions that described the probability that the considered partition structure reproduces the real data obtained for nine published food webs. Furthermore, they also used marginal likelihoods (i.e., Bayes factors) to accomplish model selection across taxonomic ranks. The major finding of the study was that models considering taxonomic

partitions (i.e., phylogenetic relationships) offered better fit to the data, and food webs are best explained by higher taxonomic units (kingdom to class). These results show that evolutionary history is important for understanding how community structures are assembled in nature.

Depraz et al. (2008) evaluated competing hypotheses about the postglacial recolonization history of the hairy land snail *Trochulus villosus* by using AIC-based model selection. They compared four refugia hypotheses (two refugia, three refugia, alpine refugia, and east–west refugia models) that could account for the phylogeographic history of 52 populations. The four hypotheses were translated into migration matrices, with maximum likelihood estimates of migration rates. These models were challenged with the data, and Akaike weights were used to make judgments about relative model support. This exercise revealed that the model considering the two refugia hypothesis overwhelmingly offered the best fit.

In a phylogenetic comparative study based on ancestral state reconstruction, Goldberg and Igić (2008) investigated ‘Dollo’s law’ which states that complex traits cannot re-evolve in the same manner after loss. When using simulated data and an NHST approach (likelihood ratio tests), they found that in most of the cases the true hypothesis about the irreversibility of characters was falsely rejected. However, when using appropriate model selection (based on AIC-based IT methods), the false rejection rate of ‘Dollo’s law’ was reduced.

Alfaro et al. (2009) developed an algorithm they called MEDUSA, which is an AIC-based stepwise approach that can detect multiple shifts in birth and death rates on an incompletely resolved phylogeny. This comparative method estimates rates of speciation and extinction by integrating information about the timing of splits along the backbone of a phylogenetic tree with known taxonomic richness. Diversification analyses are carried out by first finding the maximum likelihood for the per-lineage rates of speciation and extinction at a particular combination of phylogeny and species richness and then comparing these models across different combinations.

Further examples, e.g., for detecting convergent evolution based on stepwise AIC (Ingram and Mahler 2013) and for revealing phylogenetic paths based on the C-statistic Information Criterion (von Hardenberg and Gonzalez-Voyer 2013) can be found in Chaps. 18 and 8, respectively.

12.5 Concluding Remarks

What we have proposed here are several approaches to exploit the strengths of IT-based inference in the context of phylogenetic comparative methods. Using IT methods such as model selection in combination with phylogenetic comparative methods seems to offer the potential to elegantly solve problems which otherwise would be hard to tackle. Other IT methods such as model averaging allow dealing with phylogenetic uncertainty by explicitly incorporating it into the analysis and

exploring to what extent it compromises certainty about the results. Taken together, IT-based methods offer a great potential since they relieve researchers from the need of making arbitrary and/or poorly grounded decisions in favor of one or the other model. Instead, they allow dealing easily with such uncertainties or, at least, allow an assessment of their magnitude (among the set of potential models). Uncertainty is at the heart of our understanding about nature; thus, statistical methods are needed that appreciate this attribute instead of neglecting it.

We need to stress, though, that our propositions are based on theoretical grounds and need to be tested before they can be trusted. Particularly, simulation studies (e.g., along the design in Box 12.1) seem suitable for this purpose since they allow to investigate to what extent our propositions are able to reconstruct ‘truth’ which otherwise (i.e., in the case of using empirical data) is simply unknown. Simulation studies are warranted because the use of AIC (and other IT metrics) to non-nested models (which was largely the case here) is somewhat controversial (Schmidt and Makalic 2011). Another cautionary remark is that we refrained ourselves to suggest that only the IT-based model selection can be used to address the problems we raised. We envisage this discussion to serve as an initiative for comparative studies to consider the suggested methods as additions to the already existing toolbox, which yet await further exploitation.

Since the philosophy of IT-based inference is rather different from that of the classical NHST approach and since the two approaches are quite frequently mixed in an inappropriate way (e.g., selecting the best model using AIC and then testing it using NHST), we feel that some warnings on the use of the IT approach might be useful (particularly for those who were trained in NHST): IT-based inference does not reveal something like a ‘significance,’ and the two approaches must not be combined (Burnham and Anderson 2002; Mundry 2011). In the context of our propositions, this means that at least part of them naturally preclude the use of significance tests. This is particularly the case when sets of models with different combinations of predictors and/or sets of different phylogenetic trees are investigated. The end result of such an exercise is a number of AIC values associated with a set of models. Selecting the best model using AIC and then testing its significance is inappropriate. Rather, one could model average the estimates and their standard errors (but not the P values!) or also the fitted values and explore to what extent these vary across the different trees. Furthermore, one could use Akaike weights to infer about the relative importance of the different predictors. However, some of our proposed approaches might not necessarily and completely rule out the use of classical NHST. In fact, we do not argue against using NHST, which we regard as a scientifically sound approach if used and interpreted correctly. What we recommend is to not combine the use of significance tests with any of the approaches we suggested and draw inference solely on the basis of IT methods (e.g., Akaike weights or evidence ratios).

References

- Alfaro ME, Huelsenbeck JP (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst Biol* 55(1):89–96. doi:[10.1080/10635150500433565](https://doi.org/10.1080/10635150500433565)
- Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc Natl Acad Sci*. doi:[10.1073/pnas.0811087106](https://doi.org/10.1073/pnas.0811087106)
- Arima S, Tardella L (2012) Improved harmonic mean estimator for phylogenetic model evidence. *J Comput Biol* 19(4):418–438. doi:[10.1089/cmb.2010.0139](https://doi.org/10.1089/cmb.2010.0139)
- Arnold C, Matthews LJ, Nunn CL (2010) The 10kTrees website: a new online resource for primate phylogeny. *Evol Anthropol* 19:114–118
- Bennett PM, Harvey PH (1985) Brain size, development and metabolism in birds and mammals. *J Zool* 207:491–509
- Blomberg S, Garland TJ, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more laible. *Evolution* 57:717–745
- Bolker B (2007) *Ecological models and data in R*. Princeton University Press, Princeton and Oxford
- Burnham KP, Anderson DR (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York
- Burnham KP, Anderson DR, Huyvaert KP (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol* 65(1):23–35
- Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Nat* 164(6):683–695. doi:[10.1086/426002](https://doi.org/10.1086/426002)
- Chamberlin TC (1890) The method of multiple working hypotheses. *Science* 15:92–96
- Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism of body weight among primates. *Evolution* 39:1335–1351
- Claeskens C, Hjort NL (2008) *Model selection and model averaging*. Cambridge University Press, Cambridge
- Cohen J (1994) The earth is round ($p < .05$). *Am Psychol* 49(12):997–1003
- Collar DC, O’Meara BC, Wainwright PC, Near TJ (2009) Piscivory limits diversification of feeding morphology in centrarchid fishes. *Evolution* 63:1557–1573
- Collar DC, Schulte JA, Losos JB (2011) Evolution of extreme body size disparity in monitor lizards (*Varanus*). *Evolution* 65(9):2664–2680. doi:[10.1111/j.1558-5646.2011.01335.x](https://doi.org/10.1111/j.1558-5646.2011.01335.x)
- Congdon P (2003) *Applied bayesian modelling*. Wiley, Chichester
- Congdon P (2006) *Bayesian statistical modelling*, 2nd edn. Wiley, Chichester
- de Villemereuil P, Wells JA, Edwards RD, Blomberg SP (2012) Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evol Biol* 12. doi:[10.1186/1471-2148-12-102](https://doi.org/10.1186/1471-2148-12-102)
- Depraz A, Cordellier M, Hausser J, Pfenninger M (2008) Postglacial recolonization at a snail’s pace (*Trochulus villosus*): confronting competing refugia hypotheses using model selection. *Mol Ecol* 17(10):2449–2462. doi:[10.1111/j.1365-294X.2008.03760.x](https://doi.org/10.1111/j.1365-294X.2008.03760.x)
- Eklöf A, Helmus MR, Moore M, Allesina S (2012) Relevance of evolutionary history for food web structure. *Proc Roy Soc B-Biol Sci* 279(1733):1588–1596. doi:[10.1098/rspb.2011.2149](https://doi.org/10.1098/rspb.2011.2149)
- Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15
- Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* 160:712–726
- Gamerman D, Lopes HF (2006) *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, Boca Raton, FL
- Garamszegi LZ (2011) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. *Behav Ecol Sociobiol* 65:1–11. doi:[10.1007/s00265-010-1028-7](https://doi.org/10.1007/s00265-010-1028-7)

- Garamszegi LZ, Møller AP (2007) Prevalence of avian influenza and host ecology. *Proc R Soc B* 274:2003–2012
- Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. *Behav Ecol Sociobiol* 66:1363–1373
- Garamszegi LZ, Møller AP, Erritzøe J (2002) Coevolving avian eye size and brain size in relation to prey capture and nocturnality. *Proc R Soc B* 269:961–967
- Goldberg EE, Iqic B (2008) On phylogenetic tests of irreversible evolution. *Evolution* 62(11):2727–2741. doi:[10.1111/j.1558-5646.2008.00505.x](https://doi.org/10.1111/j.1558-5646.2008.00505.x)
- Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J Evol Biol* 23:494–508
- Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51:1341–1351
- Hansen TF, Bartoszek K (2012) Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst Biol* 61:413–425
- Harmon LJ, Losos JB, Jonathan Davies T, Gillespie RG, Gittleman JL, Bryan Jennings W, Kozak KH, McPeck MA, Moreno-Roark F, Near TJ, Purvis A, Ricklefs RE, Schluter D, Schulte II JA, Seehausen O, Sidlauskas BL, Torres-Carvajal O, Weir JT, Mooers AØ (2010) Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64(8):2385–2396. doi:[10.1111/j.1558-5646.2010.01025.x](https://doi.org/10.1111/j.1558-5646.2010.01025.x)
- Hegy G, Garamszegi LZ (2011) Using information theory as a substitute for stepwise regression in ecology and behavior. *Behav Ecol Sociobiol* 65:69–76. doi:[10.1007/s00265-010-1036-7](https://doi.org/10.1007/s00265-010-1036-7)
- Hunt G (2006) Fitting and comparing models of phyletic evolution: random walks and beyond. *Paleobiology* 32(4):578–601. doi:[10.1666/05070.1](https://doi.org/10.1666/05070.1)
- Hutcheon JM, Kirsch JW, Garland TJ (2002) A comparative analysis of brain size in relation to foraging ecology and phylogeny in the chiroptera. *Brain Behav Evol* 60:165–180
- Ingram T, Mahler DL (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. *Methods Ecol Evol* 4(5):416–425. doi:[10.1111/2041-210x.12034](https://doi.org/10.1111/2041-210x.12034)
- Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Syst Biol* 56(2):252–270
- Iwaniuk AN, Dean KM, Nelson JE (2004) Interspecific allometry of the brain and brain regions in parrots (Psittaciformes): comparisons with other birds and primates. *Brain Behav Evol* 30:40–59
- Jhweng D-C (2013) Assessing the goodness of fit of phylogenetic comparative methods: a meta-analysis and simulation study. *PLoS ONE* 8(6):e67001. doi:[10.1371/journal.pone.0067001](https://doi.org/10.1371/journal.pone.0067001)
- Johnson JB, Omland KS (2004) Model selection in ecology and evolution. *Trends Ecol Evol* 19(2):101–108
- Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New York
- Kutsukake N, Innan H (2013) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. *Evolution* 67(2):355–367
- Lajeunesse MJ (2009) Meta-analysis and the comparative phylogenetic method. *Am Nat* 174(3):369–381. doi:[10.1086/603628](https://doi.org/10.1086/603628)
- Legendre P, Lapointe FJ, Casgrain P (1994) Modeling brain evolution from behavior: a permutational regression approach. *Evolution* 48(5):1487–1499. doi:[10.2307/2410243](https://doi.org/10.2307/2410243)
- Link WA, Barker RJ (2006) Model weights and the foundations of multimodel inference. *Ecology* 87:2626–2635
- Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45(5):1065–1080
- Martins EP (1996) Conducting phylogenetic comparative analyses when phylogeny is not known. *Evolution* 50:12–22
- Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat* 149:646–667

- Massart P (ed) (2007) Concentration inequalities and model selection: école d'été de probabilités de Saint-Flour XXXIII - 2003. Springer, Berlin
- Mundry R (2011) Issues in information theory-based statistical inference—a commentary from a frequentist's perspective. *Behav Ecol Sociobiol* 65(1):57–68
- Mundry R, Nunn CL (2008) Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am Nat* 173:119–123
- Nakagawa S, Hauber ME (2011) Great challenges with few subjects: Statistical strategies for neuroscientists. *Neurosci Biobehav Rev* 35(3):462–473
- O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60(5):922–933. doi:[10.1111/j.0014-3820.2006.tb01171.x](https://doi.org/10.1111/j.0014-3820.2006.tb01171.x)
- Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401:877–884
- Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst Biol* 53(5):673–684
- Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am Nat* 167(6):808–825
- Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol* 53(5):793–808. doi:[10.1080/10635150490522304](https://doi.org/10.1080/10635150490522304)
- R Development Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>
- Rabosky DL (2006) Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* 60(6):1152–1164
- Ripplinger J, Sullivan J (2008) Does choice in model selection affect maximum likelihood analysis? *Syst Biol* 57(1):76–85. doi:[10.1080/10635150801898920](https://doi.org/10.1080/10635150801898920)
- Scales JA, King AA, Butler MA (2009) Running for your life or running for your dinner: what drives fiber-type evolution in lizard locomotor muscles? *Am Nat* 173:543–553
- Schmidt D, Makalic E (2011) The behaviour of the Akaike information criterion when applied to non-nested sequences of models. In: Li J (ed) *AI 2010: advances in artificial intelligence*, vol 6464. Lecture Notes in Computer Science. Springer, Heidelberg, pp 223–232. doi:[10.1007/978-3-642-17432-2_23](https://doi.org/10.1007/978-3-642-17432-2_23)
- Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2005) Information theory and hypothesis testing: a call for pluralism. *J Appl Ecol* 42(1):4–12
- Symonds MRE, Moussalli A (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behav Ecol Sociobiol* 65(1):13–21
- Terribile LC, Olalla-Tarraga MA, Diniz JAF, Rodriguez MA (2009) Ecological and evolutionary components of body size: geographic variation of venomous snakes at the global scale. *Biol J Linn Soc* 98(1):94–109
- Thomas GH, Freckleton RP, Székely T (2006) Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. *Proc Roy Soc B-Biol Sci* 273(1594):1619–1624
- von Hardenberg A, Gonzalez-Voyer A (2013) Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution* 67(2):378–387. doi:[10.1111/j.1558-5646.2012.01790.x](https://doi.org/10.1111/j.1558-5646.2012.01790.x)
- Whitmee S, Orme CDL (2013) Predicting dispersal distance in mammals: a trait-based approach. *J Anim Ecol* 82(1):211–221. doi:[10.1111/j.1365-2656.2012.02030.x](https://doi.org/10.1111/j.1365-2656.2012.02030.x)
- Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? *J Anim Ecol* 75:1182–1189