László Zsolt Garamszegi
Editor

Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology

Concepts and Practice

Springer

Chapter 7
Uncertainties Due to Within-Species Variation in Comparative Studies: Measurement Errors and Statistical Weights

László Zsolt Garamszegi

Abstract Comparative studies investigating evolutionary questions are generally concerned with interspecific variation of trait values, while variations observed within species are inherently assumed to be unimportant. However, beside measurement errors, several biological mechanisms (such as behaviors that flexibly change within individuals, differences between sexes or other groups of individuals, spatial, or temporal variations across populations of the same species) can generate considerable variation in the focal characters at the within-species level. Such within-species variations can raise uncertainties and biases in parameter estimates, especially when the data are hierarchically structured along a phylogeny, thus they require appropriate statistical treatment. This chapter reviews different analytical solutions that have been recently developed to account for the unwanted effect of within-species variation. However, I will also emphasize that within-species variation should not necessarily be regarded as a confounder, but in some cases, it can be subject to evolutionary forces and delineate interesting biological questions. The argumentation will be accompanied with a detailed practical material that will help users adopt the methodology to the data at hand.

7.1 Introduction

Since the days of Darwin, questions about evolution have centered around the forces of selection that have led to the extreme diversity in nature we observe today. Current diversity across species is regarded as the result of a large-scale natural experiment, in which their common ancestors had been placed in different environments, and subsequently underwent different selection regimes that

L. Z. Garamszegi (✉)
Department of Evolutionary Ecology, Estación Biológica de Doñana—CSIC,
Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: laszlo.garamszegi@ebd.csic.es

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_7,

affected their anatomy, physiology, life history, and behavior (Doughty 1996). Therefore, analyzing present-day patterns of interspecific variation enables making inferences about selective mechanisms acting in the past. Accordingly, most comparative studies rely on species as the unit of analysis, and species-specific trait values are subsequently investigated on the branches of the phylogenetic tree. By adopting this focus, most comparative studies inherently assume that species-specific means are biologically meaningful and they can be estimated without error.

In some cases, these assumptions are likely to be met. Take the example of brain size evolution, for instance. Comparing mammals based on species-specific means of relative brain size will probably reveal biologically meaningful comparisons and will allow making inferences about the evolution of cognitive capacities. Every individual within a species has a similar brain size relative to its body size when compared to the level of variation between species. Therefore, obtaining data from Usain Bolt, Albert Einstein, or even from my five-year-old daughter will equally represent the true human-specific value: each will have systematically larger brain-to-body-size ratios than any randomly chosen individual elephant or shrew. However, this assumption becomes less valid if, taking another example, running speed is the trait of interest, which varies more within species. I would bet with high confidence that Usain Bolt could beat an Asian elephant in a race, but I would be extremely worried seeing the same animal running after my daughter. How such variation occurring at the within-species level can affect the phylogenetic findings based on species-specific means?

Historically, the importance of the consideration of within-species variation has been dwarfed by another common assumption: the independence of interspecific data. Species cannot be regarded as independent observations as their shared common ancestry creates varying degrees of similarity between them in their phenotypes; this is classically regarded as the principal confounding factor that can bias interspecific patterns. The phylogenetic association of species, in fact, establishes the essence of comparative studies. Accordingly, a plethora of statistical approaches has been developed to handle such non-independence issues by incorporating the phylogenetic history of species into the analysis. However, the confounding effect of within-species variation remains somewhat neglected, and only recent developments have discovered the significance of this issue in the phylogenetic comparative context.

This chapter aims to bring these statistical developments into the focus of practicing evolutionary biologists. First, I explain what kind of mechanisms can shape variation within species. Next, I show how variation within subjects can alter observed patterns at the between-subject level in both the non-phylogenetic and phylogenetic contexts. Third, I examine how issues about within-species variance and sample size can be dealt with at the different phases of research (research design, diagnostics, core analysis, and interpretation). Following, in addition to the historical development of methods, I review recently proposed phylogenetic models that are able to account for within-species variation. This comparison will highlight the fundamental differences in the theoretical

foundations of these methods, their usefulness for different research designs and data types, and their accessibility in statistical packages. In this section, I will also point about the potential use of within-species variance to address interesting questions rather than as annoying nuisance in analyses. The description of the methods will be accompanied by illustrative biological examples, for which data files and program scripts (mostly in *R*) are made available in the corresponding Online Practical Material (hereafter OPM) at http://www.mpcm-evolution.org. As a closing remark, I discuss how the increasing recognition of within-species variation in phylogenetic comparative studies may lead to a departure from the classical philosophy of focusing on species-specific trait values as being evolutionarily relevant. Such a new direction may also open new horizons for the development of statistical methods.

7.2 Sources of Within-Species Variation

There are several sources that can generate variation, or measurement error, around the species-specific means. Although in a statistical sense, measurement error refers to deviations of any kind that appear between an observed and a true value, such variations can be at least of three types with different biological meaning. Hereafter, I refer to measurement error in a statistical sense meaning all types of within-species variation sensu lato.

First, instrument-related errors or observer effects can cause noise around the mean value of a trait of interest. For example, each equipment or molecular assay has a given precision that delineates a certain confidence range around each measurement. Similarly, estimation outcomes may vary among observers, which also raises an unwanted component of uncertainty when different people assess species-specific means. These deviations cause measurement errors in a narrow sense, since they are unlikely to be associated with the phylogeny and biology of the species at hand, if all species are measured by exactly the same method (or as similar a method as possible). Alternatively, as methods with no doubt vary, different equipment, laboratory assays, and observers may be applied randomly to different species (or at least arbitrarily across species, which is usually the case) to avoid instrument- or observer-related error to raise biases that can affect the underlying biological question. The errors that are introduced by such sources can be assessed by calculating estimates of inter-observer or inter-instrument reliability based on the repeated measure of the same subjects by different observers or via different instrument/assay conditions (e.g., Caro et al. 1979; Reed et al. 2002).

The second type of variation refers to true biological differences at the within-species level. Due to fluctuations in physiology or behavior, individuals of a given species will vary in the expression of certain traits. Even the same individual can demonstrate altering physiological states or produce different behavioral scores in different times or contexts. Moreover, individuals of different age- and/or sex-groups may have particular trait values. Such variation that appears at the within- or

between-individual levels may have biological relevance and can result in non-random fluctuations. For example, higher species-specific means can be associated with higher between-individual variations (e.g., think about body-size variations in mice and elephants), but patterns of within-species distributions may also vary due to several biological reasons and non-independently of phylogeny (e.g., higher variances are preserved in certain closely related taxa). Such non-random patterns make within-species variation due to between- or within-individual variation distinct from instrumental errors and potentially necessitate different treatments. Ideally, if variation at the between-individual level is considerable, several individuals within a species should be measured. This variation can be taken forward into the next levels of analyses (e.g., diagnostics and phylogenetic models), and, unlike in the case of instrument-related errors, should not be assumed as random.

Some type of variation may exist at a higher level. In addition to individuals, populations of the same species can differ and thus can produce deviations around the species-specific means. This is a more challenging problem than the between-individual variation. Different populations may have their own phylogeographical history, and migration between populations can play an additional role in shaping diversity within species, raising challenges in determining the true species-specific trait values (Felsenstein 2002; Ashton 2004). Therefore, to appropriately deal with between-population differences, one may not only need to collect population-specific trait data, but also take into account information about phylogeographical history and gene flow acting at the between-population context (Stone et al. 2011).

Additional complications may appear if the above sources of errors are simultaneously present and generate within-species variances in an additive or more complex manner (i.e., a trait can only be estimated with a given uncertainty due to instrumental errors, but at the same time, it also varies between individuals and populations). Since different types of error have different biological meaning, it might be desirable to treat them separately. Unfortunately, available correction methods handle measurement errors in a broad sense, thus their analytical separation remains currently impractical, and the researcher is left with needing a careful research design and targeted data collection if s/he wishes to discriminate between different types of measurement error in the comparative analysis.

Moreover, when measurement errors exist for more than one trait in the analysis, the potential correlation between these errors can be another confounding issue. For example, if the measurement of two traits relies on the same instrument or observer for a certain group of species, while another set is used for another group of species, there will be an unwanted correlation between measurement errors. Similarly, if between-individual or between-population effects cause similar variations in different traits within species, this will also result in correlated measurement errors. The statistical treatment of correlating measurement errors is achievable in some phylogenetic methods (Ives et al. 2007; Hansen and Bartoszek 2012).

Statistical properties of within-species variation (i.e., how much dispersion from the species-specific trait value exists within species due to within- or between-individual differences of variations between populations) can be

approached by different metrics that describe variations around the mean of a sample. Given that terminologies and abbreviations are used in a somewhat inconsistent manner, I highlight the most relevant definitions in Box 7.1.

Box 7.1 Metrics Describing Within-Species Variation and Sampling Effort

Variance. It is a probability descriptor that measures how far numbers within a sample are spread out. It is measured as the arithmetic mean of the squared differences from the true population mean, thus within-species variance is approached as the average of the squared differences of the within-species data points from the known species-specific trait value:

$$\sigma^2 = \frac{1}{n_i}\sum\nolimits_{i=1}^{n_i}(x_i - \mu)^2$$

where n_i is the within-species sample size (see below), x_i is the individual- or population-specific measure, and μ is the species-specific value. However, given that the true species–species trait value is unknown but is estimated from the available sample of small number of within-species repeats (populations or individuals), the above estimator introduces downward bias. Hence, the following (so-called Bessel's correction) formula should be used to describe within-species variation:

$$s^2 = \frac{1}{n_i - 1}\sum\nolimits_{i=1}^{n_i}(x_i - \bar{x})^2$$

where $\bar{x}$ is the arithmetic mean of the within-species data. Note that σ^2 and s^2 signify variance depending on whether true species-specific values or within-species mean is used for reference. The variance of a variable has units of measurement that are the square of the units of the variable. It is often impractical for interpretation, but most comparative approaches accounting for within-species variation takes data on variation in a form of s^2. For conventional reasons, I present equations based on σ^2 by using the appropriate subscripts to distinguish between within- and between-species variances.

Standard deviation. It is another probability descriptor that measures the dispersion of the distribution, but it is calculated as a square root of variance:

$$\sigma = \sqrt{\frac{1}{n_i}\sum\nolimits_{i=1}^{n_i}(x_i - \mu)^2}$$

for known species-specific values, and

$$s = \sqrt{\frac{1}{n_i - 1} \sum_{i=1}^{n_i} (x_i - \bar{x})^2}$$

for modest within-species samples. Note that although s^2 is an unbiased estimator of variance, s (so-called sample standard deviation) remains a slightly biased estimator for standard deviation. This bias can be considerable for small samples (e.g., $n_i < 10$), but becomes less important at increasing sample sizes. In spite of this, the sample standard deviation is the most commonly used formula, but unbiased sample standard deviation estimators for different distributions are also available.

Unlike for variance, the units of measurement for standard deviation are meaningful on the same scale on which the variable itself was measured. For this reason, interpreting variation within a set of data via standard deviation is more straightforward than via variance. An important aspect of standard deviation is that it is independent of sample size (unlike standard error) and it remains the same at small and large samples.

Measurement (or observational) error. It is simply the difference between a measured value of quantity and its true value. The term has theoretical importance.

Standard error of the mean. It is not a descriptive statistics like standard deviation or variance, but it estimates error bounds on a random sampling process when the true population mean μ is approached by the sample mean $\bar{x}$. By definition, standard error is the standard deviation of the mean of the within-species values and describes how accurately this mean estimate captures the true species-specific value. Note that standard deviation of a sample corresponds to the dispersion of the raw data points around their mean. Another important difference between standard error and standard deviation is that albeit they are measured on the same units (such as the original variable), the former is dependent on sample size. This is because repeated measurements reduces random measurement errors and make the estimator more accurate, thus the mean of the within-species values will be closer to the true species–species estimate as within-species sample size increases. The standard error $\mathrm{SE}_{\bar{x}}$ of the mean can be calculated as:

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n_i}}$$

The method by Ives et al. (2007) uses standard errors. In many cases, the sampling variance is assumed to be the square of the standard errors of the species-specific means (e.g., Hadfield and Nakagawa 2010; Hansen and Bartoszek 2012). However, given that standard errors are sensitive to sample

sizes, correction methods may be applied especially when sample sizes are small and vary among species.

Coefficient of variation. It is a normalized measure of dispersion and describes the spread of data as the standard deviation relative to the sample mean:

$$c_v = \frac{\sigma}{\mu} \quad \text{or} \quad c_v = \frac{s}{\bar{x}}$$

(unbiased estimators for different distributions are available). While variances and standard deviations are interpretable along the scale of the original variable, the coefficient of variation is interpretable independently of the measurement units (i.e., dimensionless) and thus comparable across different traits. This is not only important when comparing different traits, but may be an issue when different species with different species-specific means are contrasted, and larger trait values are accompanied by larger variance. As a general benchmark, distributions with $c_v < 1$ can be interpreted to include low variance, while those with $c_v > 1$ are considered to be loaded with high variance.

Within-species sample size. The number of repeats available within species (i.e., intraspecific sample size), denoted as n_i for species i, and distinguished from N, which is the number of species being compared (i.e., interspecific sample size). n_i can signify the number of populations (or other within-species groups) or the number of individuals that are sampled in a species. Therefore, the total sample size in a comparative study is $\sum_{i=1}^{N} n_i$, which equates with n_i*N if n_i is the same for all species. Within-species sample size is often used as an estimate of sampling effort, because species with large n_i can be considered as species that were studied with higher intensity than species with low n_i, thus they provide more precise estimates for the species-specific trait values. Therefore, within-species sample size can be used to adjust for heterogeneities in sampling effort through the use of statistical weights in the models. Given that $\mathrm{SE}_{\bar{x}} = s/\sqrt{n_i}$, $1/\sqrt{n_i}$ is an important component of the standard error of the mean. For example, if instrumental or between-observer error is constant across species, variation in sample size can still generate differences in measurement error between species. Accordingly, if measurement error data are not available, $1/\sqrt{n_i}$ can be used as an approximation of standard errors (see examples in the text).

7.3 The Statistical Consequences of Ignoring Within-Species Variation

7.3.1 Effects Independent of Phylogeny

7.3.1.1 Increasing statistical noise and attenuation bias

The problem caused by within-subject variation is not unique to phylogenetic comparative methods, but is a well-recognized issue in the statistical literature (Fuller 1987; Bollen 1989; Buonaccorsi 2010). In general, measurement error can introduce uncertainty in the estimates of true values even in very large samples. In a univariate case, imprecise measurements (or true within-subject variation in a broader sense) will increase the error range by which a single measurement approximates the true mean of the sample and will thus decrease our confidence in each datum (Chesher 1991; Manisha 2001). However, such effects act symmetrically on both sides of the distribution, thus measurement error raises random noise but not systematic bias if the focus is to derive estimates for a single variable. This may incur issues about statistical power when, for example, the estimated mean of a sample is contrasted against a hypothetical mean via null-hypothesis testing. Accordingly, in the case of large measurement error, we are more likely to commit type II statistical errors (failing to reject a false null hypothesis) than in the case of small or no measurement errors at the same sample size.

On the other hand, when the relationship between two or more variables is of interest, the presence of measurement error will not only affect precision and statistical significance, but can have considerable influence on parameter estimates, such as correlation or regression slopes (Judge et al. 1985; Fuller 1987; Chesher 1991; Adolph and Hardin 2007). Standard regression and correlation models assume that all variables have been measured correctly, or observed without error (to be more precise, regression models apply these predictions to the predictor but not to the response variables). When particular or all variables have been measured with errors or exhibit variations within subjects, conventional estimates of correlation coefficients (such as Pearson product-moment correlations) will perform with a bias toward zero (i.e., will underestimate true parameters, Fig. 7.1a and b). In a regression problem, such downward bias will be manifested in the underestimation of R^2 and standardized regression coefficients if within-subject variance exists within the response variable, and in the underestimation of unstandardized regression estimates in the presence of error in the predictor variables. The characteristics of this bias are even more complex in nonlinear regressions. The downward bias introduced on parameter estimates by within-subject (e.g., within-species) variability is known as *attenuation bias* and calls for statistical approaches that can account for such a bias (e.g., measurement error regression models, structural equation modeling, and unbiased estimators of correlation coefficient).

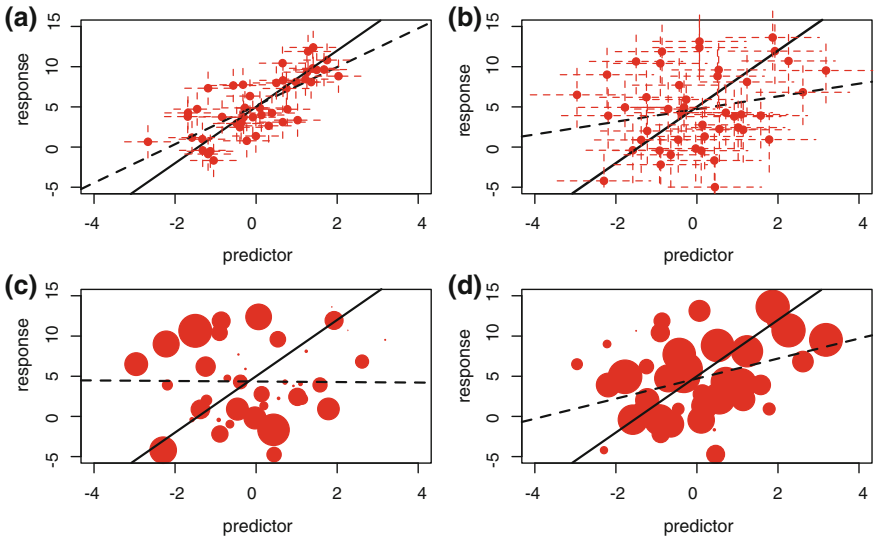

Fig. 7.1 The effect of measurement error and within-species sample size on the estimation of regression slopes from simulations that incorporate different variance components. **a** The within- and between-species variance of both the predictor and response variables are set to provide a repeatability of 0.75 (see Sect 7.4.2.1. about repeatability). **b** The within- and between-species variance of both the predictor and response variables are set to provide a repeatability of 0.40. **c** and **d** same conditions as in **b**, but the data were analyzed by weighted least square regression that relied on two different scenarios for within-species sample sizes. *Dots* represent species, *solid lines* show the regression line (same in all cases) that was used to generate species-specific means, *dashed lines* that are obtained by least square regressions (**a** and **b**: ordinary regression, **c** and **d**: weighted regression) on the simulated data loaded with measurement error, *red crosses* in **a** and **b** define error ranges around both variables that were considered during the simulation, the size of the points in **c** and **d** are proportional to the corresponding within-species sample size

7.3.1.2 Within-Subject and Between-Subject Correlations

The aforementioned situations involve the confounding effects of within-subject variances around the predictor and response that act independently of each other. However, when multiple traits are examined, issues about covariances should also be considered. That is, the apparent relationship between two traits with clustered structure can actually have two components: the between-subject and the within-subject components (Snijders and Bosker 1999, see also Fig. 7.2). The between-subject correlation or regression arises from the relationship between the two variables based on the subject- or group-specific (for us species-specific) mean trait values. The within-subject correlation, however, refers to the associations that appear at the within-subject levels (e.g., from between-individual or -population correlations or regressions). The two components can have opposite direction with effects that contradict each other (for example, in the within-species context,

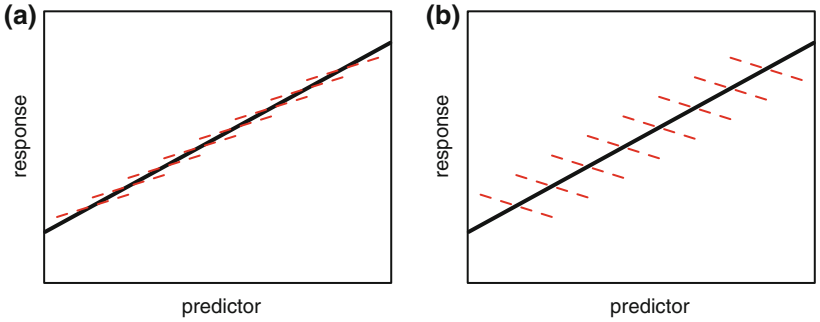

Fig. 7.2 Within-species (*red dashed lines*) and between-species (*black solid lines*) regressions when considering two possible scenarios. **a** both within-species (i.e. as estimated from regressions at the between-individual level) and between-species (i.e. as estimated from regressions using species-specific trait values) regressions are positive, but they somewhat differ in their slopes. **b** within-species regressions are negative while between-species are positive

small-sized individuals live longer than large individuals, while in the between-species context, there is a positive relationship between body size and longevity, such as in Fig. 7.2b). Moreover, the relationships within subjects can vary from subject to subject, which is likely if these represent different species with different ecology. This may be important in a comparative context, if for example, different mechanisms shape between-population patterns in different species, while other selection regimes operate and affect trait associations at the interspecific level. If data are available at the within-subject level, mathematical and statistical solutions are available to separate the within- and between-group components of correlations or regressions (Kreft et al. 1995; Gelman and Hill 2007; Bolker et al. 2009; van de Pol and Wright 2009).

7.3.1.3 Heterogeneity in Sampling Effort

An additional problem met when the unit of analysis (e.g., species) corresponds to repeated measurements of smaller units (individuals, populations) is that due to various constraints on the sampling procedure, different values will often be represented with different within-subject sample sizes (n_i). This will affect another assumption of standard statistical methods, namely that the standard deviation of the error term is constant over all values of the variables (Sokal and Rohlf 1995). This would require each data point to provide equally precise information about the deterministic part of total process variation, a condition that is likely violated if data quality is a function of sampling effort, and subject-specific estimates from small samples will be less reliable than estimates from large samples (Garamszegi and Møller 2010). Such heterogeneity in data quality can be accounted for by using statistical weights in the analyses (generally available for descriptive

statistics of univariate distributions, correlations, and regressions) that give emphasis to each data points according to the underlying sample size (Fig. 7.1c and d).

7.3.2 Effects Evoked by Phylogenetic Associations

The common ancestry of species gives another twist to the story on measurement errors. In addition to the power and attenuation issues detailed above, there are other effects of within-species variation that are called forth only in light of the specific hierarchical structure of the interspecific data. In particular, the above errors interacting with the phylogeny can often enhance the chances of detecting a spurious relationships between variables (Martins 1994; Purvis and Webster 1999; Felsenstein 2004).

7.3.2.1 Mathematical Evidence

Such types of biases are well defined mathematically for the approach based on phylogenetically independent contrasts, when uncontrolled within-species variances are differentially magnified during the standardization of contrasts (Ricklefs and Starck 1996; Felsenstein 2008). The variance of the difference between two species' means depends on the branch length involved (determining phylogenetic variation) and the measurement errors around these means (determining within-species phenotypic variation). By definition, the contrasts are standardized by a quantity proportional to the square root of their total standard deviation (Felsenstein 1985). Such standardization can have a strong influence for closely related species (i.e., when the involved branch lengths are short) if the underlying within-species sample sizes are low (i.e., when measurement errors are large), because the contrasts are divided by far too small quantity. This artificial standardization can produce outlier contrasts for all variables with considerable influence on the estimated regression or correlation coefficients.

7.3.2.2 Simulation Results

Harmon and Losos (2005) presented a simulation study to demonstrate the effect of uncontrolled within-species variation on type I error rates emerging in phylogenetic comparative studies based on independent contrasts. They generated interspecific data for two variables with an expected covariance structure between them for a modest number of species sharing a certain phylogenetic history. They also simulated within-species sampling process around these means by considering different scenarios for sample size in terms of the number of individuals and measurement error. When there is no correlation between two variables, one may

expect to find a statistically significant relationship between them only by chance in 5 % of the simulated datasets. This type I error rate could be reproduced in the non-phylogenetic context, i.e., when species-specific trait values were simulated independently of each other and the units of analysis shared no historical associations. However, the number of significant associations increased above the chance levels in models based on independent contrasts without accounting statistically for within-species variation. At large within-species variation and small sample sizes, type I error rate could increase up to 17 %. The error rates in the phylogenetic simulations could be retained at the chance (5 %) level when most of the variance occurred at the between-species level making intraspecific component negligible and when within-species sampling relied on large sample sizes. Importantly, increasing between-species sample size did not improve error rates, but, in fact, produced spurious correlations at high measurement errors more powerfully than simulations relying on fewer species. Felsenstein (2008) performed similar simulations, in which he used slightly different sampling schemes and found that a true null hypothesis was falsely rejected in about 20 % of the cases when the data were analyzed by the standard contrast method that ignores within-species variance.

If considerable within-species variation is left uncontrolled, parameter estimates become biased when the data are analyzed not only by independent contrasts, but also when the approaches rely on phylogenetic generalized least squares (PGLS). Ives et al. (2007) demonstrated the occurrence of such biases in various evolutionary test situations. For example, they designed a univariate simulation model to study the performance of ancestral state estimators at a given phylogeny and trait value at the root of the tree and with pre-defined within- and between-species variances. They considered an extreme measurement error scenario (being more than two times larger than the standard deviation of the among-species error), which provided evidence that, at a low within-species sample size, the phylogenetic generalized least squares method disregarding within-species variances reveals biased estimates for parameters of evolutionary rate (σ^2) and phylogenetic signal (K). This happens because, by ignoring the within-species component of variation, variability in the data is erroneously attributed solely to the between-species component, which results in the overestimation of the rate parameter and in the underestimation of phylogenetic signal. However, ancestral state estimation seems to be unaffected by the presence of measurement errors.

The influence of within-species variance has also been investigated in bi- or multivariate evolutionary problems. In a correlation model, Ives et al. (2007) showed that an estimator of r without accounting for measurement errors performs with downward bias even when controlling for the phylogenetic association of species. Similar patterns were found in a regression model, in which the established slope parameter was underestimated with a conventional PGLS regression approach when within-species variance was present in the data. Importantly, the degree of this bias was independent of the interspecific sample size, but the confidence intervals around the estimates were narrower when a larger number of

species was involved. Therefore, echoing the results of Harmon and Losos (2005), this indicates that large interspecific sample size when coupled with low intraspecific sample size can accentuate biases with higher statistical power.

7.3.2.3 Empirical Evidence

A meta-analysis of almost two hundred comparative studies investigated empirically if heterogeneity in within-species sample size can have an effect on research findings (Garamszegi and Møller 2010). It appears that in a ~20 % of studies, the use of statistical weights that balance for heterogeneity provides remarkably different results from that of the non-weighted model. This rate is comparable to the effect that a control for phylogenetic non-independence causes. Notably, the result of weighting was particularly strong when there was large variation in sample sizes among species, while homogenization had a minor effect when sample sizes were more balanced among species.

The above studies unanimously suggest that recent issues about within-species variance and sample size should not be taken as a false alarm, as there are situations when the variability of trait values within species can have a considerable effect on the estimation of parameters with evolutionary importance. However, most of the simulations were made under conditions in which extreme error structures were created: i.e., within-species samples consisting of few individuals and variances that are comparable with the between-species variances. Most of the comparative studies in practice might meet with more relaxed conditions when conventional phylogenetic methods can also perform with tolerable type I error rate even by ignoring within-species components of variations. This does not mean that the problem can be ignored, but highlights the importance of the consideration of the measurement error issue at the level of study design and data diagnostics, and its appropriate treatment at the analytical level if the data at hand require. Such an empirical approach to within-species variance resembles the philosophy that is similar to how we nowadays handle the confounding effects due to common descent: we estimate the phylogenetic signal in the data first (more precisely in the model residuals, see Chap. 5) and then obtain parameter estimates from the model at the estimated signal value (Freckleton 2009).

7.3.3 Phylogenetic and Other Biological Confounders

So far, within-species variances and sample sizes were treated as if they were fluctuating at random with respect to the biological question under testing. However, there might be cases when properties of between-individual or between-population distributions are shaped by evolutionary forces; thus, within-species

sample sizes and/or variances are not necessarily independent of the phylogenetic history of species or other biological predictors. For example, Garamszegi and Møller (2011) claimed that within-species sample size (also contributing to measurement errors via $1/\sqrt{n_i}$, see Box 7.1) can vary along several species–species characteristics that determines the probability of sampling of individuals. This is because some species may occur at low abundance, display trap-shy behaviors or have a life history that makes their sampling difficult, which would all decrease their probabilities of being sampled. If such probabilities are determined by certain phylogenetic or biological attributes of species, this will render within-species sample sizes to be related to the same attributes. From the empirical part of the argument (Garamszegi and Møller 2012), it was evident that sampling effort in terms of within-species sample size was consistent across different studies on birds using different sampling methods suggesting that available sample sizes are species-specific attributes. Moreover, it was dependent on abundance, body mass, and predator avoidance behavior implying that applicable sampling effort is determined by some biological properties of the species.

7.4 The Statistical Treatment of Within-Species Variation in Phylogenetic Comparative Studies

The problems posed by within-species sampling variance can be considered in each phase of research, from the design of studies to the interpretation of the results. These steps will be discussed below.

7.4.1 Study Design: Balancing Sample Sizes

When designing a phylogenetic comparative study, the observer is usually restricted to balance between the within-species and the between-species sample sizes (Harmon and Losos 2005). A certain number of species is important to reach a sufficient statistical power in the analyses, but also to make generalizable evolutionary inferences. On the other hand, simulations (Harmon and Losos 2005; Ives et al. 2007) show that low within-species sample size can bias parameter estimates even when several species are included. Thus, sample size requirements at different levels to deal with power and bias are often in conflict with each other. Therefore, some efforts should be devoted to the appropriate within-species sampling, event at the cost of decreasing between-species sample size. Ideally, these constraints can be optimized in a pilot study, in which the variances at the between- and within-species levels can be determined in a subsample of species. Such information can subsequently guide the investigator when determining sample sizes, for which the simulation results might serve some rules of thumb (Harmon and Losos 2005; Ives et al. 2007; Felsenstein 2008). In general, when within-species variance appears

negligible compared to the among-species variance in a pilot study, it may be a convincing evidence for that sampling effort at the within-species level will be less important on a wider scale. Hence, the researcher is safe to conclude that a few estimates per species provide a reliable representation of the species-specific mean trait values. However, when a considerable proportion of variance is at the within-species level (see quantitative estimation in the next section), multiple measurements are warranted for each species, and the within-species variance should be taken forward into the comparative analyses.

Due to various biological reasons, some species are easier to sample than others. Therefore, it is unrealistic to obtain the same within-species sample size in every species, which sets up additional challenges for study design. Shall we devote more research effort to sample less accessible species at the cost of lowering the sample size for a large number of more common species? Finding an answer to such questions can be a complex task and may require considerations about the biological problem and the phylogeny at hand. Some pilot studies as well as phylogenetic targeting (see Arnold and Nunn 2010) may also be of help to optimize research effort.

7.4.2 Diagnostics

7.4.2.1 Repeatability

For a comparative study of species-specific means to be meaningful, these means are required not to be confounded by measurement errors, and if this assumption is fulfilled, within-species variance could be omitted in the analyses. Although, most comparative analyses apply this omission, the potentially confounding effect of measurement error is rarely considered in practice (and even in theory). If species-specific values cannot be estimated without errors, within-species variances (or other components of errors) need to be incorporated in the phylogenetic models. In this case, the researcher needs to invest a substantial effort in collecting a sufficiently large sample for each species that would allow to capture trait variations therein and to obtain good estimates for species means (or other statistics). To make decisions about such investments and about the subsequent analytical strategy, it might be useful to have a glance about different components of the variance prior to the phylogenetic analyses. If based on the measurement of different individuals or populations, multiple data are available for, at least, a subset of species (e.g., from a pilot study), the ratio between the within-species and between-species components of variances can be assessed by calculating repeatability.

Repeatability is the intraclass correlation coefficient that can be derived from different variance components, mathematically as the proportion of the between-subject variance relative to the total variance (Sokal and Rohlf 1995):

$$R = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}, \quad (7.1)$$

where
$\sigma^2_{\text{between}}$ between-species variance and
σ^2_{within} within-species variance.

If most of the variance resides at the between-species level and within-species variance remains minuscule, the ratio will approach unity. However, if within-species variation is similar to the between-species variation, the estimator will return values that are ~0.5 and will approximate zero if within-species variation is of a major concern. Repeatability can offer a standard metric to help the observer judge about how much s/he can trust in the species-specific trait estimate as a meaningful unit for the comparative analysis. Working with a trait that has repeatability close to one indicates that a single individual will well-reflect the true species-specific value. On the other hand, low repeatability indicates that within-species variance may warrant some attention, and for an appropriate statistical treatment, a systematic within-species sampling is needed to capture such variation. Behavioral or physiological traits may often involve such a modest within-species repeatability, while morphological or life history characters may depict less variation within species incurring higher repeatability (see also Blomberg et al. 2003).

Classically, repeatability can be acquired from an ANOVA model, in which the variation in the response measurements is partitioned into components that correspond to different sources of variation (Lessells and Boag 1987). The widely used ANOVA-based repeatability takes the mean between-subject sum of squares and the mean within-subject (residual) sum of squares and also considers the species-specific sample sizes. Recently, Nakagawa and Schielzeth (2010) provided a list of functions for repeatability (together with its confidence interval, and a test statistics contrasting the estimated value against zero) based on mixed model approach that can be flexibly used for a variety of data types and distributions. These approaches can also be applied in the comparative context, where the interest is to determine whether species consistently differ in their mean trait value (some strategies are given in the OPM). However, the phylogenetic relationships of species may warrant some attention and potentially necessitate the consideration of more complex hierarchical modelling.

7.4.2.2 Independence

Another diagnostics that may be useful before entering into the core phylogenetic analyses is to assess whether within-species sample size (or variance) occurs randomly. There might be several biological reasons to why some species are systematically easier to sample resulting in larger sample sizes than other species (Garamszegi and Møller 2011). Such factors can be associated with species-specific

abundance (rare vs. common species), behavior (species that avoid traps vs. species that are attracted by them), as well as life history and ecology (species breeding in accessible vs. inaccessible habitats, solitary vs. colonial species) and even morphology (small vs. large species with different mobility and/or detectability). Furthermore, these effects may vary in non-random manner with respect to the phylogenetic associations of species. Hence, within-species sample sizes can depict a phylogenetic structure at the across-species level. These deviations from randomness can be investigated by testing if within-species sample size (or variation) is interspecifically related to biological predictors and phylogeny. (see Garamszegi and Møller (2012) for an example analysis)

7.4.2.3 Preparation of Data

Before testing the evolutionary hypotheses in the statistical models, some preparatory steps may be warranted. These may include the conventional exercises for data diagnostics and the verification of model assumptions (e.g., distribution, collinearity, balanced design, see more in Chap. 6), but also efforts to make our data suitable for the particular method that will be used to control for within-species variance. For example, we may need to decide whether we aim to work with individual (population)- or species-specific data. We can use data on individuals for the contrast and likelihood surface methods (see Table 7.1 and text below), but the calculation of repeatability discussed above also requires repeated measurement within each species and the underlying dataset needs to be tabulated accordingly. We can supply species-specific trait values for the other methods such as based on PGLS regression techniques (see Table 7.1 and text below). Means at the species level can be either calculated from the raw individual data through simple summary statistics or should already be available in this form if obtained from other sources.

When working with species-specific datasets, most of the methods assume that within-species variances are known and correspond to large samples. However, this criterion might be violated in most of the comparative studies, in which within-species samples are often limited to few individuals (or populations). It is also common that sample size at this level equals one, which makes variances mathematically inestimable. For such a case, a corrected estimate might be desired. For this purpose, Ives et al. (2007) suggest first calculating a pooled variance over the entire sample:

$$\bar{\sigma}^2 = \frac{1}{N_{\text{species}}} \sum_{i=1}^{N_{\text{species}}} \sigma^2_{\text{within}_i}, \tag{7.2}$$

where

$\bar{\sigma}^2$ weighted (pooled) within-species variance,

$\sigma^2_{\text{within}_i}$ within-species variance observed in *i*th species, and

N_{species} interspecific sample size after excluding species with a single observation.

Hansen and Bartoszek (2012) suggested an improved estimate of the pooled variance by weighing the species with their sample sizes:

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{N_{\text{species}}} \sigma^2_{\text{within}_i}(n_i - 1)}{\sum_{i=1}^{N_{\text{species}}} (n_i - 1)}, \tag{7.3}$$

where

n_i within-species sample size for the ith species, other abbreviations as in Eq. (7.2).

With these estimates, the species-specific standard errors (SE_{n_i}) as well as variances could be computed:

$$\text{SE}_{n_i} = \frac{\bar{\sigma}}{\sqrt{n_i}} \tag{7.4}$$

If the model for comparative analysis requires species-specific variances ($\bar{\sigma}^2_{n_i}$) to be entered, these could be approximated as the square of this standard error ($\text{SE}^2_{n_i}$), that gives:

$$\bar{\sigma}^2_{n_i} = \frac{\bar{\sigma}^2}{n_i}, \tag{7.5}$$

The use of this replacement procedure might be problematic, if true variances vary considerably among species due to scaling effects for example (e.g., consider body mass variation in a mouse and an elephant species). In this case, unrealistically high variances would be assigned to species, which, in reality, could actually be characterized by small variance (e.g., variance in mouse would be adjusted partially based on variance in elephant). This problem can be reduced by applying a log-transformation on variances (or other variance-stabilizing transformation) before the adjustment (see below) or by using the scale-independent coefficient variation (cv) parameter for the pooling and subsequent weighting procedure, from which error variances could be back-calculated (see Box 7.1 for calculations).

For making meaningful interpretations from interspecific patterns, tip values are often required to have a scale on a log-axis for both statistical and biological reasons. In such a case, not only species-specific means, but also the associated variances should also be transformed. Note that this transformation also normalizes within-species variation, thus diminishes the problems caused by unequal variances due to scaling effects (Revell 2010), as discussed above. Log-normal distributions require that the joint log-transformation of the mean and variance occurs according to the following approximations equations:

$$y_i = \log\left(\frac{\bar{x}_i^2}{\sqrt{\sigma^2_{\text{within}_i} + \bar{x}_i^2}}\right) \tag{7.6a}$$

$$v_i = \log\left(1 + \frac{\sigma^2_{\text{within}_i}}{\bar{x}_i^2}\right), \tag{7.6b}$$

where

$\bar{x}_i$ and $\sigma^2_{\text{within}_i}$ mean and variance, respectively, on the original scale,

y_i and v_i mean and variance, respectively, on the log-scale for species *i*.

However, these approximations are not necessary if the individual level data are accessible, when one can just compute the mean and variance on the log-scale directly.

In cases when we are interested in assessing the effect of the heterogeneity in species-specific sample sizes, it might be important to consider issues about the transformation of within-species sample sizes (n_i). Data heterogeneity is presumably more influential in cases of low sample sizes. On the other hand, after a certain level, increasing sample size has a minor effect on precision. Therefore, we may want to downweight data points with very low sample sizes, but without making too much discrimination between species-specific trait estimates that come from a reasonably large within-species samples (Garamszegi and Møller 2010). Accordingly, a logarithmic or square-root transformation on sample sizes may help dealing with this problem (see also Chap. 12).

Examples for calculating measurement errors based on pooled variances and for log-transformation are given in the OPM.

7.4.3 *Incorporating Within-Species Variation into the Phylogenetic Analysis of Species-Specific Traits*

If we cannot achieve negligible within-species variances through a careful study design, and a repeatability analysis indicates that the measurement error on any of the investigated traits is considerable and may be meaningful, we need to use phylogenetic comparative studies that can account for such variance components. Assuming that information on the dispersion of data at the within-species level is available in any form (e.g., as raw individual data or as a probability descriptor summarized in Box 7.1), different phylogenetic methods can be applied to deal with different test situations and data types with each offering different benefits. In the sections below, I will provide an overview on these methods. I start this

revision from the historical perspective, as I find it important to get a picture about the essence of the classical methods in order to understand how modern approaches discussed subsequently work.

7.4.3.1 The History of Interspecific Comparative Methods That Account for Within-Species Variation: Back to the Root

The Autoregression Model

Although the spread of comparative methods with measurement error can be witnessed in the contemporary literature, the history of the underlying methodology goes back to the beginnings of phylogenetic comparative methods. In the same year when Felsenstein's (1985) seminal paper was published, Cheverud et al. (1985) proposed an alternative approach to control for phylogenetic non-independence. This method was based on an autoregression model to accommodate the concept of phylogenetic constraints in interspecific studies through partitioning the observed total trait variance into inherited (phylogenetically determined) and taxon-specific (caused by independent evolution) components:

$$y_i = p\mathbf{W}y_i + e_i, \tag{7.7}$$

where

y_i observed trait mean for species i taken from the **y** vector of standardized trait values for N_{species} species,

p phylogenetic autocorrelation coefficient (scalar),

W phylogenetic connectivity matrix reflecting the relatedness of species (i.e., genetic correlation between species),

e_i ith element of the **e** vector of residuals.

In this equation, $p\mathbf{W}y_i$ represent the phylogenetic part, while e_i stands for the taxon-specific part. Originally, the autocorrelation model does not require the specification of an evolutionary model (such as Brownian motion), it only relies a relaxed assumption that the inherited component is similar among closely related species and different among distant species (but in principle, different evolutionary assumptions could be brought into the **W** matrix). The approach by Cheverud and coworkers (1985) includes estimation procedures based on maximum likelihood (ML) iteration for the phylogenetic autocorrelation parameter p, which can be used to make inferences about the importance of phylogeny in trait evolution. This methodology has not received as much popularity in practice as the independent contrast method (Felsenstein 1985), but its inherent promise for incorporating issues about within-species variation has been recognized in its subsequent analytical development toward greater flexibility (e.g., Gittleman and Kot 1990). Along this line, Cornillon et al. (2000) considered issues about differences between

populations and, accordingly, split the inherited variance part into inter- and the intraspecific components, while they also expressed the residual part at the level of population (and not species). These matrices were used to fit an autoregressive model through a maximum likelihood (ML). By doing so, the method is able to capture within-specific variation that occurs among populations in both univariate and multivariate test situations.

The Phylogenetic Mixed Model of Lynch and Its Extensions

The main logic of Cheverud et al. (1985) was also influential on the improvement of comparative methodologies of other types. Lynch (1991) extended mixed modeling techniques taken from quantitative genetics to decompose observed mean phenotypes into different components. His model was based on the general formula:

$$y_i = \beta_0 + a_i + e_i \tag{7.8}$$

$$\mathbf{a} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{C}\right) \tag{7.8a}$$

$$\mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^2 \mathbf{I}\right), \tag{7.8b}$$

where

- y_i observed trait mean for species i taken from the $N_{\text{species}} \times 1$ dimensioned **y** vector of standardized trait values,
- β_0 grand mean of the character over the whole phylogeny (intercept),
- a_i ith element of the **a** vector of heritable additive values with a dimension of $N_{\text{species}} \times 1$,
- e_i ith element of the **e** vector of residuals (dimension: $N_{\text{species}} \times 1$),
- $\mathcal{N}$ signifies that values of the given vector (i.e. **a**) are taken from normal distribution that is specified with a mean (i.e. 0) and variance (i.e. $0, \sigma^2\mathbf{C}$)
- σ^2 overall phylogenetically inherited variance (rate of evolution),
- σ_e^2 residual variance,
- **I** identity matrix (dimension: $N_{\text{species}} \times N_{\text{species}}$),
- **C** correlation structure defined by the phylogeny (dimension: $N_{\text{species}} \times N_{\text{species}}$).

In this approach, a_i represents the heritable phylogenetic effect sensu Cheverud et al. (1985), while e_i is Cheverud's species-specific effect that also includes sampling error beside the nonadditive genetic effects and environmental effects. Lynch suggested an iterative approach based on expectation-maximization (EM, Dempster et al. 1977) algorithm to find models with maximum likelihood (ML) that can be used to estimate parameters, such as the mean phenotypes of ancestral taxa, additive values, and residuals deviations as well as the variance-covariance structure of the components of taxon-specific means. These parameters can serve basis for computing regression coefficients and hypothesis testing. Although, in the above model the author generally assumed that within-species variation is negligible and

sampling errors on the phenotypic means can be treated as zero, he pointed that such an assumption may be violated in many cases, for which the original method could be adjusted if data on sampling variances and covariances are available.

Such a premise was further exploited by Christman et al. (1997) and Housworth et al. (2004). These authors argued that if the estimation of species-specific character states from a sample of individuals or populations is subject to considerable error due to the relatively small number of individuals sampled per species, an additional component should be added to Lynch's formula to factor out the within-species variances. The model of Housworth et al. (2004) for univariate case yields

$$y_{ij} = \beta_0 + a_i + e_i + \varepsilon_{ij} \tag{7.9}$$

$$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\mathbf{I}), \tag{7.9a}$$

where

y_{ij}	observed trait value for individual j in species i
ε_{ij}	individual error term that is associated with the measurement of individual j in species i,
$\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$	variance caused by errors when measuring a single individual, the corresponding $\mathbf{I}$ identity matrix has a dimension of dimension: $\sum_{i=1}^{N_{\text{species}}} n_i \times \sum_{i=1}^{N_{\text{species}}} n_i$,
β_0, a_i, and e_i	intercept, phylogenetic, and non-heritable residual components, respectively, as in Eq. (7.8).

Christman et al. (1997) present an illustrative analysis (and the corresponding *MATLAB* codes) on morphological characters originating from four populations of amphipods, in which they relied on the extended Lynch's model but without the non-heritable effects ($y_{ij} = \beta_0 + a_i + \varepsilon_{ij}$). Through the incorporation of the term ε_{ij}, these approaches bring the focus onto individuals as the unit of analysis assuming that each species on the phylogenetic tree is formed by a hard polytomy of individuals. The length of the within-species branches scales with the degree of measurement error and needs to be estimated in parallel with other parameters in the model.

Regression Techniques Based on Phylogenetic Generalized Least Squares (PGLS)

The autoregressive method of Cheverud et al. (1985) and the mixed model approach originating from Lynch (1991) as well as their derivates combine the phylogenetic constraint with the statistical model via a mean structure in the equation ($p\mathbf{W}y$ or a_i), while the intraspecific variance is usually lumped within the error term (Lynch 1991; Christman et al. 1997; Housworth et al. 2004). As an alternative approach, phylogenetic methods based on generalized least squares (GLS) models incorporate the phylogeny through the error structure (Grafen 1989; Martins and Hansen 1997;

Revell 2010, see Chap. 5). This solution offers a flexible combination of errors from different sources (e.g., phylogeny and within-species variance) as well as to accommodate various test situations (e.g., estimating ancestral states, rates of evolution, phylogenetic effects, correlations, and regressions). Martins and Hansen (1997) present a general linear model in the form of:

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{7.10}$$

where

- $\mathbf{y}$ vector of characters or functions of character states for extant or ancestral taxa,
- $\boldsymbol{\beta}$ vector of regression slopes,
- $\mathbf{X}$ matrix of states of other characters, environmental variables, phylogenetic distances, or a combination of these, and
- $\boldsymbol{\varepsilon}$ vector describing error structure due to various sources.

This equation can be broadly used to translate evolutionary questions into a statistical formula. By a smart definition of $\mathbf{X}$ and $\mathbf{y}$, several evolutionary problems can be tackled, while the characterization of $\boldsymbol{\varepsilon}$ allows describing the impact of confounding factors that can cause noises or biases in the estimation of $\boldsymbol{\beta}$. The error structure is composed of at least three types of error: the error due to common ancestry ($\boldsymbol{\varepsilon}_S$), the error due to within-species variation (on any character, $\boldsymbol{\varepsilon}_M$), and error due to the uncertainty in the reconstruction of phylogenetic history ($\boldsymbol{\varepsilon}_P$). These errors can be combined in the PGLS framework (see Chap. 5), which thus allows the careful definition of residuals that accommodate a complex covariance structure within the term ε.

Practical Constraints

Despite the relatively well-established statistical background for treating within-species variance in different comparative approaches (e.g., autoregressive models, phylogenetic mixed models borrowed from quantitative genetics, and phylogenetic generalized linear models), until recently, these statistical approaches were rarely exploited in practice. The reasons for such ignorance may rely on the practical intractability of the proposed algorithms, the lack of evidence pointing to the confounding role of measurement errors in the interspecific context and the scarcity of data that captures within-species patterns. For example, the expectation-maximization (EM, Dempster et al. 1977) method proposed by Lynch (1991) to fit models was very slow in practice, and even the reparameterized algorithm that remedies this problem can only be applied to uni- and bivariate cases. The first widely accessible computer software for incorporating measurement error was *Compare* (Martins 2004). When it became accessible to deal with intraspecific variance, the phylogenetic independent contrast method (that neglects such variance) was already flourishing in its renaissance epoch thanks to easy access to the

program CAIC. (Purvis and Rambaut 1995). In addition to these technical challenges, early practitioners of the comparative approach might have disregarded the importance of within-species variation because its confounding effects remained under-documented compared to the well-known biases that can be caused by common descent. Finally, various constraints during the collection of interspecific data and the assembly of phylogeny may have shifted the focus from the within- to the between-species patterns preventing the adoption of measurement error models by appropriate within-species data.

Nonetheless, even in this early phase of the history, some investigators did consider issues about measurement error in their comparative study, and their solutions deserve mentioning. In a comparative study on the relationship between population density and body size in birds, Taper and Marquet (1996) corrected estimated regression parameters for attenuation bias due to measurement error based on the ratio of the total variance and between-subject variance (Madansky 1959) and concluded that such a correction had a minimal effect on the focal relationship. However, this correction is only applicable to the parameter estimates of the ordinary least square regressions, thus the authors could only employ it in the analysis based on raw species data without adjusting for phylogeny (but see its extensions below for PGLS sensu Hansen and Bartoszek 2012). In another pioneering study, Monkkonen and Martin (2000) relied on randomization and bootstrapping procedures to investigate the influence of the between-population trait variations on the interspecific relationship between clutch size and nest excavation propensity in *Parus* tits. They found that the outcome of the analysis was largely similar across the 1,000 bootstrapped samples that randomly picked one population estimate for each trait for each species. Although such a resampling technique appears to be able to incorporate uncertainty in parameter estimates, it may not be useful as a general method to control for within-species variance, because it is not able to cope with attenuation bias (i.e., correct for the degree of underestimation of parameters). When intraspecific variance is considerable, the confidence range around the regression slope is more severely biased than the regression slope that is based on species-specific mean values (Fig. 7.3).

7.4.3.2 Recently Developed Comparative Methodology for Handling Within-Species Variance Components: Getting into Practice

Modern phylogenetic approaches began to recognize the importance of the statistical problems that can be caused by measurement errors, and such considerations gave a burst to the expansion of available softwares and also enhanced the spread of the methodology into research practice. These approaches, while taking into account within-species variance, can now accommodate a wide range of evolutionary questions about correlated trait evolution, ancestral states and phylogenetic signals (although the corresponding methodologies are not equally developed). Most of these new approaches are closely linked with classical

Table 7.1 Recently proposed phylogenetic comparative methods that can account for intraspecific variances due to measurement error or biological variations at the within- and between-individual as well as the between-population level

Approach	Reference	Test situation	Data type	Software
Independent Contrasts	Felsenstein 2008	Multivariate: Phylogenetic correlation Regression	Raw individual-specific data	Phylip R packages: *varCompPhylip (ape)* *pic.ortho (ape)*
PGLS	Ives et al. 2007 Martins & Hansen 1997	Univariate: Ancestral state estimation Rates of evolution Phylogenetic signal Multivariate: Phylogenetic correlation Regression Reduced major axis regression	Species-specific data on within-species variance or standard error Taxon-wise estimate on the universally applicable within-species variance or standard error Taxon-wise estimate on the universally applicable within-species variance or standard error that is weighted by 1/sample size Raw individual-specific data to calculate the above variance/error components	Matlab codes from authors Compare 4.6 R packages: *gls (nlme & caper)* *pgls.Ives (phytools)* *phylosig (phytools)* *fitContinuous (geiger)*
PGLS	Hansen and Bartoszek 2012	Multivariate: Regression	Species-specific data on within-species variance or standard error	R packages: *Slouch* (from author) *GLSME* (from author)
Likelihood computation (simulation-based)	Kutsukake and Innan 2012	Univariate: Ancestral state estimation Rates of evolution Estimation of evolutionary model parameters Multivariate: (could be extended)	Species-specific data on within-species variance or standard deviation Alternatives to normal distribution can be accommodated	a program written in C language (from author)
Likelihood computation (Bayesian)	Revell and Reynolds 2012	Univariate: Ancestral state estimation Rates of evolution Estimation of evolutionary model parameters	Raw individual-specific data	R packages: *fitBayes (phytools)*
Mixed models	Hadfield and Nakagawa 2010	Multivariate: Regression	Species-specific data on within-species variance or standard error	R packages: *MCMCglmm (MCMCglmm)*

methods and can be grouped into the following main (not necessarily exclusive) categories, which are also listed in Table 7.1. The performance of these methods on simulated data is demonstrated on Fig. 7.4. Worked examples can be found in the OPM.

Independent Contrasts

A new method based on independent contrasts (Felsenstein 2008) can be considered as a modified version of the Lynch's model (Lynch 1991) under the assumption that the evolution of species-specific values depicts Brownian motion with a perfect phylogenetic heritability component. This extended contrast method allows for multiple individual measurements per species resulting in within-species phenotypic variances that are greater than zero and are the same for all species. If phenotypic values are available for individuals, the contrasts can be computed at both the within- and the between-species levels. This computation assumes that individuals within a species are connected to each other with zero branch lengths to form a species-specific node on the phylogenetic tree.

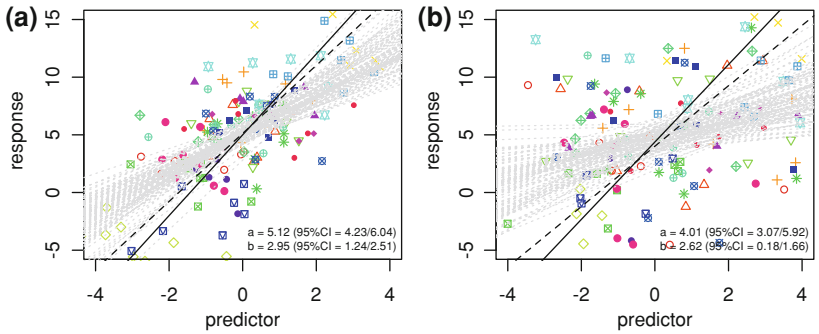

Fig. 7.3 Dealing with within-species variation in interspecific comparisons by resampling technique. Each data point represents an individual within species that are coded by different symbols. The interspecific relationship between the predictor and response is estimated by randomly taking one individual from each species to represent the species-specific character values, which are then regressed across species. This procedure can be repeated multiple times (100–1000). The resulting regression lines are given by the *gray dotted lines* (from 100 resamplings). The *dashed black line* shows the regression line that could be obtained by using the averaged individual trait values at the species level. The slope and intercept of this line are given in the legend together with the confidence intervals that could be estimated from the resampling of individual values. *Solid black line* shows the true regression line (a = 5 and b = 3.5) that originates from the generating species–species values, around which the individual-specific data were simulated under two different variance scenarios: **a** small within-species variance, **b** large within-species variance

Given such a tree structure linking all individuals of all species, the contrasts can be obtained in the classical way (e.g., Felsenstein 1985) with the exception that they are not standardized by their variances but are multiplied by coefficients that include a weight factor for the number of observations. Specifically, the modified method does not scale the contrasts to have equal variances, but it rather applies an orthonormal transformation on the original variables so that the sum of squares of the coefficients in the contrasts is forced to add up to one. Under such constraints, at the within-species level, contrasts can be written as (Felsenstein 1985; Paradis 2011):

$$\begin{aligned} c_{i_1} &= \sqrt{\frac{1}{2}}(y_{i_1} - y_{i_2}) \\ c_{i_2} &= \sqrt{\frac{2}{3}}\left(y_{i_3} - \frac{y_{i_1} + y_{i_2}}{2}\right) \\ &\vdots \\ c_{i_{n_i-1}} &= \sqrt{\frac{n_i - 1}{n_i}}\left(y_{i_{n_i}} - \frac{1}{n_i - 1}\sum\nolimits_{j=1}^{n_i-1} y_{ij}\right) \end{aligned} \tag{7.11}$$

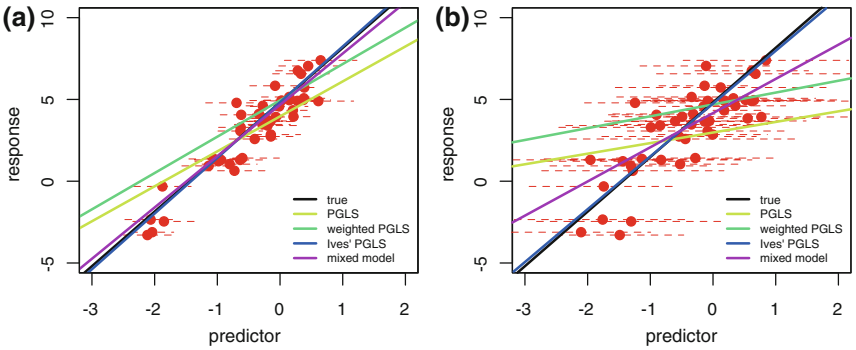

Fig. 7.4 Dealing with within-species variation in interspecific comparisons by modern phylogenetic methods (listed on Table 7.1). Each data point represents a species-specific mean that that can be measured with a certain precision as estimated by the within-species variance on the predictor (for simplicity, only errors on the predictors were considered). The interspecific relationship between the predictor and response was assessed by (1) Phylogenetic least squares methods without considering within-species variance (*yellow line*), (2) weighted phylogenetic least squares models that give emphasis on different data points according to the underlying variance (*green lines*), (3) measurement error PGLS models as was suggested by Ives et al. (2007) (*blue line*), and (4) phylogenetic mixed models (*purple line*). Solid black line shows the true regression line (a = 5 and b = 3.5) that originates from the generating species–species values, around which individual-specific data were simulated under two different variance scenarios: **a** small within-species variance (repeatability ~0.75) and **b** large within-species variance (repeatability ~0.25) on the predictor. *Dashed horizontal lines* indicate the degree of intraspecific variation within the predictor

where

c_i	$1...n_{i-1}$ within-species contrasts for species *i*,
n_i	number of individuals measured in species *i*,
y_i	$1...n_i$ individual observations for species *i*, and
$\sqrt{\frac{n_i-1}{n_i}}$	general form of the normalizing constant.

Accordingly, having four observations in a species for example, three contrasts can be derived with normalizing constants $\sqrt{1/2}$, $\sqrt{2/3}$, and $\sqrt{3/4}$, respectively. Considering alternative branching patterns within species, there can be several ways to calculate the orthonormal contrasts, each resulting in the linear combination of the original measurements taken at the individuals that defines the species-specific character states at the end. Taking the orthogonality constraints into account, Felsenstein further developed an algorithm for computing between-species contrasts recursively. In this approach, for each character, within- and between-species sets of contrasts can be derived fulfilling the constraint that different contrasts for same characters are independent. Therefore, the same contrasts for different characters will have the covariance that is equal to the covariance of the original character values.

Unlike the contrasts that are obtained by the original method relying on species means, the contrasts derived from individual data cannot directly be imported to conventional statistical approaches to estimate parameters of evolutionary importance (e.g., a slope from a regression forced through the origin). This is because covariances are composed of different components at the within- and between-species level (i.e., phenotypic covariance arising from between-individual associations and covariance due to convergent evolution, Fig. 7.2). However, the new contrast method also includes an expectation-maximization (EM) algorithm (Dempster et al. 1977) to partition observed covariances among traits, similarly to Lynch's model, into phylogenetic (i.e., the between-species covariances) and a phenotypic (i.e., the within-species covariances) components according to the model

$$\mathbf{T} \otimes \mathbf{A} + \mathbf{I} \otimes \mathrm{P}, \tag{7.12}$$

where

T phylogeny matrix (expected covariances based on the length of shared branches),
A between-species (phylogenetic) covariance matrix,
P within-species (phenotypic) covariance matrix,
I an identity matrix,
$\otimes$ Kronecker product multiplication (each element of the first matrix is multiplied by each element of the second matrix).

The elements of the **A** and **P** matrices can be estimated from the contrasts (see for example in the OPM relying on *varCompPhylip* that calls functions from program *Phylip*) and can subsequently be used for making inferences about the coevolution of traits. For example, if the aim is to challenge the null hypothesis that species-specific values of two (or more) traits vary independently of each other with regard to the phylogeny while accounting for the potential within-species covariation of traits, one can fit a model in which the elements of $\mathbf{A}_0$ are forced to be zero. This model can be compared by using likelihood ratio test with the model that is based on an $\mathbf{A}_1$ matrix that represents the true associations between species due to common ancestry. Estimated covariances from the model with the highest likelihood can also be used to obtain parameters from regressions of the variables on each other or correlation coefficients that are not confounded by phylogenetic associations and within-species covariances.

Extensions to the PGLS

Approaches based on phylogenetic generalized least squares provide a rich set of tools to study the effect of measurement error in phylogenetic comparative studies. In these models, following the logic of Martins and Hansen (1997), within-species

variance is lumped within an error term in the regression equation. Therefore, individual-specific estimates are not required, neither do assumptions about phylogenetic resolutions within species. The model can flexibly accommodate information on within-species variation. The flexibility is also prevalent in the fact that the model allows measurement errors to be same or different for different taxa, and actually, it assumes that within-species variances are available without bias and not needed to be estimated. Of further advantage is that the PGLS approach can not only applied to investigate questions about the correlated evolution of traits, but can also be tailored to various test situations.

The PGLS approach for accounting measurement error in interspecific comparative studies relying on species as the unit of analysis adheres to the following logic. Using information on the phylogeny and within-species variances (both are known from data), and considerations about how to combine different error components due to phylogenetic effects and within-species variances (and other sources), one can establish an overall variance structure to describe the expected covariance matrix in the models' residual. By the careful definition of the overall covariance structure, it is possible to handle cases when measurement errors are correlating or even have an interaction with the phylogenetic error. Then, the observer is left with a model-fitting problem, in which the task is to maximize the probability of data conditioned on the expected covariances. Therefore, parameter estimates from a best-fitting model with an error term that is composed of phylogenetic and measurement effects can be used to make inferences that are not confounded by these factors.

Such an approach has been taken forward by Ives et al. (2007), who derived statistical methods for the analysis of phylogenetically correlating data with within-species variation to investigate a broad array of evolutionary questions. Their entire methodology was built on the simple foundation that sampling variance can be added to the variance that is determined by the phylogenetic relationship of species (Martins and Hansen 1997). This scheme, on the one hand, can be applied to univariate models of evolution, when the interest is on ancestral states, rates of evolution and phylogenetic signal charactering the evolution of a single trait. Such a model can be depicted as (see also Eq. 7.10):

$$\mathbf{y} = \beta_a \mathbf{1} + \boldsymbol{\varepsilon}_{\mathbf{S}} + \boldsymbol{\varepsilon}_{\mathbf{M}} \tag{7.13}$$

$$\boldsymbol{\varepsilon}_{\mathbf{S}} \sim \mathcal{N}(0, \sigma^2 \mathbf{C}), \tag{7.13a}$$

$$\boldsymbol{\varepsilon}_{\mathbf{M}} \sim \mathcal{N}(0, \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{within}} \mathbf{I}), \tag{7.13b}$$

where

$\mathbf{y}$	vector of the observed trait values (dimension: $N_{\text{species}} \times 1$),
β_a	a scalar giving the expected value (i.e., ancestral state at the base of the tree) of the trait,
$\mathbf{1}$	vector of ones,
$\boldsymbol{\varepsilon}_{\mathbf{S}}$	vector of variances caused by the phylogeny (dimension: $N_{\text{species}} \times 1$),

$\boldsymbol{\varepsilon_M}$ vector of measurement error variances (dimension: $N_{species} \times 1$),
$\mathbf{C}$ correlation structure defined by the phylogeny (dimension:$N_{species} \times N_{species}$),
$\mathbf{I}$ identity matrix (dimension: $N_{species} \times N_{species}$),
σ^2 the overall phylogenetically inherited variance (rate of evolution),
$\boldsymbol{\sigma}^2_{within}$ measurement or within-species variance vector (dimension: $N_{species} \times 1$).

This signifies that the total error term $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon_S} + \boldsymbol{\varepsilon_M}$ depicts a multivariate normal distribution and has a covariance matrix $\sigma^2\mathbf{C} + \boldsymbol{\sigma}^2_{within}\mathbf{I}$. The phylogenetic component comes from a distribution with a zero mean and a variance that is described by the covariation matrix $\sigma^2\mathbf{C}$. This matrix is composed of σ^2 that represents the phylogenetically inherited variance (i.e., the rate of evolution), and $\mathbf{C}$ that stands for the correlation structure that can be defined by the length of shared branches on the phylogeny. On a similar vein, the measurement error component also follows a normal distribution with a zero mean and with a covariance matrix of $\boldsymbol{\sigma}^2_{within}\mathbf{I}$ that gives the diagonal matrix of measurement errors (or within-species variances). These divisions assume that the trait evolution follows the Brownian motion model and that measurement errors are uncorrelated. However, other evolutionary models can be accommodated by the appropriate translation of branch lengths into the $\mathbf{C}$ matrix, while correlated measurement errors can be treated by nonzero off diagonals in the $\boldsymbol{\sigma}^2_{within}\mathbf{I}$ (which then can be written as $\boldsymbol{\sigma}^2_{within}\mathbf{M}$). Elements of $\mathbf{C}$ and $\boldsymbol{\sigma}^2_{within}$ are given by the data (i.e., known phylogeny and standard errors around species-specific means), the only parameters that are unknown are the ancestral state (β_a) and the rate of evolution (σ^2). To estimate these parameters, Ives et al. (2007) describe a model-fitting iteration processes based on estimated generalized least squares (EGLS), maximum likelihood (ML) and restricted maximum likelihoods (REML). In addition, by determining the rate of evolution under the assumption of no phylogenetic structure in the data ($\mathbf{C} = \mathbf{I}$, assuming star phylogeny) and at the observed phylogeny, it also becomes possible to calculate the strength of the phylogenetic signal that is prevalent in the data in terms of Blombergs' K (Blomberg et al. 2003). Some procedures for characterizing univariate trait evolution by using the method by Ives et al. (2007) are given in the OPM.

Ives et al. (2007) also present a list of models for multivariate cases, when phylogenetic associations between traits are in the focus. Along this line, they make PGLS technique suitable for designs such as correlations, principal component analysis, multiple regression as well as reduced major axis regression for functional relationships (but Hansen and Bartoszek (2012) warn against this last application). The underling formula has a composition of

$$\mathbf{w} = \mathbf{b} + \boldsymbol{\varepsilon_S} + \boldsymbol{\varepsilon_M} \tag{7.14}$$

where

$\mathbf{w}$ vector of species-specific tip values for traits $\mathbf{x}$ and $\mathbf{y}$ that are placed on top of each other (dimension: $2N_{species} \times 1$),

b vector containing the ancestral states (β_a) for the two traits, with the first $N_{species}$ elements being β_{a_x}, while the second $N_{species}$ elements being β_{a_y},

$\boldsymbol{\varepsilon_S}$ phylogenetic error term vector (dimension: $2N_{species} \times 1$), in which the phylogenetic variance on **x** ($\boldsymbol{\varepsilon_{S_x}}$) is stacked on top of the phylogenetic variance on **y** ($\boldsymbol{\varepsilon_{S_y}}$),

$\boldsymbol{\varepsilon_M}$ measurement error vector (dimension: $2N_{species} \times 1$), in which the within-species variances of **x** ($\boldsymbol{\varepsilon_{M_x}}$) is stacked on top of the within-species variance vector for **y** ($\boldsymbol{\varepsilon_{M_y}}$).

The phylogenetic error term ($\boldsymbol{\varepsilon_S}$) has a joint covariance matrix, in which the diagonal blocks are $\sigma_x^2\mathbf{C}$ and $\sigma_y^2\mathbf{C}$, while the off-diagonal blocks are composed of $\sigma_x\sigma_y\mathbf{C}$ matrices that are multiplied by a linear combination of parameters that describes the association between traits **x** and **y** (i.e., r, correlation coefficient or β, regression slope). The error term $\boldsymbol{\varepsilon_M}$ is organized analogically, thus has a joint covariance matrix based on blocks of $\boldsymbol{\sigma^2_{within_x}}$ and $\boldsymbol{\sigma^2_{within_y}}$ measurement error variances in the diagonal, and matrices of $\boldsymbol{\sigma_{within_x}\sigma_{within_y}}$ representing the covariances in measurement errors scaled by a factor that is proportional to the strength of association between **x** and **y**. The unknown parameters in these models are β_{a_x}, β_{a_y}, σ_x^2, σ_y^2, and the coefficient r or β reflecting the strength of relationship between **x** and **y** originating from their correlated evolution. The procedures based on EGLS, ML, and REML approaches can be used for the estimation of these parameters and even can be flexibly extended to multivariate cases such as principal component analysis or multiple regression (i.e., when $\beta = (\beta_1, \beta_2 \ldots)$). Ives et al. (2007) recommend parametric bootstrapping (a procedure, in which estimated parameters are used to simulate a large number of datasets; then, the parameters are repeatedly re-estimated from each simulated data) to determine confidence interval around these estimates, which can be used for hypothesis testing (i.e., r or $\beta \neq 0$). (see examples in the OPM).

Emphasizing the importance of the discrimination between observation errors acting on the response and predictor variables in evolutionary regressions, Hansen and Bartoszek (2012) suggested an alternative PGLS model. In their appraisal, employing the above abbreviations, the general statistical equation can be written as:

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \cdots + \beta_m\mathbf{x}_m + \boldsymbol{\varepsilon_S} + \boldsymbol{\varepsilon_M} - \left(\beta_1\boldsymbol{\varepsilon_{M_{x_1}}} + \cdots + \beta_m\boldsymbol{\varepsilon_{M_{x_m}}}\right) \quad (7.15)$$

where

y vector of the dependent variable (dimension: $N_{species} \times 1$),

$\mathbf{x}_1 \ldots \mathbf{x}_m$ vectors for $1 \ldots m$ dependent variables (each with the dimension of $N_{species} \times 1$),

$\beta_0 \ldots \beta_m$ regression parameters including the intercept (β_0) and the slopes for each predictors ($\beta_1 \ldots \beta_m$),

$\boldsymbol{\varepsilon_S}$ residual phylogenetic error term vector,

$\boldsymbol{\varepsilon_M}$ measurement error vector for the dependent variable,
$\boldsymbol{\varepsilon_{M_{x_1}} \ldots \varepsilon_{M_{x_m}}}$ measurement error vectors for each (1…m) predictor variable.

Therefore, the residual error in the model is composed of three components; the joint covariance matrix includes (1) a matrix that describes the model of evolution (in a form of $\sigma^2\mathbf{C}$), (2) a matrix of raw observation variance in the response variable (in a form of $\boldsymbol{\sigma^2_{within_y}}$), and (3) a matrix representing the effects of measurement error in the predictor variables (a complicated variance structure relying on observation variances conditional on the observed values of the predictor variables) that is multiplied by the associated regression slopes (formally can be written as $\mathrm{Var}[\beta\boldsymbol{\sigma^2_{within_x}}|\mathbf{X}]$). The effect of measurement errors in the predictors needs to be complex, because errors in one predictor will carry over and have an influence on coefficients that pertain to other predictors.

Hansen and Bartoszek (2012) argued that applying an EGLS procedure to obtain regression parameters from the above model results in more precise confidence intervals around the estimates, but it does not remove the downward biases that are caused by measurement errors around the predictors. In classical models of measurement errors (Madansky 1959; Fuller 1987; Buonaccorsi 2010), such attenuation bias can be corrected via a reliability ratio, K. This is defined as the ratio between the true and observed sum of squares of the predictor variable, with the former being estimated by subtracting the observational variance from the observed sum of square. Hansen and Bartoszek (2012) present equations for the calculation of reliability ratio when data are correlating due to the phylogenetic structure, thus the logic of using a correction factor on the regression slope via K can be adopted to aforementioned PGLS approach. They also include these calculations in their estimation procedure, in which they combine ML-based and EGLS approaches for the estimation of different parameters in their complex model. In the OPM, I show how these functions (as implemented in GLSME) work in practice.

The main difference between the method by Ives et al. (2007) and that of Hansen and Bartoszek (2012) is that the former approach does not strategically discriminate between the effect of measurement error around the predictor and response variables (these are lumped within the $\boldsymbol{\sigma^2_{within}}$ matrix). On the other hand, Hansen and Bartoszek (2012) uses separate matrices for the observation variance for the response $\boldsymbol{\sigma^2_{within_y}}$ and predictor $\boldsymbol{\sigma^2_{within_x}}$ variables, with the latter having complicated effects. Furthermore, strategies are different with regard how the two methods correct for attenuation bias (e.g., for this purpose Hansen and Bartoszek use the reliability ratio parameter).

Further Developments on the Phylogenetic Mixed Models

Implementing Lynch's original suggestion, Housworth at al. (2004) and Hadfield and Nakagawa (2010) brought the framework based on mixed modeling for the

study the evolution of traits into practice. Along this line, Hadfield and Nakagawa (2010) emphasized that animal models used to decompose variance components in quantitative genetics are built on a mathematical basis that is very similar to what is applied in phylogenetic mixed models. This analogy arises because the matrices that define the relationship between subjects, i.e., those that represent the phylogeny and the pedigree, can be brought into a structure that is formally equivalent. Based on this relationship, the quantitative genetics toolbox can be efficiently exploited for the decomposition of error structures in phylogenetic comparative studies (see more details in Chap. 11).

Accordingly, many test situations in interspecific comparative studies can be regarded as variations to the same underlining statistical foundation, the phylogenetic meta-analysis. In this context, the general mixed model for univariate questions can be written as (see also Eqs. 7.1, 7.8 and 7.9)

$$y_i = \beta_0 + a_i + e_i + m_i, \tag{7.16}$$

$$\mathbf{m} \sim \mathcal{N}(0, \boldsymbol{\sigma}^2_{\mathbf{within}}\mathbf{I}) \tag{7.16a}$$

where

y_i	species-specific (or study-specific) effect size or trait value,
m_i	measurement error for species (or study) i (ith element of the **m** vector with dimension of $N_{\text{species}} \times 1$),
$\boldsymbol{\sigma}^2_{\mathbf{within}}$	within-species variance caused by differences between individuals or species, the corresponding identity matrix (**I**) that has a dimension of $N_{\text{species}} \times N_{\text{species}}$,
β_0, a_i, and e_i	intercept (ancestral state at the root of the phylogeny), the effect on species (or study) i that is caused by the common descent and the non-heritable residual component of variation (residual error), respectively, as in Eq. (7.8).

Chapter 11 shows how it can be extended into bi- or multivariate problems. The matrix of expected (co)variances among subjects caused by the random effects can be thus approached by a joint covariance matrix that is composed of the phylogenetic variance (again, in a form of $\sigma^2\mathbf{C}$), the measurement error variance (in a form of $\boldsymbol{\sigma}^2_{\mathbf{within}}\mathbf{I}$ if such errors are known) and the residual variance ($\sigma^2_e\mathbf{I}$ assuming that residuals are homoscedastic). The effects that are lumped in the residual component may be important if deviations from the expected species-specific means are mediated by processes that act independently of phylogeny and within-species variances. Note that most of the above models can also be fitted to this general outline. For example, the method by Ives et al. (2007) considers a covariance structure, in which the phylogenetic and measurement error components are combined in the same additive fashion, and which is equivalent with the general phylogenetic meta-analysis model under the $\sigma^2_e = 0$ scenario. When species-specific traits could be completely described by the Brownian motion of

evolution, Felseinstein's (2008) model can also be brought into a similar structure with the main difference that σ^2_{within} is common to all species and remains an unknown quantity.

To determine the unknown parameters in the model (e.g., β_0 and σ^2), previous models based on PGLS or independent contrasts use estimation procedures available in ML, EGLS, REML, or EM algorithms. However, for these approaches, it becomes challenging to deal with non-Gaussian distributions as well as with missing information regarding the species-specific trait values, within-species variances, and phylogenetic resolutions. To overcome these shortcomings, Hadfield and Nakagawa (2010) proposed a method based on Markov chain Monte Carlo (MCMC) simulation that can be accommodated to a wide range of phylogenetic questions and data types. MCMC can be used to fit the general model:

$$l = \mathbf{W}\boldsymbol{\theta} + \mathbf{e}, \tag{7.17}$$

where

l a latent variable that provides the link function (e.g., Poisson and exponential) to the values of the **y** response variable,
w design matrix of predictor variables $\mathbf{x}_1 \ldots \mathbf{x}_m$,
θ vector of fixed and random effects, and
e vector of residuals.

Estimation protocols are needed to obtain l, **θ**, the variance components σ^2 (corresponding to the distribution of phylogenetic random effect included in **θ**), and σ^2_e (corresponding to the distribution of residual error **e**). The estimation of l is achieved through a Metropolis–Hasting iteration process (Metropolis et al. 1953; Hastings 1970), while θ and the variance components are approximated by Gibbs sampling (Geman and Geman 1984). By including a matrix of within-species variances into the definition of **θ**, such sources of variation can be flexibly incorporated into models that represent various evolutionary questions. The obtained parameter estimates from such models thus can be regarded as being independent of the confounding effect of measurement error. Check the OPM for an example based on the *R* package *MCMCglmm*.

Methods Based on the Evaluation of Likelihood Surfaces

Beside the problems posed by non-Gaussian distribution and missing data, another limitation can appear in classical approaches to phylogenetic comparative questions. Specifically, they are constrained to use particular assumptions about the mode of trait evolution. The evolutionary models that are traditionally considered are the Brownian motion and the Ornstein–Uhlenbeck process, whose likelihood functions are mathematically tractable for parameter estimations. For example, due to analytical convenience, the above models generally accommodate the Brownian fashion of evolution during the translation of branch length into the

covariance matrix (i.e., elements of **C** are proportional to branch lengths). Some departures from this standard can be achieved by the appropriate adjustment of the formula, but more complicated models of evolution, such as branch-specific directional selection requiring the estimation of large number of parameters, cannot be tracked in this way.

To surmount this obstacle, Kutsukake and Innan (2012, see also Chap. 17) introduced a method that is able to deal with more complex and realistic modes of evolution. Instead of forming mathematical formulas to describe the relationship between known data and the parameters of interest, they advocate the simulation of data at different parameter settings to examine how well such simulation results coincide with the observed data. By simulating a large number of data at various parameter sets representing different hypotheses about patterns of evolutions, we only need to examine the likelihood of each set based on the joint probability of the observed and simulated data. Parameter combinations that provide the highest likelihood can be used for making evolutionary inferences. If this ML estimation is supported by approximate Bayesian computation (ABC, see Chap. 17), the algorithm can take into account prior information on the expected distributions for all parameters that can result in increased power.

An important flexibility of the above simulation-based method is that it can also incorporate intraspecific variation. Although the simulation process focuses on phenotypic trait values as estimated at the tips of the phylogeny, and the simulated and the observed data are compared at the level of species, the likelihood function can be adjusted for patterns of phenotypic variation accumulating within species. Accordingly, the joint probability of observed and simulated data can be extended to include species-specific trait values as well as standard deviations around them. In fact, not only normal distributions, but also other forms of within-species distributions can be considered. Consequently, if data are available on how the data are spread within species, such information can be efficiently incorporated in the estimation of likelihood surfaces of model parameters. Focusing on the comparison of three evolutionary models, Kutsukake and Innan (2012) present an example analysis for the estimation of ancestral statesbased on phenotypic data (mean and standard deviation) that are sorted at the level of species. However, their logic can be tailored to various evolutionary questions, and equivalent approaches can be designed for multitrait evolution, when the interest is to obtain parameters to describe patterns of correlated evolution.

Another likelihood-based method to incorporate intraspecific variation has also been developed in the Bayesian framework that relies on MCMC methods (Revell and Reynolds 2012). The main difference between this and most of the above-discussed comparative methods is that Bayesian method uses trait data that are broken down to individuals, while species-specific trait values are not needed as an input (note that merely the independent contrast approach relies on similar requirements). Therefore, only the phylogeny, individual values and evolutionary models are treated as known (Revell and Reynolds 2012 considered Brownian process for their description, but other evolutionary models can also be envisaged). Parameters that express species means and variances are estimated together with

the parameters of the evolutionary model from their joint probability distribution. Therefore, such an approach can account for the possibility that processes involved in the considered evolutionary model and its parameters can affect species means and variances. The simulations accompanying the methodology of Revell and Reynolds (2012) indicated that true species means are not necessarily the same as the arithmetic mean of individual-specific values, especially when intraspecific variance is relatively high. This suggests that accounting for the tree and the modes of trait evolution may be warranted when inferring about tip values at the species level from within-species samples. This does not only apply to the means but also to variances, as patterns of the intraspecific distribution of traits may differ between species in a phylogenetically determined fashion. In the OPM, I demonstrate the use of this Bayesian method focusing on functions available in *phytools*.

The philosophy of the two above methods (Kutsukake and Innan 2012; Revell and Reynolds 2012) is very similar in the sense that they both revolve around a maximization of a likelihood function for the proposed set of evolutionary parameters. To obtain this, one needs to derive the probability of data conditioned on the means and variances of species as well as on the parameters of the underlying evolutionary model and the phylogeny. While Kutsukake and Innan (2012) evaluate the surface of likelihoods through processes of data simulation at different combinations of parameters, Revell and Reynolds (2012) put forward an MCMC-based Bayesians algorithm to obtain the joint posterior probability distributions of parameters. In the latter procedure, propositions for parameter values are being made at each node of the chain, and these propositions are accepted proportional to their likelihood.

There are also differences in how intraspecific variance is incorporated in the likelihood function. The simulation-based method assumes that this is a known property and thus the probability function of data given the simulated values can be simply adjusted. On the other hand, in the Bayesian approach, within-species variations are unknown, thus two different probabilities are needed for the formation of likelihood. The first part gives the probability that the proposed mean data arose from the model of evolution and phylogenetic tree, while the second part describes the probability that observed individual-specific data conditioned on the proposed species-specific means and variances. The advantages of the different strategies applied in the above two methods might be exploited depending on the question and data at hand. For example, if the biological hypothesis under testing is related to the evolution of species-specific trait and within-species variance is only regarded as a confounder, the model by Kutsukake and Innan (2012) may be appropriate. However, if we have individual-specific values, we can investigate interesting questions about the evolution of within-species variances.

Adjusting for Unequal Within-Species Sample Sizes

To correct for heterogeneity in sample sizes among subjects, one can use weighted regressions (Draper and Smith. 1981; Neter et al. 1996), in which each data point

is given an emphasis according to the corresponding sample size. Such a weighting approach can also be adopted for the phylogenetic framework, but it has seen little test in that context. Given that standard error reciprocally scales with sample size (Box 7.1), the above methods (e.g., PGLS, phylogenetic mixed models) using information on error variances can be supplied with $1/n_i$ as an estimate of within-species variance component. Such an analysis will give results that are adjusted for differences in sample size. If both standard errors and sample sizes are provided, the within-species variance that is corrected for sample size can be calculated based on the regression of standard error against the sample size, which can be subsequently used in a measurement error model. Some examples are shown in the OPM.

Another way to deal with non-constant sample size is to apply data imputation method that produces estimates for cases when information is unavailable. Variation in sample size can be considered as a consequence of missing information for some individuals (Garamszegi and Møller 2011). Various approaches are available to input missing data (Nakagawa and Freckleton 2008) in order either to equalize within-species sample size by augmenting individual-specific data or even to simulate data for species for which no data are available at all. Unfortunately, such imputation methods have been rarely exploited in the phylogenetic contexts (see Fisher et al. 2003).

7.4.4 Interpreting Phylogenetic Results in Light of Within-Species Variation

Results from the above exercises should be interpreted carefully. It has been noted, for example, that different approaches may provide somewhat different outcomes (e.g., Ives et al. 2007). Hence, repeating the analyses by using alternative methods if these are available may help establish our confidence in the validity of detected patterns. If discrepancies are found, we may need to revisit the assumptions of different models to check whether these were violated. Furthermore, the visual inspection of data can also enhance the interpretations. For instance, the types of graphics presented in this chapter show error ranges or sample sizes around the species-specific estimates and also illustrate the results with and without accounting for measurement errors (Figs. 7.1 and 7.2).

How does controlling for within-species variance change the results compared to the situation when species-specific mean values are assumed to have no errors? Does this difference correspond well with the estimated trait repeatabilities (i.e., high repeatability should cause only minor difference between different outcomes)? Answers to such questions may help elucidate the validity of the findings. The inspection of the estimated parameters can also be informative. In general, as discussed above, we should expect that a control for within-species variance increases phylogenetic signal in the data, while the regression slope or correlation

coefficient strengthens after removing the attenuation bias that the measurement error causes. Finally, we can also check the model fit statistics to verify whether the model accounting for within-species variance offers better fit to the data.

7.5 Discussion

In this chapter, I investigated issues about the importance of within-species variance in phylogenetic comparative studies that generally focus on species-specific means as a unit of analysis. Evidence from simulation studies suggests that this focus on interspecific patterns should not inherently imply that intraspecific patterns are to be ignored. The potential problems posed by measurement errors around species-specific means do not only have consequences for how we analyze data, but also for how we design and collect data for comparative studies, how we interpret phylogenetic findings, and ultimately, how we think about evolution.

Warnings about the importance of the incorporation of within-species variance into analyses at the across-species level have emerged in the recent literature, which may likely demarcate avenues for the development of the methodology in the future. However, I would argue that before such statistical developments take place, empirical studies are needed that confirm the biological relevance of the appropriate control methods. Lessons about the use of phylogenetic control teach us that although interspecific data are unavoidably structured by common ancestry, the phylogenetic control is warranted only if the data at hand require so (Freckleton 2009; Revell 2010). Similarly, the smart application of measurement error models that also involves biological considerations and a closer look into the data should be preferred over the blind submission of subsequent comparative datasets to complex analyses with intraspecific variation (note that more complex models usually require fulfilling more assumptions). Accordingly, in spite of the fact that in theory it seems necessary to deal with the confounding effects arising from within-species distributions, in practice it may appear that the available data represent a case when variation below the species level is negligible, and when classical comparative methods perform with high confidence as well. I suspect that this situation will call for the performance of diagnostics statistics in most of the studies rather than full exploitation of phylogenetic approaches that account for measurement error.

Thinking based on biological motivation may also direct us into a fascinating research direction. So far, statistical considerations implied that within-species variation is a somewhat unwanted side effect that we should get rid of in the comparative analysis. This echoes the philosophy that was applied in the early days of phylogenetic methodology, i.e., when the phylogenetic structure in the data was regarded as something that should be removed from the data, for example via the use of independent contrasts. Only later progress realized that the phylogenetic trees in fact may be incorporated into the analyses in a more beneficial fashion that allows making inferences about the modes of evolution (see Chap. 1). In a similar

vein, from the current state of the art, subsequent research may also recognize that measurement error in a statistical sense may hold interesting information from the biological perspective (see Sect. 7.2) that can be exploited fruitfully. Therefore, methods that do not only control for and factor out intraspecific variation, but can also deal with its evolutionary relevance may open new dimensions. For instance, within-species variance itself can be subject to selection, thus incorporating such information into the phylogenetic comparative study as a covariate rather than handling it as a confounder may provide interesting results. As an example, in a study of flight initiation distance in birds, Møller and Garamszegi (2012) found that within-species variance of this trait can be shaped by ecological factors, suggesting that these are not just random variation around a species-specific mean. Another toolbox that holds promises toward the same direction is the phylogenetic meta-analysis (see Chap. 11), which brings effect sizes together with their confidence intervals into the focus. In such an approach, each species-specific effect size represents the strength (and direction) of a particular relationship between two traits, while confidence intervals describe the precision of the mean effect size estimate based on the underlying within-species sample size. The meta-analytic study of phylogenetically structured effect sizes is basically a comparative problem that also considers within-species variance around species-specific estimates. The exploitation of such methods for investigating evolutionary questions awaits further progress. These may be interesting, for example, when traits can show correlations at both the within- and between-species level (see Fig. 7.2), and the aim is to explain why some species display strong relationships, while others weak relationship between two phenotypic traits.

The approaches discussed in this chapter generally assume that individual- or population-specific measurements are independent of each other and thus depict no further phylogenetic or other hierarchical structure within species (i.e., they can be visualized on the phylogeny as forming a star polytomy with zero branch length). However, such an assumption may not be necessarily true, especially when within-species variance arises from variations between populations. In fact, populations of the same species can have a certain evolutionary history, thus they cannot be regarded as phylogenetically independent replicates, such as data from different individuals (see Edwards and Kot 1995 who first used phylogenetic comparative methods on intraspecific data). Moreover, populations are structured in space that has consequences for migration and gene flow so that phenotypes in one locality are affected by processes acting in other neighboring localities. Therefore, inferences made from across-population patterns need to consider statistical issues about non-independence at least due to two factors: phylogeny and gene flow. Felsenstein's group described various methods that are able to quantify evolutionary patterns across multiple populations within a single species (Stone et al. 2011).

Such methods may, however, be developed further, and combined with other comparative methods to partition variances and evolutionary patterns acting at the between-population and the between-species levels (e.g., by relying on the mixed model framework). For example, it might be straightforward to first estimate

species-specific trait values and variances over multiple populations by incorporating the effects of phylogeny and gene flow sensu Stone et al. (2011) and then subsequently use such species-specific estimates in an interspecific comparison. Moreover, it might be straightforward to make distinction between cases when populations should be treated as separate entries in the analysis and when they should be pooled as one species. This might, of course, depend on several factors such as data availably, the biological question, the relative importance of gene flow, and phylogenetic constraints.

I envisage there being great potential for further development of comparative methods incorporating measurement error along various other lines. Most of the advancements have been made so far correspond to situations when the correlated evolution of traits is of interest. However, there might be other phylogenetic problems and designs that also warrant considerations about within-species variance. In practice, we might also need specific methods that are able to deal with non-normal or skewed within-species distributions, count data as well as with missing data. Furthermore, so far there is not a strong distinction between different sources of within-species variance in the statistical approaches. Therefore, it may prove useful to derive methods that can separate or combine variance components that originate from instrumental errors, within- or between-individual variations, and fluctuations across populations. Finally, there might be a scope for amalgamating methods that implement uncertainty in estimating tip values and that consider measurement errors in a phylogenetic tree (de Villemereuil et al. 2012).

References

Adolph SC, Hardin JS (2007) Estimating phenotypic correlations: correcting for bias due to intraindividual variability. Funct Ecol 21(1):178–184

Arnold C, Nunn CL (2010) Phylogenetic targeting of research effort in evolutionary biology. Am Nat 176:601–612

Ashton KG (2004) Comparing phylogenetic signal in intraspecific and interspecific body size datasets. J Evol Biol 17:1157–1161

Blomberg S, Garland TJ, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more laible. Evolution 57:717–745

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24:127–135

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Buonaccorsi JP (2010) Measurement error: models, methods, and applications. Chapman and Hall, New York

Caro TM, Roper R, Young M, Dank GR (1979) Inter-observer reliability. Behaviour 69:303–315. doi:10.1163/156853979x00520

Chesher A (1991) The effect of measurement error. Biometrika 78(3):451–462. doi:10.1093/biomet/78.3.451

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

Christman MC, Jernigan RW, Culver D (1997) A comparison of two models for estimating phylogenetic effect on trait variation. Evolution 51(1):262–266. doi:10.2307/2410979
Cornillon PA, Pontier D, Rochet MJ (2000) Autoregressive models for estimating phylogenetic and environmental effects: accounting for within-species variations. J Theor Biol 202(4):247–256. doi:10.1006/jtbi.1999.1040
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
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B Methodol 39(1):1–38
Doughty P (1996) Statistical analysis of natural experiments in evolutionary biology: comments on recent criticisms on the use of comparative methods to study adaptation. Am Nat 148:943–956
Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York (2nd edn)
Edwards SV, Kot M (1995) Comparative methods at the species level: geographic variation in morphology and group size in Grey-crowned Babblers (*Pomatostomus temporalis)*. Evolution 49:1134–1146
Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15
Felsenstein J (2002) Contrasts for a within-species comparative method. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics. Oxford University Press, Oxford, pp 118–129
Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland
Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171(6):713–725
Fisher DO, Blomberg SP, Owens IPF (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proc R Soc Lond Ser B-Biol Sci 270(1526):1801–1808
Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22(7):1367–1375
Fuller WA (1987) Measurement error models. Wiley, New York
Garamszegi LZ, Møller AP (2010) Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. Biol Rev 85:797–805
Garamszegi LZ, Møller AP (2011) Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst Biol 60:876–880
Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. Behav Ecol Sociobiol 66:1363–1373
Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge
Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Ieee Transactions on Pattern Analysis and Machine Intelligence 6(6):721–741
Gittleman JL, Kot M (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool 39(3):227–241. doi:10.2307/2992183
Grafen A (1989) The phylogenetic regression. Philos Trans R Soc B 326:119–157
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, 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 (2005) The effect of intraspecific sample size on type I and type II error rates in comparative studies. Evolution 59:2705–2710
Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109. doi:10.2307/2334940
Housworth EA, Martins EP, Lynch M (2004) The phylogenetic mixed model. Am Nat 163:84–96
Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56(2):252–270

Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee T-C (1985) The theory and practice of econometrics. Wiley, New York
Kreft IGG, de Leeuw J, Aiken LS (1995) The effect of different forms of centering in hierarchical linear models. Multivar Behav Res 30:1–21
Kutsukake N, Innan H (2012) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution (in press)
Lessells CM, Boag PT (1987) Unrepeatable repeatabilities: a common mistake. Auk 104:116–121
Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45(5):1065–1080
Madansky A (1959) The fitting of straight lines when both variables are subject to error. J Am Stat Assoc 54:173–205
Manisha S (2001) An estimation of population mean in the presence of measurement errors. J Indian Soc Agric Stat 54:13–18
Martins EP (1994) Estimating the rate of phenotypic evolution from comparative data. Am Nat 144:193–209
Martins EP (2004) COMPARE, version 4.6b. Computer programs for the statistical analysis of comparative data. Distributed by the author at http://compare.bio.indiana.edu/, Department of Biology, Indiana University, Bloomington IN
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
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Møller AP, Garamszegi LZ (2012) Between individual variation in risk taking behavior and its life history consequences. Behav Ecol 23:843–853
Monkkonen M, Martin TE (2000) Sensitivity of comparative analyses to population variation in trait values: clutch size and cavity excavation tendencies. J Avian Biol 31(4):576–579. doi:10.1034/j.1600-048X.2000.310417.x
Nakagawa S, Freckleton R (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23:592–596. doi:10.1016/j.tree.2008.06.014
Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev 85:935–956
Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago
Paradis E (2011) Analysis of phylogenetics and evolution with R, 2nd edn. Springer, Berlin
Purvis A, Rambaut A (1995) Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comput Appl Biosci 11:247–251
Purvis A, Webster AJ (1999) Phylogenetically independent comparisons and primate phylogeny. In: Lee PC (ed) Comparative primate socioecology. Cambridge University Press, Cambridge, pp 44–70
Reed GF, Lynn F, Meade BD (2002) Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagn Lab Immunol 9(6):1235–1239. doi:10.1128/cdli.9.6.1235-1239.2002
Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1(4):319–329. doi:10.1111/j.2041-210X.2010.00044.x
Revell LJ, Reynolds RG (2012) A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution 66(9):2697–2707. doi:10.1111/j.1558-5646.2012.01645.x
Ricklefs RE, Starck JM (1996) Applications of phylogenetically independent contrasts: a mixed progress report. Oikos 77(1):167–172
Snijders TAB, Bosker RJ (1999) Multilevel analysis—an introduction to basic and advanced multilevel modelling. Sage, London
Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn. W. H. Freeman and Co, New York

Stone GN, Nee S, Felsenstein J (2011) Controlling for non-independence in comparative analysis of patterns across populations within species. Philos Trans R Soc Lond B Biol Sci 366:1410–1424

Taper ML, Marquet PA (1996) How do species really divide resources? Am Nat 147:1072–1086

van de Pol MV, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav 77(3):753–758