Syst. Biol. 56(2):252–270, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701313830

Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods

ANTHONY R. IVES,[1] PETER E. MIDFORD,[2] AND THEODORE GARLAND, JR.[3]

[1]*Department of Zoology, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA; E-mail: arives@wisc.edu*
[2]*Department of Zoology, Southern Illinois University Carbondale, Carbondale, Illinois 62901, USA*
[3]*Department of Biology, University of California, Riverside, Riverside, California 92521, USA; E-mail: tgarland@ucr.edu*

Abstract.— Most phylogenetically based statistical methods for the analysis of quantitative or continuously varying phenotypic traits assume that variation within species is absent or at least negligible, which is unrealistic for many traits. Within-species variation has several components. Differences among populations of the same species may represent either phylogenetic divergence or direct effects of environmental factors that differ among populations (phenotypic plasticity). Within-population variation also contributes to within-species variation and includes sampling variation, instrument-related error, low repeatability caused by fluctuations in behavioral or physiological state, variation related to age, sex, season, or time of day, and individual variation within such categories. Here we develop techniques for analyzing phylogenetically correlated data to include within-species variation, or "measurement error" as it is often termed in the statistical literature. We derive methods for (i) univariate analyses, including measurement of "phylogenetic signal," (ii) correlation and principal components analysis for multiple traits, (iii) multiple regression, and (iv) inference of "functional relations," such as reduced major axis (RMA) regression. The methods are capable of incorporating measurement error that differs for each data point (mean value for a species or population), but they can be modified for special cases in which less is known about measurement error (e.g., when one is willing to assume something about the ratio of measurement error in two traits). We show that failure to incorporate measurement error can lead to both biased and imprecise (unduly uncertain) parameter estimates. Even previous methods that are thought to account for measurement error, such as conventional RMA regression, can be improved by explicitly incorporating measurement error and phylogenetic correlation. We illustrate these methods with examples and simulations and provide Matlab programs. [Ancestor reconstruction; comparative methods; estimated generalized least-squares; independent contrasts; maximum likelihood; morphometrics; principal components analysis; reduced major axis; regression; restricted maximum likelihood]

Most existing phylogenetically based statistical methods, as commonly applied, assume that within-species variation is absent or negligible (see reviews by Martins and Hansen, 1996; Rohlf, 2001, 2006; Garland et al., 2005). There are two practical reasons for this. First, many published comparative data sets do not include anything like estimates of standard errors associated with the mean values for species (or populations). Second, although standard statistical methods are available for incorporating measurement error and other sources of variation (Judge et al., 1985; Fuller, 1987), they are not commonly applied (Harmon and Losos, 2005), and they have rarely been considered in the context of phylogenetic statistics in which trait values are correlated among related species (but see, for example, Harvey and Pagel, 1991: chapter 6; Martins and Lamont, 1998; Housworth et al., 2004). A related issue is how to incorporate estimates of error in the phylogenetic topology and branch lengths used for analyses (Purvis and Garland, 1993; Garland and Díaz-Uriarte, 1999; Housworth and Martins, 2001; Huelsenbeck and Rannala, 2003). Here, however, we focus on variation in trait values rather than uncertainties in phylogenies, and throughout we assume the phylogenies are known without error.

Our goal is to provide methods for incorporating within-species variation into phylogenetically based statistical methods for continuous-valued traits. Outside of biological comparative studies, the statistical literature typically refers to the problem that we address as one of "measurement error" (Fuller, 1987). Measurement error refers to any type of variation between an observed value and the "true" value of interest, such as the mean value of a trait for a species or for a given population within a species. Thus, estimates of means for whole species will be affected by differences among populations, by how many populations are sampled to compute a composite mean, and by how many and what kind of individuals are sampled from each of those populations (Pagel and Harvey, 1988b; Harvey and Pagel, 1991). Measurement error also occurs within populations, with sources including sampling variation, instrument-related error, low repeatability caused by fluctuations in behavioral or physiological state, variation related to age, sex, season, or time of day, and true individual variation within such categories.

We argue that estimates of standard errors associated with mean values for species (or populations for phylogenetic comparisons of intraspecific trait variation, e.g., Ashton, 2004), as are now commonly reported in the empirical comparative method literature, provide a convenient, useful, and statistically justified way to capture the numerous sources of measurement error. Furthermore, accounting for measurement error in this way can improve parameter estimates and tests of statistical significance for problems involving phylogenetically correlated data. Historically, many comparative studies relied largely on previously published sources for their data. As such, they rarely reported more than a mean value for each taxon under consideration. More recently, however, comparative studies are often conducted de novo, such that new data are reported and standard errors (or standard deviations and sample sizes) are becoming more commonly available. These standard errors incorporate at least part of the total measurement error. Although estimating the total measurement error (e.g., the variation among all populations of a species) is unrealistic,

incorporating the measurement error associated with the observations that are actually made to determine a species mean may provide substantial improvement to statistical methods.

In addition to addressing the case when standard errors are available for each "tip" associated with a phylogenetic tree, our techniques can also be used when less information is available. For example, in allometric studies that aim to obtain functional relations among traits, "general structural relation" models are often employed, with reduced major axis (RMA) regression a special case that is most frequently used (Rayner, 1985). General structural relation models use very little information about measurement error; for instance, RMA regression assumes simply that the ratio of within-species variance for traits is equal to the ratio of total variance of the trait values. We provide statistical models for general functional relations that incorporate phylogenetic correlation and measurement error in the most general form, with phylogenetic counterparts to general structural relation models and RMA regression as special cases.

The source of measurement error, as we have broadly defined it, will affect its statistical properties. For instance, the measurement error for one trait might be uncorrelated to the measurement error for another trait if measurement error is caused by instrument-related error and each trait is measured with a different instrument. In contrast, if measurement error is caused by among-population variation within species, then measurement errors for different traits could be correlated; for example, a functional relation between body mass and leg length observed among species might also occur among populations within each species, causing within-species variation (measurement error) of the two traits to be correlated. Although it is rare for researchers to report correlations in measurement errors among traits, we nonetheless assume that this correlation can be nonzero so that the techniques we develop can be applied under the most general circumstances.

In the literature on phylogenetically based statistical methods ("comparative methods"), several studies have explicitly considered measurement error. Lynch (1991) used a mixed models approach to phylogenetic analyses that explicitly separates components of variation due to heritable and nonheritable sources; nonheritable sources of variation include measurement error. This approach forms the basis of our work and also other previous analyses of measurement error. Christman et al. (1997), Felsenstein (2004), and Housworth et al. (2004) incorporated measurement error to estimate correlations between traits by treating individual organisms as the units of study, grafting the data from individuals onto a phylogenetic tree, with each species represented by a hard polytomy of individuals. The length of the tip branchlets (i.e., the within-species variance) is estimated simultaneously with other parameters of the overall statistical model. Our general approach is closely related, although we separate the estimation of the measurement error (i.e., the standard errors of the species values) from the estimation of parameters in the model. Although our approach does not account for the uncertainty in the estimates of the standard errors of the within-species variance, the estimates of the standard errors are unbiased, and the availability of data on species (or population) means plus standard errors is generally much greater than the availability of raw data on the measurements of all individuals (e.g., whenever data come partly from published sources). Furthermore, our approach is both easier to apply and more flexible, allowing researchers to provide known information about measurement error associated with each tip value. Thus, even if raw data on the measurement of individuals are available, it will be easier to summarize this information as standard errors and use the procedures we derive. Finally, our methods can be modified for the case when even less is known about measurement error—for example, when the standard errors averaged among species are known even though the species-specific standard errors are not.

We derive a suite of methods incorporating measurement error into frequently used statistical tests. First, we incorporate measurement error into univariate models that aim to estimate ancestral traits (e.g., Bonine et al., 2005) and quantify the magnitude of phylogenetic signal; that is, the amount of variation among species that can be attributed to phylogenetic relatedness (Blomberg et al., 2003). Second, we develop methods for calculating correlation coefficients between traits while accounting for both phylogenetic relatedness and measurement error. An extension of the correlation analysis leads to a phylogenetic version of principal components analysis (PCA) that summarizes the correlations among multiple traits. Third, we incorporate measurement error into phylogenetic regression, when there is a single dependent variable and one or more independent variables. Fourth, we derive measurement error methods for functional relation models, in which the mathematical relationship between variables is calculated without assuming that one (dependent) variable is driven by other variables, as is the case in regression. Functional relation models produce as special cases phylogenetic versions of RMA regression and other types of general structural models that include measurement error.

Each of these four problems can be analyzed using different statistical estimation approaches. Throughout this article, we consider three approaches: estimated generalized least squares (EGLS; Judge et al., 1985), maximum likelihood (ML), and restricted maximum likelihood (REML); these are described in more detail in Appendix 1. These three estimation approaches have different advantages and disadvantages. Rather than perform exhaustive comparisons among estimation methods, instead we illustrate the statistical characteristics of each method by applying them to real data and performing selective simulation studies. Our philosophy is that, when in doubt, it is best to use multiple estimation methods, and if they give different results, perform post hoc diagnostics to select the best (e.g., least biased and most precise). All data analyses and simulations were performed using programs written in Matlab that are available from TG upon request.

ANALYSES

Univariate Analyses and Phylogenetic Signal

The problem of finding the best estimator of the expectation of a random variable when there is phylogenetic correlation and measurement error is given by the statistical model

$$\begin{aligned} \mathbf{X}^* &= a + \varepsilon \\ \mathbf{X} &= \mathbf{X}^* + \eta \end{aligned} \tag{1}$$

where $\mathbf{X^*}$ is a $N\times$ 1 dimensional vector containing the true values of a trait in a sample of N species (tips), a is a scalar giving the expected value of the trait, ε is a $N\times$ 1 vector of zero-mean error terms depicting the evolutionary variance of the trait among species, $\mathbf{X}$ is a $N\times$ 1 vector containing the observed values of the trait, and $\boldsymbol{\eta}$ is the $N\times$ 1 vector of errors associated with measurement. Note that for notational convenience we have written $\mathbf{X^*} = a+ \varepsilon$ as the sum of a scalar and a vector to represent $\mathbf{X^*} = a\mathbf{1} + \varepsilon$ where $\mathbf{1}$ is the $N\times$ 1 vector of ones.

Because closely related species will likely have similar values of trait $\mathbf{x}$, values of ε will be correlated among species. Thus, we assume the covariance matrix for ε is given by $\mathrm{E}\{\varepsilon\varepsilon'\} = \sigma^2\mathbf{C}$, where σ^2 scales the overall phylogenetically inherited variance (sometimes referred to as the rate of evolution; Garland et al., 1999; Garland and Ives, 2000), and $\mathbf{C}$ gives the correlation structure created by phylogenetic relatedness. The most common assumption in phylogenetic analyses is that evolution proceeds like a "Brownian motion" process; through time, the value of a trait changes in small increments in random directions, like a random walk in continuous time (Felsenstein, 1985). Under this assumption, ε has a multivariate normal distribution in which the element c_{ij} of $\mathbf{C}$ is proportional to the length of the shared branches, from root to the last common ancestor, between species i and j (Felsenstein, 1985; Hansen and Martins, 1996; Martins and Hansen, 1997; Garland and Ives, 2000). Other models of evolutionary change are possible, such as including a nonphylogenetic component of evolutionary change (Lynch, 1991; Freckleton et al., 2002; Housworth et al., 2004) or assuming evolution follows an Ornstein-Uhlenbeck process (Hansen and Martins, 1996; Blomberg et al., 2003); each of these will lead to a different translation of branch lengths into the covariance matrix $\mathbf{C}$, but the model given by Equation 1 can be applied regardless of how $\mathbf{C}$ is selected.

The measurement error term η similarly has a covariance matrix $\sigma_m^2\mathbf{M}$. If measurement errors are uncorrelated among species, $\mathbf{M}$ is a diagonal matrix, and the variance due to measurement error of trait x for species i is $\sigma_m^2 m_{ii}$, where m_{ii} is the ith diagonal element of $\mathbf{M}$. It is possible that measurement errors are correlated among species, as might be the case if trait values for a given clade were all measured by a single researcher using the same technique that differed from the techniques used for other clades. In this case, correlation among measurement errors can be incorporated into off-diagonal elements of $\mathbf{M}$. Although we do not consider correlated measurement errors in detail, nonzero off-diagonal elements of $\mathbf{M}$ can be used in all of the methods we derive. Finally, although we will typically assume that ε and η have multivariate normal distributions, for some of the statistical procedures described below, ε and η need not be restricted to being normally distributed.

Consider first the case of no measurement error. Equation 1 can be reformulated as a phylogenetic regression problem in which the error terms are correlated, and hence can be analyzed using either independent contrasts or, as we will do here, generalized least squares (Hansen and Martins, 1996; Garland and Ives, 2000; Rohlf, 2001). Because $\mathbf{C}$ is a covariance matrix (and hence real, symmetric, and nonsingular), there exists another matrix $\mathbf{D}$ such that $\mathbf{DCD'} = \mathbf{I}$, where the apostrophe denotes transpose and $\mathbf{I}$ is the $N \times N$ identity matrix. Matrix $\mathbf{D}$ can be used to transform values of trait x by letting $\mathbf{Z} = \mathbf{DX}$, $\mathbf{U} = \mathbf{D1}$ (the $N\times$ 1 vector of 1's), and $\alpha = \mathbf{D}\varepsilon$. From Equation 1 (with $\eta = 0$), this gives

$$\mathbf{Z} = \mathbf{U}a + \alpha \tag{2}$$

The covariance matrix of α is $\mathrm{E}\{\alpha\alpha'\} = \mathrm{E}\{\mathbf{D}\varepsilon\ (\mathbf{D}\varepsilon)'\} = \mathrm{E}\{\mathbf{D}\varepsilon\varepsilon'\mathbf{D'}\} = \mathbf{D}\mathrm{E}\{\varepsilon\varepsilon'\}\mathbf{D'} = \mathbf{D}(\sigma^2\mathbf{C})\ \mathbf{D'} = \sigma^2\mathbf{I}$. Thus, no covariance terms appear in the covariance matrix of α, so the error terms α_ι are uncorrelated. Equation 2 can, therefore, be analyzed as a standard least-squares regression problem with independent errors. Specifically, the generalized least-squares (GLS) estimator of a is

$$\hat{a} = \frac{\mathbf{U'Z}}{\mathbf{U'U}} = \frac{(\mathbf{D1})'(\mathbf{DX})}{(\mathbf{D1})'(\mathbf{D1})} = \frac{\mathbf{1'D'DY}}{\mathbf{1'D'D1}} = \frac{\mathbf{1'C^{-1}X}}{\mathbf{1'C^{-1}1}} \tag{3}$$

The corresponding estimate of σ^2 is the mean squared error,

$$\hat{\sigma}^2 = \frac{1}{N-1}(\mathbf{X} - \hat{a})'\mathbf{C^{-1}}(\mathbf{X} - \hat{a}) \tag{4}$$

What advantages does the phylogenetic mean value of trait x, $\hat{a}$, have over the sample mean, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$? The expectations of both $\hat{a}$ and $\bar{x}$ are a, so both estimates are unbiased. Nonetheless, $\hat{a}$ has lower variance than $\bar{x}$; in fact, $\hat{a}$ is the minimum-variance estimator of a (Judge et al., 1985).

When there is measurement error, the expression for the observed trait values x from Equation 1 can be written

$$\mathbf{X} = a + \varepsilon + \eta \tag{5}$$

where the observed (total) error term, $\varepsilon + \eta$, has covariance matrix $\sigma_\psi^2\Psi = \sigma^2\mathbf{C} + \sigma_m^2\mathbf{M}$. Consider first the case in which the measurement error is known for each species, so the covariance matrix $\sigma_m^2\mathbf{M}$ is known. (Conventional notation separates the covariance matrix into two components, σ_m^2 and $\mathbf{M}$, and assuming the measurement

error is known implies both components are known.) Because the covariance matrix $\mathbf{C}$ is determined by the phylogeny, the only parameters that must be estimated are a and σ^2. However, unlike the case without measurement error, a simple expression like Equation 3 cannot be derived for the estimate of a, because the matrix $\sigma^2_\psi\Psi = \sigma^2\mathbf{C} + \sigma^2_m\mathbf{M}$ now contains a parameter, σ^2, that does not occur as a simple multiplicative term scaling the overall magnitude of the covariance matrix Ψ.

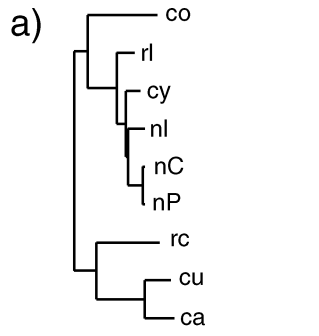
The estimation problem presented by Equation 5 is referred to in the statistical literature as a "measurement error known" problem (Fuller, 1987), because we assume that σ^2_m has been estimated independently (as reported by the standard errors of mean values for species). For nonphylogenetic analyses, corrective steps for known measurement error are fairly straightforward (Fuller, 1987). Unfortunately, these corrective steps cannot be applied when there is phylogenetic correlation (as incorrectly done by Irschick et al., 1996), and the methods we provide below are needed. However, other measurement error problems can be solved rather simply when there is phylogenetic correlation ($\mathbf{C} \neq \mathbf{I}$). Specifically, if instead of knowing the measurement error variance σ^2_m we know the ratio of measurement error variance to true variance σ^2_m/σ^2, it is possible to calculate the phylogenetic mean by replacing $\mathbf{C}$ in Equation 3 with $\Psi = \mathbf{C} + (\sigma^2_m/\sigma^2)\mathbf{M}$ and treat the problem in the usual GLS or independent contrasts fashion. Because this simple case has been addressed elsewhere (Pagel and Harvey, 1988a, 1988b; Harvey and Pagel, 1991), we do not consider it further.

Estimation.—In Equation 5, two parameters are unknown: the mean value a of trait x for all species (or, equivalently, the hypothetical ancestral value at base of tree) and the phylogenetic variance σ^2 (or, equivalently, the rate of evolution). These parameters can be estimated using an iterated version of estimated generalized least-squares (EGLS), maximum likelihood (ML), and restricted maximum likelihood (REML). To obtain ML and REML estimates, it is necessary to specify the form of the distribution of error terms ε and η; a natural assumption, and the one we use here, is that ε and η are normally distributed. Because the covariance matrix Ψ contains the parameter σ^2 that must be estimated, for all three methods the confidence intervals calculated for $\hat{a}$ are approximations. Note that the difficulties in estimation when there is measurement error disappear when there is no measurement error, in which case GLS and ML estimates are the same, and provided ε is normally distributed, the estimates of $\hat{a}$ are t-distributed.

Appendix 1 gives a full account of these methods as applied in this article. Also, univariate EGLS estimation can be implemented using independent contrasts, as done in the MS DOS program PD_SE.EXE (available from TG) and used by Bonine et al. (2005).

Example.—As an example, we analyzed data from Martins and Lamont (1998) on display duration for nine species of lizards. We chose this example because it is a real comparative data set, is small enough to depict our results graphically, and has large enough standard errors for some species that the effects of incorporating measurement error are clearly apparent. For each species, Martins and Lamont (1998) provide the standard error of the measure of the trait, which we use to compute the matrix $\sigma^2_m\mathbf{M}$ under the assumption that measurements are independent among species. For comparison, we computed parameter estimates assuming (i) no phylogenetic correlation among species ($\mathbf{C} = \mathbf{I}$; equivalent to assuming a "star phylogeny") and no measurement error ($\mathbf{M} = \mathbf{0}$), for which the estimate of a is simply the sample mean; (ii) no phylogenetic correlation but measurement error, with the measurement error variance differing among points (species); (iii) phylogenetic correlation (using as the "true" tree, Fig. 1a) and no measurement error, which gives the standard phylogenetic case analyzed by independent contrasts or GLS; and (iv) phylogenetic correlation and measurement error. For each set of assumptions, we computed 95% confidence intervals of the estimates using three approaches. First, for EGLS we used the standard GLS formulae ignoring that we estimated a parameter in the covariance matrix Ψ and the uncertainty associated with this estimate (Neter et al., 1989). Second, for ML we derived approximate confidence intervals from the log-likelihood function (Judge et al., 1985); this is a standard procedure used in ML estimation. Third, for all three estimation methods we used parametric bootstrapping under the assumption that both measurement and true errors are normally distributed. Parametric bootstrapping (Efron and Tibshirani, 1993) is a simulation procedure in which parameters are first estimated (by whatever method is being used), the statistical model with its estimated parameters is used to simulate data sets, and the parameters are estimated from the simulated data. After repeating this many (e.g., 2000) times, the resulting set of estimates approximates the distribution of the estimator (see Appendix 1 for details). The term "parametric bootstrapping" is potentially confusing, because unlike standard (nonparametric) bootstrapping, the residuals obtained from the true data are not resampled to create new data sets but are instead simulated. Parametric bootstrapping is necessary in our case, because we do not know the actual measurements for each sample used to give species values; therefore, the measurement error must be simulated from a random number generator. Although it might be less confusing to refer to parametric bootstrapping more simply as "simulation" to obtain confidence intervals, this then introduces confusion when we perform simulations to explore the statistical properties of the estimation methods. A particular advantage of parametric bootstrapping is that not only does it give confidence intervals, it also identifies bias; if, for example, the mean of the bootstrapped estimates is lower than the true estimate, then this identifies that the estimator is downward biased.

All three estimation methods incorporating measurement error gave similar estimates of a and σ^2 when phylogenetic correlation was not included (i.e., $\mathbf{C} = \mathbf{I}$, case ii). However, when assuming Brownian motion evolution along the true phylogeny (i.e., $\mathbf{C} \neq \mathbf{I}$, case iv), the



Species	Code	B. Total (s) value	B. Total (s) SE	C. Headbob (s) value	C. Headbob (s) SE
Cyclura nubila nubila (Cuba)	nC	3.22	0.406	0.28	0.058
C. n. nubila (P.R.)	nP	4.23	0.711	0.98	0.415
C. n. lewisi	nl	3.43	0.896	0.22	0.022
C. cychlura	cy	1.98	0.399	0.30	0.023
C. rileyi	rl	3.14	0.556	0.88	0.007
C. collei	co	0.98	0.953	0.24	0.058
C. ricordi	rc	3.37	0.243	0.27	0.006
C. carinata	ca	2.20	1.735	0.34	0.086
C. cornuta	cu	2.83	0.447	0.38	0.027

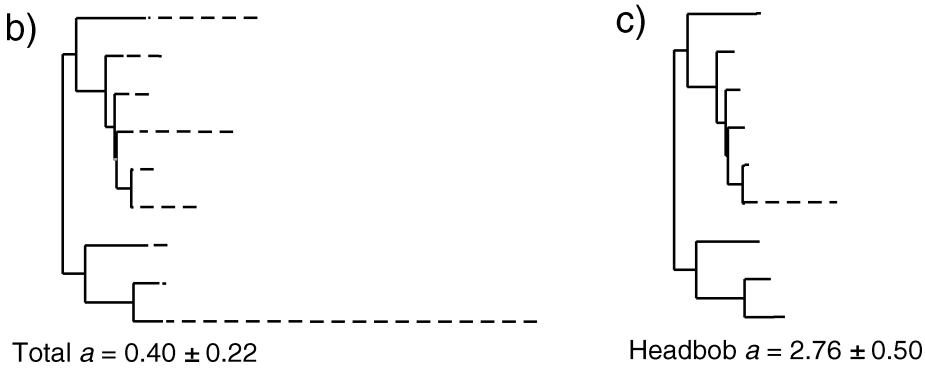


FIGURE 1. For the univariate case using data from Martins and Lamont (1998), the effects of measurement error can be visualized by constructing a tree that corresponds to the covariance structure of the data combining both phylogenetic covariance and measurement error variance (see text). The measurement error variance lengthens the terminal branch segments of the tree, with the length of the tip extension giving the measurement error variance relative to the variance of the evolutionary process. (a) The phylogenetic tree from which the covariance matrix $\sigma^2\mathbf{C}$ is calculated. (b) The phylogenetic tree with the variance associated with measurement error for total display duration graphed onto the tips of the tree, thereby giving a graphical representation of the covariance matrix $\sigma^2\mathbf{C} + \sigma^2_m\mathbf{M}$. For comparison, (c) is like (b) but with measure error for another trait, headbob duration. By increasing the expected within-species variances without changing among-species covariances, measurement error decreases the among-species correlations in the observed data. The table gives trait values and standard errors of the measurements for both traits and estimates of the phylogenetic mean a are given for each tree.

ML estimates of both parameters a and σ^2 differed sizably from the estimates obtained from EGLS and REML (Table 1). The ML estimate of σ^2 appears to be strongly downward biased; the ML estimate of σ^2 is 0.049 and the mean of the bootstrapped estimator is 0.032. Bias of ML estimates of variances is a common observation found in many types of statistical problems, and the relative lack of bias of REML estimates is a frequent justification for preferring REML over ML (Patterson and Thompson, 1971; Cooper and Thompson, 1977; Smyth and Verbyla, 1996). Unfortunately, there is no good way to predict a priori the magnitude of bias; in this particular example, strong bias in the ML estimator only occurred for cases when phylogenetic correlation was incorporated.

Comparing cases ii and iv, accounting for measurement error results in markedly lower estimates of σ^2. This occurs because part of the variability of the data is attributed to measurement error, leaving less true variability among species. The effects of measurement error can be visualized by constructing a tree that gives the covariance structure of the data combining both phylogenetic covariance and measurement error variance; this is done using the EGLS estimates σ^2 in Figure 1 for both total display duration and, for comparison, headbob duration (for another example, see Bonine et al., 2005). The effect of measurement error is to lengthen the terminal branch segments of the tree beyond the strict phylogenetic tree, with the length of the tip extension giving the measurement error variance. By increasing within-species variances without changing among-species covariances, measurement error decreases the among-species correlations in the observed data.

Accounting for measurement error will increase estimates of the strength of phylogenetic signal in data sets. Blomberg et al. (2003) derived a measure, K^*, of the strength of phylogenetic signal. The measure K^* depends on the ratio of the rate of evolution (measured by σ^2) required to explain the variability in a trait among species under the assumption of no phylogenetic correlation ($\mathbf{C} = \mathbf{I}$) to the rate of evolution required under the assumption that $\mathbf{C}$ is given by the working phylogeny. This ratio computed for the data is then compared to the

TABLE 1. Parameter estimates of the phylogenetic mean a and variance σ^2 and the measure of phylogenetic signal K^* for data on total display duration of nine species of iguanas (from Martins and Lamont, 1998: fig. 1).

Phylogeny	Method	Phylogenetic mean a	Bootstrap estimate	σ^2	Bootstrap estimate	K^*
I (star)[a]	GLS[c] (no m.e.)	2.82[1] (2.08, 3.56)	2.83 (2.22, 3.46)[4]	0.93 (0.22, 1.81)	0.94 (0.25, 2.05)[4]	
	EGLS[d]	2.95 (2.32, 3.57)[2]	2.95 (2.41, 3.47)	0.29 (0.070, 0.57)[2]	0.30 (0, 1.12)	
	ML[e]	2.95 (2.40, 3.50)[3]	2.96 (2.48, 3.43)	0.22 (0, 0.75)[3]	0.19 (0, 0.80)	
	REML[f]	2.94	2.95 (2.42, 3.46)	0.28	0.29 (0, 1.04)[3]	
C (true)[b]	GLS (no m.e.)	2.52[1] (0.10, 4.94)	2.53 (0.49, 4.63)	1.92 (0.46, 3.74)	1.91 (0.52, 4.17)	0.32 ($P < 0.05$)[5]
	EGLS	2.76 (1.61, 3.91)[2]	2.77 (1.73, 3.74)	0.35 (0.084, 0.68)	0.40 (0, 1.77)	0.53 ($P > 0.4$)
	ML	2.94 (2.16, 3.72)	2.93 (2.44, 3.41)	0.049 (0, 0.44)	0.032 (0, 0.20)	4.5 ($P > 0.5$)
	REML	2.76	2.75 (1.80, 3.72)	0.32	0.32 (0, 1.10)	0.57 ($P > 0.4$)

[a] Star phylogeny assuming no phylogenetic relatedness; covariance matrix is the identity matrix **I**.
[b] True phylogeny with covariance matrix **C**.
[c] Generalized least squares assuming no measurement error.
[d] Estimated generalized least squares incorporating measurement error.
[e] Maximum likelihood incorporating measurement error.
[f] Restricted maximum likelihood incorporating measurement error.
[1] Also implemented in the MS DOS program PD_SE.EXE, as used in Bonine et al. (2005).
[2] Approximate 95% confidence interval obtained from GLS.
[3] Approximate 95% confidence interval obtained from ML.
[4] Approximate 95% confidence interval obtained from parametric bootstrapping.
[5] Probability of rejecting the null hypothesis that K^* equals 1 (Brownian motion evolution along specified phylogeny).

theoretical expectation of the ratio to give K^*. A value of $K^* = 1$ implies that the observed pattern of covariances in the data is consistent with that expected from the working phylogeny (specified by the covariance matrix **C**), whereas values of K^* less than one imply that the strength of phylogenetic correlation is lower than expected from the phylogeny. Thus, values of K^* less than 1 imply weaker phylogenetic signal. When measurement error exists, K^* should be calculated after removing the variance caused by measurement error. Thus, K^* depends on the estimated variance σ^2 of the "true" values **X*** rather than the variance associated with the observed values **X**, which also depends on $\sigma_m^2\mathbf{M}$. (Note that Blomberg et al. [2003] also derive a measure K that is closely correlated to K^*. For technical reasons we will not discuss here, in measurement error problems K^* is a more appropriate measure of phylogenetic signal. See also Rohlf, 2006.)

The estimate of K^* for lizard display duration is statistically significantly less than 1 when assuming no measurement error (Table 1). In contrast, the value of K^* estimated, while accounting for measurement error is not statistically different from 1 (Table 1). Thus, accounting for measurement error reveals the underlying phylogenetic signal. Note that the ML estimate of K^* is greater than 1, although this is due to the same bias that produced the low ML estimate of σ^2.

Simulation.—In the example above, we know neither the true value of a nor the true phylogenetic correlation, making it impossible to study the statistical properties of the parameter estimators. To examine these properties, we simulated data using the phylogeny from Martins and Lamont (1998; see Fig. 1). We assumed that the trait evolves in a Brownian motion fashion with $\sigma^2 = 0.35$ (the EGLS estimate from the data for total display duration). To simulate measurement error, we assumed that the standard deviation of a measurement on a single animal is twice the reported standard error of total duration as reported in Martins and Lamont (1998). This gives high measurement error and hence a strong test of the estimation methods incorporating measurement error. To vary measurement error, we assumed that data from $n = 2^k$ ($k = 0, 1, \ldots, 6$) individuals were obtained for each species; increasing the sample size n decreases measurement error because the standard error of the measurement error is proportional to $1/\sqrt{n}$. For each simulated data set, we calculated the estimates of a, σ^2, and the measure of phylogenetic signal K^*. We used only EGLS estimation; REML estimation gave similar results, and ML estimation showed considerable bias, particularly in the estimates of K^*.

The simulations show that accounting for measurement error has little effect on the estimate of a, although the confidence intervals decrease (Fig. 2). In contrast, the estimate of σ^2 is greatly improved when measurement error is incorporated into the analysis. Nonetheless, when measurement error is large, even accounting for measurement error does not overcome an upwards bias in the estimate of σ^2. Similarly, when measurement error is accounted for, the estimate of K^* is less biased around its true value of 1 and has confidence intervals that are relatively insensitive to the strength of measurement error. In contrast, when measurement error is ignored, estimates of K^* are markedly low when there is large measurement error. We should point out, however, that the measurement error used in the simulations was very high; with a sample size of one, the average standard error of the measurement error was 1.4, which is more than twice the standard deviation of the true among-species error, 0.59. The measurement error reported by Martins and Lamont (1998) corresponds to our simulated sample size of $2^2 = 4$, and the bias in both σ^2 and K^* above this is minimal for the estimation methods incorporating measurement error.

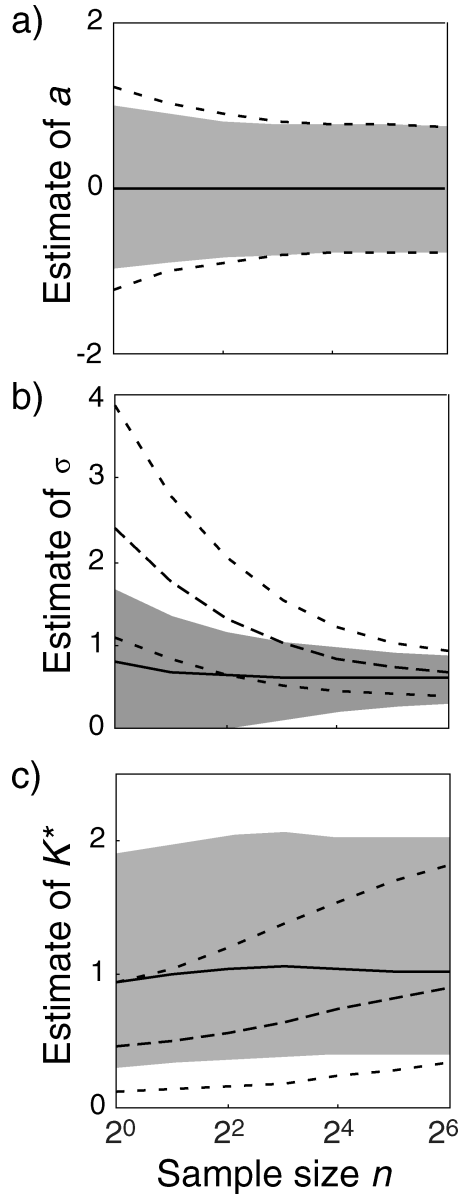


FIGURE 2. Simulation of the univariate case to provide estimates of (a) a, (b) σ, and (c) the measure of phylogenetic signal K^*. Solid lines give the EGLS estimates accounting for measurement error, and the corresponding 95% bounds of the estimate are given by the shaded region. Dashed lines give the estimate and 95% bounds of the estimate obtained without accounting for measurement error (GLS). We assume the 9-tip phylogeny presented by Martins and Lamont (1998). Trait x evolves according to Brownian motion evolution, with $a = 0$ and $\sigma^2 = 0.35$. Measurement error for measurements on single individuals is assumed to have standard error equal to 2 times the standard error provided by Martins and Lamont (1998) for total display duration. For each simulated sample size $n = 2^k$ ($k = 0, 1, \ldots, 6$), 2000 data sets were simulated, and estimates for each parameter were computed.

Correlation between Traits

When measurement error exists, the correlation coefficient between two traits x and y can be calculated from the statistical model

$$\mathbf{X}^* = a_x + \varepsilon_\mathrm{x}; \mathbf{Y}^* = a_\mathrm{y} + \varepsilon_y; \mathbf{X} = \mathbf{X}^* + \eta_\mathrm{x}; \mathbf{Y} = \mathbf{Y}^* + \eta_\mathrm{y} \quad (6)$$

where as before $\mathbf{X}^*$ and $\mathbf{Y}^*$ represent the true values of traits **x** and **y** among species with the true variation given by ε_x and ε_y, and $\mathbf{X}$ and $\mathbf{Y}$ are the values observed with measurement error η_x and η_y.

The joint covariance matrix for ε_x and ε_y is

$$E\{\varepsilon\varepsilon'\} = \begin{pmatrix} \sigma_\mathrm{x}^2\mathbf{C}_\mathrm{x} & r\sigma_\mathrm{x}\sigma_\mathrm{y}\mathbf{C}_\mathrm{xy} \\ r\sigma_\mathrm{x}\sigma_\mathrm{y}\mathbf{C}_\mathrm{xy}, & \sigma_\mathrm{y}^2\mathbf{C}_\mathrm{y} \end{pmatrix} \quad (7)$$

where ε is the $2N \times 1$ vector of error terms created by stacking ε_x on top of ε_y, and $\mathbf{C}_\mathrm{xy} = \mathbf{D}_\mathrm{x}^{-1}(\mathbf{D}_\mathrm{y}')^{-1}$ where $\mathbf{D}_\mathrm{x}^{-1}$ and $\mathbf{D}_\mathrm{y}^{-1}$ are the Cholesky decompositions of $\mathbf{C}_\mathrm{x}$ and $\mathbf{C}_\mathrm{y}$ such that $\mathbf{D}_\mathrm{x}\mathbf{C}_\mathrm{x}\mathbf{D}_\mathrm{x}' = \mathbf{D}_\mathrm{y}\mathbf{C}_\mathrm{y}\mathbf{D}_\mathrm{y}' = \mathbf{I}$. In this formulation (and in our Matlab code), the matrices $\mathbf{C}_\mathrm{x}$ and $\mathbf{C}_\mathrm{y}$ can differ, and therefore trees with different branch lengths (or even different trees) can be used for different traits. For measurement errors

$$\mathrm{E}\{\eta\eta'\} = \begin{pmatrix} \sigma_\mathrm{mx}^2\mathbf{M}_\mathrm{x} & r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathrm{xy} \\ r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathrm{xy}' & \sigma_\mathrm{my}^2\mathbf{M}_\mathrm{y} \end{pmatrix} \quad (8)$$

where η is the $2N \times 1$ vector created by stacking η_x on top of η_y, $\sigma_\mathrm{mx}^2\mathbf{M}_\mathrm{x}$ and $\sigma_\mathrm{my}^2\mathbf{M}_\mathrm{y}$ are matrices containing the measurement error variances, and $r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathbf{xy}$ is the matrix containing covariances in measurement errors between traits for each species. If measurement errors for each trait are independent among species, then $\mathbf{M}_\mathrm{x}$, $\mathbf{M}_\mathrm{y}$, and $\mathbf{M}_\mathbf{xy}$ will be diagonal matrices (i.e., all off-diagonal elements will be zero). If measurement errors for the two traits within species are correlated (e.g., the measurements of traits x and y for a given species tend to err either high or low in unison), then this correlation is given by $r_\mathrm{m}\mathbf{M}_\mathbf{xy}$.

As in the univariate case, the observed values of traits x and y can be expressed in terms of both ε and η as

$$\mathbf{W} = \mathbf{A} + \varepsilon + \eta \quad (9)$$

where $\mathbf{W}$ is the $2N \times 1$ vector created by stacking $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{A}$ is the $2N \times 1$ vector whose first N elements are a_x and second N elements are a_y. The resulting covariance matrix $\mathrm{E}\{(\mathbf{W}–\mathbf{A})(\mathbf{W}–\mathbf{A})'\} = \sigma^2\Psi$ is

$$\sigma^2\Psi = \begin{pmatrix} \sigma_\mathrm{x}^2\mathbf{C}_\mathrm{x} + \sigma_\mathrm{mx}^2\mathbf{M}_\mathrm{x} & r\sigma_\mathrm{x}\sigma_\mathrm{y}\mathbf{C}_\mathrm{xy} + r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathrm{xy} \\ r\sigma_\mathrm{x}\sigma_\mathrm{y}\mathbf{C}_\mathrm{xy}' + r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathbf{xy}' & \sigma_\mathrm{y}^2\mathbf{C}_\mathrm{y} + \sigma_\mathrm{my}^2\mathbf{M}_\mathrm{y} \end{pmatrix}. \quad (10)$$

The case for more than two variables is similar: $\mathbf{W}$ is created by stacking vectors of trait values, and $\sigma^2\Psi$ is constructed with diagonal blocks $\sigma_\mathrm{x}^2\mathbf{C}_\mathrm{x} + \sigma_\mathrm{mx}^2\mathbf{M}_\mathrm{x}$ and off-diagonal blocks $r\sigma_\mathrm{x}\sigma_\mathrm{y}\mathbf{C}_\mathbf{xy} + r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathrm{xy}$ for any pair of traits x and y.

Estimation.—When estimates are available for the standard errors of the trait values for each species, these give the values of $\sigma_\mathrm{mx}^2\mathbf{M}_\mathrm{x}$, $\sigma_\mathrm{my}^2\mathbf{M}_\mathrm{y}$, and $r_\mathrm{m}\sigma_\mathrm{mx}\sigma_\mathrm{my}\mathbf{M}_\mathbf{xy}$.

TABLE 2. GLS, EGLS, and REML estimates of the correlation coefficient (r) between log body size and log sprint speed from Bauwens et al. (1995).

Phylogeny	GLS	GLS bootstrap	EGLS	EGLS bootstrap	REML	REML bootstrap
I (star)	0.466	0.454[1] (−0.11, 0.81)[2]	0.478	0.465[1] (−0.15, 0.85)[2]	0.497	0.486[1] (−0.098, 0.85)[2]
C (true)	0.022	0.017[1] (−0.57, 0.56)[2]	0.025	0.033[1] (−0.60, 0.65)[2]	0.341	0.331[1] (−0.28, 0.81)[2]

[1]Mean of the parametric bootstrap distribution of r.
[2]95% parametric bootstrap confidence intervals from 2000 replication data sets.

Furthermore, the phylogeny and associated assumption about evolutionary change give $\mathbf{C}_x$ and $\mathbf{C}_y$. Therefore, the only parameters that must be estimated are a_x, a_y, σ_x^2, σ_y^2, and r for the case of bivariate correlation. As with the univariate case, multiple methods can be used to estimate the parameters for the model given by Equations 9 and 10. Here, we illustrate EGLS and REML (Appendix 1), although we also provide Matlab programs for ML. EGLS has the advantage that it can be formally applied when the true variation and/or measurement error variation are not normally distributed. As we show below, REML has the advantage of having almost no bias, compared to a slight bias shown by EGLS. Furthermore, when calculating the correlation between multiple pairs of traits, REML (and ML) uses data from all of the traits in estimating each pairwise correlation; this leads to the best estimates when performing multivariate analyses such as PCA. Although there are multiple methods for obtaining confidence intervals for the estimates (see "Univariate Analyses" above), we restricted attention to parametric bootstrapping; for small sample sizes typical of many phylogenetic studies, estimators of the correlation coefficients are often biased, and therefore parametric bootstrapping is often the most robust approach for obtaining confidence intervals.

Example.—We analyzed data from Bauwens et al. (1995) on the body mass, hind-limb length, and sprint speed of 13 species of lizards using phylogeny A from their figure 2. Their table 1 provides means and standard errors for these traits on the arithmetic scale. We chose this example because it includes traits that might be subjected to a number of different statistical analyses (correlation, regression, and functional relation models) and because it is of a size (13 species, $n = 4$ to 20 individuals measured per species), which is not atypical of "small" comparative studies (e.g., see compilations in Ricklefs and Starck, 1996; Freckleton et al., 2002; Blomberg et al., 2003). We log-transformed all traits, which reduced skew in the distribution of trait values (analyses not presented). When log-transforming values that are measured with variation, both the mean and variance of the log-transformed data depend on the variance of the measurement error; thus, we assumed that a given trait value for a given species was log-normally distributed and performed the log-transformation accordingly (Appendix 2). Finally, we assumed that measurement errors are not correlated among traits, so $r_m = 0$ in Equation 10.

In this example, as is likely to be common (e.g., Martins and Lamont, 1998; Bonine et al., 2005), the sample sizes for some species values were small ($n = 4$). When sample sizes are small, the standard errors themselves are imprecise estimates of the measurement error. In practice, this issue is often inconsequential, because the estimates of measurement error, while imprecise, are nonetheless unbiased. A possible approach when there are small sample sizes, or if some species are represented by a single individual (e.g., Langerhans et al., 2006), is to compute the average per sample measurement error and from this calculate the measurement error for each species based on its corresponding sample size (Appendix 3). For the analyses below, we used both the standard errors provided in Bauwens et al. (1995) and the measurement error obtained by averaging across species; both procedures gave quantitatively very close results and so we present only the results using the standard errors for each species.

For a bivariate example, we computed estimates of r between body mass and sprint speed using GLS (i.e., with no measurement error), EGLS, and REML assuming either no phylogenetic relatedness among species (a star phylogeny, $\mathbf{C}_x = \mathbf{C}_y = \mathbf{I}$) or phylogenetic relatedness given by the true phylogeny under Brownian motion evolution (Table 2). For the star phylogeny, EGLS and REML estimates of r were similar and did not differ greatly from the GLS estimate. However, for the true phylogeny, the EGLS estimate (0.025) was similar to the GLS estimate (0.022), and both were much lower than the REML estimate (0.341). The mean of the REML bootstrapped estimates of r (0.327) was lower than the REML estimate, suggesting that if anything, the REML estimate is biased downwards. This suggests that the EGLS estimate (0.025) is even more severely biased than the REML estimate. Despite the large difference between the EGLS and REML estimates, the confidence intervals for both are large, and in neither case is the estimate of r statistically different from zero.

To investigate correlations between multiple pairs of traits, we estimated r for the three pairs of traits: body mass, sprint speed, and hind-limb length using GLS, EGLS, and REML. To implement REML, we estimated correlations in both a pairwise fashion (pairwise REML) and simultaneously for all three traits (joint REML). Joint REML is the correct REML procedure, because REML estimation is based on the likelihood of the entire data set. (Our Matlab program automatically implements joint REML.) Thus, information about the correlation between traits x and y, and between traits y and z, is used in the estimation of the correlation between traits x and z. Stated another way, the estimates of the pairwise correlations

TABLE 3. Estimates of correlation coefficients and loadings on the first principal component (PC 1) for traits log body mass (x), log sprint speed (y), and log hind-limb length (z) provided by Bauwens et al. (1995) for 13 lizard species.

Method	r_{xy}	r_{xz}	r_{yz}	% Variance PC 1	Loading x	Loading y	Loading z
GLS (no m.e.)	0.022	0.845	0.491	0.66	0.36	0.22	0.42
EGLS	0.025	0.867	0.550	—[1]			
Pairwise REML	0.341	0.899	0.799	—[1]			
Joint REML	0.257	0.887	0.635	0.74	0.34	0.27	0.39

[1]The covariance matrix obtained from pairwise analyses was not positive definite.

among traits are not independent. This differs from the case without measurement error, where estimates of pairwise correlations are independent. Researchers might be tempted to calculate correlation coefficients separately in a pairwise fashion, particularly when large numbers of pairwise correlations are desired. We calculated pairwise REML estimates, even though this is not a correct procedure, to illustrate the problems that this can cause.

The estimated correlations for all three pairs of traits (body mass, sprint speed, hind-limb length) using pairwise REML tended to be larger than the GLS and EGLS estimates (Table 3). The joint REML estimates are slightly less high. The three-species correlation matrices obtained from both EGLS and pairwise REML are not valid, because they are not positive definite. The requirement that correlation matrices be positive definite is equivalent to the requirement that correlation coefficients be between -1 and $+1$; just as it makes no sense for correlation coefficients to be greater than $+1$, it makes no sense for a correlation matrix not to be positive definite. The failure of the correlation matrices obtained from EGLS and pairwise REML to be positive definite is caused by the low estimated correlation between body size and sprint speed, r_{xy}. Because traits x (log body size) and z (log hind-limb length) are highly correlated, and traits y (log sprint speed) and z are highly correlated, traits x and y must also be highly correlated for the correlation matrix to be positive definite, but this condition is not met for EGLS and pairwise REML. In contrast, the correlation matrices obtained from GLS (for which pairwise estimates of correlation coefficients are independent) and joint REML (which estimates all correlation coefficients simultaneously) are positive definite. This particular data set is prone to the problem of estimated correlation matrices not being positive definite, because the sample size is small and the correlations between traits are high. Nonetheless, this problem is likely to arise frequently in similar data sets.

Using the estimated correlation matrices, we performed a PCA (Sokal and Rohlf, 1981), calculating the first PC axis and corresponding loadings (Table 3). The high correlations obtained from the joint REML caused 74% of the correlation to be captured by the first principal components axis (PC1). In contrast, the PC1 using the GLS estimates was 66%. Thus, incorporating measurement error reveals a stronger correlation structure in the data. Because the correlation matrices obtained from EGLS and pairwise REML are not positive definite, the resulting PCAs are invalid.

Simulation.—To investigate the properties of the estimators of r, we performed a simulation study based on the example. Specifically, we simulated data for 13 species having the phylogeny of the 13 species studied by Bauwens et al. (1995). We assumed that two traits x and y followed Brownian motion evolution up the phylogenetic tree with rates $\sigma_x = 0.86$, $\sigma_y = 0.28$, and $r = 0.83$. We assumed that the standard deviation of the measurement on a single animal is 4 times the standard error reported by Bauwens et al. (1995) for body mass and sprint speed, and that data from $n = 2^k$ ($k = 0, 1, \ldots, 6$) individuals were obtained for each species. Thus, when $k = 4$ ($n = 16$), the measurement error variance is equal to that reported in Bauwens et al. (1995), and higher variance occurs for smaller sample sizes. For each of 2000 simulated data sets at each sample size n, we estimated parameters using both EGLS and REML.

After accounting for measurement error, REML estimates of r had only slight downward bias, with the approximate expectation ranging between 0.813 and 0.821 for a true value of $r = 0.83$ (Fig. 3). The EGLS estimates had greater downward bias for small sample sizes. In contrast, the GLS estimates that ignore measurement error had much greater downward bias. In addition to having less bias than the EGLS estimates, the REML estimates were also consistently more precise, with narrower 95% inclusion bounds. Note also that the distribution of the REML estimates is highly skewed; the upper 95% inclusion bound never exceeds 0.88, while the lower inclusion bound drops to almost zero. This is expected, as r is constrained to be less than or equal to one (see also Martins and Garland, 1991).

Regression

A statistical model for regression with phylogenetic relatedness and measurement error is given by

$$\mathbf{X}^* = a_x + \varepsilon_x; \mathbf{Y}^* = b_0 + b_1 \mathbf{X}^* + \varepsilon_y; \mathbf{X} = \mathbf{X}^* + \eta_x; \mathbf{Y} = \mathbf{Y}^* + \eta_y \quad (11)$$

where, as before, $\mathbf{X}^*$ and $\mathbf{Y}^*$ represent the true values of traits x and y among species, and $\mathbf{X}$ and $\mathbf{Y}$ are the values observed with measurement error η_x and η_y. Because values of the independent trait x will likely be phylogenetically correlated, we assume ε_x has covariance matrix $\mathbf{C}_x$, and because residual variance in $\mathbf{Y}^*$ given by ε_y is also likely to be phylogenetically correlated, we assume ε_y has covariance matrix $\mathbf{C}_y$. Following standard regression, we assume that variations in $\mathbf{X}^*$ and ε_y are independent. As before, we assume that measurement error may be correlated for traits x and y within species, leading to

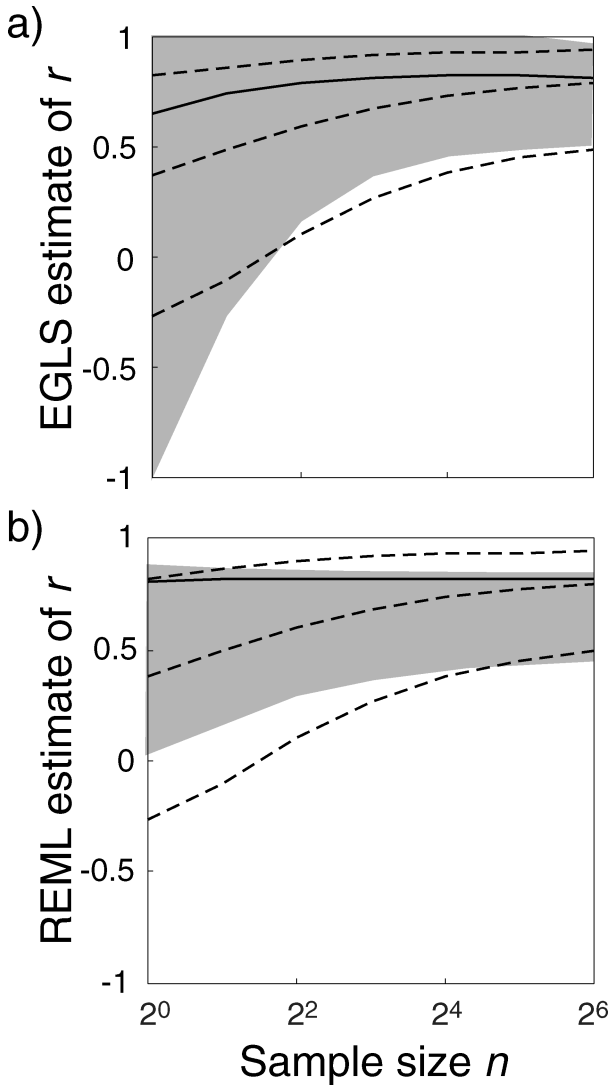


FIGURE 3. (a) EGLS and (b) REML estimates of the correlation coefficient r from simulated data sets based on Bauwens et al. (1995). Solid lines give estimates accounting for measurement error, and the corresponding 95% bounds of the estimate are given by the shaded region. Dashed lines give the estimate and 95% bounds of the estimate obtained without accounting for measurement error (GLS). We assume there are 13 species with phylogeny given by the phylogeny for the 13 lizards analyzed by Bauwens et al. (1995). Both traits evolve according to Brownian motion evolution, with $\sigma_x = 0.86$, $\sigma_y = 0.28$, and $r = 0.83$. Measurement error for measurements on single individuals is assumed to have standard error equal to 4 times the species standard errors provided by Bauwens et al. (1995). For each simulated sample size $n = 2^k$ ($k = 0, 1, \ldots, 6$), 2000 data sets were simulated, and estimates for each parameter were computed.

the measurement error covariance matrix given in Equation 8. If there is no measurement error ($\mathbf{M}_x = \mathbf{M}_y = 0$), then this problem reduces to phylogenetic regression that can be solved with GLS (e.g., as implemented in the Matlab REGRESSION.M program of Blomberg et al., 2003) or independent contrasts (Garland and Ives, 2000).

Just like the correlation model, the regression model can be written in terms of $\mathbf{W}$ given by Equation 9, leading to the covariance matrix

$$\sigma^2\Psi = \begin{pmatrix} \sigma_x^2\mathbf{C}_x + \sigma_{mx}^2\mathbf{M}_x & b_1\sigma_x^2\mathbf{C}_x + r_m\sigma_{mx}\sigma_{my}\mathbf{M}_{xy} \\ b_1\sigma_x^2\mathbf{C}_x + r_m\sigma_{mx}\sigma_{my}\mathbf{M}'_{\mathbf{xy}} & b_1^2\sigma_x^2\mathbf{C}_x + \sigma_y^2\mathbf{C}_y + \sigma_{my}^2\mathbf{M}_y \end{pmatrix}. \quad (12)$$

Thus, the regression model can be analyzed like the correlation model, with the covariance matrix for $\sigma^2\Psi$ given in Equation 12 replacing that in Equation 10. More than one independent variable (multiple regression) can be incorporated in a similar manner and different branch lengths for different traits can be used, as done in our Matlab programs.

Estimation.—Like univariate analyses and correlation, EGLS, ML, and REML can be used for estimation (Appendix 1). Here we consider all three in analyzing an example, and study REML in more detail with a simulation.

Example.—As in the example of correlation, we analyzed the data from Bauwens et al. (1995). Table 4 gives GLS, EGLS, ML, and REML estimates of the slope b_1 for the regression of log hind-limb length on log body size. The estimates under the assumption of no phylogeny relatedness ($\mathbf{C}_x = \mathbf{C}_y = \mathbf{I}$) are similar for all three methods incorporating measurement error. Furthermore, the parametric bootstrap confidence intervals are similar to the approximate confidence intervals obtained for EGLS and ML. The only differences among the statistical analyses is the relatively low estimate of b_1 obtained for EGLS when the true phylogeny of the species is used.

In this example incorporating phylogenetic relatedness caused a large decrease in the estimates of b_1, whereas measurement error had relatively little effect. Interestingly, the 95% confidence intervals obtained with phylogenetic information excluded the slope of 1/3 that would be expected for geometric similarity when EGLS was used, but not with ML and REML. In this case, selecting an estimation approach does make a difference in interpreting the results, at least if a confidence level of 95% is strictly adhered to. Unfortunately, in this case there is no ground to statistically prefer one estimation method over another, because all methods showed little bias. In rare situations like this, all we can recommend is to report the results cautiously.

Simulation.—We designed simulations similar to previous simulations (Figs. 2 and 3) to investigate the effect of measurement error by varying the sample sizes of individuals measured per species. We also wanted to compare data sets with different numbers of species; increasing the number of species will not decrease the measurement error, but it should decrease the variance of the parameter estimates by providing more information about the relationship between the two traits. Thus, we wanted to compare the variance in the estimate of b_1 when the number of individuals sampled from the same species is increased versus when the number of species sampled is increased. We only consider REML estimation, because EGLS and ML give similar results.

TABLE 4. Estimates of regression slope b_1 for log hind-limb length regressed on log body mass for 13 species of lizards from Bauwens et al. (1995).

Phylogeny	GLS (no m.e.)	EGLS	EGLS bootstrap	ML	ML bootstrap	REML	REML bootstrap
I (star)	0.305 (0.19, 0.42)[1]	0.307 (0.20, 0.41)[2]	0.307 (0.21, 0.41)[3]	0.310 (0.21, 0.41)[4]	0.312 (0.21, 0.42)[3]	0.309	0.310 (0.21, 0.41)[3]
C (true)	0.224 (0.13, 0.32)[1]	0.232 (0.14, 0.33)[2]	0.231 (0.14, 0.32)[3]	0.263 (0.17, 0.36)[4]	0.265 (0.19, 0.35)[3]	0.260	0.261 (0.18, 0.35)[3]

[1]95% confidence interval from GLS.
[2]95% confidence interval using the approximate standard error obtained from the GLS formulae.
[3]95% confidence interval from parametric bootstrapping.
[4]95% confidence interval from a *t*-distribution ML.

For the simulation, we assumed that there were either 13 or 49 species. For the 13-species case, we used the phylogeny for 13 lizards given by Bauwens et al. (1995). For the 49-species case, we used the phylogeny for 49 Carnivora and ungulates from Garland et al. (1993). We set the true value of $b_1 = 1/3$ as would be expected if the dependent variable was the log of a linear dimension (e.g., leg length), the independent variable was the log of body mass, and species of different body size were geometrically similar. We set the other parameters equal to the REML estimates from the Bauwens et al. (1995) data using the full measurement error model with the true phylogeny (Table 4). For the 13-species phylogeny, we assumed that the standard deviation of measurement error for a single individual was 9 times greater than the standard error reported by Bauwens et al. (1995); we used such a large measurement error because the true measurement error did not have a strong effect on the analyses of the true data. For the 49-species phylogeny, we assigned standard errors to the species by randomly selecting from the 13 standard error values used in the 13-species simulation. We assumed that measurement errors between traits were independent.

In both 13- and 49-species cases, the REML estimate of b_1 incorporating measurement error was at most slightly biased, whereas the GLS estimate of b_1 without measurement error was badly biased when the number of individuals sampled per species was small (Fig. 4). The bias of the GLS estimates was nearly the same for both the 13- and 49-species data sets, illustrating that bias due to measurement error does not depend on the number of species sampled, only on the precision of the measurements for each species (and hence the number of individuals sampled per species). Nonetheless, the confidence intervals of the estimates of b_1 become narrower with increasing numbers of species. This accentuates the statistical problems that can arise from bias. In the 49-species case, the true value of b_1, 1/3, is excluded from the 95% inclusion interval of the estimates when sample sizes n are small, so the hypothesis that $b_1 = 1/3$ would be rejected even though we know that the true value of b_1 is 1/3! Harmon and Losos (2005) discuss more generally the effect of measurement error on type I and type II errors in phylogenetic analyses.

Although increasing sample sizes n per species will reduce measurement error and therefore give more precise estimates of b_1, precision of the estimates of b_1 is also limited by the number of species in the data set. For these examples, increasing sample sizes per species

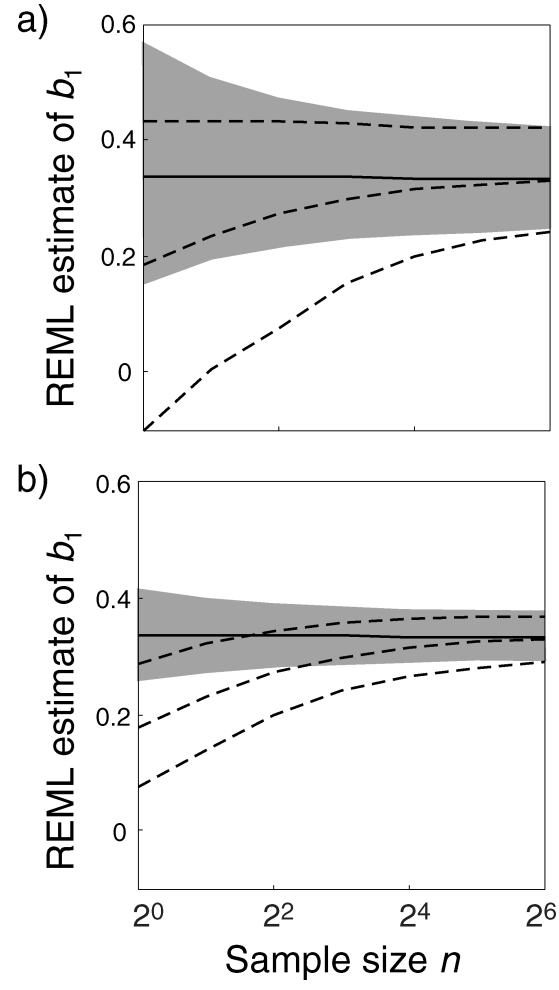


FIGURE 4. REML estimates of the slope of a regression of simulated data sets. In (a) simulated data consisted of 13 species having phylogeny given by Bauwens et al. (1995), and in (b) there are 49 species with the phylogeny given by Garland et al. (1993). Solid lines give estimates accounting for measurement error, and the corresponding 95% bounds of the estimate are given by the shaded region. Dashed lines give the estimate and 95% bounds of the estimate obtained without accounting for measurement error (GLS). Both traits evolve according to Brownian motion evolution, with $b_1 = 1/3$, $\sigma_x = 0.8$, and $\sigma_y = 0.1$. In (a) the measurement error for measurements on a single individual is assumed to have standard error equal to 9 times the species standard errors provided by Bauwens et al. (1995), whereas in (b) measurement errors are selected at random from these 13 values. For each sample size $n = 2^k$ ($k = 0, 1, \ldots, 6$), 2000 data sets were simulated. Comparable figures using EGLS and ML estimation were not qualitatively different.

provides only moderate improvement in the precision in the estimates of b_1 once the sample size exceeds 4. However, the standard deviation of the estimates of b_1 is decreased by roughly 50% when there are 49 species relative to 13 species. This is because increasing the number of species increases information about the relationship between traits x and y; this information is contained in the variance among species.

Functional Relation Models

Regression is the appropriate statistical model to apply when one variable is causally determined by another, or if one variable is to be predicted by the value of another variable. However, in many biological questions the goal is to understand how two traits are functionally related without assigning a direction of causality. For example, one might be interested in the relationship between tail length and leg length among a group of species. The problem of functional relations has been addressed with "general structured relations" models (Rayner, 1985), of which reduced major axis (RMA) regression is the most frequently used special case. Here we develop functional relation models that incorporate phylogenetic correlation and measurement error in a flexible way. We then present some of the special cases that are derived for the non-phylogenetic case by Rayner (1985). Our goal in presenting these cases is to show that the models derived by Rayner (1985) can be easily expanded to include phylogenies and measurement error.

Our functional relation model is

$$\begin{aligned} &\mathbf{X}^* = a_x + \gamma_x;\ \mathbf{Y}^* = b_0 + b_1\mathbf{X}^* \\ &\mathbf{X} = \mathbf{X}^* + \varepsilon_x + \eta_x;\ \mathbf{Y} = \mathbf{Y}^* + \varepsilon_y + \eta_y. \end{aligned} \tag{13}$$

Variation in the true value of x among species, $\mathbf{X}^*$, is given by γ_x having covariance matrix $\sigma^2_{\gamma x}\mathbf{C}_{\gamma x}$. Variation in the observed values $\mathbf{X}$ and $\mathbf{Y}$ is divided into two components. The measurement errors are given by η_x and η_x with covariance matrices $\sigma^2_{mx}\mathbf{M}_x$, $\sigma^2_{my}\mathbf{M}_y$, and $r_m\sigma_{mx}\sigma_{my}\mathbf{M}_{xy}$, which we assume are known (see Equation 8). The observed values of $\mathbf{X}$ and $\mathbf{Y}$ also depend on unknown sources of variation given by ε_x and ε_y with covariance matrices $\sigma^2_x\mathbf{C}_x$, and $\sigma^2_y\mathbf{C}_y$, and cross-covariance matrix $r\sigma_x\sigma_y\mathbf{C}_{xy}$. The errors ε_x and ε_y represent biological variation in $\mathbf{X}$ and $\mathbf{Y}$ that may have a phylogenetic component, and may additionally contain unknown measurement error not captured by η_x and η_y.

The model given by Equation 13 contains seven parameters: a_x, b_0, b_1, $\sigma^2_{\gamma x}$, σ^2_x, σ^2_y, and r. However, the information available from a data set allows only five parameters to be estimated. This can be explained heuristically by noting that a data set provides five pieces of information about the distribution of x and y: the means of x and y, their variances, and the covariance between them. This presents the problem of statistical identifiability, which appears frequently in measurement error models (Fuller, 1987). If no additional information is available, then the only solution to the identifiability problem is to make assumptions about the values of parameters or the mathematical relationships among them. For example, if it is assumed that there is no unknown variation in x ($\sigma^2_x = r = 0$), then Equation 13 reduces to the regression model of y on x given by Equation 11. Conversely, if there is no unknown variation in y ($\sigma^2_y = r = 0$), then Equation 13 reduces to a regression of x on y. Finally, if it is assumed that there is no variation in $\mathbf{X}^*$ ($\sigma^2_{\gamma x} = 0$, and hence $b_1 = 0$), then the model reduces to the correlation model given by Equation 6.

Additional special cases can be derived by recognizing that Equation 13 is a generalization of the general structural relation model of Rayner (1985); specifically, the general structural equation model is obtained when there is no phylogenetic correlation, $\mathbf{C}_{\gamma x} = \mathbf{C}_x = \mathbf{C}_y = \mathbf{I}$, and the measurement error is zero, $\mathbf{M}_x = \mathbf{M}_y = \mathbf{M}_{xy} = \mathbf{0}$; a proof is given in Appendix 4. RMA regression is then derived as a special case assuming that there is no correlation between ε_x and ε_y ($r = 0$), and the ratio of standard deviations of ε_x and ε_y satisfies $\sigma_y/\sigma_x = b_1$ (Rayner, 1985). A specific property of RMA regression is that there is no causal directionality between traits x and y, because the RMA regression of trait y on x is equivalent to the RMA regression of x on y. Specifically, the first two lines of Equation 13 could be written equivalently as $\mathbf{Y}^* = a_y + \gamma_y$ and $\mathbf{X}^* = \beta_0 + \beta_1\mathbf{Y}^*$, where $a_y = b_0 + b_1a_x$, $\beta_0 = -b_0/b_1$, and $\beta_1 = 1/b_1$. Furthermore, because $\sigma_y/\sigma_x = b_1$ for RMA regression, $\sigma_x/\sigma_y = 1/b_1 = \beta_1$. Therefore, in RMA regression either x or y can be treated as the "dependent variable."

Phylogenetic relatedness can be incorporated into RMA regression by removing the assumption that $\mathbf{C}_{\gamma x} = \mathbf{C}_x = \mathbf{C}_y = \mathbf{I}$. In this case, the estimate of b_1 (i.e., the RMA slope) in the absence of measurement error is

$$\left[\frac{(\mathbf{Y} - \hat{a}_y)'\mathbf{C}_y^{-1}(\mathbf{Y} - \hat{a}_y)}{(\mathbf{X} - \hat{a}_x)'\mathbf{C}_x^{-1}(\mathbf{X} - \hat{a}_x)}\right]^{1/2} \tag{14}$$

where $\hat{a}_x$ and $\hat{a}_y$ are the phylogenetically correct means of x and y given by Equation 3. In the phylogenetic version of the general structural relation model, all statistical tests and confidence intervals can be calculated by modifying standard formulae (Rayner, 1985). In particular, the $(1-\alpha)$ % confidence intervals of the estimate of b_1 are

$$\hat{b}_1 sign(\rho)\left(\frac{1 \pm q}{1 \mp q}\right)^{1/2}, \tag{15}$$

where

$$\rho = \frac{(\mathbf{X} - \hat{a}_x)'\mathbf{C}_{xy}^{-1}(\mathbf{Y} - \hat{a}_y)}{\left[(\mathbf{X} - \hat{a}_x)'\mathbf{C}_x^{-1}(\mathbf{X} - \hat{a}_x) * (\mathbf{Y} - \hat{a}_y)'\mathbf{C}_y^{-1}(\mathbf{Y} - \hat{a}_y)\right]^{1/2}}$$

$$q = t_{\alpha,n-2}\left[\frac{1-\rho^2}{(n-2)\rho^2}\right]^{1/2}$$

Another special case can be derived by assuming that the relative magnitudes of variances in ε_x and ε_y are known to be a constant k ($k = \sigma_y/\sigma_x$), and ε_x and ε_y are uncorrelated ($r = 0$). This might be a reasonable model when a researcher has limited information about variation in ε_x and ε_y—enough to assign relative but not absolute magnitudes of variance between traits, as was the case for Pagel and Harvey (1989). We refer to this case as VRF (variance ratio fixed) regression. With the assumptions that $k = \sigma_y/\sigma_x$ and $r = 0$, parameters a_x, b_0, b_1, $\sigma^2_{\gamma x}$, and σ^2_x can be estimated from Equation 13.

Measurement error can be incorporated into special cases of the general structural relation model (in particular RMA and VRF regression) by explicitly considering the covariance matrix for the observed values of traits x and y. Specifically, for the bivariate case parameters can be estimated from the covariance matrix $\sigma^2\Psi$ for the vector $\mathbf{W}$ constructed by placing $\mathbf{X}$ on top of $\mathbf{Y}$:

$$\sigma^2\Psi = \begin{pmatrix} \sigma^2_{\gamma x}\mathbf{C}_{\gamma x} + \sigma^2_x\mathbf{C}_x + \sigma^2_{mx}\mathbf{M}_x & \\ \quad b_1\sigma^2_x\mathbf{C}_{\gamma x} + r\sigma_x\sigma_y\mathbf{C}_{\mathbf{xy}} + r_m\sigma_{mx}\sigma_{my}\mathbf{M}_{\mathbf{xy}} & \\ b_1\sigma^2_x\mathbf{C}_{\gamma x} + r\sigma_x\sigma_y\mathbf{C}'_{\mathbf{xy}} + r_m\sigma_{mx}\sigma_{my}\mathbf{M}'_{\mathbf{xy}} & \\ \quad b_1^2\sigma^2_{\gamma x}\mathbf{C}_x + \sigma^2_y\mathbf{C}_y + \sigma^2_{my}\mathbf{M}_y & \end{pmatrix} \quad (16)$$

Different special cases are given by imposing different restrictions on the parameters; for example, for RMA regression, $r = 0$ and $\sigma_y/\sigma_x = b_1$.

From a statistical perspective, the numerous different general structural relation models that can be derived as special cases from Equation 13 might all be considered equally valid. Thus, correlation, regression, and RMA regression are all similarly well defined statistically. Nonetheless, different statistical models are prone to different mistakes in interpretation. For example, if RMA regression is applied to two traits that are independent, the expectation for the estimated slope will be 1, even though there is no relationship between traits. In this case, the way to guard against misinterpreting the RMA slope is to pay attention to confidence intervals; if the traits are independent, the confidence intervals will be wide. Also, a correlation analysis will reveal lack of correlation between the traits. Here, we do not want to condone or condemn the use of RMA regression and other structural relation models, but we do believe in caution when interpreting their results.

Estimation.—For the general case with measurement error (Equation 16), EGLS, ML, and REML estimation can be used. Here we restrict attention to ML estimation. An advantage of ML over EGLS estimation is that likelihoods can be used to compare different formulations of the model. Mixing and matching different assumptions, for example whether or not the error terms ε_x and ε_y contain phylogenetic correlations, lead to multiple possible models, and ML can be used to sort models and find the best. It is also possible to use REML estimation, although interpreting likelihoods computed from REML is not as straightforward as ML, so we prefer ML estimation. For structural relation models, we found little bias in ML estimates, so this major limitation of ML estimation found for other problems (especially those involving estimates of variances and correlations) did not occur.

We provide Matlab programs for RMA and VRF regression.

Example.—For the example, we analyzed the same log body mass and log hind-limb length data from the 13 lizard species in Bauwens et al. (1995) used in the regression example. We consider three pairs of models. First, we use RMA regression and its phylogenetic counterpart in which estimates of b_1 and confidence intervals are given by Equations 14 and 15 when there is no measurement error ($\mathbf{M}_x = \mathbf{M}_y = \mathbf{M}_{\mathbf{xy}} = \mathbf{0}$). We refer to these models as rma(I) and rma(C), respectively. Second, we consider the nonphylogenetic and phylogenetic pair of VRF regressions in which $k = \sigma_y/\sigma_x$ and $r = 0$, and there is no measurement error. For the value of k, we use the simple average standard errors of log-transformed traits x (0.0495) and y (0.0177) reported by Bauwens et al. (1995). This mimics the case in which a researcher has only rough information about the variation in traits and assumes that the total unexplained variation in traits x and y is proportional to their within-species variation estimated from the standard error. We refer to these models as vrf(I) and vrf(C) models, respectively. Third, we consider the pair of nonphylogenetic and phylogenetic models that make the assumptions $b_1 = \sigma_y/\sigma_x$ and $r = 0$ as in RMA regression, but $\mathbf{M}_x$ and $\mathbf{M}_y$ are derived from the standard errors reported by Bauwens et al. (1995). We refer to these as rmaM(I) and rmaM(C), respectively. We do not consider measurement error in the VRF regression model, because we have already used information about the measurement error to select a value of k.

The nonphylogenetic versions of all three models give very similar estimates of the functional relation slope b_1 (Table 5). The similarity between the RMA and VRF models is due to the fact that for the standard errors reported in Bauwens et al. (1995), $k = 0.36$ which, by coincidence, is very close to the value of $b_1 = 0.347$, which is assumed to equal $k = \sigma_y/\sigma_x$ in RMA regression. The estimates of b_1 for all phylogenetic versions of the models are lower than the nonphylogenetic versions. However, the lower log-likelihoods for the phylogenetic versions indicate that they fit the data more poorly than the nonphylogenetic models. Thus, on statistical grounds the estimates from the nonphylogenetic versions are preferred. To formally arbitrate between phylogenetic and nonphylogenetic models, the best approach is to introduce a parameter that explicitly governs the strength of phylogenetic correlation. For example, Blomberg et al. (2003) derive a transform from an Ornstein-Uhlenbeck process, which introduces a parameter d into the covariance matrix $\mathbf{C}$ that dictates the strength of phylogenetic correlation; incorporating this into the models (as done for the case of no measurement error in REGRESSIONv2.m) would allow tests of phylogenetic strength and selection of the best estimate of b_1 (see also Grafen, 1989; Freckleton et al., 2002). In our experience with real data sets, it is often the case that even a slight distortion of the

TABLE 5. Estimates of functional relation slope b_1 for log body mass and log hind-limb length for 13 species of lizards from Bauwens et al. (1995).

Model	Assumptions	Phylogeny	ML estimate of b_1	Bootstrap ML estimate of b_1	Log-likelihood
RMA[a]	$r = 0$	**I** (star)	0.347 (0.25, 0.45)[1] (0.24, 0.51)[3]	0.349 (0.25, 0.46)[2]	8.64
	$\sigma_y/\sigma_x = b_1$	**C** (true)	0.265 (0.18, 0.35)[1] (0.17, 0.42)[3]	0.267 (0.19, 0.36)[2]	3.72
VRF[b]	$r = 0$	**I** (star)	0.346 (0.23, 0.46)[1]	0.349 (0.24, 0.49)[2]	8.64
	$\sigma_y/\sigma_x = k = 0.36$	**C** (true)	0.251 (0.15, 0.35)[1]	0.253 (0.16, 0.36)[2]	3.72
RMA with measurement error	$r = 0$	**I** (star)	0.349 (0.25, 0.45)[1]	0.352 (0.26, 0.47)[2]	8.69
	$\sigma_y/\sigma_x = b_1$	**C** (true)	0.292 (0.20, 0.38)[1]	0.295 (0.22, 0.39)[2]	5.13

[a] Reduced major axis regression.
[b] Variance ratio fixed regression.
[1] 95% confidence interval from a *t*-distribution using the approximate standard error obtained from the information matrix in ML estimation.
[2] 95% confidence interval from parametric bootstrapping.
[3] 95% confidence interval from Equation 15.

phylogenetic tree to make it somewhat more star-like yields substantially improved fits.

Note that the maximum log-likelihoods for the RMA and VRF models with the same phylogenetic assumptions are the same. This is a result of the identifiability problem of having seven parameters; different ways of constraining the model to give five parameters (the maximum that can be estimated) will all give the same maximum likelihoods. The ML approximate confidence intervals for b_1 are close to the parametric bootstrap confidence intervals, demonstrating that the approximation is accurate. Finally, the ML approximate confidence intervals are better (using the bootstrap confidence intervals as the gold standard) than confidence intervals for RMA regression given by Equation 15.

Simulation.—To explore the statistical properties of the estimators for the different models, we simulated data using the rmaM(C) model—the model with $r = 0$, $b_1 = \sigma_y/\sigma_x$, and phylogeny and measurement errors given by Bauwens et al's. (1995) example of 13 lizards—using parameter values obtained from the fit of the model to the data (Table 5). For comparison, we also simulated the same model after increasing the standard deviations of the measurement errors by a factor of 4 (4× simulations). For 2000 simulated data sets, we fit the same six models as illustrated in the example (Table 5).

All estimators of b_1 were unbiased in the 1× and 4× measurement error simulations (Fig. 5a and c). Heuristically, this can be explained by noting that the estimate of functional relation of y on x, b_1, is the inverse of the functional relation of x on y, $1/b_1$ for the RMA and VRF models. If there were, for example, consistent downward bias in the estimate of the functional relationship, then the estimates of both b_1 and $1/b_1$ would have to be downwards biased, which clearly is not simultaneously possible.

Despite absence of bias, there is considerable variability among models in the precision of the estimates of b_1, as revealed by their 95% inclusion intervals. Not surprisingly, the greatest precision (smallest inclusion interval) was achieved by the model used to simulate the data, rmaM(C). The precision of the rma(C) model, however, was almost identical. The VRF models both had poor precision, particularly when there was large measurement error (Fig. 5c).

Surprisingly, the rmaM(C) model used to generate the data did not always fit the simulated data best (Fig. 5b, d); even in the high measurement error case (Fig. 5d), it was the best-fitting model for only a little over 50% of the simulated data sets, and in the low measurement error case (Fig. 5b) the rma(C) model was selected as the best-fitting model more frequently than the rmaM(C) model.

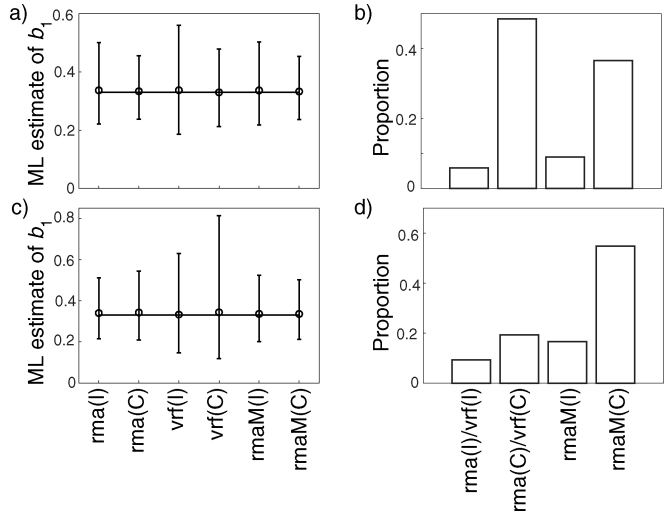


FIGURE 5. ML estimates of b_1 in the functional relations model given by Equation 13 for simulated data when the standard deviations in measurement error are (a) those reported by Bauwens et al. (1995) and (c) four times these values. For each of 2000 simulated data sets, ML estimates of b_1 were obtained for 6 model variants: rma(I) and rma(C), reduced major axis regression ($b_1 = \sigma_x/\sigma_y, r = 0$) with no phylogenetic correlation and phylogenetic correlation given by the lizard phylogeny of Bauwens (1995) under Brownian motion evolution; vrf(I) and vrf(C), the functional relations model having measurement error variance ratio fixed ($k = \sigma_x/\sigma_y, r = 0$) with and without phylogenetic correlation; and rmaM(I) and rmaM(C), reduced major axis regression incorporating measurement error with and without phylogenetic correlation. Error bars give 95% inclusion intervals for estimates. Numerical convergence to the ML estimate did not occur for 1.2% of the data set estimation method combinations; nonconvergent cases were included in the 95% inclusion intervals but not in the mean estimates of b_1. (b and d) The proportion of the 2000 simulated data sets corresponding to (a) and (c), respectively, in which a given model had the highest likelihood. Because the likelihoods of the rma and vrf models are the same, these are combined.

Furthermore, the nonphylogenetic models were selected fairly frequently. In part this is due to the small sample size of 13 species; using simulation studies, Blomberg et al. (2003) showed that the reliable detection of phylogenetic signal (nonstar phylogenies) using univariate data sets requires at least 20 species. This example provides caution about the statistical ability to identify the correct statistical model from small data sets.

Discussion

We have developed phylogenetic models that incorporate measurement error to investigate four statistical problems: (i) univariate estimates of the mean, variance, and phylogenetic signal of a trait; (ii) correlation and principal component analysis; (iii) multiple regression; and (iv) bivariate functional relation models, of which reduced major axis regression is a special case. Measurement error is ubiquitous in comparative studies, as it encompasses a wide variety of sources of variation, including variation caused by instrumentation and techniques, variation among individuals within a given population, and variation among populations of the same species. We have treated measurement error as variability that is given by the standard error of mean trait values for species (or populations); when this information is available, it can be used to improve statistical tests in many cases.

When measurement error is large, it can obscure patterns of variation and covariation in data. In the univariate case, measurement error obscures the pattern of phylogenetic correlation in trait values among species, thereby leading to underestimates of the strength of phylogenetic signal measured for a trait. It can also adversely affect estimates of ancestral values. For example, in a study of muscle fiber type composition of 24 species of lizards, in which four individuals were measured in each species, Bonine et al. (2005) found that although the point estimates for ancestral values were virtually unaffected by incorporating standard errors, the confidence intervals about them were narrowed, which would enhance statistical power for testing hypotheses about particular nodes.

Measurement error similarly obscures covariation between traits, leading to underestimates of correlation coefficients. The consequences of ignoring measurement error are likely to be severe for univariate estimates of phylogenetic signal and bivariate correlations. This is because measurement error increases the variance among observed trait values for species, thus diminishing the observed correlation among trait values. A particular concern arises when comparing the strength of phylogenetic signal among numerous traits. For example, Blomberg et al. (2003) compared 121 traits corresponding to 35 phylogenetic trees, finding that behavioral traits exhibit lower phylogenetic signal than body size, morphological, life history, and physiological traits. Although this result matches the long-standing idea that behavioral traits are evolutionarily labile, it could also be caused by behavioral traits having higher measurement error, as Blomberg et al. (2003) duly noted. Not all of the data sets analyzed by Blomberg et al. (2003) included standard errors of the species trait values, but future analyses could test the hypothesis that behavioral traits show lower phylogenetic signal simply because they show greater within-species standard errors.

In regression, measurement error leads to bias, generally downward, in the estimates of slopes. The magnitude of this bias obviously depends on the magnitude of the measurement error. However, the consequences of this bias for statistical tests depends also on the number of species analyzed and hence the precision of the estimates of the slope. When the number of species is large, the slope estimates have lower standard errors. This heightens the danger of bias caused by measurement error, because the wrong estimate is "known" with greater precision. This argument also applies to the other statistical models we have investigated, leading to the somewhat counter-intuitive recommendation that it is more important to incorporate measurement error when analyzing large data sets (see also Harmon and Losos, 2005). This recommendation makes more sense, however, when realizing that increasing the number of species will increase information about regression slopes, correlations, etc., thereby reducing the variation in parameter estimates that is not associated with measurement error and hence making the variation due to measurement error relatively greater. The greater the proportion of variation in parameter estimates caused by measurement error, the greater is the impact of not accounting for measurement error.

RMA regression and other forms of bivariate general structural relation models are often recommended as replacements for simple regression when there is variation in both x and y variables; thus, they are recommended when measurement error exists. We prefer to think of the broad suite of functional relation models explicitly in terms of Equation 13, in which there is a linear relationship between traits x and y that is modified by variation in x and y, ε_x and ε_y, that might or might not depend on phylogeny. Measurement error is then additional within-species variation (from any number of sources) that is known to the researcher (i.e., standard errors). The explicit accounting of sources of variation given by Equation 13 makes clearer the assumptions of RMA regression and other functional relation models. For the special cases of functional relation models in which the estimate of the slope b_1 of y on x is the reciprocal of the slope $1/b_1$ of x on y (such as RMA regression), the estimates of the slope are unbiased even when measurement error is not explicitly accounted for. This is a reason to use RMA regression when a researcher knows that measurement error exists but does not know its magnitude. (Of course, one must always be cautious when the traits are actually independent, because RMA regression gives the nonsensical result that the slope is unity for independent traits.) Despite being unbiased, however, incorporating phylogeny and measurement error makes the estimates of the slope b_1 more precise. Unfortunately, the increase in precision requires knowing which model

best fits a data set, and when data sets contain few species, identifying the best-fitting model may be statistically difficult (Fig. 5b and d). Thus, our recommendation is to approach functional relation problems with a variety of plausible models, examine each, and determine the robustness of any conclusions to the choice of models.

It is important to note that RMA regression is a bivariate technique and that it is not actually a formal "measurement error method" (see also Rayner, 1985:417–418); for example, we analyzed RMA regression models with and without the formal inclusion of measurement error. This raises important issues when one wishes to estimate a functional relations slope but also needs to control statistically for nuisance variables and/or other factors. As an example of the former, some studies report whole body mass while others report fat-free body mass. As an example of the latter, whole clades may vary in general body shape, thus causing a "grade shift" (e.g., see Garland et al., 1993, 2005) in the relation between, say, leg length and body size. A bivariate slope fitted to such heterogeneous data will be inappropriate. Unfortunately, RMA regression for more than two traits has yet to be developed, although it could be derived from a generalization of Equation 13. Until such methods are developed, practitioners might adopt a strategy of using (multiple) regression to remove effects of other covariates and factors, saving residuals, and analyzing residuals with bivariate RMA regression.

Throughout this article we have performed statistical tests assuming that interspecific correlations in trait values are known with certainty, given either by a star phylogeny or by the true phylogeny (which we assume in known without error) under the assumption of Brownian motion character evolution. These two assumptions about interspecific correlations can be considered as extremes in a continuum of phylogenetic signal (Garland et al., 2005). Rather than pick one of the extremes, the analyses can be performed by explicitly incorporating the strength of phylogenetic signal as a parameter in the models (Grafen, 1989; Freckleton et al., 2002; Huey et al., 2006). In this case, the phylogenetic covariance matrix $\mathbf{C}$ (or $\mathbf{C}_x$ and $\mathbf{C}_y$, etc.) contains parameters that must be estimated. For example, Hansen (1997), Blomberg et al. (2003), and Butler and King (2004) provide branch-length transforms that correspond to evolution occurring as an Ornstein-Uhlenbeck process, with a parameter measuring the strength of phylogenetic signal (although the exact parameterization of the OU process differs among those sources). This parameter can be incorporated into the statistical models by letting the phylogenetic correlation matrices $\mathbf{C}$ depend on this parameter (Huey et al., 2006; Ives and Godfray, 2006). The parameter can then be estimated simultaneously with all of the other parameters of the model. A particular advantage of this approach is that it overcomes the need to identify whether the model using the true phylogeny fits the data better than the model with a star phylogeny (e.g., Table 5); arbitration between these extremes is made while estimating the Ornstein-Uhlenbeck parameter simultaneously with the other parameters. In general, incorporating parameters that control the phylogenetic covariance matrix $\mathbf{C}$ is straightforward for all of the different problems we have considered in this article, and the estimation techniques can be applied with little modification, as we intend to provide in future versions of our Matlab programs.

Throughout our analyses, we have assumed that information on standard errors for mean species trait values are available. However, even more limited information can be incorporated into the analyses. For example, incorporating an estimate of the average among-species measurement error for a trait can be done by simply giving all species the same standard error for a trait, or by weighting measurement errors according to sample sizes (Appendix 3). Furthermore, if measurement errors are assumed to be the same for all species, then measurement errors can themselves be estimated along with the other parameters in the model. This involves estimating the variance σ_m^2 in the models used throughout this article. This approach is analogous to models that separate among-species variation into phylogenetic and nonphylogenetic components (Lynch, 1991; Freckleton et al., 2002; Housworth et al., 2004; Rochet et al., 2006); in this case, the nonphylogenetic component represents measurement error.

Throughout this article we have used EGLS, ML, and REML estimation methods. For most problems and data sets, we suspect that they will all give similar results, provided measurement error is not too large. For most of the simulations we analyzed, REML outperformed EGLS, which in turn outperformed ML estimation. This does not mean, however, that ML estimation should not be used. An advantage of ML estimation is that the likelihoods from different models can be compared using a likelihood-ratio test, or the likelihood can be used to compute the Akaike information criterion (AIC) for model comparisons (Burnham and Anderson, 1998). In general, we suggest multiple methods be used; TG will provide Matlab (MathWorks, 1996) code that performs all three estimation methods for most of the models we have presented.

Much of the work developing phylogenetically based comparative methods over the last two decades has been done largely in isolation from a large body of statistical literature dealing with correlated data (e.g., Ives and Zhu, 2005). Nonetheless, there is much to be gained by applying relatively off-the-shelf approaches to comparative problems. For example, Paradis and Claude (2002) recently described how generalized estimating equations can be used to apply phylogenetic comparative methods to noncontinuous data, opening up comparison of discrete traits to phylogenetic analyses. The models presented throughout this manuscript are applicable to a broad range of comparative problems, and they easily surrender to well-worn statistical approaches. Answers to a wide range of additional problems in comparative analyses probably await within the statistical literature.

ACKNOWLEDGMENTS

We thank Matt Helmus, Todd Oakley, Roderic Page, Jun Zhu, and two anonymous reviewers for helpful comments on the manuscript. This work was supported by NSF DEB-0196384 to TG and ARI.

REFERENCES

Ashton, K. G. 2004. Comparing phylogenetic signal in intraspecific and interspecific body size datasets. J. Evol. Biol. 17:1157–1161.

Bauwens, D., T. Garland, Jr., A. M. Castilla, and R. Vandamme. 1995. Evolution of sprint speed in lacertid lizards—Morphological, physiological, and behavioral covariation. Evolution 49:848–863.

Blomberg, S. P., T. Garland, Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:171–745.

Bonine, K. E., T. T. Gleeson, and T. Garland, Jr. 2005. Muscle fibre-type variation in lizards (Squamata) and phylogenetic reconstruction of hypothesized ancestral states. J. Exp. Biol. 208:4529–4547.

Burnham, K. T., and D. R. Anderson. 1998. Model selection and inference: A practical information-theoretic approach. Springer, New York.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Christman, M. C., R. W. Jernigan, and D. Culver. 1997. A comparison of two models for estimating phylogenetic effect on trait variation. Evolution 51:262–266.

Cooper, D. M., and R. Thompson. 1977. Note on estimation of parameters of autoregressive-moving average process. Biometrika 64:625–628.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. Am. Nat.160:712–726.

Fuller, W. A. 1987. Measurement error models. John Wiley & Sons, New York.

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

Garland, T., Jr., and A. R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346–364.

Garland, T., Jr., P. E. Midford, and A. R. Ives. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. Am. Zool. 39:374–388.

Garland, T., Jr., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. J. Exp. Biol. 208:3015–3055.

Garland, T., Jr., and R. Díaz-Uriarte. 1999. Polytomies and phylogenetically independent contrasts: An examination of the bounded degrees of freedom approach. Syst. Biol. 48:547–558.

Grafen, A. 1989. The phylogenetic regression. Trans. R. Soc. Lond. B Biol. Sci. 326:119–157.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. Evolution 50:1404–1417.

Harmon, L. J., and J. B. Losos. 2005. The effect of intraspecific sample size on type I and type II error rates in comparative studies. Evolution 59:2705–2710.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford, UK.

Harville, D. A. 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61:383–385.

Housworth, E. A., and E. P. Martins. 2001. Random sampling of constrained phylogenies: Conducting phylogenetic analyses when the phylogeny is partially known. Syst. Biol. 50:628–639.

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

Huelsenbeck, J. P., and B. Rannala. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57:1237–1247.

Huey, R. B., B. Moreteau, J. C. Moreteau, P. Gibert, G. W. Gilchrist, A. R. Ives, T. Garland Jr., and J. R. David. 2006. Evolution of sexual size dimorphism in a *Drosophila* clade, the *D. obscura* group. Zoology 109:318–330.

Irschick, D. J., C. C. Austin, K. Petren, R. N. Fisher, J. B. Losos, and O. Ellers. 1996. A comparative analysis of clinging ability among pad-bearing lizards. Biol. J. Linn. Soc. 59:21–35.

Ives, A. R., and H. C. J. Godfray. 2006. Phylogenetic analysis of trophic associations. Am. Nat. 168:E1–E14.

Ives, A. R., and J. Zhu. 2005. Statistics for correlated data: Phylogenies, space, and time. Ecol. Appl. 16:20–32.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee. 1985. The theory and practice of econometrics, 2nd edition. John Wiley and Sons, New York.

Langerhans, R. B., J. H. Knouft, and J. B. Losos. 2006. Shared and unique features of evolutionary diversification in Greater Antillean Anolis ecomorphs. Evolution 60:362–369.

Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.

Martins, E. P., and T. F. Hansen. 1996. The statistical analysis of interspecific data: A review and evaluation of comparative methods. Pages 22–75 *in* Phylogenies and the comparative method in animal behavior (E. P. Martins, ed.). Oxford University Press, Oxford, UK.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667; erratum 153:448.

Martins, E. P., and J. Lamont. 1998. Estimating ancestral states of a communicative display: A comparative study of *Cyclura* rock iguanas. Anim. Behav. 55:1685–1706.

MathWorks. 1996. Matlab, version 5.0. The MathWorks, Inc., Natick, MA.

Neter, J., W. Wasserman, and M. H. Kutner. 1989. Applied linear regression models. Richard D. Irwin, Homewood, Illinois.

Pagel, M. D., and P. H. Harvey. 1988a. How mammals produce large-brained offspring. Evolution 42:948–957.

Pagel, M. D., and P. H. Harvey. 1988b. The taxon-level problem in the evolution of mammalian brain size—Facts and artifacts. Am. Nat. 132:344–359.

Pagel, M. D., and P. H. Harvey. 1989. Taxonomic differences in the scaling of brain on body-weight among mammals. Science 244:1589–1593.

Paradis, E., and J. Claude. 2002. Analysis of comparative data using generalized estimating equations. J. Theor. Biol. 218:175–185.

Patterson, H. D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–564.

Purvis, A., and T. Garland, Jr. 1993. Polytomies in comparative analyses of continuous characters. Syst. Biol. 42:569–575.

Rayner, J. M. V. 1985. Linear relations in biomechanics: The statistics of scaling functions. J. Zool. Lond. A 206:415–439.

Ricklefs, R. E., and J. M. Starck. 1996. Applications of phylogenetically independent contrasts: A mixed progress report. Oikos 77:167–172.

Rochet, M. J., P. A. Cornillon, R. Sabatier, and D. Pontier. 2006. Comparative analysis of phylogenetic and fishing effects in life history patterns of teleost fishes. Oikos 91:255–270.

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

Rohlf, F. J. 2006. A comment on phylogenetic correction. Evolution 60:1509–1515.

Smyth, G. K., and A. P. Verbyla. 1996. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. J. R. Stat. Soc. Ser. B Methodol. 58:565–572.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry, 2nd edition. W. M. Freeman & Company, New York.

First submitted 27 October 2005; reviews returned 26 February 2006; final acceptance 16 October 2006
Associate Editor: Todd Oakley

APPENDIX 1: ESTIMATION METHODS

This appendix summarizes the estimation techniques we use: iterative EGLS, ML, and REML. We also give an overview of our parametric bootstrapping approach. We have structured the models throughout the article so that the estimation techniques can be applied in a similar way for all models.

Iterative EGLS Estimation

Estimated generalized least-squares is an extension of GLS for the case in which unknown parameters occur in the covariance matrix (Judge et al., 1985:169–187). For the univariate case given in Equation 5, the problem is that $\Psi(\theta) = \mathbf{C} + (\sigma_m^2/\sigma^2)\mathbf{M}$ has an unknown parameter $\theta = (\sigma_m^2/\sigma^2)$ that rules out application of GLS. EGLS is a standard procedure, and our iterative approach is a simple extension in which EGLS is repeated to obtain successively better estimates.

The iterative EGLS is initiated by choosing a value of θ; zero is an obvious choice in most cases. Conditioned on this value of θ, GLS can be used to estimate the mean a and variance σ^2 from Equations 3 and 4. Because σ_m^2 is known, a new estimate of θ is $\sigma_m^2[(N-1/N)\hat{\sigma}_m^2]^{-1}$; here, the estimate of σ^2 is multiplied by $(N-1)/N$ to give the maximum likelihood estimate, rather than the unbiased estimate given by Equation 4. This new value of θ is used to update $\Psi(\theta) = \mathbf{C} + \theta\mathbf{M}$, and the procedure is iterated until estimates of a and σ^2 converge.

The multivariate models are slightly more complex, and they employ a method-of-moments approach. For the case of regression (Equation 12), let $\Psi_x(\theta_x) = \mathbf{C}_x + \theta_x\mathbf{M}_x$ and $\Psi_y(\theta_{xy}, \theta_y) = \mathbf{C}_y + \theta_{xy}\mathbf{C}_x + \theta_y\mathbf{M}_y$, where $\theta_x = \sigma_{mx}^2/\sigma_x^2$, $\theta_{xy} = b_1^2\sigma_x^2/\sigma_y^2$, and $\theta_y = \sigma_{my}^2/\sigma_y^2$. With initial values of θ_x, θ_{xy}, and θ_y to give $\Psi_x(\theta_x)$ and $\Psi_y(\theta_{xy}, \theta_y)$ estimates of a_x and σ_x^2, and a_y and σ_y^2 are computed from Equations 3 and 4. Using the method of moments, an estimate of b_1 is given by

$$\hat{b}_1 = \frac{1}{\hat{\sigma}_x^2}\left[\frac{(\mathbf{X}-\hat{a}_x)'\mathbf{C}_{xy}^{-1}(\mathbf{Y}-\hat{a}_y)}{N-1} - r_m\sigma_{mx}\sigma_{my}\mathbf{M}_{xy}\right]. \tag{A1}$$

These estimates of σ_x^2, σ_y^2, and b_1 are then used to update θ_x, θ_{xy}, and θ_y, and the procedure is iterated until the estimates converge.

Approximate confidence intervals for the parameter estimates can be obtained using standard GLS formulae, treating $\Psi(\theta)$ as known (i.e., ignoring the uncertainty arising from estimating parameters θ). In addition, parametric bootstrapping can be used, which is particularly convenient for iterative EGLS because EGLS takes relatively little computing time. The statistical properties of EGLS estimators can be complex, but in general the EGLS estimates of a and b_1 will be consistent, meaning that for large enough sample size the probability distributions of the estimators converge to the true parameter values (Judge et al., 1985:175–177). In practice, we have found that the iterative EGLS estimates are often similar to ML and REML estimates, and the iterative EGLS outperforms ML and REML for some tasks, such as estimating the phylogenetic signal K*.

ML Estimation

Maximum likelihood estimation consists of using the matrix $\Psi(\theta)$, containing parameters θ that need to be estimated, to construct the likelihood function, and then maximizing the likelihood function to obtain those parameter values for which the observed data set is the most likely. The log-likelihood function is

$$\begin{aligned} L(\mathbf{a}, \sigma^2, \theta|\mathbf{X}) = &-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\{\det[\sigma^2\Psi(\theta)]\} \\ &- \frac{1}{2}(\mathbf{X}-\mathbf{Z}\mathbf{a})'[\sigma^2\Psi(\theta)]^{-1}(\mathbf{X}-\mathbf{Z}\mathbf{a}) \end{aligned} \tag{A2}$$

For the univariate case, $\mathbf{a} = a$ is a scalar giving the phylogenetic mean, and $\mathbf{Z}$ is the $N \times 1$ vector of ones. For correlation of p traits, $\mathbf{a}$ is the $1 \times p$ vector containing the means of the traits, and $\mathbf{Z}$ is the $pN \times p$ matrix given by $\mathbf{I}_{p\times p} \otimes \mathbf{1}_{N\times 1}$ where $\mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix, $\mathbf{1}_{N\times 1}$ is a $N \times 1$ vector of ones, and $\otimes$ denotes the Kronecker (inner) product. For regression with $p-1$ independent variables, $\mathbf{a}$ is the $1 \times p$ vector containing the intercept and slopes, and $\mathbf{Z}$ is the $N \times p$ matrix containing ones in the first column and the independent variables in the remaining columns. To speed the numerical maximization of the log-likelihood function, it is possible to concentrate the likelihood function by replacing $\mathbf{a}$ and σ^2 by their GLS estimates conditioned on θ and maximizing over θ alone. To account for the known bias of ML estimates of variances, we multiplied the ML estimates of variances by $N/(N-p)$ to obtain a less biased estimate. Finally, maximization must be restricted so that estimates of variances are greater than zero.

Special considerations are necessary for the case of correlation to guarantee that the covariance matrix $\sigma^2\Psi(\theta)$ remains positive definite while the log-likelihood function is maximized. Rather than maximize over the correlation coefficients for the multivariate case, we instead maximized over the parameters q_{ij} defined such that the correlation matrix $\mathbf{R}(\theta) = \mathbf{Q}(\theta)^2$, where $\mathbf{Q}(\theta)$ is symmetric having values of 1 along the diagonal and values of q_{ij} in the off-diagonals. Because $\mathbf{R}(\theta)$ is the square of a symmetric matrix, it is necessarily positive definite. After maximizing over values of q_{ij} constrained to be between -1 and 1, estimates of the correlation coefficients are given as the off-diagonal elements of $\mathbf{R}(\theta)$.

An advantage of ML estimation is that it can be used to provide approximate confidence intervals for the parameter estimates. In particular, let $I(\varphi) = -\frac{\partial^2}{\partial\varphi^2}L(\varphi|\mathbf{X})$ be the observed information matrix calculated at the ML parameter estimates φ. Then the approximate covariance matrix for the parameter estimates φ is given by $I(\varphi)^{-1}$. From this, standard errors and asymptotic confidence intervals can be calculated (Judge et al. 1985:177–182). Unfortunately, this is only an asymptotic result, and the accuracy of the approximation for small samples is generally unknown.

REML Estimation

Restricted maximum likelihood estimation is a variant of ML estimation in which the likelihood function is partitioned into components, allowing estimation of variance parameters in the model independently from the parameters involving means (Patterson and Thompson, 1971; Cooper and Thompson, 1977; Smyth and Verbyla, 1996). The marginal log-likelihood function from which variance parameters are estimated is (Harville, 1974)

$$\begin{aligned} L(\sigma^2, \theta|\mathbf{X}) = &-\frac{N-p}{2}\ln(2\pi) + \frac{1}{2}\ln[\det(\mathbf{Z}'\mathbf{Z})] \\ &- \frac{1}{2}\ln\{\det[\sigma^2\Psi(\theta)]\} \\ &- \frac{1}{2}\ln(\det\{\mathbf{Z}'[\sigma^2\Psi(\theta)]^{-1}\mathbf{Z}\}) \\ &- \frac{1}{2}(\mathbf{X}-\mathbf{Z}\hat{\mathbf{a}}_{GLS})'[\sigma^2\Psi(\theta)]^{-1}(\mathbf{X}-\mathbf{Z}\hat{\mathbf{a}}_{GLS}) \end{aligned} \tag{A3}$$

where $\hat{\mathbf{a}}_{GLS}$ is the GLS estimate of $\mathbf{a}$, and the remaining terms are as defined for ML estimation.

In general, we found that REML estimates of the variance parameters were less biased than those obtained from ML and iterative EGLS. Estimates of the parameters involving means (e.g., a, b_0, and b_1), however, were very similar to those obtained from ML and iterative EGLS. Confidence intervals for parameters involving means can be approximated using the formulae for standard GLS with the fitted covariance matrix $\sigma^2\Psi(\theta)$ or asymptotically using the information matrix.

Parametric Bootstrapping

For all estimation approaches, we used parametric bootstrapping to obtain confidence intervals. Parametric bootstrapping is useful because it not only produces confidence intervals but also identifies bias in the estimates. We performed parametric bootstrapping for a given model by first estimating parameters and then using the model fitted with

these estimates to simulate 2000 data sets. For each of the simulated data sets, we estimated the parameters of the model. The resulting 2000 sets of parameter estimates approximate the distribution of the parameter estimators under the assumption that the model (with its fitted parameter values) is correct. Thus, the 95% inclusion intervals for the 2000 estimates give the approximate 95% confidence intervals for the parameter estimates. Bias in the estimates is revealed if the mean of the 2000 simulated parameter estimates differs substantially from the parameter values estimated from the data and used to construct the simulated data sets.

APPENDIX 2: LOG-TRANSFORMING DATA WITH MEASUREMENT ERROR

Most empirical comparative studies report means and standard errors for species on the arithmetic scale. However, for many problems, such as allometric analyses, data need to be log-transformed before analysis. When measurement error exists, log-transformation requires not only transforming the standard error of the measured values but also the mean of the measured values. Assuming that the observed values follow a log-normal distribution with mean m and variance v, the mean and variance of the log-transformed data are $\bar{x} = \log(m) - \frac{v}{2}$ and $\sigma^2 = \log(\frac{v}{m^2} + 1)$. These relationship can be derived directly from the probability density function of a log-normal distribution.

APPENDIX 3: MEASUREMENT ERROR WITH SMALL SAMPLE SIZES

When sample sizes used to estimate trait values for some or all species are small, measurement error can be obtained by averaging among species. Suppose there are N species with standard errors SE_i ($i = 1, \ldots, N$) taken from n_i individuals. Assume that each observation on each individual is measured with the same error. Let $\hat{\sigma}_i = \sqrt{n_i} SE_i$, and let $\bar{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}_i^2$. Then the standard error for a species with n_i individuals estimated by pooling species is $\overline{SE}(n_i) = \bar{\sigma}/\sqrt{n_i}$. For data sets in which one or more species are represented by a single individual ($n_i = 1$), these species can be excluded from the calculations and then assigned standard errors $\overline{SE}(1) = \bar{\sigma}$.

As an example, suppose there are two species with standard errors 2 and 3, and sample sizes 16 and 9. Then $\bar{\sigma}^2 = \frac{1}{2}(\sqrt{16} \times 2 + \sqrt{9} \times 3) = \frac{17}{2}$, and the standard errors estimated from pooling species are $\overline{SE}(16) = 2.125$ and $\overline{SE}(9) = 2.833$.

APPENDIX 4: DEMONSTRATION THAT EQUATION 13 IS THE MODEL OF RAYNER (1985)

In the general structural equation model of Rayner (1985), the slope b_1 of the relationship between two traits x and y is the value that minimizes the residual sum-of-squares given by

$$S_R^2 = \frac{b_1^2 S_{xx} - 2b_1 S_{\mathbf{xy}} + S_{yy}}{b_1^2 S_{\xi\xi} - 2b_1 S_{\xi\eta} + S_{\eta\eta}} \tag{A4}$$

where S_{xx}, $S_{\mathbf{xy}}$, and S_{yy} are the sums-of-squares of the values $\mathbf{X}$ and $\mathbf{Y}$, and $S_{\xi\xi}$, $S_{\xi\eta}$, and $S_{\eta\eta}$ are the variances and covariances of the error terms. Here we derive the same expression from Equation 13 for the case when there is no phylogenetic correlation ($\mathbf{C}_{\gamma\mathbf{x}} = \mathbf{C}_{\mathbf{x}} = \mathbf{C}_{\mathbf{y}} = \mathbf{I}$) and measurement error is zero ($\mathbf{M}_{\mathbf{x}} = \mathbf{M}_{\mathbf{y}} = \mathbf{M}_{\mathbf{xy}} = \mathbf{0}$). In this case, $S_{\xi\xi} = \sigma_x^2 S_{\xi\eta} = r\sigma_x\sigma_y$, and $S_{\eta\eta} = \sigma_y^2$.

Let $\mathbf{Z}$ be the $N \times 2$ matrix whose first and second columns consist of $\mathbf{X} - \bar{x}$ and $\mathbf{Y} - \bar{y}$. From Equation 13, the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\mathbf{V} = E\left\{\mathbf{Z}'\mathbf{Z}\right\} = \begin{pmatrix} \sigma_\gamma^2 + \sigma_x^2 & b_1\sigma_\gamma^2 + r\sigma_x\sigma_y \\ b_1\sigma_\gamma^2 + r\sigma_x\sigma_y & b_1^2\sigma_\gamma^2 + \sigma_y^2 \end{pmatrix} \tag{A5}$$

From this, the residual sum-of-squares, after some algebra, is

$$\begin{aligned} S^2 &= \mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} \\ &= \frac{S_{xx}\left(b_1^2\sigma_\gamma^2 + \sigma_y^2\right) - 2S_{\mathbf{xy}}\left(b_1\sigma_\gamma^2 + r\sigma_x\sigma_y\right) + S_{yy}\left(\sigma_\gamma^2 + \sigma_x^2\right)}{b_1^2\sigma_\gamma^2\sigma_x^2 - 2b_1\sigma_\gamma^2 r\sigma_x\sigma_y + \sigma_\gamma^2\sigma_x^2 + \sigma_x^2\sigma_y^2 - (r\sigma_x\sigma_y)^2} \end{aligned} \tag{A6}$$

The least squares estimates of b_1 and σ_γ^2 are those values that minimize the value of S^2. Taking the derivatives of S^2 with respect to b_1 and σ_γ^2 and setting them equal to zero,

$$\begin{aligned} \frac{\partial S^2}{\partial b_1} &= \frac{b_1 S_{xx} - S_{\mathbf{xy}}}{b_1\sigma_x^2 - r\sigma_x\sigma_y} - S^2 = 0 \\ \frac{\partial S^2}{\partial \sigma_\gamma^2} &= \frac{b_1^2 S_{xx} - 2b_1 S_{\mathbf{xy}} + S_{yy}}{b_1^2\sigma_x^2 - 2b_1 r\sigma_x\sigma_y + \sigma_x^2} - S^2 = 0 \end{aligned} \tag{A7}$$

The value of b_1 that minimizes the sum-of-squares given by Rayner (Equation A4) is

$$\begin{aligned} \frac{\partial S_R^2}{\partial b_1} &= \frac{b_1 S_{xx} - S_{\mathbf{xy}}}{b_1\sigma_x^2 - r\sigma_x\sigma_y} - \frac{b_1^2 S_{xx} - 2b_1 S_{\mathbf{xy}} + S_{yy}}{b_1^2\sigma_x^2 - 2b_1 r\sigma_x\sigma_y + \sigma_x^2} \\ &= \frac{\partial S^2}{\partial b_1} - \frac{\partial S^2}{\partial \sigma_\gamma^2} = 0 \end{aligned} \tag{A8}$$

Thus, the least-squares estimate of b_1 computed for the model given by Equation 13 in the text equals the least-squares estimate of b_1 computed from Rayner's function (Equation A4).