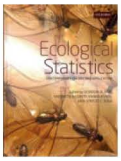


Lajeunesse, M.J. and Fox, G.A. (2015) Statistical approaches to the problem of phylogenetically correlated data. In G.A. Fox, S. Negrete-Yankelevitch, and V.J. Sosa, editors. Ecological Statistics: Contemporary Theory and Application (pp. 261–283). Oxford University Press, Oxford, UK.

CHAPTER 11

Statistical approaches to the problem of phylogenetically correlated data

Marc J. Lajeunesse and Gordon A. Fox

11.1 Introduction to phylogenetically correlated data

Multi-species data sets violate some of the most basic assumptions of traditional statistics, and so present an important challenge for data analysis. For example, ecologists might want to use regression to explore the relationship between body mass and aerobic capacity. Given the considerable variation in body mass across mammals, from shrews to whales, there is extensive opportunity to explore trends in these two quantities. However, linear regression assumes that each observation is independent of the other (Stuart and Ord 1994). What is at risk here is that data from closely related mammals will not adequately form independent pieces of information: the shared evolutionary history of these taxa will introduce correlations, or dependencies, in data (Felsenstein 1985). A conventional regression would treat data from multiple species of canines, cats, and weasels as independent, despite the potential correlations in their characteristics due to their shared ancestry as Carnivora.

Because of these potential dependencies in data, and their effects on statistical assumptions, serious inferential errors can emerge when analyzing and comparing data from multiple species using conventional regression methods. For example, there is no longer a guarantee that statistical hypothesis tests remain valid. Standard tests, like those asking whether the slope of the regression is significantly non-zero, are no longer valid (Diaz-Uriarte and Garland 1996). To minimize this problem with interspecific (multi-species) data, techniques based on the *phylogenetic comparative method* are used to improve the reliability of inferences with regression (e.g., Felsenstein 1985; Grafen 1989; Pagel 1993). These statistics use information on phylogenetic evolutionary history to hypothesize the strength of correlations across taxa. These *phylogenetic correlations* are then applied to generalized least squares (GLS) models—a statistical framework less rigid to violations of assumptions like independence of data. GLS modeling is also used in other areas of ecology, including studies in which *serial correlations* or *spatial correlations* occur in the data (see chapters 10 and 13). Thus the principles of regression modeling used in this chapter

Ecological Statistics: Contemporary Theory and Application. First Edition. Edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa.
Published in 2015 by Oxford University Press.

are rather general, and we hope that readers can glean some broad lessons even if they never use multi-species data sets.

The basic principles underlying most of these phylogenetic comparative statistics are also straightforward, and share two common themes. The first is to enable a phylogenetic framework to test hypotheses on evolutionary processes (e.g., Hansen 1997; Pagel 1999). The second, and perhaps more germane to many ecologists, is to offer some assurance that inferences drawn from multi-species analyses are valid and statistically sound (e.g., Price 1997; Schluter 2000). The goals of this chapter are to introduce the basic principles of phylogenetic comparative methods, and to demonstrate *why* it is important to apply these methods when analyzing interspecific data. We emphasize *why* here because this is not often clearly addressed in introductory texts. What gets addressed typically is *how* to apply these techniques. For example, there are already several published reviews, surveys, and how-to guides on comparative methods (e.g., Harvey and Pagel 1991; Martins 1996; Martins 2000; Blomberg and Garland 2002; Felsenstein 2004; Garland et al. 2005; Nunn 2011; Paradis 2011; O'Meara 2012). By focusing here on *why* rather than *how*, we aim to provide a unique view of the risks of not applying these statistical tools, as well as insight on the limitations for what they can accomplish.

To achieve these goals, we center the chapter on a series of *Monte Carlo experiments* that aim to answer the following: *Why is it risky to use regression with multi-species data? When do you expect the greatest risk? What are phylogenetic correlations, and how are they used in regression models? What happens when the incorrect model of evolution is assumed?* Monte Carlo simulations use randomly generated data to investigate the conditions for when statistical tests, such as regression, provide reliable outcomes (Rubinstein and Kroese 2007)–or equally when they fail to provide reliable outcomes. This simulation approach has been crucial to the development of comparative phylogenetic methods and the way they are practiced (e.g., Martins and Garland 1991; Freckleton et al. 2002; Martins et al. 2002; Revell 2010; Freckleton et al. 2011). Our intention is to use simulations to: (1) reveal the underlying principles on why it is important to apply phylogenetic correlations to regression models by simulating interspecific data sets, and (2) introduce several of the powerful and diverse statistical functionalities offered in *R*. These include the widely used `ape` (Paradis et al. 2004) and `geiger` library (Harmon et al. 2008), useful for manipulating and applying phylogenies for regression modeling.

We focus exclusively on the analysis of interspecific data using linear regression, which historically has received the most attention (Felsenstein 1985; Pagel 1999), and for the purposes of this chapter, serves as an accessible introduction to more advanced statistical models and practices covered elsewhere (Martins and Hansen 1997; Pagel 1999; Revell 2010). Our aim is to channel the reader from simple to more complex topics using linear regression, assuming that the reader has little to no familiarity with comparative methods, by introducing key concepts as they emerge. We hope that this stepwise exposition helps readers gain insight on the power of these statistical tools, as well as practical information on how to interpret results from analyses with interspecific data.

11.2 Statistical assumptions and the comparative phylogenetic method

Because closely related species tend to be more similar than distantly related species, methods like regression and ANOVA—which assume normality, homogeneity of variance, and independent (uncorrelated) errors (Stuart et al. 1999)—are not generally valid for

multi-species data. We use Monte Carlo experiments and regression analyses to explore the consequences of violating these assumptions for a simple reason: because we have simulated the data, we know the right answers, so we can see how different methods affect our conclusions. We begin by introducing a simple linear regression model, and then extend this model to include phylogenetic correlations. Because we will need to refer to bits of R script repeatedly, we label them somewhat like equations: the *j*th piece of R script we use is labeled R.j.

11.2.1 *The assumptions of conventional linear regression*

Let's start by decomposing a simple linear regression model. This will provide insight on how multi-species data sets can challenge inferences with this model. The goal of regression is to estimate the linear relationship between a dependent variable (y) and an independent (explanatory or predictor) variable (x). For example, can mass (x) predict aerobic capacity (y), or body size predict fitness? Perhaps the simplest way to ask this question is to model a straight line in $x - y$ coordinates. This line can be described in equation form as:

$$y_i = a + bx_i + \varepsilon_i. \tag{11.1}$$

Here the subscript i indexes (refers to) individual data points (or samples) of these variables, and it can take on the values $i = 1, 2, \ldots, N$ where N is the total number of observations. Equation (11.1) contains the y-intercept (a) and the slope (b) of the line. These are unknown, and what we want to estimate with regression. An important aspect of modeling the relationship between y and x is the random error (or residual) variable ε. This term is the stochastic noise in this relationship. It is often interpreted as all the variation in y that isn't explained by x. Under the simplest linear models, ε is assumed to be independent across all observations; a further assumption is that these observations will be Normally distributed (N) around a mean of zero and have a common (or homogenous) variance (σ^2). These assumptions on the distribution of ε can be summarized with:

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2). \tag{11.2}$$

Given these assumptions, let's simulate some data that fit the relationship described in equation (11.1) and see how regression can estimate the intercept and slope of this linear model. For our simulation, and given equation (11.1), let's arbitrarily set the slope to 0.5 ($b = 0.5$), and the intercept to 1 ($a = 1.0$). What remains is to simulate ε and the explanatory variable x. Equation (11.2) describes the distribution of ε, and so for this term we will simulate random errors that are Normally distributed with a mean of zero and a variance of 1 ($\sigma^2 = 1.0$). Likewise, data for x will be generated by randomly sampling a Normal distribution with a mean of 0 and variance of 1. Finally, we will generate 30 data points for y_i and x_i ($N = 30$) and analyze these with regression. Applying equation (11.1) generates a correlation between y and x; our aim is to detect this relationship with regression. Following these parameters, we can quickly simulate random y_i and x_i in *R* as follows:

```
N <- 30; x_mean <- 0; x_variance <- 1; a <- 1; b <- 0.5
e <- rnorm(N, 0, 1) # sample 30 standard Normal deviates
x <- x_mean + sqrt(x_variance) * rnorm(N, 0, 1) # random x's     (R.1)
y <- a + b * x + e # see eq. (11.1)
```

Using a linear regression model in *R* to estimate the slope and intercept from these simulated y_i and x_i:

```
library(nlme) # load R library
# get regression results
results <- summary(gls(y ~ x, method="ML"))
# print only regression coefficients
results$tTable
# print only residual error
paste("residual error = ", results$sigma)
```

(R.2)

we get the following *R* output from this `gls` function (bold our emphasis):

```
                Value Std.Error t-value p-value
(Intercept)  1.0292094 0.2212227 4.652369  0.0001
x            0.4728596 0.2209420 2.140198  0.0412

"residual error = 1.167178"
```

Given the sample size and the method used to simulate ε to add a lot of stochastic noise to the model (i.e., with large variance of $\sigma^2 = 1.0$), the regression model provides a reasonable estimation of the residual error ε ($1.167 \approx 1.0$), and is able to detect that the intercept ($1.029 \approx 1.0$) and slope ($0.473 \approx 0.5$) were significantly non-zero via *t*-scores (i.e., *p*-values < 0.05). (Well, in fact, it took several runs of the R.1 and R.2 scripts to get these nice results! More on this later when we explore how sampling error can influence how well regression can detect this slope.) The variances of the regression coefficients are also close to what we simulated. For example, the regression output reported the standard error (S.E.) of the intercept to be 0.22; this translates to a variance of approximately 1 (i.e., $\sigma^2 = \text{S.E.}\sqrt{N} = 0.22\sqrt{30} = 1.21 \approx 1$). The slope's variance was also approximately 1. However, note that the expected variance of the slope in this model is 1 only because $\sigma_b^2 = \sigma_s^2/\sigma_x^2 = 1/1 = 1$. These results make sense given the way we randomly generated our data.

However, given that it took multiple runs of R.1 and R.2 to get these nice results, let's explore the error rates of this regression model in a more rigorous way. To achieve this, we will perform a Monte Carlo experiment to investigate how sampling error influences the way regression can detect a pre-defined relationship between y and x. In other words, we'll try to determine why we needed to run our previous example multiple times to get the results predicted by our linear model from equation (11.1) parameterized as: $y = 1 + 0.5x + \text{N}(0, 1)$. More specifically, we will assess the *Type II error* rates of linear regression at differing sample sizes—that is, estimate the probability of regression statistics failing to detect the intercept and slope for a given N. For this simulation, we will generate random data following the linear model in equation (11.1), analyze these with regression, and repeat this process 1,000 times with increasing sample sizes (N). For each iteration, we will count when the *p*-value of each *t*-score was greater than 0.05 (our significance level) for each regression coefficient. Finally, we divide this count by the number of simulation replications (1,000 per N). This will give us the proportion of analyses that concluded incorrectly that the regression coefficients did not differ from zero (i.e., Type II error). The *R* script for this simulation is found in appendix 11.A.

Our simulation of regression analyses with $N = 5$ to 50 revealed that with small sample sizes linear regression performs poorly and has large Type II error rates for detecting non-zero regression coefficients (figure 11.1). Generally, any results based on a regression with

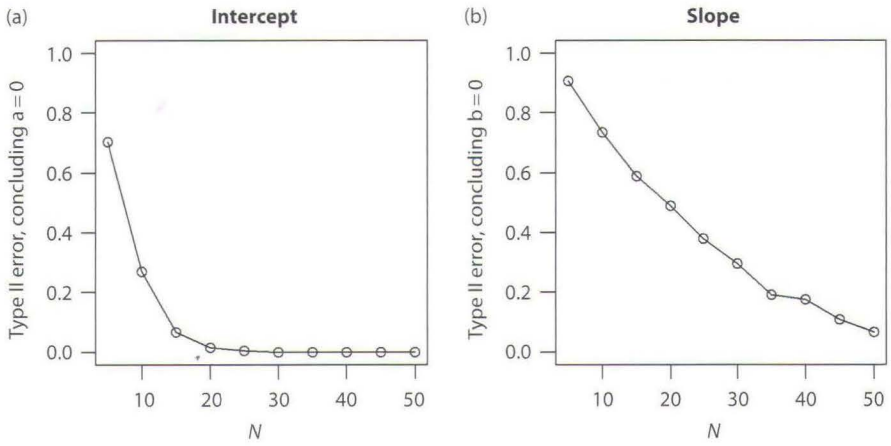


Fig. 11.1 The risk of concluding null results when few data are used in regression analyses. Presented are results from a Monte Carlo experiment exploring the Type II error rates (i.e., false negative outcomes) of ordinary least square (OLS) regression with small to large sample sizes (*N*). Error rates are based on the proportion of 1,000 regression analyses incorrectly concluding that the intercept (*a*) and slope (*b*) were zero. Here regression analyses have improved ability to detect a non-zero slope and intercept with larger sample sizes (*N*). The R script for this simulation is in appendix 11.A.

$N < 30$ should be interpreted with caution. This is because there are too few data sampled to generate enough variation for our regression analyses to properly estimate the slope and intercept of these simulated data. There is also substantial difficulty in detecting our non-zero slope; this is because we modeled the residual error (ε) to add a lot of stochastic noise to our model.

In summary, regression is a tool that aims to estimate the relationship between two variables, but the ability for regression statistics to detect this relationship is often largely dependent on the sample size (N). Our next goal is to repeat this simulation but with interspecific (multi-species) data and explore how these multi-species data can invalidate the assumptions of how ε is modeled, and why this can further impact the outcome of regression analyses.

11.2.2 *The assumption of independence and phylogenetic correlations*

Interspecific data sets generally violate the assumption of independence, because species form a nested hierarchy of phylogenetic relationships. This shared history introduces phylogenetic correlations among related species, and as a result, data from related species may not be statistically independent (Felsenstein 1985). Let us emphasize this in another way: data from related species may not form independent pieces of information and may be correlated—they share a common ancestor and therefore may also have common characteristics. However, we can use phylogenetic information to predict these correlations, and these predictions can then be applied to improve regression estimation and statistical inferences with interspecific data.

But how do phylogenetic correlations come into play with regression analyses? To answer this, we will first need to expand our regression model. Phylogenetic correlations are a problem for the linear model defined in equations (11.1) and (11.2) because the residual errors (ε)

are assumed to be mutually uncorrelated. This linear model is in fact a simplification of a more general way to model ε based on *ordinary least squares* (OLS):

$$\varepsilon_i^{\text{OLS}} \sim \text{MVN}(0, \sigma^2\mathbf{I}). \tag{11.3}$$

This formulation is a different way of writing the linear regression model that we have been discussing. Writing it this way will allow us to relax assumptions about independence of data points and homoscedasticity. But first, let's understand the model in equation (11.3). Here ε has a multivariate (MV) Normal distribution, with a mean of zero and variance equal to $\sigma^2\mathbf{I}$. The idea is that instead of a single variance σ^2 that holds for all values of y, we have a variance–covariance matrix of dimension $N \times N$. This matrix gives the variances for each value of y on its diagonal, and has important properties for linear modeling because it contains information describing the dependency between each pair of data points. These dependencies are modeled by the covariances between pairs of values in all off-diagonals of the matrix. In this case, the matrix $\mathbf{I}$ is the identity matrix (1's on the diagonal, and 0's everywhere else), so $\sigma^2\mathbf{I}$ tells us that, indeed, every point has the same variance and they are all independent of one another.

Given this variance–covariance matrix, let's relax the assumptions of homoscedasticity and independence. We need to do this when the residual error (ε) is not distributed according to equation (11.3), such as when data are phylogenetically correlated. This is because under these conditions, OLS models may no longer provide unbiased estimates of regression coefficients, and statistical tests used for null hypothesis testing may no longer be valid (Diaz-Uriarte and Garland 1996). If phylogenetic correlations are known (or hypothesized), as is the case when we have a hypothesis on the phylogenetic history of taxa (section 11.2.3), then analyzing interspecific data now becomes a generalized least squares (GLS) problem (Pagel 1993; Revell 2010). The error term of this GLS model is defined as:

$$\varepsilon_i^{\text{GLS}} \sim \text{MVN}(0, \sigma^2\mathbf{C}). \tag{11.4}$$

Here, $\varepsilon_i^{\text{GLS}}$ models the variance-covariance matrix ($\mathbf{C}$) to have off-diagonal covariance among data from different but related species. The next section describes exactly how we hypothesize these covariances using phylogenies.

11.2.3 *What are phylogenetic correlations and how do they affect data?*

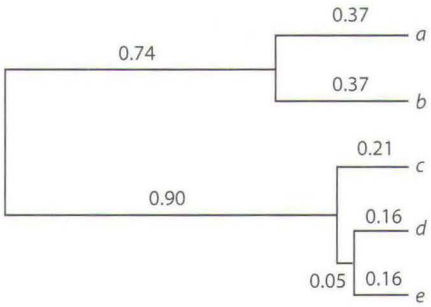
Before exploring how regression can be modified to analyze interspecific data, we need to know a little more about phylogenies and how to extract phylogenetic correlations. Phylogenies are statistical hypotheses on the shared history of taxa (Felsenstein 2004), and for our purposes they contain information on the relative phylogenetic distances of species. The sources of phylogenies are diverse; for example, molecular or morphological information can be used to statistically group related species. The methods used to construct trees are beyond the scope of this chapter (but see Felsenstein 2004); instead we will generate a simple random tree by simulating lineages "branching-out" or diverging randomly with time. This random branching model is called a *Yule birth–death process*. Here is some *R* script that uses the `geiger` library (Harmon et al. 2008) to simulate this birth–death process to generate a small random phylogeny with 5 species (also often described as a phylogeny with five "tips"):

```
# Simulate random phylogenetic tree with 5 species using geiger
# (see Harmon 2008).
library(ape); library(geiger);
```

```
K <- 5 # five species or tips on the simulated tree
# random tree from birth-death model
tree <- sim.bdtree(b=1, d=0, stop="taxa", K)                    (R.3)
# assign letter names to tips
tree$tip.label <- paste(letters[K:1], sep="")
# plots random tree graphically plot(tree)
plot(tree)
# outputs tree in Newick text format
write.tree(tree, digits=2)
```

Running R.3, we generated the random phylogeny shown in the left of figure 11.2.

There are several characteristics of this tree that are notable in terms of predicting phylogenetic correlations. First, the branching pattern of this phylogeny, known as its *topology*, has two major lineages: one with species *a* and *b*, and a second with *c*, *d*, and *e*. One way to interpret these two lineages is to think of them as two distantly related taxonomic groups (e.g., families or orders). These broad groupings are important because they help predict which species will be correlated with one another. For example, these two lineages will translate into two clusters of phylogenetic correlations: data from species *a* and *b* will be correlated with one another, but not with *c*, *d*, and *e*. There is no correlation between these two groups because they stem from the *root* of the tree. The root is the hypothesized ancestral divergence of the entire lineage. Second, note that the nodes of the tree, which designate historic divergence or speciation events, are clustered near the tips of the tree for each group. This tight grouping will create strong correlations among species within these groups; if they were positioned closer to the root (i.e., designating more ancient divergences), then correlations would be weaker.



	a	b	c	d	e
a	1.11	0.74	0.00	0.00	0.00
b	0.74	1.11	0.00	0.00	0.00
c	0.00	0.00	1.11	0.90	0.90
d	0.00	0.00	0.90	1.11	0.95
e	0.00	0.00	0.90	0.95	1.11

Fig. 11.2 Left, a phylogenetic tree generated from a random birth–death process (R.3). There are two major lineages: one with species *a* and *b*, and a second with *c*, *d*, and *e*. Right, the variance–covariance matrix corresponding to this phylogeny. All off-diagonals of this matrix contain the sum of the shared pairwise distance between species; for example, *a* and *b* share an internode distance of 0.74. This is not the branch-length distance from tip *a* to tip *b* (which coincidently sums to 0.74); rather it is only the internode distance shared by *a* and *b*. Also, note that the distance from root to tip for species *a* equals 1.11. This is because it shares an internode distance (common history) with *b* from the root of the tree of length 0.74, followed with a divergence period after a speciation event of length 0.37. Branch lengths calculated from this variance–covariance matrix are plotted along each branch on the left. The Newick (a computer-readable notation) version of this tree is "(((e:0.16,d:0.16):0.05,c:0.21):0.90,(b:0.37,a:0.37):0.74);".

Comparative phylogenetic methods use these characteristics of phylogenetic trees, such as the topology and the distances between each node (known as the internode distance), to predict which species will be correlated with one another and to quantify the strength of these correlations between species. Our next step is to extract these correlations from our random tree. We can quickly calculate its phylogenetic correlations using the matrix functions available in the APE library (Paradis et al. 2004), beginning with the raw phylogenetic distance matrix (*VCV*) whose elements are the sums of branch-length internode distances:

```
# Convert the phylogenetic tree into a variance-covariance matrix.
# Note that we reorder the columns of this matrix to make it easy
# to compare with the topology of the phylogeny in Figure 11.2.
# calculate matrix from phylogeny
VCV <- vcv(tree)                                                    (R.4)
# assign non-number names to tips
order <- paste(letters[1:K], sep="")
# round-down numbers, and order matrix by name
round(VCV[order,order], 2)
```

The distance matrix of our phylogeny from the left of figure 11.2 is shown in the right of figure 11.2. Note how there are essentially two submatrices (with non-zero values) in this matrix. These two submatrices designate the major lineages of our tree (i.e., group *a* and *b*, and group *c*, *d*, and *e*). Also note that all the main diagonals of this matrix equal 1.11. This is the sum of all the branch-length distances (internode distances) from the root to tip for each species.

The left of figure 11.2 also shows the internode branch-length distances. We want to emphasize that all the main diagonals of the matrix in the right of figure 11.2 are equal to 1.11 because all the tips (species, or terminal taxa) are aligned contemporaneously—that is, the distance from the root to each tip is the same. Trees with this alignment are described as having an *ultrametric* shape and are called dendrograms because they depict evolutionary time; there is a chronological ordering of nodes that hypothesize the historic divergences among lineages (Felsenstein 2004). In effect, this is how we generated our tree, by simulating random speciation events and divergences of taxa through time (via a Yule process; see Harmon et al. 2008). Trees estimated from genetic information, such as maximum likelihood trees based on nucleotide sequence data, can also generate dendrograms by assuming constant rates of random molecular change (e.g., a molecular clock). In fact, this time component of dendrograms is a crucial aspect of the comparative phylogenetic method, and later we will describe how it is used to hypothesize evolutionary processes (section 11.2.5).

The elements of the matrix in the right of figure 11.2 are still a little abstract given that they are in terms of branch-length distances; remember our goal is to use the phylogeny to estimate correlations among species. These correlations are meant to quantify the predicted relationship between interspecific variation and the phylogeny for which taxa evolved (Martins and Hansen 1997). Luckily this is straightforward, and we can quickly convert all these distance values into correlations by dividing each element in the variance–covariance matrix with the total branch-length distance from root to tip of the tree (i.e., 1.11). This is only possible because our tree is ultrametric. Dividing all the elements of the matrix by 1.11 (*R* script: `1/1.11 * vcv(tree)`) yields the correlation matrix shown in table 11.1. You can also extract this matrix by using the `cov2cor(vcv(tree))` function in *R*. Now the main diagonals in table 11.1 equal 1

Table 11.1 Correlation matrix (**C**) of our simulated phylogeny (see figure 11.2). Numbers in bold are meant to emphasize the two major subgroups *a–b* and *c–d–e* of phylogenetic correlations in this phylogeny

	a	b	c	d	e
a	**1.00**	**0.67**	0.00	0.00	0.00
b	**0.67**	**1.00**	0.00	0.00	0.00
c	0.00	0.00	**1.00**	**0.81**	**0.81**
d	0.00	0.00	**0.81**	**1.00**	**0.86**
e	0.00	0.00	**0.81**	**0.86**	**1.00**

because taxa are perfectly correlated with themselves, and off-diagonals have the pairwise correlations among taxa. For example, the correlation between *e* and *c* equals 0.81 (e.g., 0.90/1.11 = 0.81) since they only "recently" diverged—that is, recent relative to all other divergences on the tree.

Given this tree and its phylogenetic correlation matrix, our next step is to update our regression analysis and apply these correlations to model potential dependencies in interspecific data. Our matrix from table 11.1 will become the phylogenetic correlation matrix **C** used in GLS regression models [equation (11.4)]. We will also use **C** to generate random interspecific data in Monte Carlo experiments. There are many packages available to simulate correlated data in *R* (see Harmon et al. 2008), but here we will generate these directly using the *Cholesky decomposition* method (Rubinstein and Kroese 2008). The Cholesky method (described more fully following this example) takes random data and transforms it into new, correlated data. Our aim here is to generate random but correlated data with the covariance properties modeled in equation (11.4) and defined by our phylogenetic correlations in table 11.1. To start, let's first randomly generate and plot some independent (`ind`) and correlated (`cor`) *y*'s and *x* using the Cholesky method with this *R* script:

```
K <- 5; x_mean <- 0; x_variance <- 1; a <- 1.0; b <- 0.5
e_rand <- rnorm(K, 0, 1)
x_rand <- rnorm(K, 0, 1)                                                    (R.5)
# independent (uncorrelated) data following the I matrix
I <- diag(K) # creates identity matrix for OLS model
e_ind <- t(chol(I)) %*% e_rand # as modeled in eq. 11.1.3
x_ind <- x_mean + sqrt(x_variance) * t(chol(I)) %*% x_rand
y_ind <- a + b * x_ind + e_ind
# correlated data following the C matrix
C <- cov2cor(vcv(tree))
e_cor <- t(chol(C)) %*% e_rand # as modeled in eq. 11.1.4
x_cor <- x_mean + sqrt(x_variance) * t(chol(C)) %*% x_rand
y_cor <- a + b * x_cor + e_cor
# now organize three scatter-plots of these data
par(mfrow=c(1,3), xpd=TRUE);
plot(x_ind, y_ind, xlim=c(-2.5,2.5), ylim=c(-1.5,3.5), main="random
data")
text(x_ind, y_ind, tree$tip.label, cex=1, pos=1, font=4)
```

```
plot(x_ind, y_ind, xlim=c(-2.5,2.5), ylim=c(-1.5,3.5), main="phylo-
transformation")
text(x_ind, y_ind, tree$tip.label, cex=1, pos=1, font=4)
arrows(x_ind, y_ind, x_cor, y_cor, length=0.05)
plot(x_cor, y_cor, xlim=c(-2.5,2.5), ylim=c(-1.5,3.5),
main="transformed data")
text(x_cor, y_cor, tree$tip.label, cex=1, pos=1, font=4)
```

Figure 11.3 contains the *R* output of two plots where each data point is labeled by its species; the left panel has independent random data and the right panel has the same data but transformed via the correlation matrix (**C**). First note that the random data, once phylogenetically transformed, are now clustered more tightly among groups *a–b* and *c–d–e*. The phylogenetic transformation had the effect of making data more similar relative to their correlations. For example, *b* and *a* are now much closer together; they are no longer independent points and therefore have some similarity due to their shared phylogenetic history.

As an aside, note that the (x, y) positioning of species *b* and *e* remained the same in both data sets (see center panel in figure 11.3). This is a property of the way we transformed our random data phylogenetically. Our transformation method finds an upper triangular matrix (**U**) or Cholesky matrix that satisfies the condition $\mathbf{C} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$ (with the superscript T indicating the transposition of a matrix). Multiplying $\mathbf{U}^{\mathrm{T}}$ to a collection of random data will transform them following the correlations in **C**. However, if **C** is a proper correlation matrix, where diagonals all equal 1, and off-diagonals have correlations that range from zero to almost (but not) one, then the first (upper) element in the vector of transformed data will remain untransformed. In fact, what happens is that the phylogenetic transformation will rotate and shear the other data relative to this untransformed data point. In our case, because we have two independent groups in our correlation matrix (e.g., groups *a–b* and *c–d–e*), the transformation method will rotate and shear data relative to the way

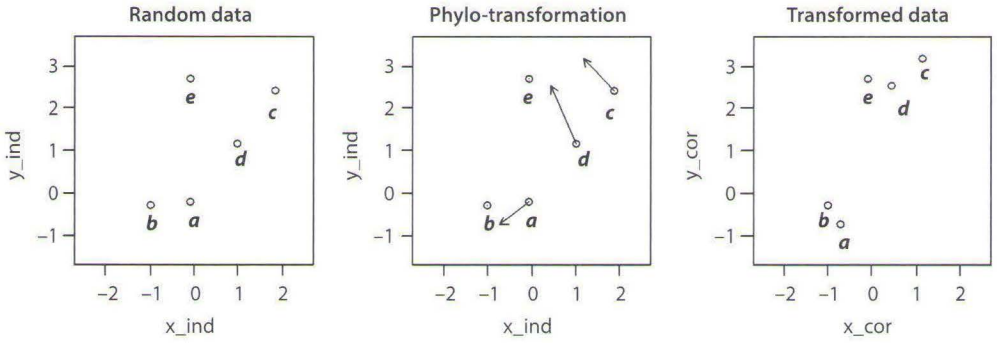


Fig. 11.3 The effects of phylogenetic correlations on randomly generated *x* and *y* data. The leftmost panel has randomly generated (independent) data for each species (*a*, *b*, *c*, *d*, *e*; see figure 11.2), the center panel depicts the direction of the phylogenetic transformation on these random data, and the rightmost panel shows the correlated random data after a phylogenetic transformation (based on **C**; see figure 11.2). The random data within groups *a–b* and *c–d–e* are sheared and rotated closer to one other because they belong to two independent groups of related species (see topology of the phylogeny in figure 11.2). The phylogenetic transformation was achieved using a Cholesky decomposition method, and the modeled relationship between *x* and *y* is defined in equation (11.1).

they are correlated with *b* and *e*. *But why species b and e?* The original correlation matrix calculated from the phylogeny with the `vcv` function (Paradis et al. 2004) actually had an order of *e*, *d*, *c*, *b*, *a*. We re-ordered this matrix as *a*, *b*, *c*, *d*, *e* to simplify comparisons with the phylogeny shown in the left of figure 11.2. The original un-ordered **C** had *e* and *b* occupying the first (upper) elements of each submatrix.

The random and correlated data in figure 11.3 provide a nice visualization of the effects of phylogenetic transformations. However, unless the predicted means of *y*'s and *x*'s for each species differ, or the means among groups *a*–*b* and *c*–*d*–*e* differ, then it is nearly impossible to predict how the phylogenetic transformation will position random data (especially with large *K*). This is because we are simulating all the *x*'s of each species to be centered around zero—what really gets affected by phylogenetic correlations is the residual error (ε) of these data, relative to *y*. For example, if we repeat R.5 30 times and plot these 30 data sets together, we can see the effects of random sampling and the unreliability of visually diagnosing phylogenetic effects in interspecific data. These results are in figure 11.4 (*R* script for this simulation is found in appendix 11.B with $N = 30$). Visually, we can barely see a positive correlation between *x* and *y*, and we can only see that because we modeled the relationship between the dependent (*y*) and predictor (*x*) variables to have a moderately strong slope (see equation (11.1)). To see why it is so difficult to visualize phylogenetic correlations in interspecific data sets, note that random data from species *a* can potentially occupy any part of that scatter plotted in figure 11.4.

Despite this large scatter in figure 11.4, the phylogenetic correlations do exist. In fact, we can recover the correlation matrix **C** and the means across all species for *x* and *y* quite

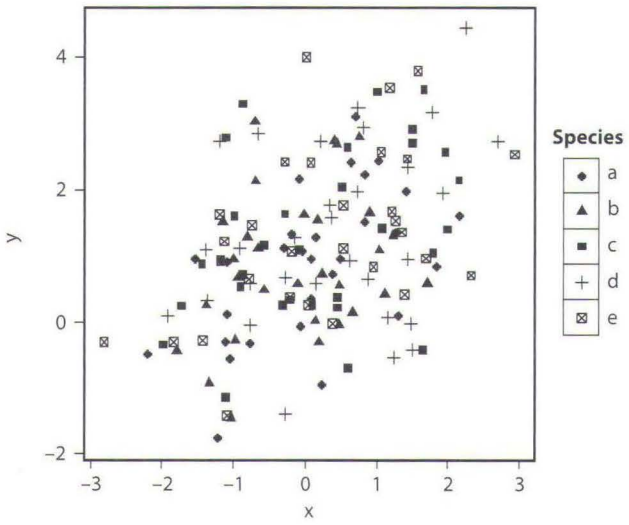


Fig. 11.4 A simulation on the unreliability of visualizing phylogenetic correlations in interspecific data. Here for each species (*a*, *b*, *c*, *d*, *e*; see figure 11.2), 30 *x* and *y* pairs were randomly generated using the Cholesky decomposition method (based on **C**, defined in table 11.1). This plot is equivalent to figure 11.2, but overlaid 30 times. Note that the random data for a single species can occupy nearly any region on the plot, and the only discernible pattern is the modeled relationship between *x* and *y* (defined in equation (11.1)). The R script for this simulation is in appendix 11.C.

easily in *R*. Again using the script found in appendix 11.B, but now with $N = 1,000$, and estimating the correlations between the random data generated for each species, we can recover the correlation matrix and means of each species for the x-variable; see table 11.2. With a minor modification to the script in appendix 11.B, we can also estimate the correlations and means of y (see table 11.2). The correlations are not perfect, but they are very close to **C** for both x and y (as in table 11.1). The means of each species are also very close—the x's of all species are near 0, and all y's are near 1. Had we simulated data with a larger N, our estimates would have converged to the expected **C** (table 11.1) and to our (expected) simulated means. This simulation emphasizes the importance of having precise species-level estimates of characteristics or traits for cross-species comparisons. Here, sampling error within species traits makes it harder for regression models to detect the underlying linear relationship between traits. However, only recently have comparative phylogenetic methods have been able to include within-species variation in regression analyses (Ives et al. 2007; Felsenstein 2008; Hansen and Bartoszek 2012).

11.2.4 *Why are phylogenetic correlations important for regression?*

Now let's return to our original regression model [equation (11.1)] and consider the case where y_i and x_i are two traits to be compared across multiple taxa—that is, the ith data point represents a characteristic from species i. In our previous simulation, the ith observation could be considered as N samples from a single species, but here let's use K rather than N to denote the total number of species analyzed. Again, N is the number of samples within species, and K is the number of species. Our goal is to repeat our previous simulation using OLS regression, but now with interspecific data—here we will assess the error rates of this regression model when the condition of independence assumed by ε is violated [equation (11.2)].

Let's start by analyzing our data set from figure 11.2 to assess how OLS and GLS models perform with our phylogenetically correlated data. Here are these data along with the GLS regression analysis including the phylogenetic correlations:

```
library(ape); library(nlme);
# raw phylogenetically correlated data from figure 11.3
x <- c(-0.07684503, 0.44569569, 1.15961757, -1.00146522, -0.71858873)
y <- c(2.7098214, 2.5464312, 3.1840059, -0.2871652, -0.7509973)
# using ape to load our Newick phylogeny; see Figure 11.2
tree <-
read.tree(text="(((e:0.16,d:0.16):0.05,c:0.21):0.90,
   (b:0.37,a:0.37):0.74);")                                    (R.6)
# The gls function of the nmle library requires a correlation
# matrix in the form of a corStruct object class.
VCV <- cov2cor(vcv(tree))
# convert matrix to corStruct object
C <- corSymm(VCV[lower.tri(VCV)], fixed=T)
# extract only coefficients
summary(gls(y ~ x, method="ML", correlation=C))$tTable
```

When a phylogenetic correlation matrix is included in a GLS model like this, it is commonly referred to as a phylogenetic generalized least squares (PGLS) regression (Martins

Table 11.2 Correlations (top) and means (bottom) for the *x* and *y* variables (left and right, respectively) estimated from random phylogenetically correlated data. Estimates are based on the script in appendix 11.B, using $N = 1,000$

Estimated correlation matrix for *x*.						Estimated correlation matrix for *y*.					
	a	b	c	d	e		a	b	c	d	e
a	**1.000**	**0.672**	-0.004	0.007	-0.002	a	**1.000**	**0.666**	-0.014	-0.012	-0.015
b	**0.672**	**1.000**	-0.009	0.001	-0.006	b	**0.666**	**1.000**	-0.022	-0.024	-0.022
c	-0.004	-0.009	**1.000**	**0.810**	**0.811**	c	-0.014	-0.022	**1.000**	**0.817**	**0.813**
d	0.007	0.001	**0.810**	**1.000**	**0.856**	d	-0.012	-0.024	**0.817**	**1.000**	**0.863**
e	-0.002	-0.006	**0.811**	**0.856**	**1.000**	e	-0.015	-0.022	**0.813**	**0.863**	**1.000**

Estimated means for *x*.						Estimated means for *y*.					
	a	b	c	d	e		a	b	c	d	e
x	-0.007	-0.012	0.011	0.000	0.004	**y**	0.996	0.997	0.999	1.001	1.00

and Hansen 1997; Pagel 1997, 1999; Garland et al. 1999). Our PGLS analysis estimated the following regression coefficients:

```
               Value Std.Error  t-value   p-value
(Intercept) 1.2424182 0.8446677 1.470896 0.2376872
x           0.7582988 0.5545423 1.367432 0.2649284
```

For comparison, let's also look at how a conventional OLS regression (without phylogenetic correlations) estimated the same coefficients (*R* script of regression without the phylogenetic correlation matrix: `gls(y ~ x, method="ML")`):

```
               Value Std.Error   t-value   p-value
(Intercept) 1.552139 0.4424145 3.508336 0.03924480
x           1.871734 0.5647376 3.314343 0.04524494
```

It may be useful to visualize these coefficients:

```
plot(x, y, xlim=c(-2.5,2.5), ylim=c(-1.5,3.5))
text(x, y, tree$tip.label, cex=0.7, pos=3, font=4)
abline(gls(y ~ x, method="ML")) # regression line from OLS
# regression line from PGLS
abline(gls(y ~ x, method="ML", correlation=C), lty=2)
legend(0.5, 0, c("OLS","PGLS"), cex=0.8, lty=1:2)
```

The results are shown in figure 11.5. The OLS regression line seems to have a much nicer fit to our species data than the PGLS model—it passes right through our simulated data (figure 11.5). The *t*-scores of the OLS estimate also concluded the slope and intercept to be non-zero (*p*–values are just above 0.04 for both estimates). In contrast, the regression line of the PGLS estimator does not look like a very robust fit (figure 11.5), and in fact, its slope and intercept were not significant ($p > 0.05$).

These regression results are counter-intuitive; OLS seems to provide a better fit than PGLS to the phylogenetically correlated data. However, contrasting the results from PGLS and OLS regressions underlines the importance of including phylogenetic correlations when analyzing interspecific data. Had we relied solely on the OLS regression, we would have concluded that there is a strong positive relationship between x and y. However, our PGLS analysis reveals that much of this relationship between x and y is due to their shared phylogenetic history—which is true given the way we phylogenetically transformed our data. Without the PGLS analysis, the findings of the OLS regression are at risk of making a Type II error (Diaz-Uriarte and Garland 1996; Harvey and Rambaut 1998). If we knew nothing about the underlying properties of our data, then we would have to conclude that there is no evidence for a positive linear relationship between x and y given our PGLS results. Although this is a conservative way to interpret results, it is appropriate given that only the PGLS analysis accounted for the potential phylogenetic correlations among species.

However, we simulated these data, and we know a positive relationship between x and y exists. *So what happened? Why was the OLS able to detect an effect while PGLS was not?* We do not typically have the luxury of knowing the true underlying relationships among species prior to analyses; however, our simulation approach provides us an opportunity to explore other explanations for why disparities among analyses exist. One explanation, which we will consider in section 11.2.5, is that we might have used an inappropriate model of evolution in our PGLS analysis. Another explanation, and the primary scourge of all analyses, is sampling error. It doesn't matter if you have the best phylogenetic hypothesis, or

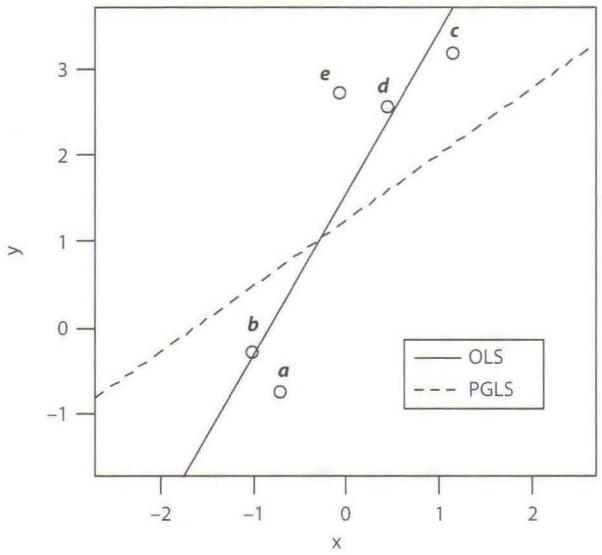


Fig. 11.5 Regression lines fit to randomly data generated with phylogenetic correlations (see figure 11.3). Regression lines were estimated with either ordinary least squares (OLS) or a phylogenetic generalized least squares (PGLS) model. These phylogenetic correlations were based on the **C** matrix from table 11.1.

the most precise trait data for your species—if you do not have enough data you will not be able to correctly estimate (with confidence) the underlying relationship (effect) with regression. In our case, this effect is the correlation (r) between x and y. Random sampling alone will generate data sets with strong positive or negative correlations—what we need is some assurance that our observed effect is true and did not emerge because of sampling error. We explored this issue of sampling error and low sample sizes previously with our simulations with conventional regression on independent data (section 11.2.1).

One way to assess the reliability of our regression analysis is to estimate its predicted false negative rate (Type II error rate, or β); that is, determine its probability of failing to detect a non-zero correlation. We can do this directly by first estimating our predicted effect, which is the expected correlation (r) between x and y, or more explicitly:

$$r = b\sqrt{\sigma_x^2/\sigma_y^2}. \tag{11.5}$$

In our simulation, $b = 0.5$ and $\sigma_x^2 = 1$ (see R.1). We also need to know the predicted variance of y, and given the way we modeled x and ε in equation (11.1), the predicted distribution of y is:

$$y_i \sim \mathrm{N}(a + bx + \varepsilon, b^2\sigma_x^2 + \sigma_\varepsilon^2) = \mathrm{N}(1, 1.25). \tag{11.6}$$

Thus y has a mean of 1 (i.e., $1 + 0.5 \times 0 + 0 = a + bx + \varepsilon$) and a variance of $\sigma_y^2 = b^2\sigma_x^2 + \sigma_\varepsilon^2 = 0.25 \times 1 + 1 = 1.25$. Given these values, the predicted correlation between x and y in our simulations is $0.447 \approx 0.5\sqrt{1/1.25} = r$. This is a large effect (Cohen 1988), and a strong relationship between x and y. Following our simulation conditions for x and y, we do not expect r to differ much between the raw and the phylogenetically transformed versions

of x and y (Garland and Ives 2000). Finally, using the tabulated estimates of statistical power (which equal $1-\beta$) reported by Cohen (1988), we find that regression analyses with sample sizes of $K = 5$ will have a Type II error rate of 92% for detecting a correlation of approximately 0.45.

Given this large error rate, interpreting any regression results with such a small sample size is very risky. In our case, it is impossible to determine whether our OLS regression detected the true underlying effect or found a strong positive effect because of sampling error. Examining the magnitude of the estimated slope ($a = 1.87$) from OLS when the true slope equals 0.5 may provide evidence for the latter. Likewise, our PGLS regression could not detect the true effect (although the slope and intercept were very near the predicted values of 1 and 0.5, respectively). This was because the variances of regression coefficients were too large. Again, these large variances are a consequence of small sample size. This is exactly how we want our regression analyses to behave, and why the PGLS model properly estimated the variances of our random data: these variances were broad, as predicted, given our small sample size. We do not want our variances to be biased, as can potentially occur with OLS, as these will increase our chances of making wrong conclusions with our data.

Let's explore the interaction between sample size and phylogenetic correlations in more detail, and compare the Type II error rates of OLS and PGLS by simulating interspecific data. This will allow us to assess the error rates of concluding that the slope and intercept are non-zero. The *R* script for this simulation is in appendix 11.C. Briefly, we generated a random phylogeny with 100 species, and then randomly subsampled this phylogeny to generate subtrees of size K. We then phylogenetically transformed K random data (following R.5), and analyzed these with both OLS and PGLS. Repeating this 1,000 times for each K, we counted the number of times the p-values of t-scores for the regression coefficients were not significant (i.e., concluding that they were zero). Recall that for our linear model, the intercept and slope *are* non-zero (section 11.2.3). figure 11.6 has the simulation results of the error rates from our two regression models. Generally, increasing the sample size improves the ability to detect non-zero effects. More notably, however, the real benefits of including phylogenetic correlations in GLS models (i.e., PGLS) only emerge at larger sample sizes—given that they have significant improved ability to detect non-zero effects relative to OLS.

This is typically as far as comparative analyses can take us, and the best we can glean from a simple PGLS regression is whether non-zero effects exist given our interspecific data. But with our simulated data, we know the true underlying effects, and so we can extend our Monte Carlo experiments to investigate how close OLS and PGLS were in estimating the correct intercept and slope of our linear model. In our previous simulation, statistical tests (t-scores) assessed whether there is any evidence that $a \neq 0$ and $b \neq 0$. Now we will adjust the null hypotheses of these tests to investigate whether $a \neq 1$ and $b \neq 0.5$, and count the number of cases when t-scores incorrectly rejected our simulated regression coefficients (i.e., $a = 1$ and $b = 0.5$). This type of error is referred to as false positive, or Type I error. The simulation results are in figure 11.7, and the *R* script in appendix 11.D. Note that the OLS regression has a fairly high probability of incorrectly concluding that the intercept and slope were different from $a = 1$ and $b = 0.5$, and that this probability increases with larger sample sizes. This is clear evidence that the OLS estimator is not optimal for analyzing interspecific data (Martins and Garland 1991; Diaz-Uriarte and Garland 1996; Harvey and Rambaut 1998; Freckleton et al. 2002; Revell 2010). These findings also counter the seemingly amazing ability for OLS to detect a non-zero intercept (see Type II errors in figure 11.5), since OLS analyses will likely estimate significant non-zero yet erroneous

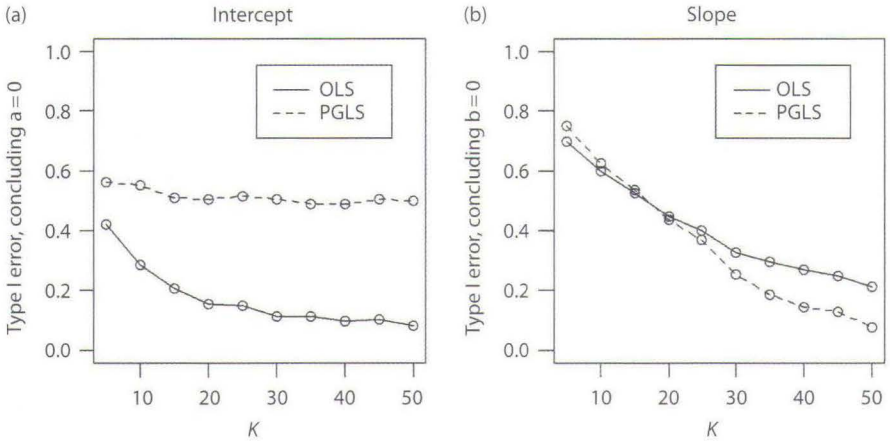


Fig. 11.6 The risk of incorrectly concluding null results with regression analyses of interspecific data. Presented are results from a Monte Carlo experiment exploring the Type I error rates (i.e., false positive outcomes) of OLS and PGLS regression with the number of species (*K*) varying from few to many. Error rates are based on the proportion of 1,000 regression analyses concluding that the intercept (*a*) and slope (*b*) were zero when data are phylogenetically correlated. PGLS analyses are more likely to correctly conclude that the slope is non-zero with larger *K*. The R script for this simulation is in appendix 11.D.

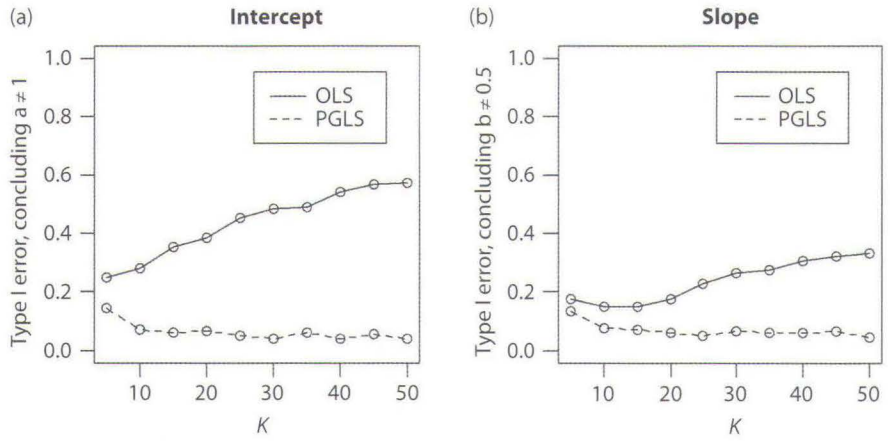


Fig. 11.7 The risk of concluding non-zero but erroneous intercept and slope estimates when using regression to analyze interspecific data. Presented are results from a Monte Carlo experiment exploring the Type I error rates (i.e., false positive outcomes) of OLS and PGLS regression with small to large number of species (*K*). Error rates are based on the proportion of 1,000 regression analyses concluding that the intercept (*a*) and slope (*b*) did not equal their true simulated values (i.e., $a = 0.5$ and $b = 1$). Data were simulated to have phylogenetic correlations. Here OLS analyses are more likely to incorrectly conclude significant erroneous intercept and slope values irrespective of K. The R script for this simulation is in appendix 11.D.

regression coefficients. There are also clearly issues with how PGLS estimates the intercept (see Type II errors in figure 11.5), but at least it tends to be conservative when estimating the intercept's standard error (i.e., it tends to be large). This favors the null hypothesis (see figures 11.5 and 11.6). More practically, however, evaluating whether the intercept is non-zero is typically not the focus of analyses. Generally the aim is to determine if the slope is non-zero, and PGLS seems optimal for this estimation goal with interspecific data. In fact, one of the original regression approaches to analyzing interspecific data excluded the intercept entirely from analyses (Felsenstein 1985).

11.2.5 *The assumption of homoscedasticity and evolutionary models*

Phylogenetic correlations arise because of the similarities between ancestors and their descendants, and we estimated these correlations using the pairwise phylogenetic distances between species (see table 11.1). Here, we assume that the strength of these correlations predict similarity among related taxa: the stronger the correlation, the greater the similarity of data measured between two taxa. Including these correlations in GLS models is meant to improve the way we model stochastic errors (ε) in linear regression [equation (11.4)]. However, when we apply phylogenetic correlations to GLS, we are also making an important assumption about the stochastic nature of evolution and how this process can shape variation in the characteristics of species (Martins and Hansen 1997).

For example, the way we model the residual error ε actually has an important biological interpretation regarding the variances of traits and how they are predicted to evolve along the branches of a phylogeny. Implicit in the way we quantified our phylogenetic correlations is a time component: we expect that the strength of correlations (and therefore also similarity among traits) will erode linearly with time as taxa evolve independently from a common ancestor. This type of stochastic erosion is called *Brownian motion evolution*—a model of evolutionary change where random genetic drift is the primary process resulting in the loss of similarity from ancestral characteristics (Martins and Garland 1991). As traits follow the paths along each branch of a phylogeny, random drift results in independent shifts of magnitude and direction in these characteristics, and the total change accrued is proportional to time (O'Meara et al. 2006).

Another way to think about Brownian motion (BM) evolution is that it is a hypothesis on the predicted distribution of characteristics among related species. With this in mind, we can interpret how we modeled the stochastic error (ε) in our linear model for interspecific data [equation (11.4)] as the expected variance and covariance that is proportional to shared phylogenetic history (i.e., $\sigma^2\mathbf{C}$). Here $\mathbf{C}$ (as defined earlier in R.4), but now more precisely $\mathbf{C}^{\mathrm{BM}}$ because we know now that it has a Brownian motion structure, quantifies the correlations among species based on the pattern and timing of their phylogenetic history. Further, σ^2 becomes the phylogenetic variance or evolutionary rate for x and y. This rate of change is an important property of BM evolution as it is a process that acts equally (i.e., has the same σ^2 rate) among the traits of evolving taxa. This satisfies the assumption of homogeneity of variances (homoscedasticity) of our GLS model as applied via PGLS.

Brownian motion is by far the most commonly assumed model of phenotypic evolution by comparative phylogenetic methods (see Edwards et al. 1963; Lynch 1991), and is the model explicitly assumed when Felsenstein's (1985) widely applied phylogenetically independent contrasts (PIC) are used to analyze interspecific data. Our PGLS analyses are a generalization of this PIC approach, as both will yield similar conclusions when assumptions of BM are met (Rohlf 2001). Interestingly, we can only recently say this

confidently now, as it took nearly 30 years after the introduction of PICs to develop a mathematical proof that PGLS and PICs were equivalent under certain conditions (see Blomberg et al. 2012). Essentially, whenever phylogenetic correlations are defined by **C** (see table 11.1), and are applied to regression analyses, then the evolutionary model is a Brownian motion process.

However, random drift through Brownian motion is a rather a simplistic view of how evolution can shape the covariances among related taxa. Other processes like natural selection, along with random drift, can work together to generate very different phylogenetic correlations (Hansen 1997; O'Meara 2012). Therefore there is always the risk that the phylogenetic correlations derived from Brownian motion will not adequately model the covariances of interspecific data. Let's explore this issue by comparing the performance of PGLS when the evolutionary covariance structure differs from BM evolution. Evolution via an *Ornstein–Uhlenbeck* (OU) process is another stochastic model of evolution that is increasingly being investigated by comparative biologists (Uhlenbeck and Ornstein 1930; Lande 1976; Martins and Hansen 1997). Under the OU model, stabilizing selection acts to keep phenotypes near an optimum by removing extreme values in characters. This process works in conjunction with random genetic drift to erode phylogenetic correlations among the phenotypes of related taxa, and the process of keeping phenotypes at an optimum is what erodes phylogenetic correlations. However, because of this added selection component, phylogenetic correlations are no longer predicted to decay proportionally with time as in BM, but instead decay exponentially (i.e., at a much quicker rate) as species become more distantly related (Hansen 1997).

To visualize how stabilizing selection effects the magnitude of phylogenetic correlations, we can simulate phylogenetic correlations derived from BM and OU processes, and compare how their rates of change differ relative to the same time since divergence. The predicted phylogenetic correlations under this exponential model of evolution are estimated as:

$$\mathbf{C}^{\mathrm{OU}} = 1/2\alpha \left\{ e^{-2\alpha\left[\mathrm{diag}(\mathbf{C}^{\mathrm{BM}})-\mathbf{C}^{\mathrm{BM}}\right]} - e^{-2\alpha\left[\mathrm{diag}(\mathbf{C}^{\mathrm{BM}})\right]} \right\}, \tag{11.7}$$

where α is the stabilizing selection parameter that can range from zero (no selection) to infinity (very high selection), and where "diag" indicates a vector containing only the main diagonals of $\mathbf{C}^{\mathrm{BM}}$. By manipulating the strength of stabilizing selection (α) in $\mathbf{C}^{\mathrm{OU}}$, we can visualize the effects of selection eroding phylogenetic correlations by simulating a random tree and plotting the divergence time versus the correlations found in $\mathbf{C}^{\mathrm{BM}}$ and $\mathbf{C}^{\mathrm{OU}}$. The script for this simulation is in appendix 11.E, and the phylogenetic correlations derived from these two models of phenotypic evolution are shown in figure 11.8. As expected under BM, the phylogenetic correlations are linearly proportional with the time since divergence (figure 11.8); they form a straight line between the time of divergence (i.e., shared phylogenetic branch-length distance) and the correlations. However, under the OU model, increasing intensity of stabilizing selection (i.e., larger values of α), the magnitudes of correlations erode exponentially. Taxa far apart in a phylogeny quickly achieve evolutionary independence relative to those under BM. When α is near 0, the phylogenetic correlations of an OU model ($\mathbf{C}^{\mathrm{OU}}$) are equivalent to $\mathbf{C}^{\mathrm{BM}}$. However, as selection (α) increases in intensity, $\mathbf{C}^{\mathrm{OU}}$ approaches $\mathbf{I}$; species become nearly independent (section 11.2.2; also see Lajeunesse 2009). In the latter case, selection is so strong that it quickly erases all phylogenetic correlations among related taxa.

11.2.6 *What happens when the incorrect model of evolution is assumed?*

With the predicted phylogenetic correlations $\mathbf{C}^{OU}$ and $\mathbf{C}^{BM}$ described above, we can simulate interspecific data derived from these two models of phenotypic evolution, and then compare the performance of our PGLS analyses using interspecific data that do not fit the BM model of evolution. Thus, we will assess the Type II error rates of PGLS analyses assuming a variance–covariance structure of ε based on BM evolution. This will provide some insight as how PGLS performs when an incorrect model of evolution is assumed with interspecific data. We simulated interspecific data evolving via an OU model with strong selection ($\alpha = 3$; see figure 11.8) using the *R* script in appendix 11.E. We then analyzed these data using PGLS assuming BM evolution [equation (11.6)]. The results (figure 11.9), revealed a slight loss of efficiency when estimating the slope with a *PGLS* assuming *BM* with data simulated under an *OU* model. There is also a lack of efficiency when estimating the intercept. In terms of the slope parameter, this may seem like a trivial amount of error when the incorrect model of evolution is applied to interspecific data. However, these results more likely reflect the relative similarity between BM and OU models, rather an apparent robustness of BM when analyzing data from a different evolutionary model. For example, even though the OU data were modeled with strong stabilizing selection ($\alpha = 3$), these two models still preserve groups of correlations based on the topology of the tree; such as within the two major groups a–b and c–d–e described earlier (although at smaller magnitudes; figure 11.4). Assuming BM under our simulation conditions provides (albeit

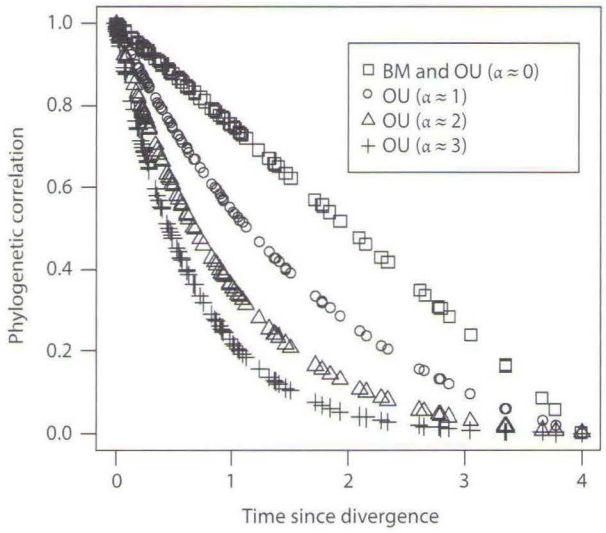


Fig. 11.8 The change in magnitude among phylogenetic correlations when they are based on different models of evolution. Brownian motion (BM) assumes a linear decay of correlations with time; whereas the Ornstein–Uhlenbeck (OU) model assumes an increasingly exponential decay with rising intensities of stabilizing selection, $\alpha = \{1, 2, 3\}$. Note that when stabilizing selection is near zero, the OU model converges to a BM model. The R script of this simulation is in appendix 11.E.

very coarsely) some useful correlational structure to assist the linear regression with *OU* data. Nonetheless, the potential risk of incorrectly concluding a null result still exists, and fitting the appropriate model of evolution to your data will help minimize this risk—even if in our case with simulated data the improvement was only about 5%.

11.3 Establishing confidence with the comparative phylogenetic method

Using Monte Carlo experiments, we were able to investigate the challenges of analyzing interspecific data and ask how applications of the comparative phylogenetic method can improve inferences with these data. An advantage of our simulation approach is that we knew the underlying relationships in our simulated data. With real (observed) interspecific data, little to no information will be known with any certainty about such underlying processes. This suggests that one should make a great deal of effort to approach interspecific data with a robust statistical framework, and to present results in a way that provides confidence that the observed relationships are biologically meaningful and not statistical artifacts. Below, we sketch a few guidelines on how to approach and present your comparative analyses to achieve these goals (for more extensive guidelines see also Garland et al. 2005; Freckleton 2009).

Nearly all comparative phylogenetic methods assume that the phylogenies used to estimate phylogenetic correlations are known without error (Rohlf 2001). However, phylogenies are only statistical hypotheses on the evolutionary history of taxa. They vary

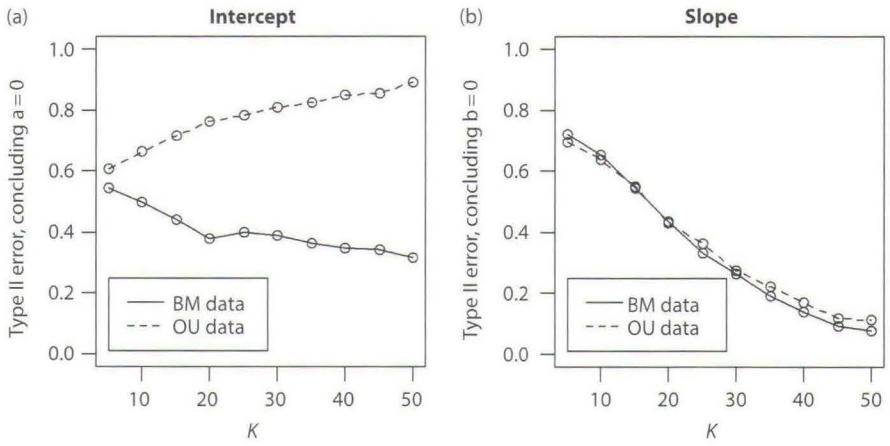


Fig. 11.9 The risk of concluding null results when PGLS assumes the incorrect model of evolution. Presented are results from a Monte Carlo experiment exploring the Type II error rates (i.e., false negative outcomes) of PGLS with the number of species (*K*) varying from few to many. Error rates are based on the proportion of 1,000 regression analyses incorrectly concluding that the intercept (*a*) and slope (*b*) were zero. Data were simulated to have phylogenetic correlations derived from an Ornstein–Uhlenbeck (OU) model of evolution with a stabilizing selection parameter set to $\alpha = 3$; these data were analyzed with a PGLS assuming a model of Brownian motion (BM) evolution. The *R* script for this simulation is in appendix 11.F.

tremendously in availability and uncertainty for distinguishing both deep and recent divergences, as well as their relative timing (Felsenstein 2004). Our phylogenetic correlation matrix (**C**) is at best a hypothesis of the expected true correlations (i.e., variance–covariances) that may or may not exist among the traits of related taxa. It is therefore always important to ask how uncertainty in **C** can influence the performance and statistical outcomes of PGLS analyses. Incorrectly specifying **C**, either by using an incorrect tree topology or an incorrect model of evolution, can result in PGLS models performing more poorly than OLS models (Mittelhammer et al. 2000). Deep topological errors near the root of the tree are also expected to have more strongly negative effects on the performance of PGLS, compared with errors in the positioning of nodes near the tips of phylogenetic trees (Martins and Housworth 2002). Therefore, if multiple (alternative) hypotheses of divergences are available for a collection of taxa, it is good practice to incorporate each of these phylogenetic hypotheses in PGLS analyses (see Donoghue and Ackerly 1996). Doing so allows one to compare regression results based on alternative phylogenetic hypotheses if they are biologically meaningful. Alternatively, multiple separate regression analyses can be averaged to provide an aggregate view based on different phylogenetic hypotheses. Model selection criteria (e.g., AIC scores) may also be useful in assessing the relative fit of competing phylogenetic hypotheses (Lajeunesse et al. 2013).

A more common challenge with phylogenetic trees is a lack of information needed to connect the divergences among taxa. Complete phylogenetic information (e.g., a phylogenetic tree that is completely bifurcated) can help minimize the Type I error rates of comparative analyses (see Purvis and Garland 1993). Several solutions to this problem of missing topologies within trees are available. For example, a sophisticated approach applies birth–death models to simulate random divergences among taxa with missing phylogenetic information (Kuhn et al. 2011). This imputation approach (see chapter 4) is not too different from the way we simulated our random phylogenic tree (section 11.2.3). The aim of these imputations are to fill gaps of information about the topology (and therefore correlations in **C**) by randomly resolving polytomies (nodes that specify unresolved divergences among lineages or taxa (Maddison 1989). Models of evolution can also be assumed to make the internode branch-length distances (i.e., simulated divergence times) less arbitrary (Kuhn et al. 2011). Analyses are then repeated several times with these randomly resolved topologies to minimize the risk of the method itself introducing bias to PGLS results. Alternatively, a coarse hypothesis on phylogenetic history, such as estimating **C** with a tree based on Linnaean rankings (e.g., grouped by class or order), can also help improve the performance of PGLS models, as long as the overall topology is correct and matches the true major divergence events (e.g., Freckleton et al. 2002). The disadvantage of this coarse approach is a lack of information about relative divergence times; these are useful for making predictions regarding the evolutionary basis for phenotypic change and their predicted phylogenetic correlations (i.e., **c**). Several online resources are also available that can help supplement phylogenetic information to generate fully bifurcated trees. For example, the widely used *phylomatic* by Webb and Donoghue (2005) is an important tool for generating phylogenetic trees for PGLS. The massive tree of life project called *timetree* by Hedges et al. (2006) is very helpful for determining the divergence time between distantly related taxonomic groups.

Another problem with PGLS analyses is the risk of over-fitting phylogenetic correlations to data that are weakly or not phylogenetically correlated. One common solution to this problem is to apply a transformation parameter such as Pagel's λ to the phylogenetic correlations in **C** (Pagel 1999). The idea is to make the correlation structure of analyses flexible relative to the observed phylogenetic signal (λ) of the interspecific data. A phylogenetic

signal is a measure that quantifies the overall statistical dependence among species traits, relative to their phylogenetic relationships. Applying this transformation is meant to relax the assumption of Brownian motion as the primary evolutionary model for phenotypic change, and therefore help minimize the potential of over-specifying **C** for interspecific data that are not actually phylogenetically correlated (Garland et al. 2005). For example, λ can first be estimated via maximum likelihood with a PGLS model, and then λ multiplies all the off-diagonals of **C** (i.e., all the correlations). If λ is estimated to be near zero, then $\lambda\mathbf{C}$ approaches **I**. Therefore, when no phylogenetic signal is detected, the PGLS will converge to OLS, which is more efficient at estimating regression coefficients if the data are independent (Revell 2010). Likewise, other evolutionary models can be used to adjust phylogenetic correlations following different models of evolution (e.g., the selection parameter of the OU model; Hansen 1997). It is also important to note that the accuracy of estimating evolutionary parameters like Pagel's λ is largely dependent on the number of species included in the analysis. Generally, phylogenies with fewer than 30 species will provide unreliable estimates of λ (Revell 2010).

11.4 Conclusions

By focusing solely on simple linear regression and Monte Carlo experiments, we hope that this chapter provides some clarity to why it is important to apply this statistical framework to interspecific data. However, it is important to note that the same statistical issues and interpretive problems outlined here are equally relevant to *any* analyses using phylogenetic correlations to model dependencies in interspecific data. These include more elaborate phylogenetic analogues such as GLS modeling to perform ANOVA or ANCOVA, principle component analysis (Revell 2009), and meta-analysis (Lajeunesse et al. 2013). Finally, we urge readers interested in applying these methods to think beyond treating phylogenetic correlations as a nuisance to be controlled in analyses. Phylogenetic dependence in our data is not just another pitfall to avoid, like pseudoreplication. Much more can be gleaned from these analyses if one adopts an evolutionary framework and compares multiple evolutionary models (e.g., BM vs. OU) with an aim of providing insight into how and why phenotypic and ecological data are phylogenetically correlated (Butler and King 2004). Dobzhansky (1973) famously commented that "Nothing in biology makes sense except in the light of evolution;" this also applies to ecological problems!

Acknowledgments

Support for this work was funded by the College of Arts and Sciences, University of South Florida.

References

Blomquist, N. S. 1980. A note on the use of the coefficient of determination. *Scandinavian Journal of Economics* 82: 409–412.

Blomberg, S. P., and T. Garland. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology* 15:899–910.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* 164:683–695.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Diaz-Uriarte, R., and T. Garland. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Systematic Biology* 45:27–47.

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* 126:1–25.

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Mass.

Felsenstein, J. 2008. Comparative methods with sampling error and within species variation: contrasts revisited and revised. *American Naturalist* 171:713–725.

Freckleton, R. P., and P. H. Harvey. 2006. Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biology* 4:2104–2111.

Freckleton, R. P., Cooper, N. and W. Jetz. 2011. Comparative methods as a statistical fix: the danger of ignoring an evolutionary model. American Naturalist 178: E10–E17.

Freckleton, R. P., Harvey, P. H., and M. Pagel. 2002. Phylogenetic analysis of comparative data: a test and review of evidence. *American Naturalist* 160:712–726.

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

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

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

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

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

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

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

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Vol. 1. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, Orlando, FL.

Hedges, S.B., Dudley, J. and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.

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

Kuhn, T.S., Mooers, A.Ø., and G.H. Thomas. 2011. A simple polytomy resolver for dated phylogenies. *Methods in Ecology and Evolution* 2:427–436.

Lajeunesse, M. J. 2009. Meta-analysis and the comparative phylogenetic method. *American Naturalist* 174:369−381.

Lajeunesse, M.J., Rosenberg, M.R. and Jennions, M.D. 2013. Phylogenetic nonindependence and meta-analysis. In J. Koricheva, J. Gurevitch, and K. Mengersen, editors. *Handbook of meta-analysis in ecology and evolution* (pp. 284−299). Princeton University Press, Princeton, New Jersey, USA.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30:314–334.

Martins, E. P. 1996. *Phylogenies and the comparative method in animal behavior.* Oxford University Press.

Martins, E. P. 2000. Adaptation and the comparative method. *Trends in Ecology & Evolution* 15:296–299.

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

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

Martins, E. P., Diniz-Filho, J. A. F. and E. A. Housworth. 2002. Adaptive Constraints and the Phylogenetic Comparative Method: A Computer Simulation Test. *Evolution* 56:1–13.

Nunn, C. L. 2011. *The comparative approach in evolutionary anthropology and biology*. University of Chicago Press.

O'Meara, B. C. 2012. Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics* 43:267–285.

Pagel, M. 1993. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. *Journal of Theoretical Biology* 164:191–205.

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

Pagel, M. D., and P. H. Harvey. 1989. Comparative methods for examining adaptation depend on evolutionary models. *Folia Primatologica* 53:203–220.

Paradis, E., Claude, J., and K. Strimmer. 2004. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Paradis, E. 2012. *Analysis of Phylogenetics and Evolution with R.* Springer.

Pinheiro, J., Bates, D., DebRoy, and S., Sarkar, D. 2004. NLME: Linear and nonlinear mixed effects models. R package version 3.1-53.

Price, T. 1997. Correlated evolution and independent contrasts. *Philosophical Transactions of the Royal Society of London B*, Biological Sciences 352:519–529.

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

Rubinstein, R. Y. and D. K. Kroese 2008. Simulation and the Monte Carlo Method, 2nd Ed. John Wiley & Sons.

Schluter, D. 2000. The ecology of adaptive radiations. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford.

Stuart, A., and J. K. Ord. 1994. Kendall's advanced theory of statistics. Volume 2A: Classical inference and the linear model. Griffin, London, UK.

Uhlenbeck, G. E., and L. S. Ornstein. 1930. On the theory of Brownian motion. *Physical Review* 36:823–841.

Westoby, M., M. R. Leishman, and J. M. Lord. 1995. On misinterpreting the "phylogenetic correction." *Journal of Ecology* 83:531–534.

```
# Appendix 11.A: The risk of concluding null results when few data are used
#                in regression analyses.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for a Monte Carlo experiment exploring the Type II error
# rates (i.e. false negative outcomes) of ordinary least-square ("OLS" ) regression with
# small to large sample sizes (N).  Error rates are based on the proportion of 10,000 regression
# analyses incorrectly concluding that the intercept (a) and slope (b) were zero.
# Example of simulation results are found in Figure 11.1.
#
# Inputs: none
# Outputs: a_error_II (vector of Type II error rates of intercept [a] with increasing N)
#          b_error_II (vector of Type II error rates of regression slope [b] with increasing N)


library(utils); library(nlme);

s <- 1000; s_pb <- 0;
pb <- txtProgressBar(min=1, max=(s * 10), style=3)
x_mean <- 0; x_variance <- 1; a <- 1.0; b <- 0.5;
for(i_N in seq(5, 50, by=5)) {
        a_null <- 0; b_null <- 0;
        for(i_s in 1:s) {
                e <- rnorm(i_N, 0, 1)
                x <- x_mean + sqrt(x_variance) * rnorm(i_N, 0, 1)
                y <- a + b * x + e
                OLS_results <- summary(gls(y ~ x, method="ML"))
                if(OLS_results$tTable[1,4] > 0.05) a_null <- a_null + 1
                if(OLS_results$tTable[2,4] > 0.05) b_null <- b_null + 1
                setTxtProgressBar(pb, i_s + s_pb)
        }
        s_pb <- s_pb + i_s
        if (i_N == 5) { a_error_II <- a_null / s; b_error_II <- b_null / s; N <- i_N; }
        else {
                a_error_II <- c(a_error_II, a_null / s)
                b_error_II <- c(b_error_II, b_null / s)
                N <- c(N, i_N)
        }
}

par(mfrow=c(1,2)); x_label <- bquote(italic(N));
y_a_label <- "Type II error, concluding a = 0"; y_b_label <- "Type II error, concluding b = 0";
plot(N, a_error_II, ylim=c(0,1), type="o", xlab=x_label, ylab=y_a_label, main="intercept (a)")
plot(N, b_error_II, ylim=c(0,1), type="o", xlab=x_label, ylab=y_b_label, main="slope (b)")
```

```
# Appendix 11.B: A simulation on the unreliability of visualizing phylogenetic
#                correlations in interspecific data.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for a simulation where for each species (a, b, c, d, e)
# thirty x and y pairs were randomly generated using the Cholesky decomposition
# method (based on C; see Chapter 11).  This plot is equivalent to Figure 11.2,
# but overlaid 10000 times.  Note that the random data for a single species can occupy
# nearly any region on the plot, and the only discernible pattern is the modeled
# relationship between x and y (defined in equation 11.1).
#
# Inputs: none
# Outputs: Two correlation matrices, with one estimating the correlations between species
#          across the x variable of the regression, and the other matrix among the y
#          variable.  Also outputed are the estimated means of the y variable across species.


library(ape); library(ggplot2); library(utils);

# Generate correlated data using the Cholesky decomposition method
# using a phylogenetic correlation matrix extracted from the NEWICK tree below.
tree <- read.tree(text="(((e:0.16,d:0.16):0.05,c:0.21):0.90,(b:0.37,a:0.37):0.74);")
C.Cholesky <- t(chol(cov2cor(vcv(tree))))
N <- 10000; K <- 5; a <- 1.0; b <- 0.5;
pb <- txtProgressBar(min=1, max=N, style=3)
for(i_N in 1:N) {
        e_cor <- C.Cholesky %*% rnorm(K, 0, 1)
        x_cor <- C.Cholesky %*% rnorm(K, 0, 1)
        y_cor <- a + b * x_cor + e_cor
        if(i_N == 1) { x <- x_cor; y <- y_cor; species <- tree$tip.label; }
        else { x <- c(x, x_cor); y <- c(y, y_cor); species  <- c(species, tree$tip.label); }
        setTxtProgressBar(pb, i_N)
}

# Plot the scatter of the phylogenetically correlated random samples.
data_xy <- data.frame(x, y, species)
ggplot(data_xy, aes(x, y, shape=species)) +
        geom_point() + theme_bw() +
        theme(panel.grid.major=element_blank(), panel.grid.minor=element_blank() )

# Estimate the means and correlations from the random data generated above.
# (Note the script below only estimates these for x).
for(i_sp in tree$tip.label) {
        x_sub <- subset(data_xy$x, data_xy$species == i_sp)
        y_sub <- subset(data_xy$y, data_xy$species == i_sp)
        if(i_sp == "e") { x_new <- data.frame(x_sub); y_new <- data.frame(y_sub); }
        else { x_new <- data.frame(x_new, x_sub); y_new <- data.frame(y_new, y_sub); }
}
colnames(x_new) <- tree$tip.label; colnames(y_new) <- tree$tip.label;
order <- paste(letters[1:K], sep="") # replaces names with letters
round(cor(x_new)[order, order], 3) # estimate correlations between the species x's
round(colMeans(x_new), 3) # estimate means between species
round(cor(y_new)[order, order], 3) # estimate correlations between the species y's
round(colMeans(y_new), 3) # estimate means between species
```

```
# Appendix 11.C: The risk of incorrectly concluding null results with
#                regression analyses of interspecific data.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for a Monte Carlo experiment exploring Type I error rates
# (i.e. false positive outcomes) of "OLS"  and "PGLS"  regression with small
# to large number of species (K).  Error rates are based on the proportion
# of 10,000 regression analyses concluding that the intercept (a) and slope (b)
# were zero when data are phylogenetically correlated.
#
# Inputs: none
# Outputs: a_OLS (vector of Type I errors in intercept estimation with ordinary-regressions & increasing K)
#          a_PGLS (vector of Type I errors in intercept estimation with phylogenetic-regressions & increasing K)
#          b_OLS (vector of Type I errors in slope estimation with ordinary-regressions & increasing K)
#          b_PGLS (vector of Type I errors in slope estimation with phylogenetic-regressions & increasing K)



library(ape); library(geiger); library(utils); library(nlme);

s <- 1000; s_pb <- 0;
K <- 100; x_mean <- 0; x_variance <- 1; a <- 1.0; b <- 0.5; a_NULL <- 0.0; b_NULL <- 0.0;
pb <- txtProgressBar(min=1, max=(s * 10), style=3)
tree <- sim.bdtree(b=1, d=0, stop="taxa", K)
for(i_N in seq(5, 50, by=5)) {
        a_OLS_error <- 0; b_OLS_error <- 0; a_PGLS_error <- 0; b_PGLS_error <- 0;
        for(i_s in 1:s) {
                subtree <- drop.tip(tree, sample(1:K, K - i_N))
                e <- t(chol(cov2cor(vcv(subtree)))) %*% rnorm(i_N, 0, 1)
                x <- x_mean + sqrt(x_variance) * t(chol(cov2cor(vcv(subtree)))) %*% rnorm(i_N, 0, 1)
                y <- a + b * x + e
                OLS_results <- summary(gls(y ~ x, method="ML"))
                        GLS_a_t <- (OLS_results$tTable[1,1] - a_NULL) / OLS_results$tTable[1,2]
                        if(2 * pt(-abs(GLS_a_t), df=K-2) > 0.05) a_OLS_error <- a_OLS_error + 1
                        GLS_b_t <- (OLS_results$tTable[2,1] - b_NULL) / OLS_results$tTable[2,2]
                        if(2 * pt(-abs(GLS_b_t), df=K-2) > 0.05) b_OLS_error  <- b_OLS_error + 1
                PGLS_results <- summary(gls(y ~ x, correlation=corBrownian(phy=subtree), method="ML"))
                        PGLS_a_t <- (PGLS_results$tTable[1,1] - a_NULL) / PGLS_results$tTable[1,2]
                        if(2 * pt(-abs(PGLS_a_t), df=K-2) > 0.05) a_PGLS_error <- a_PGLS_error + 1
                        PGLS_b_t <- (PGLS_results$tTable[2,1] - b_NULL) / PGLS_results$tTable[2,2]
                        if(2 * pt(-abs(PGLS_b_t), df=K-2) > 0.05) b_PGLS_error <- b_PGLS_error + 1
                setTxtProgressBar(pb, i_s + s_pb)
        }
        s_pb <- s_pb + i_s
        if (i_N == 5) {
                a_OLS <- a_OLS_error / s; b_OLS <- b_OLS_error / s;
                a_PGLS <- a_PGLS_error / s; b_PGLS <- b_PGLS_error / s;
                N <- i_N;
        }
        else {
                a_OLS <- c(a_OLS, a_OLS_error / s); b_OLS <- c(b_OLS, b_OLS_error / s);
                a_PGLS <- c(a_PGLS, a_PGLS_error / s); b_PGLS <- c(b_PGLS, b_PGLS_error / s);
                N <- c(N, i_N)
        }
}

par(mfrow=c(1,2)); x_name <- bquote(italic(K));
y_a_name <- bquote(paste("Type I error, concluding   ", a==.(a_NULL)))
y_b_name <- bquote(paste("Type I error, concluding   ", b==.(b_NULL)))
plot(N, a_OLS, ylim=c(0,1), type="o", xlab=x_name, ylab=y_a_name, main="intercept (a)")
lines(N, a_PGLS, type="o", lwd=1.5, lty=2)
legend(27, 0.95, c("OLS","PGLS"), cex=0.8, lty=1:2)
plot(N, b_OLS, ylim=c(0,1), type="o",  xlab=x_name, ylab=y_b_name, main="slope (b)")
lines(N, b_PGLS, type="o", lwd=1.5, lty=2)
legend(27, 0.95, c("OLS","PGLS"), cex=0.8, lty=1:2)
```

```
# Appendix 11.D: The risk of concluding non-zero but erroneous intercept and
#                slope estimates when using regression to analyze interspecific data.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for a Monte Carlo experiment experiment the Type I error rates
# (i.e. false positive outcomes) of "OLS"  and "PGLS"  regression with small to large
# number of species (K).  Error rates are based on the proportion of 10,000
# regression analyses concluding that the intercept (a) and slope (b) did not
# equal their true simulated values (i.e. a=0.5 and b=1).  Data were simulated
# to have phylogenetic correlations.
#
# Inputs: none
# Outputs: a_OLS (vector of Type I errors in intercept estimation with ordinary-regressions & increasing K)
#          a_PGLS (vector of Type I errors in intercept estimation with phylogenetic-regressions & increasing K)
#          b_OLS (vector of Type I errors in slope estimation with ordinary-regressions & increasing K)
#          b_PGLS (vector of Type I errors in slope estimation with phylogenetic-regressions & increasing K)


library(ape); library(geiger); library(utils); library(nlme);

s <- 1000; s_pb <- 0;
K <- 100; x_mean <- 0; x_variance <- 1; a <- 1.0; b <- 0.5; a_NULL <- 1.0; b_NULL <- 0.5;
pb <- txtProgressBar(min=1, max=(s * 10), style=3)
tree <- sim.bdtree(b=1, d=0, stop="taxa", K)
for(i_N in seq(5, 50, by=5)) {
        a_OLS_error <- 0; b_OLS_error <- 0; a_PGLS_error <- 0; b_PGLS_error <- 0;
        for(i_s in 1:s) {
                subtree <- drop.tip(tree, sample(1:K, K - i_N))
                e <- t(chol(cov2cor(vcv(subtree)))) %*% rnorm(i_N, 0, 1)
                x <- x_mean + sqrt(x_variance) * t(chol(cov2cor(vcv(subtree)))) %*% rnorm(i_N, 0, 1)
                y <- a + b * x + e
                OLS_results <- summary(gls(y ~ x, method="ML"))
                        GLS_a_t <- (OLS_results$tTable[1,1] - a_NULL) / OLS_results$tTable[1,2]
                        if(2 * pt(-abs(GLS_a_t), df=K-2) <= 0.05) a_OLS_error <- a_OLS_error + 1
                        GLS_b_t <- (OLS_results$tTable[2,1] - b_NULL) / OLS_results$tTable[2,2]
                        if(2 * pt(-abs(GLS_b_t), df=K-2) <= 0.05) b_OLS_error  <- b_OLS_error + 1
                PGLS_results <- summary(gls(y ~ x, correlation=corBrownian(phy=subtree), method="ML"))
                        PGLS_a_t <- (PGLS_results$tTable[1,1] - a_NULL) / PGLS_results$tTable[1,2]
                        if(2 * pt(-abs(PGLS_a_t), df=K-2) <= 0.05) a_PGLS_error <- a_PGLS_error + 1
                        PGLS_b_t <- (PGLS_results$tTable[2,1] - b_NULL) / PGLS_results$tTable[2,2]
                        if(2 * pt(-abs(PGLS_b_t), df=K-2) <= 0.05) b_PGLS_error <- b_PGLS_error + 1
                setTxtProgressBar(pb, i_s + s_pb)
        }
        s_pb <- s_pb + i_s
        if (i_N == 5) {
                a_OLS <- a_OLS_error / s; b_OLS <- b_OLS_error / s;
                a_PGLS <- a_PGLS_error / s; b_PGLS <- b_PGLS_error / s;
                N <- i_N;
        }
        else {
                a_OLS <- c(a_OLS, a_OLS_error / s); b_OLS <- c(b_OLS, b_OLS_error / s);
                a_PGLS <- c(a_PGLS, a_PGLS_error / s); b_PGLS <- c(b_PGLS, b_PGLS_error / s);
                N <- c(N, i_N)
        }
}

par(mfrow=c(1,2)); x_name <- bquote(italic(K));
y_a_name <- bquote(paste("Type I error, concluding   ", a!=.(a_NULL)))
y_b_name <- bquote(paste("Type I error, concluding   ", b!=.(b_NULL)))
plot(N, a_OLS, ylim=c(0,1), type="o", xlab=x_name, ylab=y_a_name, main="intercept (a)")
lines(N, a_PGLS, type="o", lwd=1.5, lty=2)
legend(27, 0.95, c("OLS","PGLS"), cex=0.8, lty=1:2)
plot(N, b_OLS, ylim=c(0,1), type="o",  xlab=x_name, ylab=y_b_name, main="slope (b)")
lines(N, b_PGLS, type="o", lwd=1.5, lty=2)
legend(27, 0.95, c("OLS","PGLS"), cex=0.8, lty=1:2)
```

```
# Appendix 11.E: The change in magnitude among phylogenetic correlations
#                when they are based on different models of evolution.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for simulating phylogenetic correlations based
# on Brownian motion (BM) and Ornstein–Uhlenbeck (OU) models of evolution.
# Brownian motion assumes a linear decay of correlations with
# time; whereas the Ornstein–Uhlenbeck model assumes an increasingly exponential
# decay with rising intensities of stabilizing selection: alpha={1,2,3}.  Note that
# when stabilizing selection is near zero, the OU model converges to a BM model.
#
# Inputs: none
# Outputs: BM (vector of random BM phylogenetic correlations)
#          OU_1 (vector of the same random phylogenetic correlations assuming OU with alpha=1)
#          OU_2 (vector of the same random phylogenetic correlations assuming OU with alpha=2)
#          OU_3 (vector of the same random phylogenetic correlations assuming OU with alpha=3)


library(ape); library(geiger);

K <- 100; plot_symbol <- 0;
x_name <- "Time Since Divergence"; y_name <- "Phylogenetic Correlation"
tree <- sim.bdtree(b=1, d=0, stop="taxa", K)
for(alpha in seq(0, 3, by=1)) {
        raw_BL <- diag(vcv(tree)) - vcv(tree)
        C_BM <- cov2cor(vcv(tree))
        C_OU <- (exp(-2 * alpha * (diag(C_BM) - C_BM)) - exp(-2 * alpha * diag(C_BM)))/(2 * alpha)
        if(alpha == 0) { plot(raw_BL, C_BM, xlab=x_name, ylab=y_name, pch=plot_symbol); }
        else { points(raw_BL, C_OU * (1 / C_OU[1,1]), pch=plot_symbol); }
        plot_symbol <- plot_symbol + 1
}

BM <- bquote(paste("BM and OU ", (alpha%~~%0)))
OU_1 <- bquote(paste("OU ", (alpha==1)))
OU_2 <- bquote(paste("OU ", (alpha==2)))
OU_3 <- bquote(paste("OU ", (alpha==3)))
models <- c(BM, OU_1, OU_2, OU_3)
legend(vcv(tree)[1,1] * 0.5, 0.95, sapply(models, as.expression), cex=0.8, pch=0:4)
```

```
# Appendix 11.F: The risk of concluding null results when phylogenetic regressions
#                assume the incorrect model of evolution.
#
# Lajeunesse, M.J., University of South Florida, lajeunesse@usf.edu, 5/5/14
#
# Presented is the R script for a Monte Carlo experiment exploring the Type II error rates (i.e. false
# negative outcomes) of PGLS with the number of species (K) varying from few to many.  Error rates are based
# on the proportion of 10,000 regression analyses incorrectly concluding that the intercept (a) and slope (b)
# were zero.  Data were simulated to have phylogenetic correlations derived from an Ornstein–Uhlenbeck (OU) model
# of evolution with a stabilizing selection parameter set to alpha=3; these data were analyzed with a PGLS
# assuming a model of Brownian motion (BM) evolution.
#
# Inputs: none
# Outputs: a_BM (vector of Type II errors in intercept estimation with PGLS assuming BM)
#          a_OU (vector of Type II errors in intercept estimation with PGLS assuming OU)
#          b_BM (vector of Type II errors in slope estimation with PGLS assuming BM)
#          b_OU (vector of Type II errors in slope estimation with PGLS assuming OU)

library(ape); library(geiger); library(utils); library(nlme);

s <- 1000; s_pb <- 0;
K <- 100; x_mean <- 0; x_variance <- 1; a <- 1.0; b <- 0.5; a_NULL <- 0.0; b_NULL <- 0.0;
alpha <- 3
pb <- txtProgressBar(min=1, max=(s * 10), style=3)
tree <- sim.bdtree(b=1, d=0, stop="taxa", K)
for(i_N in seq(5, 50, by=5)) {
        a_BM_error <- 0; b_BM_error <- 0; a_OU_error <- 0; b_OU_error <- 0;
        for(i_s in 1:s) {
                subtree <- drop.tip(tree, sample(1:K, K - i_N))
                e_rand <- rnorm(i_N, 0, 1)
                x_rand <- rnorm(i_N, 0, 1)
                C_BM <- cov2cor(vcv(subtree))
                e_BM <- t(chol(C_BM)) %*% e_rand
                x_BM <- x_mean + sqrt(x_variance) * t(chol(C_BM)) %*% x_rand
                y_BM <- a + b * x_BM + e_BM
                C_OU <- (exp(-2 * alpha * (diag(C_BM) - C_BM)) - exp(-2 * alpha * diag(C_BM))) / (2 * alpha)
                C_OU <- C_OU * (1 / C_OU[1,1])
                e_OU <- t(chol(C_OU)) %*% e_rand
                x_OU <- x_mean + sqrt(x_variance) * t(chol(C_OU)) %*% x_rand
                y_OU <- a + b * x_OU + e_OU
                BM_results <- summary(gls(y_BM ~ x_BM, correlation=corBrownian(phy=subtree), method="ML"))
                        BM_a_t <- (BM_results$tTable[1,1] - a_NULL) / BM_results$tTable[1,2]
                        if(2 * pt(-abs(BM_a_t), df=K-2) > 0.05) a_BM_error <- a_BM_error + 1
                        BM_b_t <- (BM_results$tTable[2,1] - b_NULL) / BM_results$tTable[2,2]
                        if(2 * pt(-abs(BM_b_t), df=K-2) > 0.05) b_BM_error <- b_BM_error + 1
                OU_results <- summary(gls(y_OU ~ x_OU, correlation=corBrownian(phy=subtree), method="ML"))
                        OU_a_t <- (OU_results$tTable[1,1] - a_NULL) / OU_results$tTable[1,2]
                        if(2 * pt(-abs(OU_a_t), df=K-2) > 0.05) a_OU_error <- a_OU_error + 1
                        OU_b_t <- (OU_results$tTable[2,1] - b_NULL) / OU_results$tTable[2,2]
                        if(2 * pt(-abs(OU_b_t), df=K-2) > 0.05) b_OU_error <- b_OU_error + 1
                setTxtProgressBar(pb, i_s + s_pb)
        }
        s_pb <- s_pb + i_s
        if (i_N == 5) {
                a_BM <- a_BM_error / s; b_BM <- b_BM_error / s;
                a_OU <- a_OU_error / s; b_OU <- b_OU_error / s;
                N <- i_N;
        }
        else {
                a_BM <- c(a_BM, a_BM_error / s); b_BM <- c(b_BM, b_BM_error / s);
                a_OU <- c(a_OU, a_OU_error / s); b_OU <- c(b_OU, b_OU_error / s);
                N <- c(N, i_N)
        }
}

par(mfrow=c(1,2)); x_name <- bquote(italic(K));
y_a_name <- bquote(paste("Type II error, concluding   ", a==.(a_NULL)))
y_b_name <- bquote(paste("Type II error, concluding   ", b==.(b_NULL)))
plot(N, a_BM, ylim=c(0,1), type="o", xlab=x_name, ylab=y_a_name, main="intercept (a)")
lines(N, a_OU, type="o", lwd=1.5, lty=2)
legend(5, 0.25, c("BM data","OU data"), cex=0.8, lty=1:2)
plot(N, b_BM, ylim=c(0,1), type="o",  xlab=x_name, ylab=y_b_name, main="slope (b)")
lines(N, b_OU, type="o", lwd=1.5, lty=2)
legend(5, 0.25, c("BM data","OU data"), cex=0.8, lty=1:2)
```