Phylogenetic generalized least squares (PGLS)

Simon Joly

Fall 2016

Contents

This document contains information and practical examples in R on Phylogenetic generalized least squares (PGLS). It was develop for a half-day workshop that consists in a short presentation followed by the exercises of this document. Some of the information given in the presentation are repeated here so that this document could stand by itself and contains the necessary background information to understand the examples.

I assume that the readers are "reasonably" familiar with R as well as with linear regression and its assumptions. There are a lot of good R introductory tutorials on the web and for linear models, Zuur et al. (2007) provide a good introduction.

To perform the examples of this document, you will need to load the following R packages.

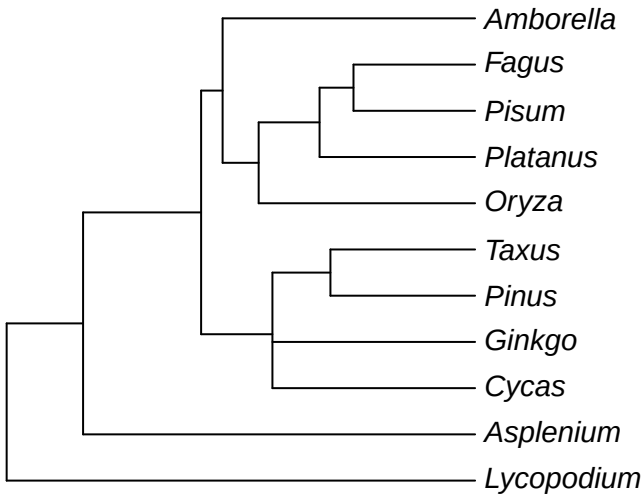
```
library(nlme)
library(ape)
library(RColorBrewer)
```

If some of these are not yet installed on your computer, you will have to install them using the function `install.packages()`. Also note that if you are using both the packages `nlme` and `ape`, `nlme` should be loaded first. If you don't do this, you might get errors; you could then restart R and start over.

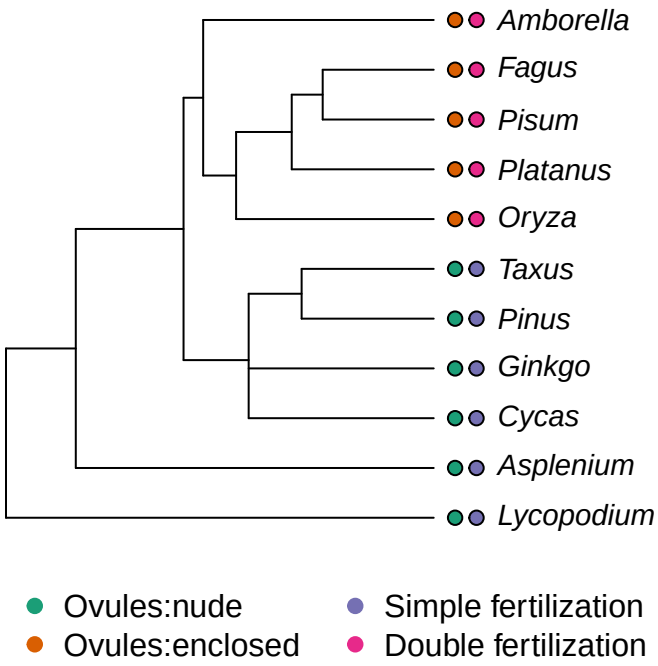
I first introduce comparative methods more generally before introducing PGLS, and I finish with slightly more advanced topics such as model testing with PGLS.

Phylogenetic Comparative Methods

Phylogenetic comparative methods were introduced by Joseph Felsenstein in 1985. The idea of phylogenetic comparative methods was to correct for the non-independence of species in statistical tests because of their shared evolutionary histories. Indeed, two species may look similar, not because they have been given the same *treatment*, but rather because they are closely related. For instance, considering the following angiosperm phylogeny.



It is clear that *Fagus* (Beech) and *Pisum* (pea) are more likely to share similar characteristics compared to *Asplenium* (a fern), because they share a more recent common ancestor. In other words, their evolutionary histories are shared over a longer period then with *Asplenium*. As such, they have more chance to have more similar traits (and in fact they do). For instance, take two characters, ovule and fertilization type, within this group.



Ignoring the phylogeny, we might be tempted to see a strong correlation between these two characters. Indeed, the states between the two characters show a perfect correspondence. Using standard contingency table statistics, we could do a Fisher exact test:

```
fisher.test(matrix(c(5,0,0,6),ncol=2))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matrix(c(5, 0, 0, 6), ncol = 2)
## p-value = 0.002165
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.842809      Inf
## sample estimates:
## odds ratio
##        Inf
```

This would suggest that the assotiation is significant. However, we know that the comparisons made are not completely independent. Actually, both characters evolved only once, and this along the same branch.

A more appropriate question would be "what is the probability that two characters evolves along the same branch?". This can be calculated using a contingency table, but this time taking the observations along the branches of the phylogeny. In the example, there are 18 branches and the two characters evolved only once and on the same branch.

```
fisher.test(matrix(c(1,0,0,17),ncol=2))
```

```
##
##  Fisher's Exact Test for Count Data
##
```

```
## data:  matrix(c(1, 0, 0, 17), ncol = 2)
## p-value = 0.05556
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4358974       Inf
## sample estimates:
## odds ratio
##        Inf
```

You can see that the result is no longer significant. While this approach is correct, more powerful comparative methods have been developped. Clearly, the most powerful and flexible approach is the Phylogenetic Generalized Least Squares (PGLS) and it is the one that will will introduce here.

The linear regression model

Let's start by doing a bit of revision on linear models. The linear model has the following form:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Y_i is the response (or dependent) variable, X_i is the explanatory (or inde- pendent) variable, and ϵ_i is the residual of observation i and represents the unexplained information. The parameters α and β are the population intercept and slope, respectively, and are unknown. In practive, you take a sample of size N and you get estimates for a and b for the intercept and the slope, respectively. When the linear regression is standardly fitted using ordinary least squares (OLS), the residuals ϵ_i are assumed to be normally distributed with expectation 0 and variance σ^2. In mathematic terms, $\epsilon_i \sim N(0, \sigma^2)$.

Obtaining reliable estimates with a linear regression implies that the data meets reveral assumptions, amongst which are normality, homogeneity, fixed X, independence, and correct model specification. We won't review all these here, but we will focus on one that is often violated when the data are phylogenetically structured, which is **independence**. This assumption is important as a lack of independence invalidates important tests such as the F-test and the t-test.

You get a violation of independece when the Y value at X_i is influenced by other X_i. Obviously, this can happen with phylogenetically structured data as a response variable can be more likely to react similarly to an explanatory variable if they are closely related species.

One example

To provide applied examples of PGLS in this document, we will use a dataset of tree functional traits from the province of Quebec, published by Paquette, Joly and Messier (2015). The dataset consists in a number of plant functional traits and in a molecular phylogeny built using the plant barcode markers. You can download the two required file by clicking on the links below and save them in your R working directory.

seedplants.tre

seedplants.csv

Before analysing the data, we will start by opening the data and the phylogenetic tree and clean them to keep only the species present in both the tree and the trait table. This is necessary because some additional species were included in the phylogenetic tree reconstruction to get a good topology.

```
require(ape)
# Open the documents; it assumes that they are in the working directory of R
```

```
seedplantstree <- read.nexus("./data/seedplants.tre")
seedplantsdata <- read.csv2("./data/seedplants.csv")
# Remove species for which we don't have complete data
seedplantsdata <- na.omit(seedplantsdata)
# Remove species in the tree that are not in the data matrix
species.to.exclude <- seedplantstree$tip.label[!(seedplantstree$tip.label %in%
                                                   seedplantsdata$Code)]
seedplantstree <- drop.tip(seedplantstree,species.to.exclude)
rm(species.to.exclude)
```

Now, we can have a look at the data, and then order the plant trait to be in the same order as the species in the tree.

```
# Here is what the loaded data looks like
head(seedplantsdata)
```

```
##   Code       Species.name Occurrence maxH   Wd    Sm Shade    N
## 1 ABBA     Abies balsamea       7759   25 0.34   7.6   5.0 1.66
## 2 ACNE       Acer negundo          0   20 0.44  34.0   3.5 2.50
## 3 ACNI        Acer nigrum          1   30 0.52  65.0   3.0 1.83
## 4 ACPE Acer pensylvanicum        665   10 0.44  41.0   3.5 2.22
## 5 ACPL   Acer platanoides          0   15 0.51 172.0   4.2 1.99
## 6 ACRU        Acer rubrum       3669   25 0.49  20.0   3.4 1.91
```

```
# Order tree to make it nicer when plotting
seedplantstree <- ladderize(seedplantstree, right = FALSE)
# Name the rows of the data.frame with the species codes used as tree labels
#  and remove the obsolete column with species codes.
rownames(seedplantsdata) <- seedplantsdata$Code
seedplantsdata <- seedplantsdata[,-1]
# Order the data in the same order as the tip.label of the tree. In the present
#  example, this was already the case, but it is an important step for
#  any analysis.
seedplantsdata <- seedplantsdata[seedplantstree$tip.label,]
```

Now that the data is ready, let's fit a linear model and try to explain shade tolerance (Shade) of trees using wood density (Wd).

```
# Fit a linear model using OLS
shade.lm <- lm(Shade ~ Wd, data = seedplantsdata)
summary(shade.lm)
```

```
##
## Call:
## lm(formula = Shade ~ Wd, data = seedplantsdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87120 -1.02501  0.05628  0.70132  2.38261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

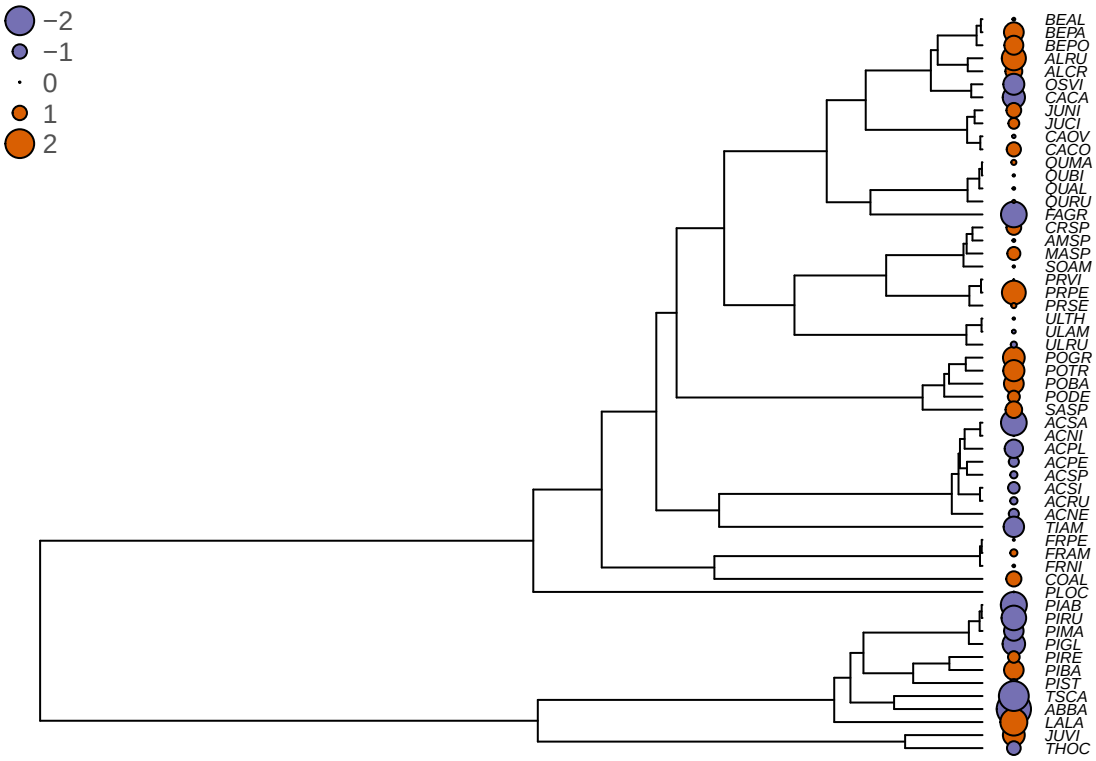
```
## (Intercept)   2.0010     0.7501   2.668    0.010 *
## Wd            1.8130     1.5676   1.157    0.252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.146 on 55 degrees of freedom
## Multiple R-squared:  0.02374,    Adjusted R-squared:  0.005992
## F-statistic: 1.338 on 1 and 55 DF,  p-value: 0.2525
```

You can see that the slope of coefficient 1.81 is not significant (p=0.252). The standard descriptive plots obtained with `plot(shade.lm)` show that there is slightly greater variation in the residuals for low fitted values, but these are not extreme. However, what is important here is to investigate whether the residuals are independent from the phylogeny. Let's see what this gives.

```
# Extract the residuals
shade.res <- residuals(shade.lm)

###
# Plot the residuals beside the phylogeny

# The following command changes the graphical parameters for nicer tree output
op <- par(mar=c(1,1,1,1))
# Colors for the tree plotting
cols <- c("#7570b3","#d95f02")
# The next three commands will plot the tree, then circles that reflect
#  the residuals values at the tips of the tree, and will finally
#  add a legend.
plot(seedplantstree,type="p",TRUE,label.offset=0.01,cex=0.5,no.margin=FALSE)
tiplabels(pch=21,bg=cols[ifelse(shade.res>0,1,2)],col="black",cex=abs(shade.res),adj=0.505)
legend("topleft",legend=c("-2","-1","0","1","2"),pch=21,
       pt.bg=cols[c(1,1,1,2,2)],bty="n",
       text.col="gray32",cex=0.8,pt.cex=c(2,1,0.1,1,2))
```




```
# Reset graphical parameters to defaults
par(op)
```

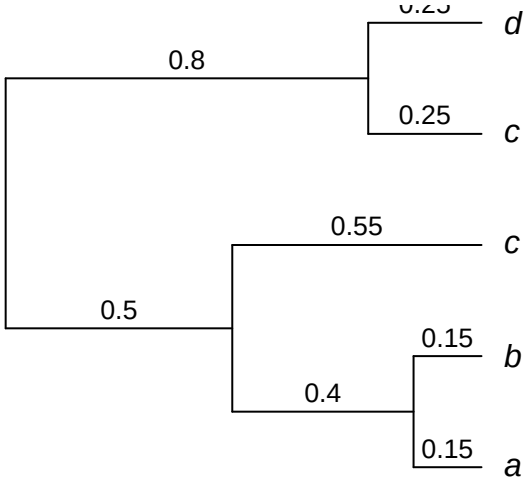
You can see that in several cases, closely related species tend to have similar residuals, which is problematic. In such cases, the assumption of independence of the ordinary least squares (OLS) no longer holds and the statistical tests for the null hypotheses are no longer valid. We will now see how PGLS can correct this.

Phylogenetic generalized least squares (PGLS)

Phylogenetic generalized least squares (PGLS) is just a precise application of the larger method called generalized least squares (GLS). Generalized least squares relaxe the assumption that the error of the model need not be correlated with each other. Indeed, they allow the user to specify the structure of that residual correlation. This is used, for instance, to correct for spatial correlation, time series, or phylogenetic correlation, the topic of interest here. But to be able to account for this phylogenetic correlation, we need to be able to describe it.

Phylogenetic correlation structure

Phylogenetic relationships can be described using a correlation structure. Below, you have phylogenetic tree with branch lengths indicated above the branches, followed by a variance-covariance matrix that perfectly represents the tree.




```
##      a    b    c    c    d
## a 1.05 0.90 0.50 0.00 0.00
## b 0.90 1.05 0.50 0.00 0.00
## c 0.50 0.50 1.05 0.00 0.00
## c 0.00 0.00 0.00 1.05 0.80
## d 0.00 0.00 0.00 0.80 1.05
```

The matrix diagonal of the variance-covariance matrix represents the species variances. This is the distance of the tips of the tree from the root and it determines how much the tips have evolved from the root. The off-diagonal values of the matrix are the covariances between the species. They indicate the proportion of the time that the species have evolved together. This corresponds to the length of the branches that two species share, starting from the root of the tree. For instance, species *a* and *c* have shared a common history for 0.5 units of time; hence they have a covariance of 0.5.

> Note that all the tips are equidistant from the root. When trees have this property, they are said to be ***ultrametric***. Most phylogenetic comparative methods require the trees to be ultrametric, although there are sometimes ways to relax this assumption. If you do not have an ultrametric tree, it is possible to make it ultrametric using the function `chronos` of the `ape` package, although this approach is not ideal.

The variance-covariance matric of a phylogenetic tree can be obtained from a tree using the function `vcv` from the `ape` package.

```
# 'atree' corresponds to the phylogenetic tree shown above
atree <- "(((a:0.15,b:0.15):0.4,c:0.55):0.5,(c:0.25,d:0.25):0.8);"
atree <- read.tree(text=atree)
# Extract the variance-covariance matrix
varcovar <- vcv(atree)
varcovar
```

```
##      a    b    c    c    d
## a 1.05 0.90 0.50 0.00 0.00
## b 0.90 1.05 0.50 0.00 0.00
## c 0.50 0.50 1.05 0.00 0.00
## c 0.00 0.00 0.00 1.05 0.80
## d 0.00 0.00 0.00 0.80 1.05
```

This is great, but we mentioned above that it is a correlation matric that we need in a GLS to account for the correlation in the residuals. To obtain a correlation matrix from the variance-covariance matrix shown above, you only need to divide the variance-covariance matrix by the length of the tree, or the distance from the root to the tips. It can also be obtained using the R function `cov2cor`.

```
corrmat <- cov2cor(varcovar)
round(corrmat,3)
```

```
##       a     b     c     c     d
## a 1.000 0.857 0.476 0.000 0.000
## b 0.857 1.000 0.476 0.000 0.000
## c 0.476 0.476 1.000 0.000 0.000
## c 0.000 0.000 0.000 1.000 0.762
## d 0.000 0.000 0.000 0.762 1.000
```

Now, the diagonal elements equal to 1, indicating that the species are perfectly correlated to themselves. Note that it is also possible to obtain directly the correlation matrix from the function `vcv` by using the `corr=TRUE` option.

```
corrmat <- vcv(atree,corr=TRUE)
round(corrmat,3)
```

```
##       a     b     c     c     d
## a 1.000 0.857 0.476 0.000 0.000
## b 0.857 1.000 0.476 0.000 0.000
## c 0.476 0.476 1.000 0.000 0.000
## c 0.000 0.000 0.000 1.000 0.762
## d 0.000 0.000 0.000 0.762 1.000
```

We are now ready to run a PGLS. There are several ways to do this in R. For instance, the package `caper` is a very well known package for PGLS. However, we will use the function `gls` here from the `nlme` package, which comes with the `base` packages in R. This function is robust and has the advantage to be very flexible. Indeed, it allows to easily use more complex models such as mixed effect models, although this will not be discussed here.

Before we run the PGLS, let's run the basic model with the function `gls` as a reference. Running the standard linear model with the package `nlme` will allow to run model comparison functions in R (see below), which would not be possible is different models are fitted using different packages.

```
require(nlme)
shade.gls <- gls(Shade ~ Wd, data = seedplantsdata)
summary(shade.gls)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd
##   Data: seedplantsdata
##       AIC     BIC    logLik
##   180.472 186.494 -87.23602
##
## Coefficients:
##               Value Std.Error  t-value p-value
## (Intercept) 2.00098 0.7500707 2.667722  0.0100
```

```
## Wd          1.81296 1.5675668 1.156544  0.2525
##
##  Correlation:
##    (Intr)
## Wd -0.979
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.63307700 -0.89457443  0.04911902  0.61207032  2.07940955
##
## Residual standard error: 1.145813
## Degrees of freedom: 57 total; 55 residual
```

You can see that the output is essentially identical to that of the `lm` function. Now, let's run a PGLS model. To assign the correlation matrix to the `gls` function, you can use the correlation function called `corSymm` that assumes that the correlation matrix is symmetric. This is the case with phylogenetic trees; the correlation between species a and b is the same as between b ad a. Only the lower triangular part of the matrix has to be passed to the `corSymm` structure. If `mat` is the correlation matrix, this is done using the command `mat[lower.tri(mat)]`. Then you pass the correlation matrix to `gls` using the `correlation` argument.

```
# Calculate the correlation matrix from the tree
mat <- vcv(seedplantstree,corr=TRUE)
# Create the correlation structure for gls
corr.struct <- corSymm(mat[lower.tri(mat)],fixed=TRUE)
# Run the pgls
shade.pgls1 <- gls(Shade ~ Wd, data = seedplantsdata, correlation=corr.struct)
summary(shade.pgls1)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd
##   Data: seedplantsdata
##        AIC      BIC    logLik
##   214.3762 220.3982 -104.1881
##
## Correlation Structure: General
##  Formula: ~1
##  Parameter estimate(s):
##  Correlation:
##    1     2     3     4     5     6     7     8     9     10    11    12
## 2  0.000
## 3  0.000 0.967
## 4  0.000 0.967 0.976
## 5  0.000 0.967 0.981 0.976
## 6  0.000 0.967 0.974 0.974 0.974
## 7  0.000 0.967 0.997 0.976 0.981 0.974
## 8  0.000 0.967 0.974 0.974 0.974 0.997 0.974
## 9  0.000 0.967 0.976 0.983 0.976 0.974 0.976 0.974
## 10 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654
## 11 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.984
## 12 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726
## 13 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.952 0.952 0.726
## 14 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.952 0.952 0.726
## 15 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.952 0.952 0.726
```

```
## 16 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.945 0.945 0.726
## 17 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.876 0.876 0.726
## 18 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.876 0.876 0.726
## 19 0.000 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596
## 20 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.989
## 21 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.835 0.835 0.726
## 22 0.000 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596
## 23 0.000 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596
## 24 0.000 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596 0.596
## 25 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.876 0.876 0.726
## 26 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.876 0.876 0.726
## 27 0.528 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 28 0.843 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 29 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.983
## 30 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.945 0.945 0.726
## 31 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 32 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 33 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 34 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 35 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 36 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 37 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 38 0.000 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523
## 39 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.675 0.675 0.675
## 40 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.675 0.675 0.675
## 41 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.675 0.675 0.675
## 42 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.675 0.675 0.675
## 43 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.898
## 44 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.898
## 45 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.898
## 46 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.835 0.835 0.726
## 47 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.835 0.835 0.726
## 48 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.835 0.835 0.726
## 49 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.835 0.835 0.726
## 50 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.675 0.675 0.675
## 51 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.980
## 52 0.528 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 53 0.000 0.720 0.720 0.720 0.720 0.720 0.720 0.720 0.720 0.654 0.654 0.654
## 54 0.906 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 55 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.800
## 56 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.800
## 57 0.000 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.726 0.726 0.800
##    13    14    15    16    17    18    19    20    21    22    23    24
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
```

```
## 13
## 14 0.998
## 15 0.994 0.994
## 16 0.945 0.945 0.945
## 17 0.876 0.876 0.876 0.876
## 18 0.876 0.876 0.876 0.876 0.998
## 19 0.596 0.596 0.596 0.596 0.596 0.596
## 20 0.726 0.726 0.726 0.726 0.726 0.726 0.596
## 21 0.835 0.835 0.835 0.835 0.835 0.835 0.596 0.726
## 22 0.596 0.596 0.596 0.596 0.596 0.596 0.715 0.596 0.596
## 23 0.596 0.596 0.596 0.596 0.596 0.596 0.715 0.596 0.596 0.997
## 24 0.596 0.596 0.596 0.596 0.596 0.596 0.715 0.596 0.596 0.999 0.997
## 25 0.876 0.876 0.876 0.876 0.984 0.984 0.596 0.726 0.835 0.596 0.596 0.596
## 26 0.876 0.876 0.876 0.876 0.984 0.984 0.596 0.726 0.835 0.596 0.596 0.596
## 27 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 28 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 29 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.983 0.726 0.596 0.596 0.596
## 30 0.945 0.945 0.945 0.988 0.876 0.876 0.596 0.726 0.835 0.596 0.596 0.596
## 31 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 32 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 33 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 34 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 35 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 36 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 37 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 38 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523 0.523
## 39 0.675 0.675 0.675 0.675 0.675 0.675 0.596 0.675 0.675 0.596 0.596 0.596
## 40 0.675 0.675 0.675 0.675 0.675 0.675 0.596 0.675 0.675 0.596 0.596 0.596
## 41 0.675 0.675 0.675 0.675 0.675 0.675 0.596 0.675 0.675 0.596 0.596 0.596
## 42 0.675 0.675 0.675 0.675 0.675 0.675 0.596 0.675 0.675 0.596 0.596 0.596
## 43 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.898 0.726 0.596 0.596 0.596
## 44 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.898 0.726 0.596 0.596 0.596
## 45 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.898 0.726 0.596 0.596 0.596
## 46 0.835 0.835 0.835 0.835 0.835 0.835 0.596 0.726 0.881 0.596 0.596 0.596
## 47 0.835 0.835 0.835 0.835 0.835 0.835 0.596 0.726 0.881 0.596 0.596 0.596
## 48 0.835 0.835 0.835 0.835 0.835 0.835 0.596 0.726 0.881 0.596 0.596 0.596
## 49 0.835 0.835 0.835 0.835 0.835 0.835 0.596 0.726 0.881 0.596 0.596 0.596
## 50 0.675 0.675 0.675 0.675 0.675 0.675 0.596 0.675 0.675 0.596 0.596 0.596
## 51 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.980 0.726 0.596 0.596 0.596
## 52 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 53 0.654 0.654 0.654 0.654 0.654 0.654 0.596 0.654 0.654 0.596 0.596 0.596
## 54 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 55 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.800 0.726 0.596 0.596 0.596
## 56 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.800 0.726 0.596 0.596 0.596
## 57 0.726 0.726 0.726 0.726 0.726 0.726 0.596 0.800 0.726 0.596 0.596 0.596
##    25    26    27    28    29    30    31    32    33    34    35    36
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
```

```
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26 0.992
## 27 0.000 0.000
## 28 0.000 0.000 0.528
## 29 0.726 0.726 0.000 0.000
## 30 0.876 0.876 0.000 0.000 0.726
## 31 0.000 0.000 0.528 0.843 0.000 0.000
## 32 0.000 0.000 0.528 0.843 0.000 0.000 0.874
## 33 0.000 0.000 0.528 0.843 0.000 0.000 0.985 0.874
## 34 0.000 0.000 0.528 0.843 0.000 0.000 0.997 0.874 0.985
## 35 0.000 0.000 0.528 0.843 0.000 0.000 0.874 0.965 0.874 0.874
## 36 0.000 0.000 0.528 0.843 0.000 0.000 0.999 0.874 0.985 0.997 0.874
## 37 0.000 0.000 0.528 0.843 0.000 0.000 0.874 0.926 0.874 0.874 0.926 0.874
## 38 0.523 0.523 0.000 0.000 0.523 0.523 0.000 0.000 0.000 0.000 0.000 0.000
## 39 0.675 0.675 0.000 0.000 0.675 0.675 0.000 0.000 0.000 0.000 0.000 0.000
## 40 0.675 0.675 0.000 0.000 0.675 0.675 0.000 0.000 0.000 0.000 0.000 0.000
## 41 0.675 0.675 0.000 0.000 0.675 0.675 0.000 0.000 0.000 0.000 0.000 0.000
## 42 0.675 0.675 0.000 0.000 0.675 0.675 0.000 0.000 0.000 0.000 0.000 0.000
## 43 0.726 0.726 0.000 0.000 0.898 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 44 0.726 0.726 0.000 0.000 0.898 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 45 0.726 0.726 0.000 0.000 0.898 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 46 0.835 0.835 0.000 0.000 0.726 0.835 0.000 0.000 0.000 0.000 0.000 0.000
## 47 0.835 0.835 0.000 0.000 0.726 0.835 0.000 0.000 0.000 0.000 0.000 0.000
## 48 0.835 0.835 0.000 0.000 0.726 0.835 0.000 0.000 0.000 0.000 0.000 0.000
## 49 0.835 0.835 0.000 0.000 0.726 0.835 0.000 0.000 0.000 0.000 0.000 0.000
## 50 0.675 0.675 0.000 0.000 0.675 0.675 0.000 0.000 0.000 0.000 0.000 0.000
## 51 0.726 0.726 0.000 0.000 0.980 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 52 0.000 0.000 0.918 0.528 0.000 0.000 0.528 0.528 0.528 0.528 0.528 0.528
## 53 0.654 0.654 0.000 0.000 0.654 0.654 0.000 0.000 0.000 0.000 0.000 0.000
## 54 0.000 0.000 0.528 0.843 0.000 0.000 0.860 0.860 0.860 0.860 0.860 0.860
## 55 0.726 0.726 0.000 0.000 0.800 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 56 0.726 0.726 0.000 0.000 0.800 0.726 0.000 0.000 0.000 0.000 0.000 0.000
## 57 0.726 0.726 0.000 0.000 0.800 0.726 0.000 0.000 0.000 0.000 0.000 0.000
##    37    38    39    40    41    42    43    44    45    46    47    48
## 2
## 3
## 4
## 5
## 6
```

```
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34
## 35
## 36
## 37
## 38 0.000
## 39 0.000 0.523
## 40 0.000 0.523 0.959
## 41 0.000 0.523 0.964 0.959
## 42 0.000 0.523 0.964 0.959 0.982
## 43 0.000 0.523 0.675 0.675 0.675 0.675
## 44 0.000 0.523 0.675 0.675 0.675 0.675 0.986
## 45 0.000 0.523 0.675 0.675 0.675 0.675 0.998 0.986
## 46 0.000 0.523 0.675 0.675 0.675 0.675 0.726 0.726 0.726
## 47 0.000 0.523 0.675 0.675 0.675 0.675 0.726 0.726 0.726 0.997
## 48 0.000 0.523 0.675 0.675 0.675 0.675 0.726 0.726 0.726 0.997 0.999
## 49 0.000 0.523 0.675 0.675 0.675 0.675 0.726 0.726 0.726 0.984 0.984 0.984
## 50 0.000 0.523 0.936 0.936 0.936 0.936 0.675 0.675 0.675 0.675 0.675 0.675
## 51 0.000 0.523 0.675 0.675 0.675 0.675 0.898 0.898 0.898 0.726 0.726 0.726
## 52 0.528 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 53 0.000 0.523 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654 0.654
## 54 0.860 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
## 55 0.000 0.523 0.675 0.675 0.675 0.675 0.800 0.800 0.800 0.726 0.726 0.726
## 56 0.000 0.523 0.675 0.675 0.675 0.675 0.800 0.800 0.800 0.726 0.726 0.726
## 57 0.000 0.523 0.675 0.675 0.675 0.675 0.800 0.800 0.800 0.726 0.726 0.726
##    49    50    51    52    53    54    55    56
## 2
## 3
```

```
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
## 32
## 33
## 34
## 35
## 36
## 37
## 38
## 39
## 40
## 41
## 42
## 43
## 44
## 45
## 46
## 47
## 48
## 49
## 50 0.675
## 51 0.726 0.675
## 52 0.000 0.000 0.000
## 53 0.654 0.654 0.654 0.000
## 54 0.000 0.000 0.000 0.528 0.000
## 55 0.726 0.675 0.800 0.000 0.654 0.000
## 56 0.726 0.675 0.800 0.000 0.654 0.000 0.983
## 57 0.726 0.675 0.800 0.000 0.654 0.000 0.999 0.983
```

```
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 0.911433  4.409058 0.2067184  0.8370
## Wd          4.361028  1.693349 2.5753865  0.0127
##
##  Correlation:
##    (Intr)
## Wd -0.166
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.26890642 -0.16431866 -0.02645422  0.09638984  0.34953444
##
## Residual standard error: 7.455109
## Degrees of freedom: 57 total; 55 residual
```

Note that the term `fixed=TRUE` in the corSymm structure indicates that the correlation structure is fixed during the parameter optimization.

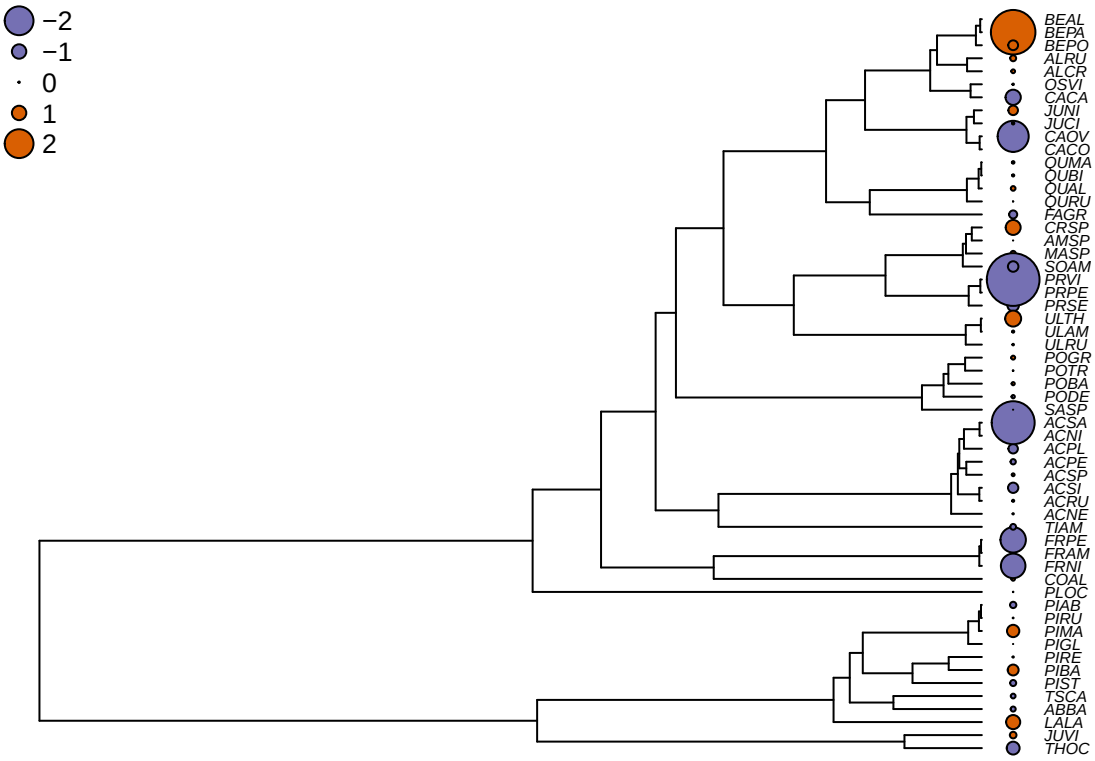
The output is similar to that of the model without the correlation. However, there are some differences. The item "Correlation:" gives the correlation among the estimated parameters. The "Standardized residuals" are the raw residuals divided by the residual standard error (the raw residuals can be output with `residuals(shade.pgls1,"response")`).

Interestingly, you can see that the coefficient estimate for the slope is greater (4.361) than with standard regression and also significant (p=0.0127). This is a positive exmple of PGLS. Indeed, the relationship between shade tolerance and wood density was obscured by the phylogenetic correlation of the residuals. Once this correlation is accounted for, the significant relationship is revealed.

A significant relationship between shade tolerance and wood density actually make sense, even though this relationship is most likely not causal. Indeed, shade tolerant trees are generally sucessional species and often grow slower, partly because of the limited light availability, and thus tend to develop higher density woods.

Now, let's have a look at the residuals of the model. To extract residuals corrected by the correlation structure, you need to ask for the normalized residuals.

```
# Extract the residuals corrected by the correlation structure
pgls1.res <- residuals(shade.pgls1,type="normalized")
# Change the graphical parameters
op <- par(mar=c(1,1,1,1))
# Same plotting as above
plot(seedplantstree,type="p",TRUE,label.offset=0.01,cex=0.5,no.margin=FALSE)
tiplabels(pch=21,bg=cols[ifelse(pgls1.res>0,1,2)],col="black",cex=abs(pgls1.res),adj=0.505)
legend("topleft",legend=c("-2","-1","0","1","2"),pch=21,
       pt.bg=cols[c(1,1,1,2,2)],bty="n",
       text.col="black",cex=0.8,pt.cex=c(2,1,0.1,1,2))
```




```
# Reset graphical parameters to defaults
par(op)
```

If you compare with the ordinary least squares optimization, the residuals are much less phylogenetically correlated.

Other correlation structures

In the previous PGLS, we have used the corSymm structure to pass the phylogenetic correlation structure to the gls. This is perfectly fine, but there are more simple ways. Julien Dutheil has developped phylogenetic structures to be used especially in PGLS.

The one we used above is equivalent to the `corBrownian` structure of `ape`. This approach is easier and you just have to pass the tree to the correlation structure. Here is the same example used with `corBrownian`.

```
# Get the correlation structure
bm.corr <- corBrownian(phy=seedplantstree)
# PGLS
shade.pgls1b <- gls(Shade ~ Wd, data = seedplantsdata, correlation=bm.corr)
summary(shade.pgls1b)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd
##   Data: seedplantsdata
##        AIC      BIC    logLik
##   214.3762 220.3982 -104.1881
##
## Correlation Structure: corBrownian
```

```
##  Formula: ~1
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 0.911433  4.409058 0.2067184  0.8370
## Wd          4.361028  1.693349 2.5753865  0.0127
##
##  Correlation:
##    (Intr)
## Wd -0.166
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.26890642 -0.16431866 -0.02645422  0.09638984  0.34953444
##
## Residual standard error: 7.455109
## Degrees of freedom: 57 total; 55 residual
```

You can see that the results are identical. The only difference is that the correlation structure is not output in the summary. The `numeric(0)` means that no parameter was estimated during the optimization (it is fixed).

Now, you might wonder why the correlation structure is called corBrownian. This is because is uses a Brownian model model to model the evolution along the branch of the tree. This is often refferred as a neutral model. Other models are also available. But before looking into more models, it is a good idea to have a closer look at the Brownian motion model.

The Brownian Motion (BM) model

When we want to account for the non-independence of species due to their evolutionary histories in statistical analyses, a model of evolution is necessarily implied. Indeed, we assume that traits evolved through time (along the phylogeny) and that closely related species are more likely to be more similar on average at a given trait than distantly related species. In evolutionary biologogy, the more basic model (often used as a null model in many analyses) is the Brownian motion model. This model of evolution is named after Robert Brown, a celeb botanist that published an important Flora of Australia in 1810. He was also the first to distinguish gymnosperms from angiosperms. His discovery of the Brownian motion is due to the observation that small particules in solution have the tendency to move in any direction, an observation first made while observing *Clarkia* pollen under a microscope. The explanation would come later, in terms of random molecular impacts.

Mathematicians have constructed a stochastic process that is intended to approximate the Brownian motion. In this model, each step is independent from the others and can go in any direction. The mean displacement is zero and the variance is uniform across the parameter space. The displacements can be summed, which means that the variances of the independent displacements can be added up. If σ^2 is the variance of a single displacement, the variance after time t will be $\sigma^2 t$. When the number of steps is large, as in a phylogenetic context, the result is normally distributed.

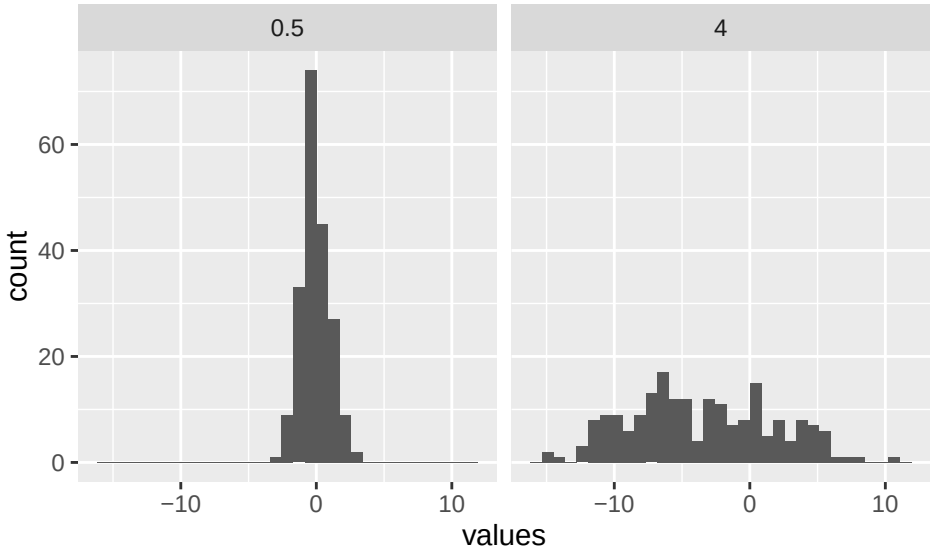
When applied to phylogenies, the Brownian motion model is kind of applied indepenpenty to each branch of the phylogeny. That allows to model the amount of change that occured along a given branch. If the variance of the Brownian motion model is σ^2 per unit of time t, then the net change along a branch of time t is drawn from a normal distribution with mean 0 and variance $\sigma^2 t$. This model can also be represented mathematically the following way, such as the amount of change for character X over the infinitesimal time in the interval between time t and $t + dt$ is:

$$dX(t) = \sigma^2 dB(t),$$

where $dB(t)$ is the gaussian distribution. Importantly, this model assumes that:

1. Evolution occuring in each branch of the phylogeny is independent of that occuring in other branches.
2. Evolution is completely random (i.e., no selection).

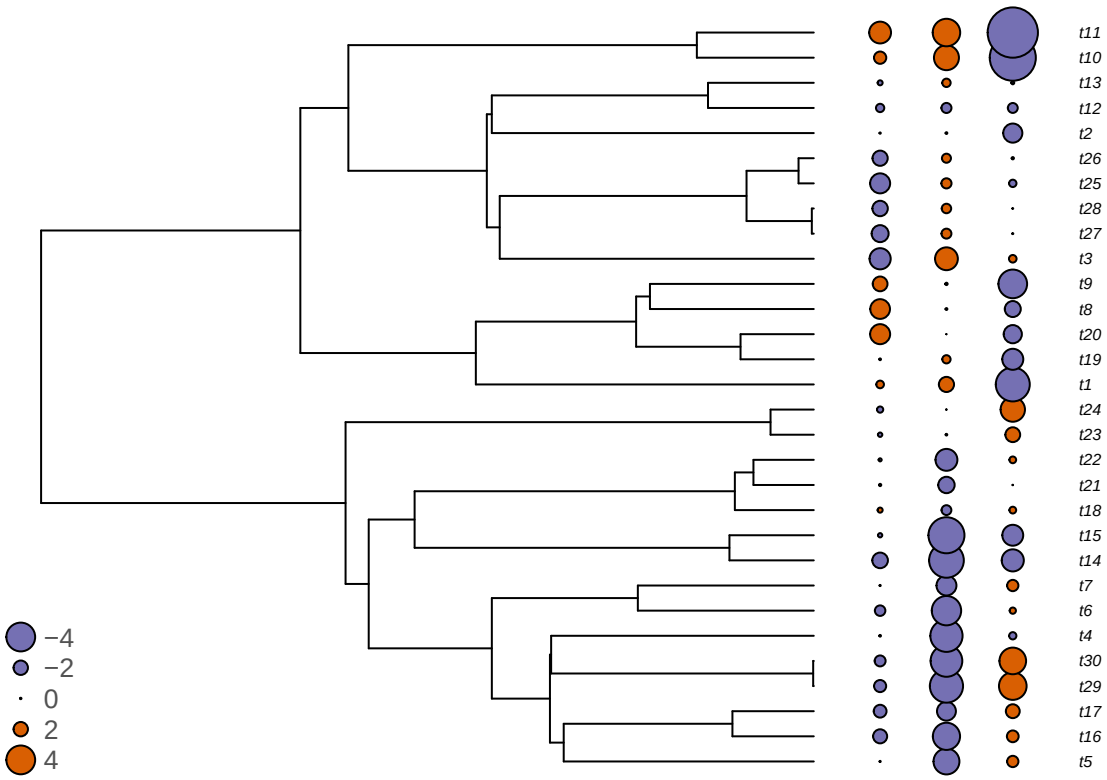
The parameter σ^2 in the model gives the variance, or in other word the speed of evolution. The higher the variance, the faster the character will evolve. Here are two examples of simulated characters on a tree of 200 species with $\sigma^2 = 0.5$ and $\sigma^2 = 4$.



A more thorough introduction to the Brownian Motion model can be found in Felsenstein (2004, chapter 23).

The Brownian motion model is often said to model neutral drift, although a good fit to this model does not necessarily means that the data evolved via random drifts as other processes can also result in BM-like patterns (Hansen and Martins, 1996).

Note also that the model is stochastic. That is, even if two closely related species are more likely to share similar character states than a distant one, this is only true on average. For any given simulated character, closely related species can sometimes be more different than to a distant species. Look at the following figure, that shows three traits simulated under the Brownian motion.



Other correlation structures (or evolutionary models)

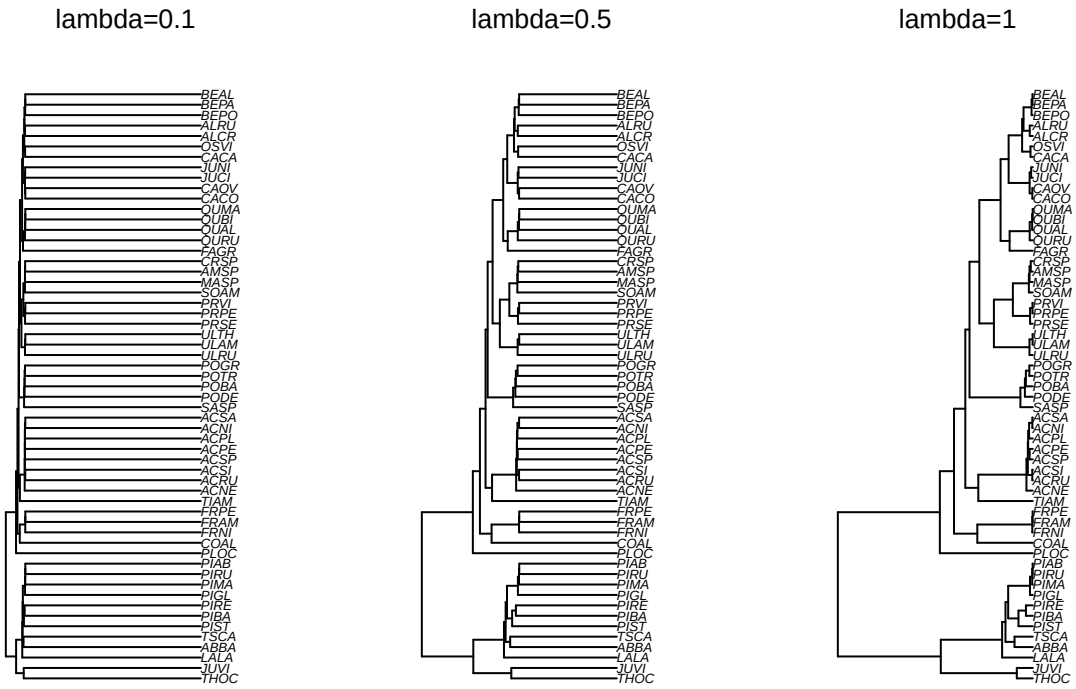
The correlation structures available in the package `ape` offer other alternatives for the assumed model of character evolution. For instance, the `corMartins` correlation structure models selection using the Ornstein-Uhlenbeck or Hansen model. It has a parameter α that determines the strength of the selection. Also, `corBlomberg` models accelerating or decelerating Brownian evolution. That is, the evolutionary rate of the Brownian motion is either accelerating or decelerating with time with this model. We won't look into these correlation structure here, but we will look at another one, called `corPagel`, in the next section.

Pagel's correlation structure

When controling for phylogenetic relationships with phylogenetic generalized least squares, we assume that the residuals are perfectly correlated according to the correlation structure. In practice, it might not be always the case and it is difficult to really know how important it is to control for the phylogenetic relationship in a specific case. For instance, for a given study, the correlation in the residuals might not be highly phylogenetically correlated. This is possible to model using the parameter λ of Pagel (1999). The idea is to multiply the off-diagonal of the correlation matrix (essentially the branch lengths of the phylogeny) by a parameter λ, but not the diagonal values. This essentially leads to a modification of branch lengths of the phylogeny. A λ value near zero gives very shorts branch lengths to the branches of the phylogenies, leaving only long tip branches. This, in effect, reduces the phylogenetic correlation. At the opposite, if λ is close to 1, then the modified phylogeny ressembles the true phylogeny. Indeed, the parameter λ is often interpreted as a parameter of phylogenetic signal; as such, a greater λ value implies a stronger phylogenetic signal.

The following figure shows how different lambda values affect the shape of the Quebec trees phylogeny.

```
## Loading required package: geiger
```



Pagel's λ model can be used in PGLS using the `corPagel` correlation structure. The usage of this correlation structure is similar to that of the `corBrownian` structure, except that you need to provide a starting parameter value for λ. The `gls` function will then optimize this parameter if you choose the option `fixed=FALSE`, which is the default.

```
# Get the correlation structure
pagel.corr <- corPagel(0.3, phy=seedplantstree, fixed=FALSE)
# PGLS
shade.pgls2 <- gls(Shade ~ Wd, data = seedplantsdata, correlation=pagel.corr)
summary(shade.pgls2)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd
##   Data: seedplantsdata
##        AIC     BIC    logLik
##   163.3967 171.426 -77.69833
##
## Correlation Structure: corPagel
##  Formula: ~1
##  Parameter estimate(s):
##    lambda
## 0.9581665
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 1.254987  1.636575 0.7668377  0.4465
## Wd          3.573527  1.497808 2.3858381  0.0205
##
##  Correlation:
##    (Intr)
## Wd -0.397
```

```
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.75145692 -0.44908843 -0.05417524  0.25655008  0.96493685
##
## Residual standard error: 2.621947
## Degrees of freedom: 57 total; 55 residual
```

You can see that gls has estimated the λ parameter, which is 0.958 here. Because the estimated λ is very close to 1, we can conclude that residuals of the model were highly phylogenetically correlated. This, in turns, thus confirms the importance of using a PGLS with this model. If the λ estimated would have been close to 0, it would have suggested that the PGLS is not necessary.

Extending PGLS... phylogenetic ANOVA

The great thing with PGLS as implemented with the `gls` function is that it can easily be adapted to testing many different types of models. To give just one example here, it is easy to implement a phylogenetic ANOVA in R. Indeed, you just need to give `gls` a categorical trait as independent variable.

Because there is no categorical variable in the plant functional trait dataset, we will create one by dividing the wood density category in two categories, light and dense wood.

```
# Make categorical variable
seedplantsdata$Wd.cat<-cut(seedplantsdata$Wd,breaks=2,labels=c("light","dense"))
# Phylogenetic ANOVA
shade.pgls3 <- gls(Shade ~ Wd.cat, data = seedplantsdata, correlation=pagel.corr)
summary(shade.pgls3)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd.cat
##   Data: seedplantsdata
##        AIC      BIC    logLik
##   166.7352 174.7646 -79.36762
##
## Correlation Structure: corPagel
##  Formula: ~1
##  Parameter estimate(s):
##    lambda
## 0.9439646
##
## Coefficients:
##                 Value Std.Error  t-value p-value
## (Intercept) 2.6826723 1.3844404 1.937730  0.0578
## Wd.catdense 0.6179855 0.2526902 2.445626  0.0177
##
##  Correlation:
##             (Intr)
## Wd.catdense -0.037
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.69257567 -0.48677930 -0.04143001  0.33640615  0.95379525
##
```

```
## Residual standard error: 2.429586
## Degrees of freedom: 57 total; 55 residual
```

You can see that the wood density, even when transformed in a categorical variable, has a significant effect on shade tolerance.

Advanced topic: model testing

You might be interested in comparing different models, which is a common approach to modelisation in biology. However, there is a slight twist that you need to be aware of with PGLS.

The default method for model fitting with `gls` is restricted maximum likelihood estimation (REML), obtained by `method="REML"`. This is different than standard maximum likelihood estimation (ML), which can be obtained with `method="ML"`. The difference between these is complex, but suffice to say that they differ in the way the variance parameters are estimated. REML provides less biased parameter estimates and is the prefered method to report the parameter coefficients in a publication. It is also the method of choice if you want to compare models with different correlation (or variance) structures. For example, if you want to test whether a PGLS model with an optimized Pagel's λ fits the data better than a model with no phylogenetic correlation (that is, with Pagel $\lambda = 0$):

```
pagel.0 <- gls(Shade ~ Wd, data = seedplantsdata,
               correlation=corPagel(0,phy=seedplantstree, fixed=TRUE),
               method="REML")
pagel.fit <- gls(Shade ~ Wd, data = seedplantsdata,
                 correlation=corPagel(0.8,phy=seedplantstree, fixed=FALSE),
                 method="REML")
anova(pagel.0,pagel.fit)
```

```
##           Model df      AIC     BIC    logLik   Test  L.Ratio p-value
## pagel.0       1  3 180.4720 186.494 -87.23602
## pagel.fit     2  4 163.3967 171.426 -77.69833 1 vs 2 19.07537  <.0001
```

You can use the AIC or BIC to compare the model, or the likelihood ratio test. You can see here that the PGLS model with a fitted Pagel λ has a better fit than the one with a $\lambda = 0$. This is also a test of whether a PGLS model is better than a standard regression model.

Now, if you are interested in testing the fixed parameters in the model, you need to use maximum likelihood fitting. For instance, if you want to use a likelihood ratio test to test the model with wood density as independent variable versus a null model with just the intercept, you can do the following.

```
wd <- gls(Shade ~ Wd, data = seedplantsdata,
          correlation=corBrownian(phy=seedplantstree), method="ML")
null <- gls(Shade ~ 1, data = seedplantsdata,
            correlation=corBrownian(phy=seedplantstree),method="ML")
anova(wd,null)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## wd       1  3 222.0088 228.1380 -108.0044
## null     2  2 226.4988 230.5848 -111.2494 1 vs 2 6.489907  0.0108
```

You can see the model with the wood density variable is better than the model with only the intercept. However, as mentionned above, because the REML fitting provides better parameter estimates, you would have to refit the model using REML to present the results.

```
wd.final <- gls(Shade ~ Wd, data = seedplantsdata,
                correlation=corBrownian(phy=seedplantstree), method="REML")
summary(wd.final)
```

```
## Generalized least squares fit by REML
##   Model: Shade ~ Wd
##   Data: seedplantsdata
##        AIC      BIC    logLik
##   214.3762 220.3982 -104.1881
##
## Correlation Structure: corBrownian
##  Formula: ~1
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 0.911433  4.409058 0.2067184  0.8370
## Wd          4.361028  1.693349 2.5753865  0.0127
##
##  Correlation:
##    (Intr)
## Wd -0.166
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -0.26890642 -0.16431866 -0.02645422  0.09638984  0.34953444
##
## Residual standard error: 7.455109
## Degrees of freedom: 57 total; 55 residual
```

When should we use PGLS?

A very common mistake made when someone considers to use PGLS is to test for phylogenetic signal in Y or X using either Pagel's λ or Blomberg's K, and if they observe some phylogenetic signal, they use a PGLS to analyse their data. This is a ***big mistake***. As we saw earlier, PGLS corrects for phylogenetic correlation in the residuals and not in the variables. Therefore, the presence of phylogenetic signal in the variables does not necessarily mean that the residuals are phylogenetically correlated.

So what should we do then? It might be tempted to say to always use PGLS in every case. However, previous studies have shown that using PGLS when the residuals are not phylogenetically correlated results in poor statistical performance and inflated type I error (e.g., Revell 2010). One approach proposed by Revell (2010) is to always fit Pagel's λ with the PGLS model. Consequently, if the residuals are not phylogenetically correlated, λ will be close to 0 and the model will essentially be non-monophyletic. And when there is phylogenetic signal in the residuals, the model will be statistically correct. Therefore, this is a win-win situation! An alternative would be to use model testing, as we did above, to decide whether it is worth it to use PGLS despite the extra parameters to fit.

Your turn

Now that we have gone through PGLS using the example of explaining shade tolerance with wood density, you can do the same with by trying to exaplain seed mass (Sm) by with leaf nitrogen content (N), using the

same dataset. How do the results compare between normal linear regression and PGLS?

Further readings

To undertand well a new research field, it is always advisable to read a lot on it. Here are some references that you might find useful. The different sources also sometimes explain the theory in different ways or use different examples, which might help you understand better.

- Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist* 125, 1-15. **The classic initial paper that launched the field of comparative analyses. The phylogenetic independent contrasts are introduced here**
- Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer Associates, Inc. Sunderland, MA. **A thorough reference on phylogenies, from reconstruction to phylogenetic methods**
- Paradis, E. (2012). *Analysis of phylogenetics and evolution with R*. New York, USA: Springer. **This is the book that explains the analyses available in the R package APE. It is also a great reference on many phylogenetic analyses, including the comparative method. This is a classic and a must for users of phylogenies in R.**
- Revell, L J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* 1: 319-329. **A great paper on PGLS. It uses simulations to show when it is important to use PGLS.**
- Zuur, A.F., E.N. Ieno, N. Walker, A. A. Saveliev, G.M. Smith. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer New York. **This is not a book on phylogenetic methods, but it is a great book on the analysis of ecological data with examples in R. Its chapter 6 and 7 discuss correlation structures and although they are not about phylogenies, they are very instructive on how to deal with them and how to compare models and analyse complex data. It also has tons of information on how to deal with more complex data, along with correlation structure. A very good read!**

References

Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist* 125: 1-15.

Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer Associates, Inc. Sunderland, MA.

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

Paradis, E. (2012). *Analysis of phylogenetics and evolution with R*. New York, USA: Springer.

Revell, L J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* 1: 319-329.

Rohlf, F.J. (2001). Comparative Methods for the Analysis of Continuous Variables: Geometric Interpretations. *Evolution* 55: 2143-2160

Zuur, A.F., E.N. Ieno, G.M. Smith. (2007) *Analysing Ecological Data*. Springer. 680 p.

Zuur, A.F., E.N. Ieno, N. Walker, A. A. Saveliev, G.M. Smith. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer New York.
