# An introduction to phylogenetic comparative methods (in R)

Simon Joly

Winter 2021

## Contents

# 1 Preface

This document consist in an introduction to the comparative methods. It contains theory as well as practical examples in R on Phylogenetic Generalized Least Squares (PGLS). It was developed for a half-day workshop that consists in short presentations followed by R exercises. Note that the present document should pretty much stand by itself because most of the theory given in the presentations are incorporated into the theory sections. Therefore, this document should contain all the necessary 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 linear models, mixed-effects models and model comparison. Good introductions to model fitting in R can also be found on Dolph Schluter's webpage and among the QCBS workshops.

# 2 Getting ready for the workshop

## 2.1 Install R and the required packages

To perform the examples of this document, you will need to have the R software installed on your computer. I also recommend that you install RStudio. If R Studio is not required, it facilitates interactions between scripts and the R console and provides many great tools.

After installing R, you will have to install some packages. For this specific tutorial, we will need to load the following R packages.

```r
library(nlme)
library(ape)
library(RColorBrewer)
library(ggplot2)
```

To execute the code of this tutorial in R, you can just copy and paste the code in the boxes in your R console. This will replicate the analyses presented in the tutorial.

If some of the packages above are not yet installed on your computer, you get error messages when trying to load them. If this is the case, you will have to install them using the function `install.packages()`. You only have to install them only once.

```r
install.packages('nlme')
install.packages('ape')
install.packages('RColorBrewer')
install.packages('ggplot2')
```

Once the packages are installed, you can load the packages using the `library()` function. 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.

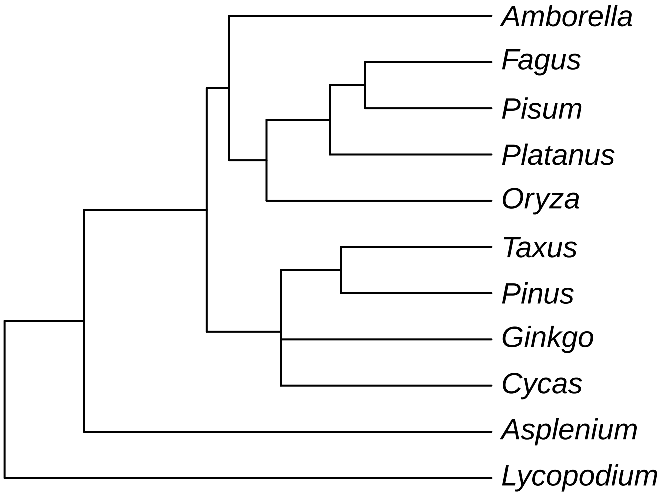
## 2.2 Working with github

All the data of the workshop are deposited in a github repository. The easiest way to access them is to go to the Workshop repository and clone it on your computer. The option to clone the directory is in a green box called "code".

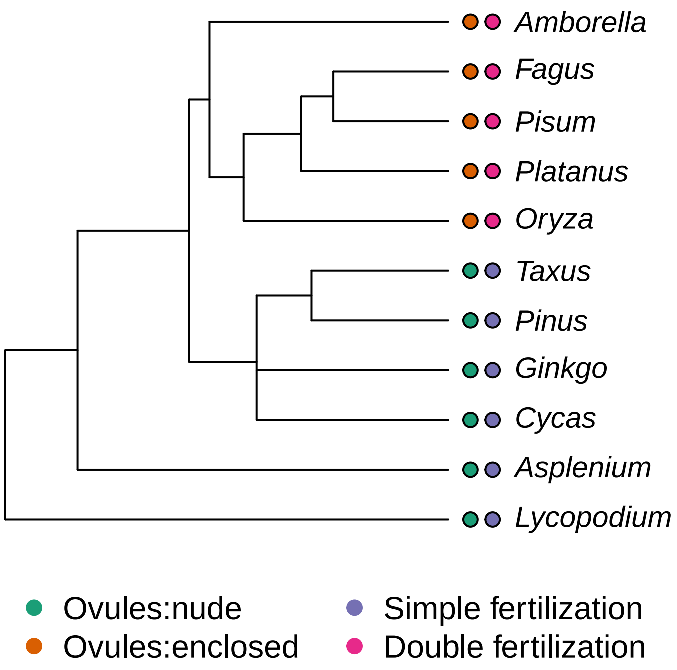
Once you have cloned the repository, you need to open R (R studio) and set the working directory to the repository folder ("/ComparativeMethods-HalfDayWorkshop"). In RStudio, to change the working directory, you choose `>Session>Set Working directory>Choose Directory`.You are now ready for the workshop.

## 3 An introduction to 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 live in the same environment but 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 than 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
```

The test suggests that the assotiation is highly 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 evolved 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. The contingency table when considering the changes along the branches looks like this:

| | Change in trait 2 | No change in trait 2 |
|---|---:|---:|
| **Change in trait 1** | 1 | 0 |
| **No change in trait 1** | 0 | 17 |

With this table, Fisher’s exact test will give the following result:

```
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. One useful and powerful approach is the Phylogenetic Generalized Least Squares (PGLS) and it is the one that will be introduced next. But first, let's do some revision and look briefly at the standard regression to better understand PGLS.

# 4 The linear regression model

## 4.1 Theory

The linear model has the following form:

$$\mathbf{y} = \alpha + \beta\mathbf{x} + \mathbf{e}$$

$\mathbf{y}$ is the response (or dependent) variable, $\mathbf{x}$ is the explanatory (or independent) variable, and $\mathbf{e}$ represent the residuals or in other words the information unexplained by the model. For a simple regression model, this represents the distance along the y axis between the observations and the regression line. The parameters $\alpha$ and $\beta$ are the population intercept and slope, respectively, and are unknown. In practive, you take a sample of size $N$ and you get estimates $\hat{\alpha}$ and $\hat{\beta}$ for the intercept and the slope, respectively. When the linear regression is standardly fitted using ordinary least squares (OLS), the residuals $\mathbf{e}$ are assumed to be normally distributed with expectation 0 and variance $\sigma^2$. In mathematic terms, $\mathbf{e} \sim N(0, \sigma^2)$.

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 $\mathbf{y}_i$ value at $\mathbf{x}_i$ is influenced by other $\mathbf{x}_i$. Obviously, this can happen with phylogenetically structured data as a response variable is be more likely to react similarly to an explanatory variable if they are closely related species. In other words, the y value for a species in not completely independent from the y value of a closely related species: theirs y are correlated. We'll illustrate this in an example below.

## 4.2 Practice

To provide pratical examples in this workshop, 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 *rbc*L and *mat*K. The dataset you need to run the examples are already in the /data/ folder of the github repository. However, you can also download them by clicking on the links below.

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 you are in the main directory of the workshop folder
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)
# Remove unnecessary object
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). In R, a very simple way to do a regression is to use the function 'lm', which stands for linear model. To fit a linear model, you need to tell the `lm` function which variable is the response variable and which one is the explanatory variable. This is done using formulas in the form `Shade ~ Wd`. The variable at the left of the tilde ('~') is the response variable (`Shade`) whereas the explanatory variale (1 or more) are at the right of the tilde.

```
# Fit a linear model using Ordinary Least Squares (OLS)
shade.lm <- lm(Shade ~ Wd, data = seedplantsdata)
# Print the results
summary(shade.lm)
```

```
##
## Call:
## lm(formula = Shade ~ Wd, data = seedplantsdata)
```

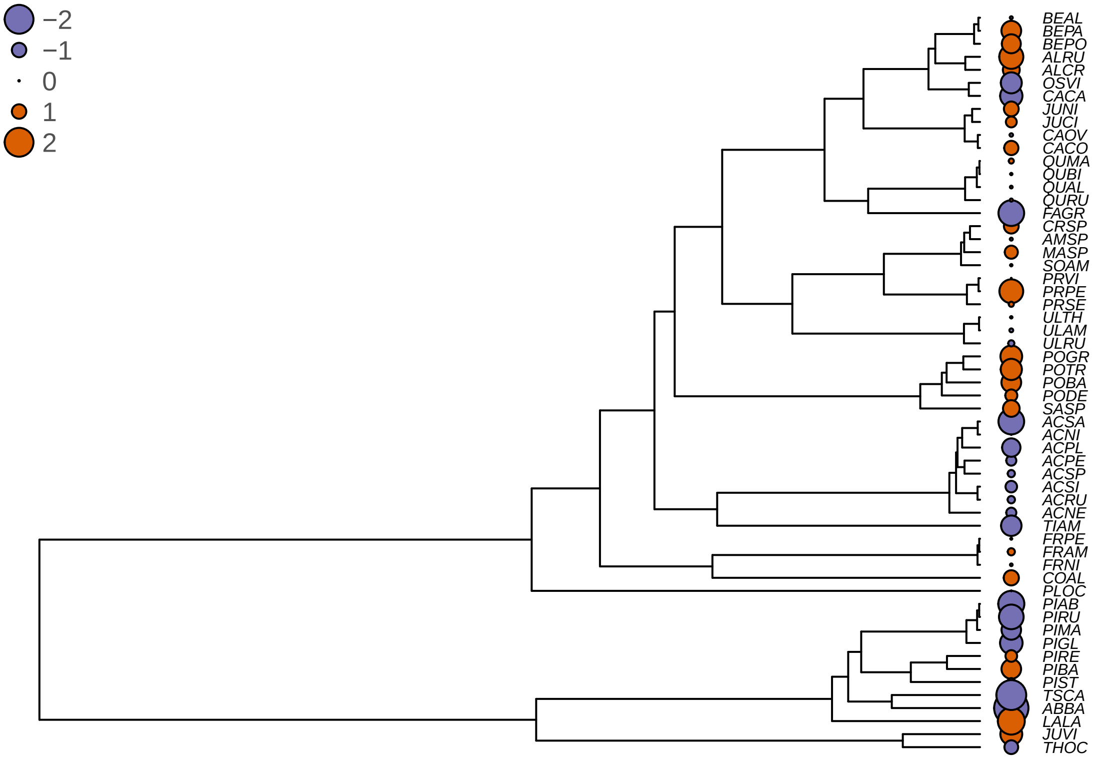
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87120 -1.02501  0.05628  0.70132  2.38261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0010     0.7501   2.668    0.010 *
## Wd            1.8130     1.5676   1.157    0.252
## ---
## 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 estimate (here the parameter `Wd`) is 1.81 and that 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, another way that the assumption of independence can be violated is if the residuals are phylogenetically correlated. One way to test this is to plot the residuals at the tips of 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))
# Vector of 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.
# The plot command plots the tree and leaves some space to plot the
#  residuals at the tips with the 'label.offset=0.01' option
plot(seedplantstree,type="p",TRUE,label.offset=0.01,cex=0.5,no.margin=FALSE)
# The next command plots the residuals. the 'bg' option is for the background color. If
#  the residuals are greater than 0 (shade.res>0), it will print the first colour (1) of
#  the 'cols' array and if it is below zero, it prints the second color (2). The the size
#  of the circle (the 'cex' option) is relative to the absolute value of the residuals
#  (abs(shade.res). To plot other values, just replace the 'shade.res' vector by another one.
tiplabels(pch=21,bg=cols[ifelse(shade.res>0,1,2)],col="black",cex=abs(shade.res),adj=0.505)
# Print the legend
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 (they are of the same color, which means that they are of the same side of the regression slope). This is problematic. Indeed, it shows that the assumption of independence of the ordinary least squares (OLS) regression no longer holds and the statistical tests for the null hypotheses are no longer valid. We will see next how phylogenetic generalized least squares can correct this.

### 4.3 Challenge 1

In the `seedplantsdata` data frame, there were many different traits. Try to fit a regression of tree shade tolerance (`shade`) on the seed mass (`Sm`). In other words, test if shade tolerance can be explained by the seed mass of the trees. Then, try to see if the residuals are phylogenetically correlated.

## 5 Phylogenetic generalized least squares (PGLS)

### 5.1 Theory

Phylogenetic generalized least squares (PGLS) is just a specific application of the broader method called generalized least squares (GLS). Generalized least squares relax the assumption that the error of the model has to be uncorrelated. 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.

GLS have the same structure as Ordinary Least Squares (OLS):

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

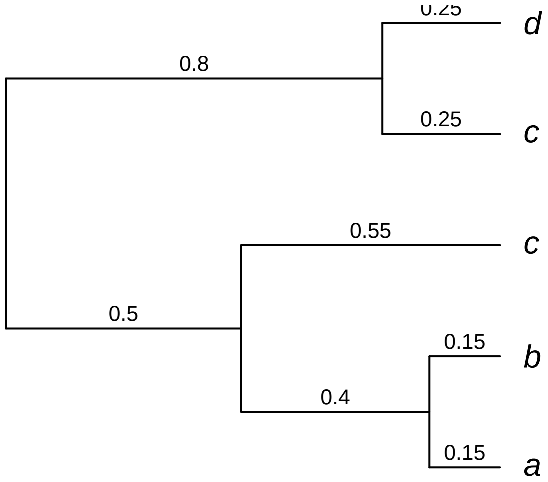
The only difference is that the residuals are correlated with each other according to a correlation structure **C**:

$$\epsilon_i \sim N(0, \sigma^2\mathbf{C})$$

Here, **C** is a correlation matrix that describes how the residuals are correlated with each other. To be able to account for phylogenetic relationships in a PGLS, we thus need to be able to express the phylogenetic relationships in the form of a correlation matrix.

### 5.1.1 Phylogenetic correlation structure

Phylogenetic relationships can be described using a correlation structure. Below, you have phylogenetic tree with branch lengths indicated above the branches.



Now, this tree can be perfectly represented by a variance-covariance matrix.

```
##      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 diagonal elements of the matrix are the species variances; these numbers represent the total distance from the root of the tree to the tips. It determines how much the tips have evolved from the root. The off-diagonal elements 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. The greater the covariance, the longer the two species have shared the same evolutionary history.

> 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 `chronopl` 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 in newick format
atree <- "(((a:0.15,b:0.15):0.4,c:0.55):0.5,(c:0.25,d:0.25):0.8);"
# Let's now read this tree and store it as a phylogenetic tree object in R
atree <- read.tree(text=atree)
# Extract the variance-covariance matrix
varcovar <- vcv(atree)
# Print the variance-covariance matrix
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`.

```
# Convert the covariance matrix to a correlation matrix
corrmat <- cov2cor(varcovar)
# Print the matrix, rounding the numbers to three decimals
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.

```
# Obtaining a correlation matrix using the 'vcv' function
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
```

Now that we know how to obtain a correlation matrix from a phylogenetic tree, we are ready to run a PGLS.

## 5.2 Challenge 2

Can you get the covariance matrix and the correlation matrix for the seed plants phylogenetic tree from the example above (`seedplantstree`)?

## 5.3 Practicals

There are several ways to run PGLS 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 were fitted using different packages.

```
require(nlme)
shade.pgls0 <- gls(Shade ~ Wd, data = seedplantsdata)
summary(shade.pgls0)
```

```
## 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. However, there are some differences. One is the presence of the item "Correlation:" that gives the correlation among the estimated parameters. Also, the "Standardized residuals" are the raw residuals divided by the residual standard error (the raw residuals can be output with residuals(shade.gls,"response")).

Now, let's run a PGLS model. To assign the correlation matrix to the gls function, you simply need to use the corr option of the gls function. You need to pass a specific correlation function so that R can calculate the model correctly. There are several different types of correlation structures that are available in R. We will start by using one of the simplest one, 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
```

```r
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.

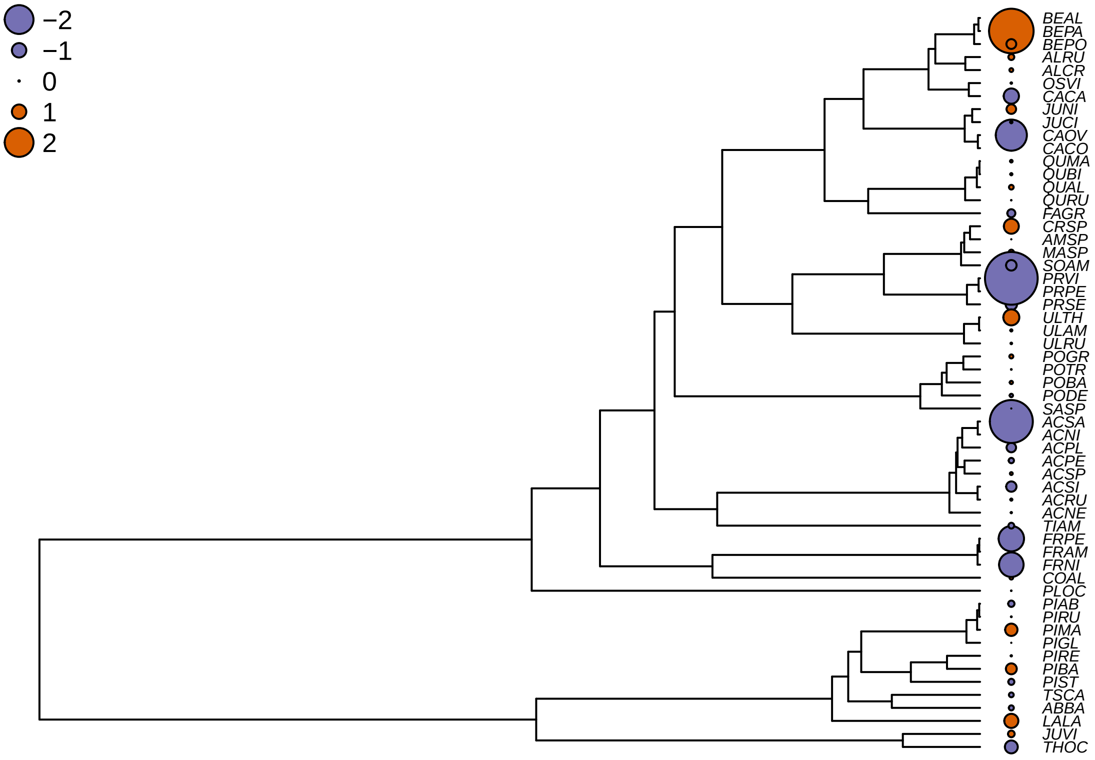
The output is similar to that of the model without the correlation, except for the output of the correlation matrix.

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 except for using pgls1.res as residuals
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.

### 5.3.1 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 using the corBrownian structure.

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

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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 outputed 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 Brownian motion to model the evolution along the branch of the tree. This is often refferred as a neutral model. If you want to know more about the Brownian Motion model, you can look at the section on this model at the end of the tutorial.

## 5.4 Challenge 3

Fit a PGLS model to see whether the seed mass (Sm) explains shade tolerance (Shade) with the seedplantdataset. How does it compare to the results from the standard regression.

# 6 Phylogenetic Independent Contrasts

## 6.1 Practicals

Let's make a digression to look at Phylogenetic Independent Contrasts (PIC). PIC were the first comparative approach proposed to deal with phylogenetic correlations (Felsenstein, 1985). Although they are less powerful than PGLS, they give the same results. Let's see how they can be used.

Phylogenetic independent contrast are estimated one trait at a time. They essentially transform the observed trait in contrasts that are not correlated with the phylogeny. This can be done in R using the `pic` function of the `ape` package.

```
# Estimate PIC for shade tolerance
Shade.pic <- pic(seedplantsdata$Shade, phy=seedplantstree)
# Estimate PIC for Wood density
Wd.pic <- pic(seedplantsdata$Wd, phy=seedplantstree)
```

Once this is done, the only thing to do is to fit a regression between these contrast. Note that it is important that the intercept is fixed to 0 in the model. This is done by adding `- 1` to the right of the formula.

```
# Estimate PIC for shade tolerance
pic.results <- lm(Shade.pic ~ Wd.pic - 1)
summary(pic.results)
```

```
##
## Call:
## lm(formula = Shade.pic ~ Wd.pic - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.943  -4.106   1.013   5.679  21.614
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## Wd.pic    4.361      1.693   2.575   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.21 on 55 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.09139
## F-statistic: 6.633 on 1 and 55 DF,  p-value: 0.01273
```

You can see that the slope estimate, `4.361`, it identical to the slope estimate obtained with PGLS. Same thing for the p-value. The main retriction with PIC is that you are limited in always comparing two variables. Much more flexibility is possible with PGLS.

# 7 Relaxing the assumption that all residuals need to be phylogenetically correlated

Phylogenetic Generalized Least Squares assume that the residuals are all phylogenetically correlated. This is relatively constraining because it means that other sources of errors that are not phylogenetically correlated are not allowed by the model. Morever, if these exist, they can bias the results of the PGLS.

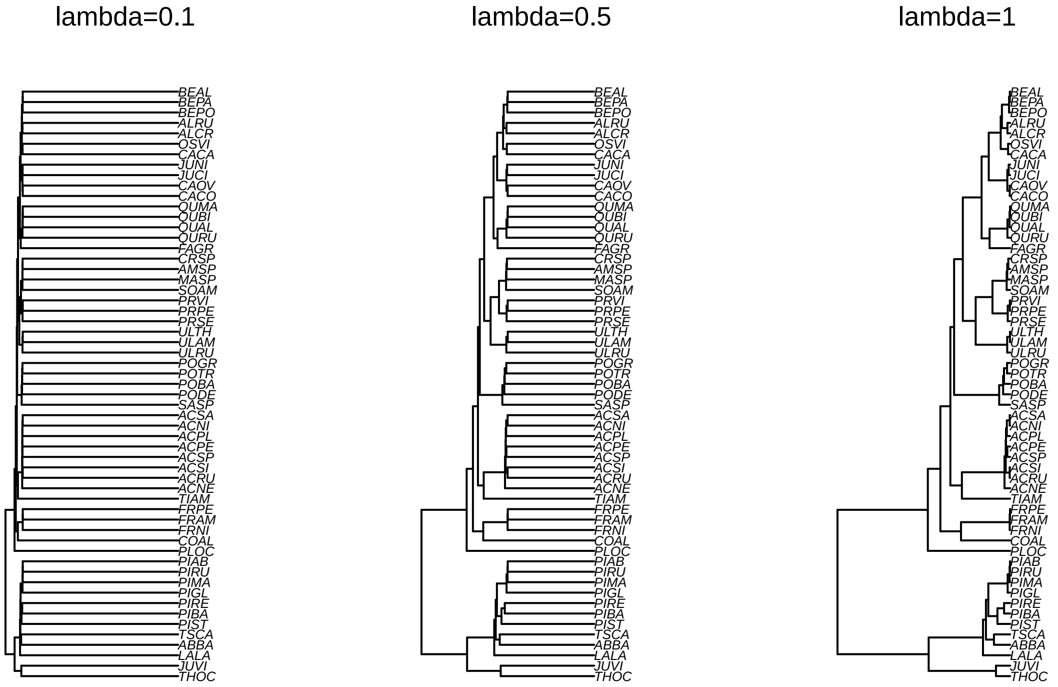
There are ways to relax this assumption, and one of this is to use a type of correlation structure that allows to relax this assumption.

## 7.1 Theory: 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 account for this using the $\lambda$ model 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 $\lambda$, but not the diagonal values. This essentially leads to a modification of branch lengths of the phylogeny. A $\lambda$ 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 (the correlations are reduced). At the opposite, if $\lambda$ is close to 1, then the modified phylogeny ressembles the true phylogeny. Indeed, the parameter $\lambda$ is often interpreted as a parameter of phylogenetic signal; as such, a greater $\lambda$ value implies a stronger phylogenetic signal.

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



You can see that with small values of lambda, the weight given to the shared history (the phylogeny) are greatly reduced and leaves long terminal branches on the tree. These long terminal branches can be seen as errors that are present in the estimates for each species but that are independent of the phylogeny (such as measurement errors for instance).

## 7.2 Practicals

Pagel's $\lambda$ 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 $\lambda$.

```
# Get the correlation structure
pagel.corr <- corPagel(0.3, phy=seedplantstree, fixed=FALSE)
```

The value given to `corPagel` is the starting value for the $\lambda$ parameter. Also, note that the option `fixed=` is set to `FALSE` This means that the $\lambda$ parameter will be optimized using generalized least squares. If it was set to TRUE, then the model would be fitted with the starting parameter, here 0.3.

Let's now fit the PGLS with this correlation structure.

```
# PGLS with coraPagel
shade.pgls2 <- gls(Shade ~ Wd, data = seedplantsdata, correlation=pagel.corr)
```

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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 $\lambda$ parameter, which is 0.958 here. Because the estimated $\lambda$ 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 $\lambda$ estimated would have been close to 0, it would have suggested that the PGLS is not necessary.

## 7.3 Challenge 4

Try to fit a PGLS with a Pagel correlation structure when regressing Shade tolerance on seed mass. Are the residuals as phylogenetically correlated than in the previous regression with wood density?

## 7.4 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 with parameter $\alpha$ 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.

# 8 Extending PGLS... phylogenetic ANOVA

## 8.1 Practicals

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"))
# Look at the result
seedplantsdata$Wd.cat
```

```
##  [1] light light dense light dense dense dense light light light light dense
## [13] dense light light dense dense dense dense dense dense dense light dense
## [25] light dense light light dense dense light light light light light light
## [37] light light light light light light light light light dense dense dense
## [49] dense light light light light light light light dense
## Levels: light dense
```

We can now fit a phylogenetic ANOVA.

```
# Phylogenetic ANOVA
shade.pgls3 <- gls(Shade ~ Wd.cat, data = seedplantsdata, correlation=pagel.corr)
```

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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.

# 9 Model testing

## 9.1 Practicals

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 $\lambda$ 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")
```

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

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

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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 $\lambda$ 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")
```

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

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

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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")
```

```
## Warning in Initialize.corPhyl(X[[i]], ...): No covariate specified, species
## will be taken as ordered in the data frame. To avoid this message, specify a
## covariate containing the species names with the 'form' argument.
```

```
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
```

# 10 When should we use comparative methods?

Comparative methods should always be used when working with datasets that comprise multiple species. A good advice though is to use a method that allows the residuals of the model not to be all phylogenetically correlated, as when using the PGLS with the corPagelstructure or using the Phylogenetic Mixed Model. Previous strudies have shown that using such comparative methods results in more precise and accurate fixed effect estimation, lower type I error, and greater statistical power. Therefore, it is always advantageous to use these methods.

A common mistake made when someone considers to use PGLS is to test for phylogenetic signal in $Y$ or $X$ using either Pagel's $\lambda$ 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.

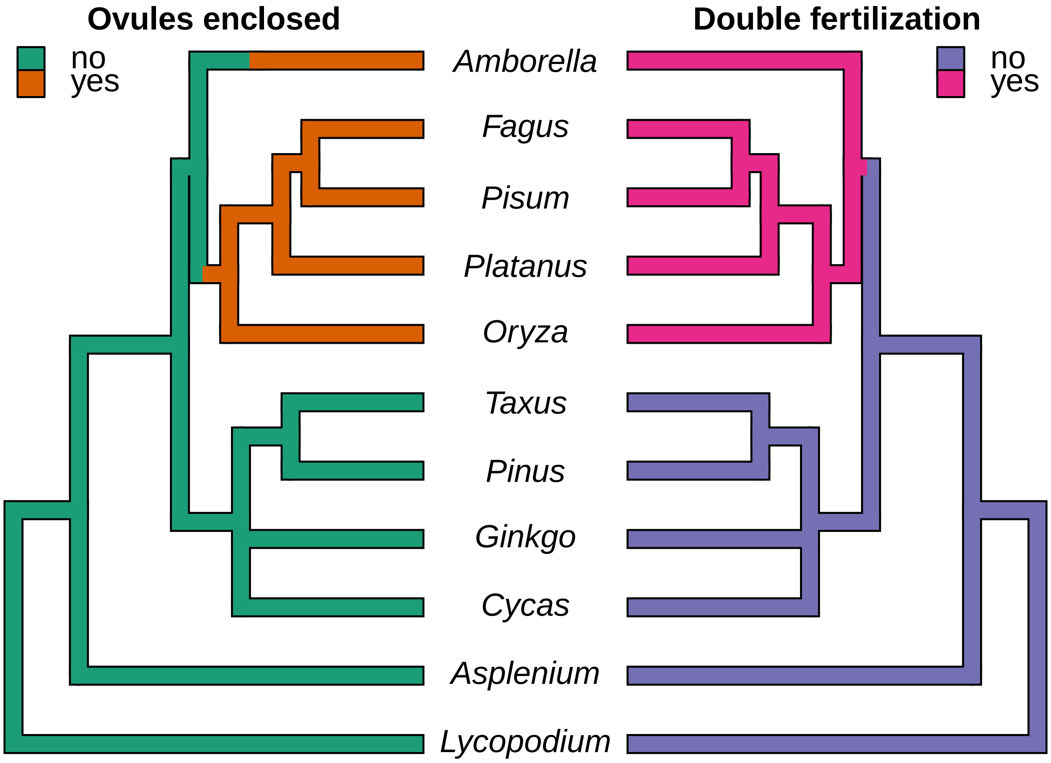
Another common misconception of comparative methods is that it removes all variation in the data related to the phylogeny and that this could affect the interpretation of the variable of interest. This was true of old methods like phylogenetic autoregression that first removed the phylogenetic signal from the data before analysing them. These approaches were indeed problematic. But the methods presented here to not suffer from these problems. They account for the phylogenetic structure and quantify it, but it does not removes variation from the model.

# 11 A final word: the problem of replication

Every biologist is well aware of the importance of replicating their experiments in orer to be confident in their conclusions. This is a lot more tricky when we consider evolution. To test our hypotheses on evolution, the ideal approach would be to rewind the "tape of evolution" (S. J. Gould) and let the story repeats itself several times to see what happens. This is unfortunately not possible, although some experimental evolution studies do manage to replicate evolutionary experiments.

The phylogenetic comparative method introduced in this tutorial is one appropriate approach to protect ourselves from reaching concusions that are not strongly supported when considered in an evolutionary context. However, even these approach can sometimes fail. This is why extra care is needed in such studies.

When interpreting their results, biologists should first ask whether they have sufficient replication in their data to allow strong conclusions to be made. And by replication, I mean evolutionary replication. Consider the seed plant example introduced above.



If there are multiple species with each state of the characters, the most parsimonious scenario is that each character evolved once along the branch of the tree. Therefore, the transition between the states of each character occured only once in the evolution of this group. In other words, there is no replication for these events in the evolution of these species. So even if the likelihood that those two events occured on the same branch is very small and even if a contingency test to calculate the likelihood of such an event is significant, this is a little bit like an experiment with one replicate. Therefore, even when a test that accounts for the phylogeny is significant, a lot of caution is needed when interpreting these results.

Ideally, before planning an experiment, one should make sure that there is sufficient replication in the evolution of the traits under study among the species considered to have greater confidence in the results. For instance, it would be much better if each character would have evolved 5-6 times each in the previous example, especially if the two characters were always evolving simultaneously!

## 12 Appendix 1: 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 $\sigma^2$ 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.

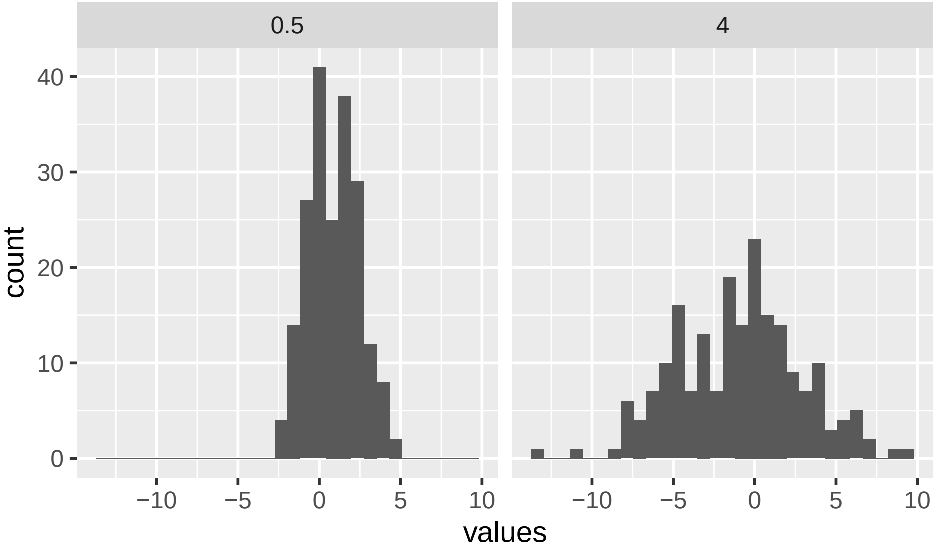
When applied to phylogenies, the Brownian motion model is kind of applied indepenpenty to each branch of the phylogeny. That allows to model the amount of change that occured along a given branch. If the variance of the Brownian motion model is $\sigma^2$ per unit of time $t$, then the net change along 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).

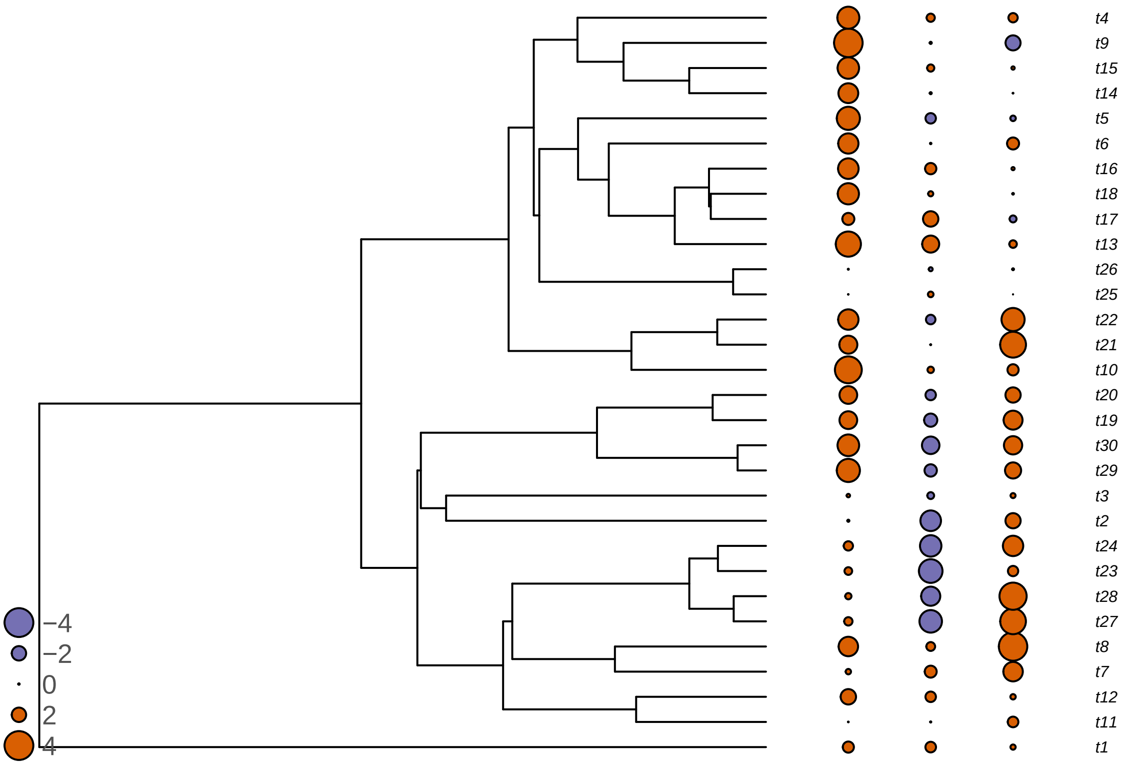
The parameter $\sigma^2$ in the model gives the variance, or in other word the speed of evolution. The higher the variance, the faster the character will evolve. Here are two examples of simulated characters on a tree of 200 species with $\sigma^2 = 0.5$ and $\sigma^2 = 4$.



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.



## 13 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**
- Hadfield, J. D., S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23:494–508. **This paper describes the phylogenetic mixed model and its implementation in MCMCglmm. It is a very important paper**
- Housworth, E.A., E.P. Martins, M. Lynch. 2004. The phylogenetic mixed model. *The American Naturalist* 163:84–96. **Excellent paper on the Phylogenetic Mixed Model**
- 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.**
- Villemereuil, P., S. Nakagawa. 2014. General quantitative genetic methods for comparative biology. Pp. 287–303 in L. Z. Garamszegi, ed. *Modern phylogenetic comparative methods and their application in evolutionary biology.* Springer-Verlag, Berlin, Heidelberg. **Nice book chapter explaining the phylogenetic mixed model**

- 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!**

# 14 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.

Hadfield, J. D., S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23:494–508.

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

Housworth, E.A., E.P. Martins, M. Lynch. 2004. The phylogenetic mixed model. *The American Naturalist* 163:84–96.

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

Villemereuil, P., S. Nakagawa. 2014. General quantitative genetic methods for comparative biology. Pp. 287–303 in L. Z. Garamszegi, ed. *Modern phylogenetic comparative methods and their application in evolutionary biology.* Springer-Verlag, Berlin, Heidelberg.

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.