# Introduction to Phylogenetic Comparative Methods

*Presented by Maxwell J. Farrell*

*May 18 2018*

Presented as part of the *IDEAS Computational Modeling Workshop 2018* http://ideas.ecology.uga.edu/computational-summer-workshop.html The material presented here is largely comprised of workshop materials graciously published by leading researchers on phylogenetic comparative methods. The original materials are:

- Simon Joly's course on Comparative Methods
- Simon Joly's full day workshop on phylogenies and statistics developed for the QCBS
- Graham Slater's tutorial on macroevolutionary models for AnthroTree 2014
- Workshop on Comparative Methods in R - Ilhabela with exercises by Liam Revell and Luke Harmon
- Joseph Uyeda's SSB-PCM tutorial
- Steven Kembel's workshop on Biodiversity in R
- Unpublished code by William D. Pearse

### Libraries

```r
require(nlme)
require(ape)
require(picante)
require(geiger)
require(phytools)
require(mvtnorm)
require(brms)
```

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.

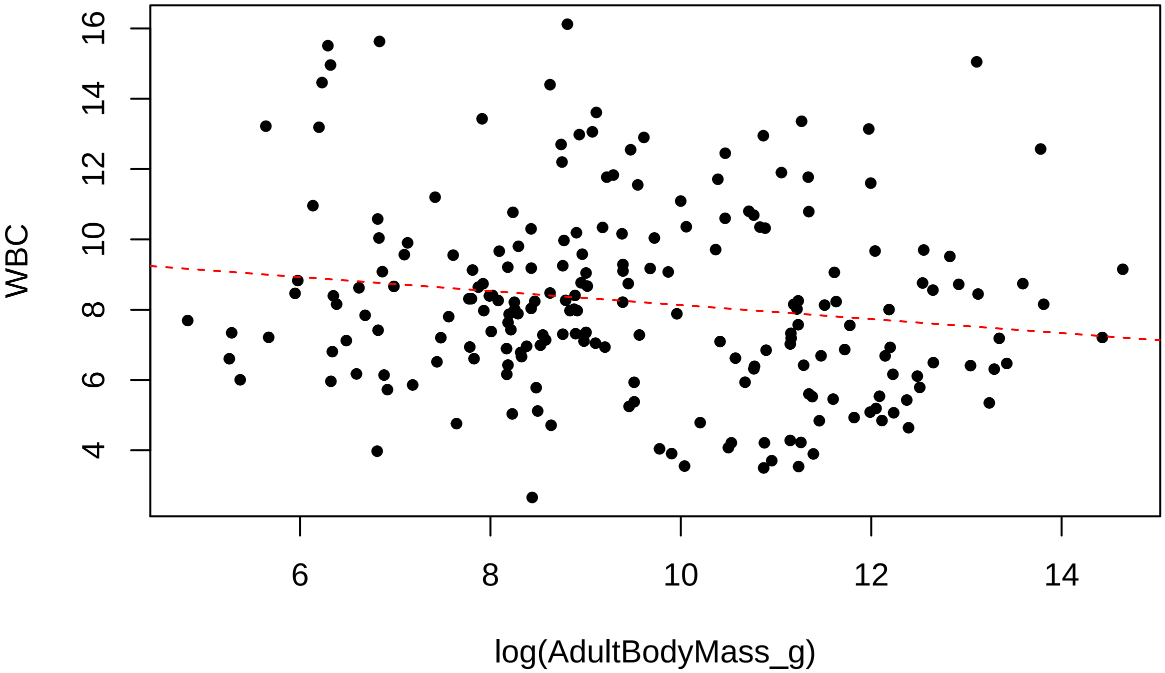
I first introduce comparative methods more generally before introducing PGLS, and then finish with slightly more flexible approaches for model fitting.

## Phylogenetic Comparative Methods

Phylogenetic comparative methods were introduced by Joseph Felsenstein in 1985. The idea of phylogenetic comparative methods is 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 relationship between body size and white blood cell count for a subset of Mammals from Cooper et al. 2012:

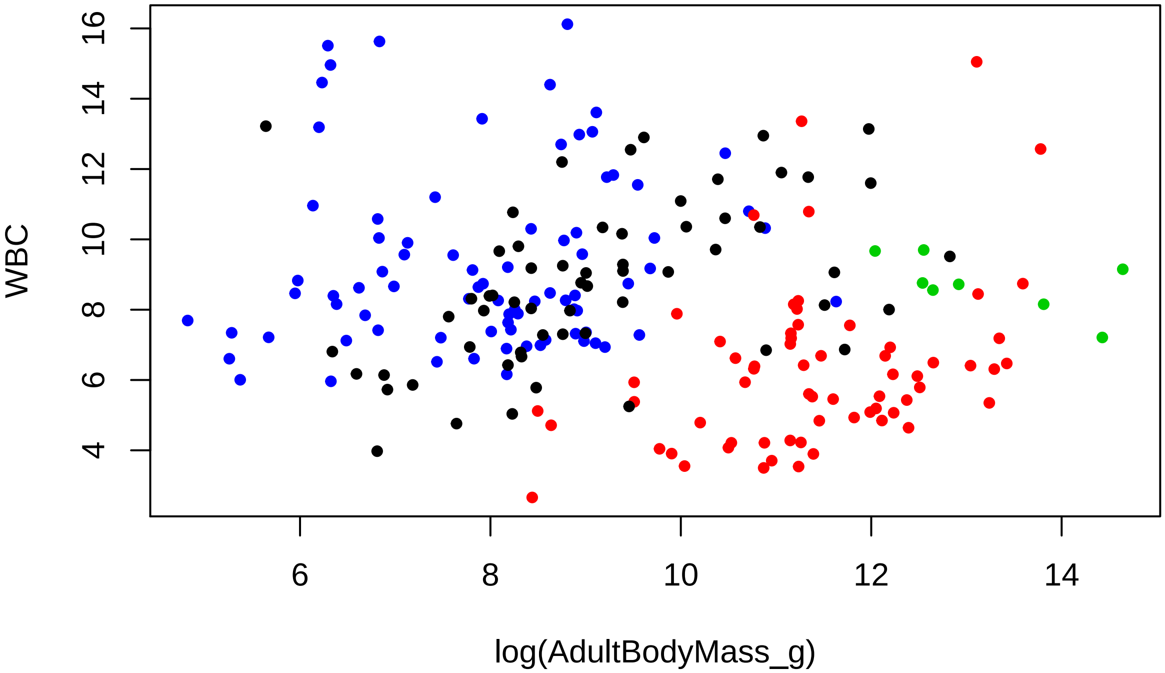
```
##
## Call:
## lm(formula = WBC ~ log(AdultBodyMass_g), data = dat)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7826 -1.7107 -0.3401  1.3251  7.7510
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          10.12399    0.83857  12.073   <2e-16 ***
## log(AdultBodyMass_g) -0.19924    0.08604  -2.316   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.646 on 211 degrees of freedom
## Multiple R-squared:  0.02478,    Adjusted R-squared:  0.02016
## F-statistic: 5.362 on 1 and 211 DF,  p-value: 0.02154
```



This would suggest that the association is significant. However, we know that the comparisons made are not completely independent because each point represents a different species.

Let's look at the points colored by Order:



> What do you notice? How do you think trend lines would be fit if we ran separate linear models for each order?

One useful and powerful approach to account for this problem of non-independence is Phylogenetic Generalized Least Squares (PGLS), but first let's get a sense of a simple model of trait evoultion and look briefly at the history of phylogenetic comparative methods.

## 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, and often refferred as a neutral model) is the Brownian motion model. This model of evolution is named after Robert Brown, a 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 applied independently to each branch of the phylogeny. This allows us to model the amount of change that has 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 $\sigma^2 t$.

Mathematically, if we let $X(t)$ be the value of the character at time $t$, then:

$$E[X(t)] = X(0)$$

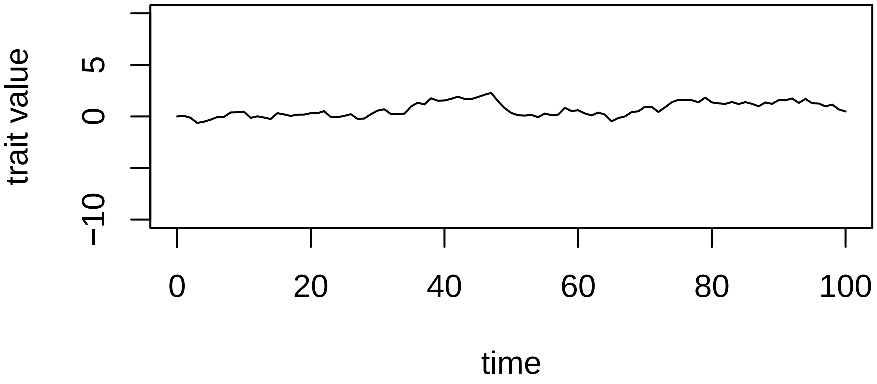
$$X(t) \sim N(X(0), \sigma^2 t)$$

Importantly, this model assumes that:

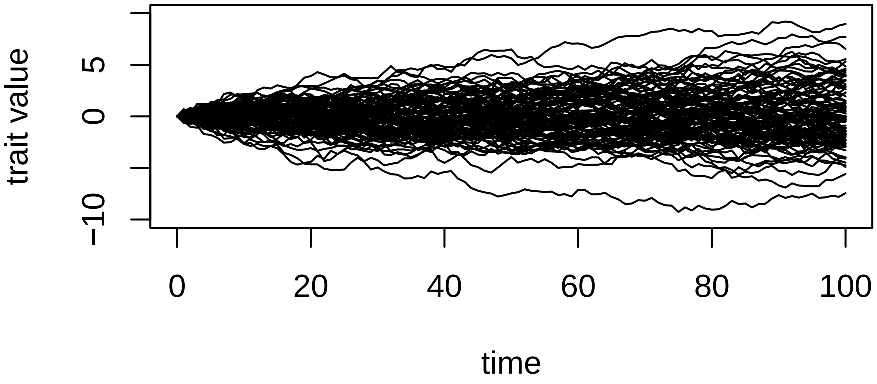
1. Evolution occuring in each branch of the phylogeny is independent of that occuring in other branches.
2. Evolution is completely random (i.e., no selection).

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 is an example of a simulated character with $\sigma^2 = 0.1$.

```r
interval.length <- 1
times <- seq(0, 100, interval.length);

rate <- 0.1
root <- 0

normal.dev <- c(root, rnorm(n=(length(times)-1), mean = 0, sd = sqrt(rate * interval.length)))

# take the cumulative sum and add in the root state to get trait values through time

traits <- cumsum(normal.dev);

plot(times, traits, type = "l", xlab = "time", ylab = "trait value", ylim = c(-10,10))
```



That's just one realization. But we can generate lots of realizations of BM and see what happens:



The shape looks a lot like a normal distribution - in fact the outcomes of a simple bm process are normally distributed. In this case, we drew our evolutionary changes from a normal distribution; however it's worth noting that (due to the central limit theorem) regardless of the distribution, evolution will proceed by Brownian motion as the width of our timesteps decrease towards zero!

---

**CHALLENGE**

With 1000 Brownian Motion simulations (root=0 and rate = 0.1) over 100 timesteps each, plot the distribution of trait values at time 100 and calculate the mean and variance.

Bonus: Try re-doing this with evolutionary changes drawn from a non-normal distribution.

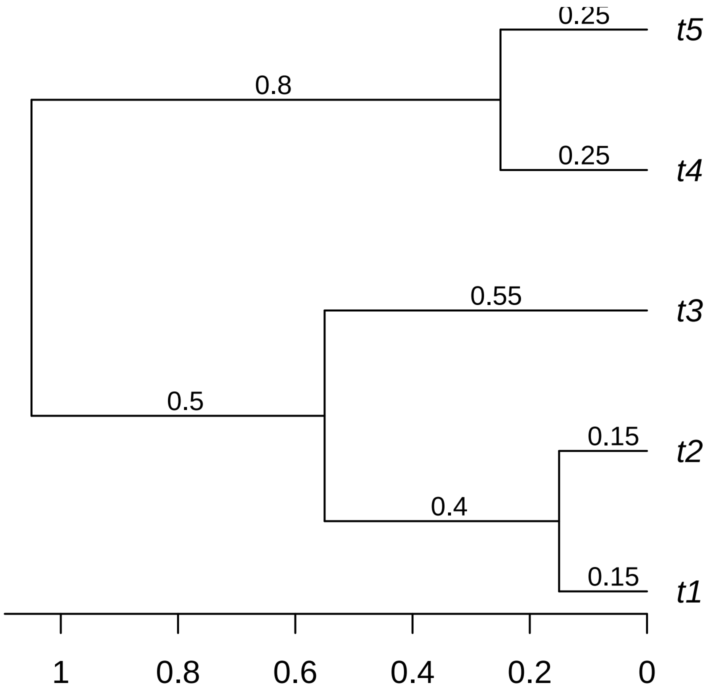
---

## BM on a phylogenetic tree

For evolution along a single branch of a phylogeny, trait values are draws from a normal distribution. But on the complete phylogeny, trait values at the tips may not be independent of one another due to shared evolutionary history.

The amount of shared evolution among species can be represented using variance-covariance matrix. Below, we generate a phylogenetic tree with branch lengths indicated above the branches, followed by a variance-covariance matrix that perfectly represents the tree.

```
phy <- "(((t1:0.15,t2:0.15):0.4,t3:0.55):0.5,(t4:0.25,t5:0.25):0.8);"
phy <- read.tree(text=phy)
plot(phy, label.offset=0.05)
edgelabels(c(0.5,0.4,0.15,0.15,0.55,0.8,0.25,0.25),adj=c(0.3,-0.3),frame="none",bg="",cex=0.8)
axisPhylo() # put up a scale bar
```



```
vcv(phy) # this is ape's vcv function
```

```
##      t1   t2   t3   t4   t5
## t1 1.05 0.90 0.50 0.00 0.00
## t2 0.90 1.05 0.50 0.00 0.00
## t3 0.50 0.50 1.05 0.00 0.00
## t4 0.00 0.00 0.00 1.05 0.80
## t5 0.00 0.00 0.00 0.80 1.05
```

The 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 amount 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 $t1$ and $t3$ 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. (You can test if your tree is ultrametric with `is.ultrametric`).

We can also use `cophenetic()` to generate a matrix that returns the phylogenetic distances among pairs of species (we'll see this again in section on community phylogenetics). If we divide by 2, we return the distances from species to their most recent common ancestor:

```
cophenetic(phy)/2
```

```
##      t1   t2   t3   t4   t5
## t1 0.00 0.15 0.55 1.05 1.05
## t2 0.15 0.00 0.55 1.05 1.05
## t3 0.55 0.55 0.00 1.05 1.05
## t4 1.05 1.05 1.05 0.00 0.25
## t5 1.05 1.05 1.05 0.25 0.00
```

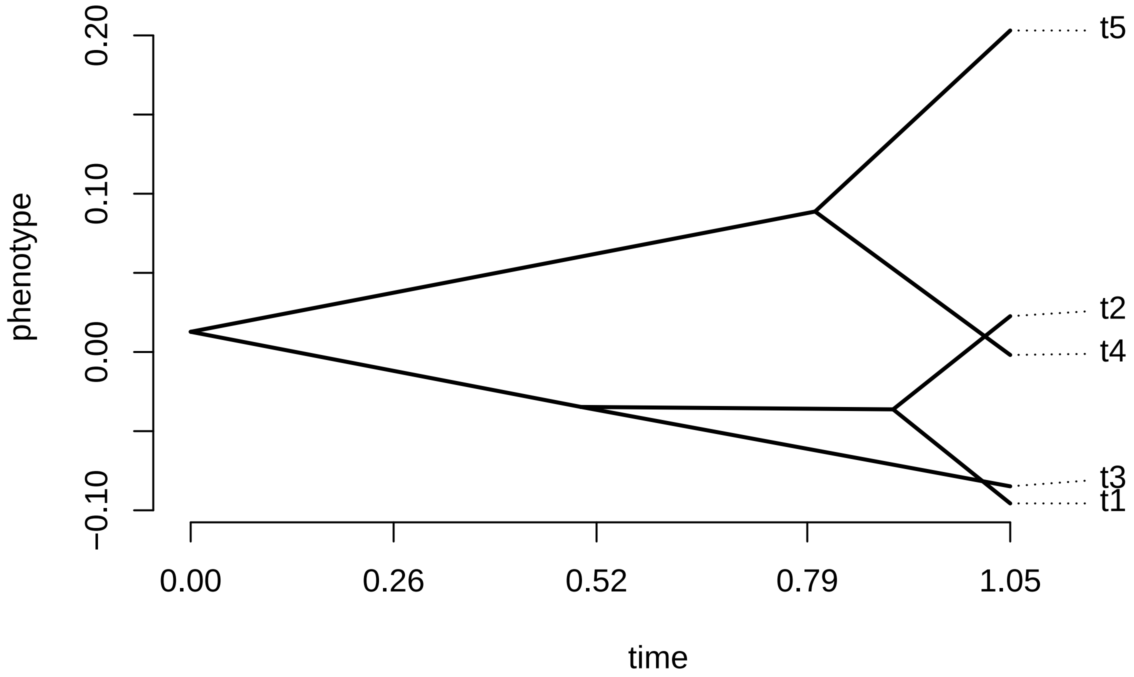
Remeber that the BM model is stochastic. That is, even if two closely related species are more likely to share similar character states than a distant one, this is only true on average. For any given simulated character, closely related species can sometimes be more different than to a distant species.

Let's simulate a continuous trait under BM on this tree using the root and rate paramters from our earlier simulation:

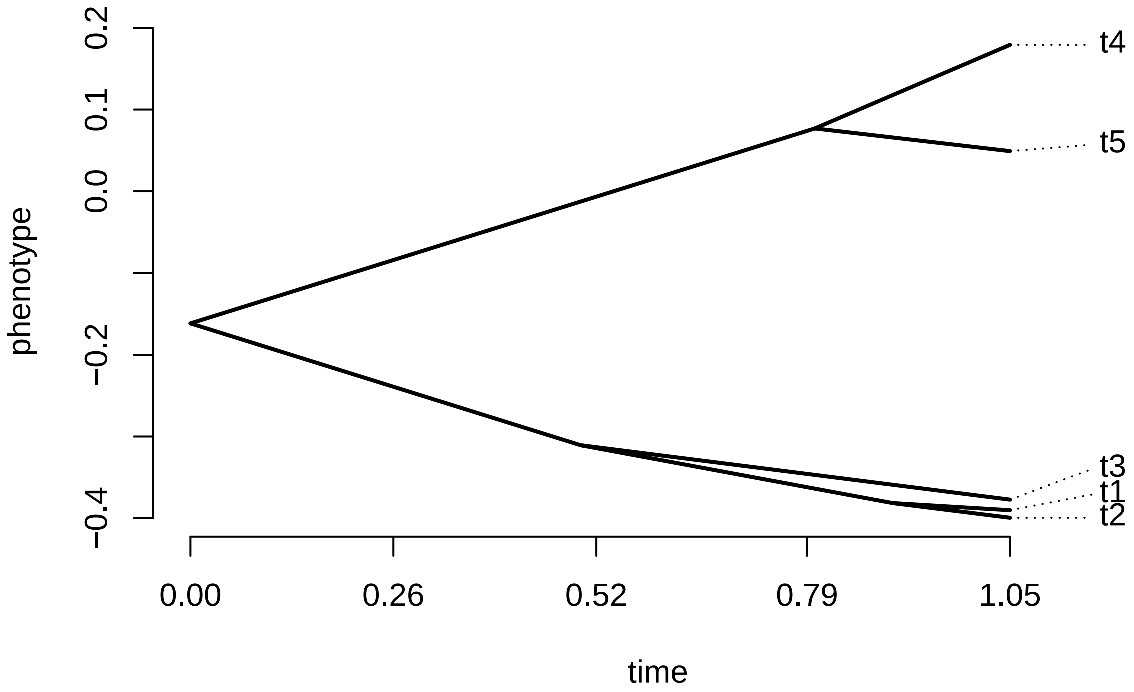
```
require(phytools)
set.seed(100)
d <- fastBM(tree=phy, a=root, sig2=rate)
d # take a look - we have five trait values
```

```
##           t1           t2           t3           t4           t5
## -0.095652720  0.022621124 -0.084861429 -0.001866986  0.203099740
```

```
# visualize trait evolution on the tree
phenogram(phy,d,spread.labels=TRUE)
```



```r
# Let's simulate another trait and compare to the first
d2 <- fastBM(tree=phy, a=root, sig2=rate)
phenogram(phy,d2,spread.labels=TRUE)
```



## CHALLENGE

For any given pair of roots and rates, we can compute the likelihood of observing our trait data using the multivariate normal density. The multivariate normal is similar to the normal but allows for a covariance structure among observations and / or different mean values.

Using the `dmvnorm` function from the `mvtnorm` package, find the maximum likelihood estimates for the root and rate parameters used to generate the traits in `d`.

```
# d was already generated above
d
```

```
##           t1           t2           t3           t4           t5 
## -0.095652720  0.022621124 -0.084861429 -0.001866986  0.203099740
```

```
# we will also need the variance covariance matrix
v <- vcv(phy)

require(mvtnorm)
# we can use dmvnorm to compute the likelihood of getting our data with our actual values
dmvnorm(x=d, mean = rep(root, length(phy$tip.label)), sigma = v * rate, log = T)
```

```
## [1] 1.535187
```

*Hint: we can optimize the likelihood (find the pair of parameters that give us the highest likelihood of observing our data) by trying different possible values of root and rate. You'll need to define some bounds for the search - examining all possible values of the root and rate would be unreasonable.*

---

## Fitting evolutionary models

Brownian Motion is just one of many evolutionary models (others include Ornstein-Uhlenbeck, early burst, speciational models). If you are interested comparing the likelihood of alternative models, we can use the `fitContinuous` function in the `geiger` package:

```
require(geiger)

bm <- fitContinuous(phy=phy, dat=d, model="BM")
bm
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'BM' model parameters:
##  sigsq = 0.029617
##  z0 = 0.012770
## 
##  model summary:
##  log-likelihood = 2.819652
##  AIC = -1.639304
##  AICc = 4.360696
##  free parameters = 2
## 
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  frequency of best fit = 1.00
## 
##  object summary:
##  'lik' -- likelihood function
##  'bnd' -- bounds for likelihood search
##  'res' -- optimization iteration summary
##  'opt' -- maximum likelihood parameter estimates
```

The resulting ML parameters can be accessed via `bm$opt`.

- `bm$opt$sigsq` is the rate
- `bm$opt$z0` is the root state
- `bm$opt$lnL` is the log likelihood
- `bm$opt$k` is the number of free (estimated) parameters. For BM this is just 2
- `bm$opt$aic` and `bm$opt$aicc` are the AIC and small-sample corrected AIC scores, which can be used to perform model selection among different evolutionary models.

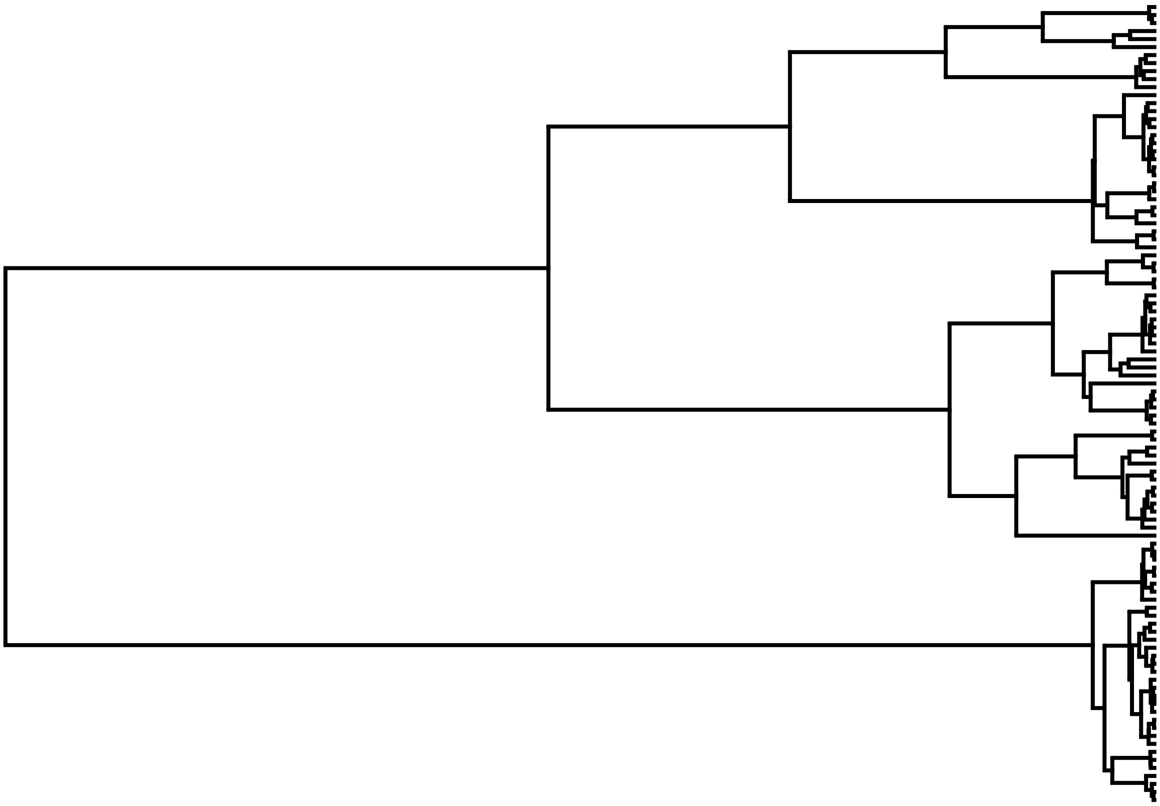
# Phylogenetic Independent Contrasts (PICs)

Phylogenetic independent contrasts (PIC) were introduced by Joseph Felsenstein in 1985. They were the first comparative method proposed, have been used many times since, and form the basis for many comparative methods.

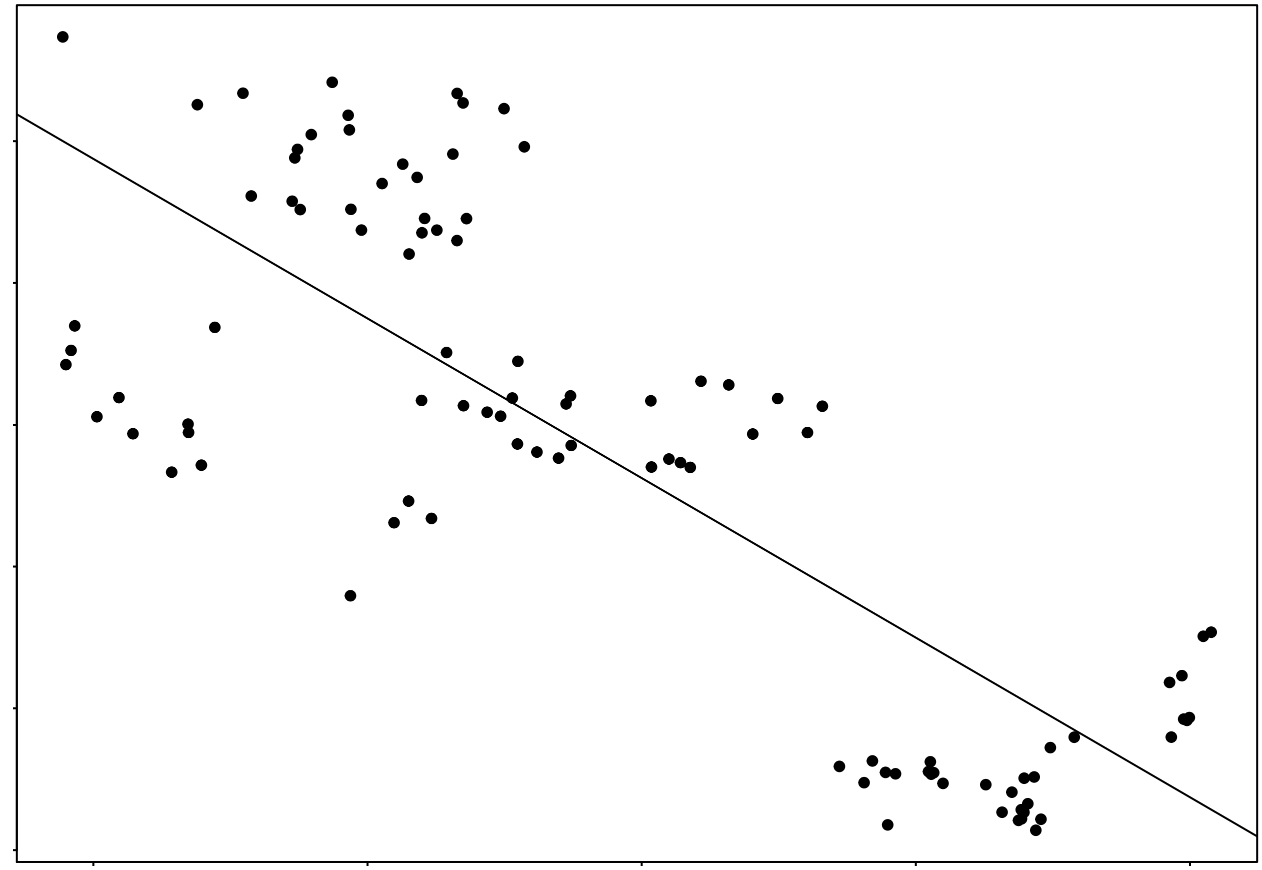
> With standard correlation we ask: can we predict $Y$ from $X$?
>
> We might be able to do this for two reasons: species are related, or $X$ and $Y$ tend to evolve together.
>
> With evolutionary correlation, $X$ and $Y$ evolve in a correlated fashion and when $X$ changes, $Y$ tends to change in a predictable way.

Let's examine the relationship between two independent BM traits using ordinary least squares (OLS):

```r
library(phytools)
set.seed(999)

## simulate a coalescent shaped tree
tree<-rcoal(n=100)
plotTree(tree,ftype="off")
```



```
## simulate uncorrelated Brownian evolution
x<-fastBM(tree, a=0, sig2=1)
y<-fastBM(tree, a=0, sig2=1)
plot(x,y,pch=20)
fit<-lm(y~x)
abline(fit)
```



```r
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0257 -0.8694  0.1011  0.8518  2.0404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.74971    0.12978   13.48   <2e-16 ***
## x           -1.12556    0.08575  -13.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 98 degrees of freedom
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6337
## F-statistic: 172.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

We can see from this example, that it is not difficult for phylogenic relationships to induce a type I error.

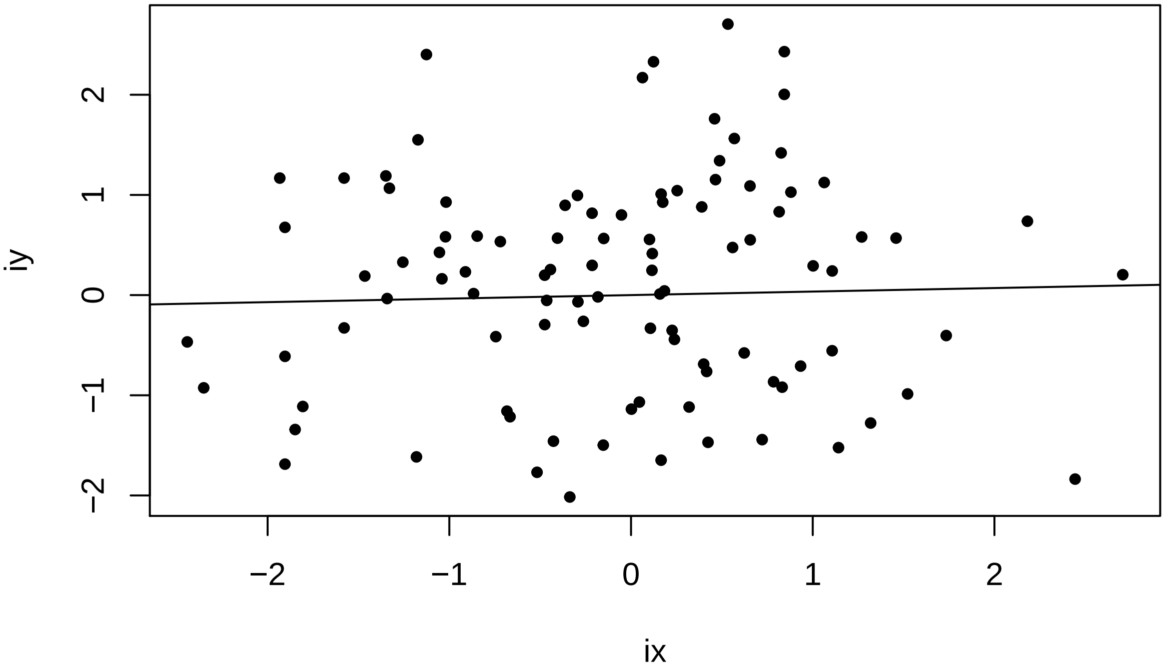
Here it is because clusters of closely related taxa have highly similar phenotypes. In other words, they are not independent data points about the evolutionary process for x and y on the tree.

Contrasts are essentially weighted averges of trait values for adjacent tips in the tree. They are then standardized by dividing the raw contrast by its variance.

Therefore, for each trait, a contrast will be calculated for each node in the tree. So if there are $n$ species in your tree, $n-1$ contrasts will be estimated. Note that contrasts are estimated for each character individually.

We will now calculate independent contrasts using the `pic` function in `ape`.

```
ix<-pic(x, tree, scaled=TRUE)
iy<-pic(y, tree, scaled=TRUE)
plot(ix,iy,pch=20)
fit<-lm(iy ~ ix - 1) ## we have to fit the model without an intercept term (this treats the contrasts a
abline(fit)
```



```
summary(fit)
```

```
##
## Call:
## lm(formula = iy ~ ix - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0037 -0.7224  0.2417  0.8879  2.6863
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## ix   0.0352     0.1056   0.333     0.74
##
## Residual standard error: 1.09 on 98 degrees of freedom
## Multiple R-squared:  0.001133,   Adjusted R-squared:  -0.00906
## F-statistic: 0.1111 on 1 and 98 DF,  p-value: 0.7396
```

If you have more of less traits in your own data, you will have to play with the `adj = c(-0.1, -2)` option to get a nice graphical representation. –>

We see that p > 0.05, so we do not reject the null hypothesis of no evolutionary correlation (i.e. we find there is no evidence for evolutionary correlation between these traits).

When taking phylogenetic information into account, these traits are not significantly related to eachother anymore. This means that the apparent correlation observed on the raw data was an artefact of their evolutionary histories. It is important to note that the application of PICs does not always make relationships less significant. Sometimes, it helps highlight significant relationships that were obscured by the evolutionary history of species.

---

**CHALLENGE**

Each standardized contrast is telling us something about the RATE of evolution. The contrasts have a close relationship with $\sigma^2$, the rate parameter from BM.

Show that the sum of the squared contrasts divided by $n$ gives the ML estimate of $\sigma^2$.

---

# Phylogenetic Generalized Least Squares

## OLS

We fit an ordinary least squares (OLS) regression model above to show how ignoring phylogenetic relationships can lead to type I error, but let's quickly review the theory behind linear models. A linear model can be written in 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 independent) variable, and $\epsilon_i$ is the residual of observation $i$, which represents unexplained variation.

The parameters $\alpha$ and $\beta$ are the population intercept and slope, respectively, and are unknown. In practice, 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 fit using ordinary least squares (OLS), the residuals $\epsilon_i$ are assumed to be normally distributed with expectation 0 and variance $\sigma^2$:

$$\epsilon_i \sim N(0, \sigma^2)$$

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.

Independence is violated when the $Y_i$ value at $X_i$ is influenced by other values of $X$. 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.

---

**CHALLENGES**

Show that OLS regression for BM traits simulated on a tree results in elevated type I error. Use 200 pure-birth trees. *HINT: mapply works well for this!*

Show that PIC regression results in correct type I error.

### CHALLENGE

The following code attempts to load the Cooper 2012 data and corresponding mammal phylogeny, then calculate independent contrasts for WBC and Mass from, however it causes multiple errors - correct them.

*NOTE: YOU MUST COMPLETE THIS CHALLENGE BEFORE MOVING ONWARDS*

```r
dat <- read.csv("data/Cooper_2012.csv", as.is=T)
fritz_tree <- read.nexus("data/Fritz_2009.tre")[[1]]

# Remove species for which we don't have complete data
dat <- na.omit(dat)

# Match data to tree names
species.to.exclude <- fritz_tree$tip.label[!(fritz_tree$tip.label %in% dat$Species_W.R05)]
tree <- drop.tip(fritz_tree,species.to.exclude)

# Order tree to make it nicer when plotting
tree <- ladderize(tree, right = FALSE)

# Name the rows of dat with the species codes remove obsolete columns
rownames(dat) <- dat$Species_W.R93
dat <- subset(dat, select=-c(Species_W.R05,Species_W.R93))

# Check that the order of the tree and data match
name.check(tree, dat)

# Great! Time for analysis!

wbcPic <- pic(wbc, tree, scaled=TRUE)
massPic <- pic(log(mass), tree, scaled=TRUE)
```

To verify, your mean values should be:

```
## [1] "mean wbcPic: 0.0212494457120913"
```

```
## [1] "mean massPic: -0.0168820061961508"
```

*Hint!* For many comparative methods, you must use a fully bifurcating tree with the traits in the same order as the tips on the tree, and with no missing data. With real data it is important to assign names to your character vectors. These will be used to match the names of the tips of the phylogeny.

## PGLS

Phylogenetic generalized least squares are very similar to PICs. The idea is the same, that is to remove the effect of the evolutionary relationships among species when fitting a regression between two variables. Phylogenetic generalized least squares (PGLS) is just a specific application of a more general method called generalized least squares (GLS). GLS relaxes the assumption that the error of the model has to be uncorrelated. They allow the user to specify the structure of the residual correlation. This is used, for instance, to correct for spatial correlation, time series, or phylogenetic correlation.

With GLS, 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.

The trick with PGLS is to give a covariance matrix that represents the evolutionary relationships between species. Depending on the model of evolution of the characters, the covariance matrix can be scaled using different approaches. For instance, one might assume that the character evolves under the Brownian motion model, or under other models of evolution, such as an Ornstein-Uhlenbeck model where the co-variance between two species decreases exponentially according to a parameter *alpha*. There are several correlation structures available in `ape`.

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

This is great, but for GLS we need a correlation matrix account for the correlation in the residuals. To obtain a correlation matrix from the variance-covariance matrix, we 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`, or directly from the function `vcv` by using the `corr=TRUE` option.

```
vcv(phy)
```

```
##      t1   t2   t3   t4   t5
## t1 1.05 0.90 0.50 0.00 0.00
## t2 0.90 1.05 0.50 0.00 0.00
## t3 0.50 0.50 1.05 0.00 0.00
## t4 0.00 0.00 0.00 1.05 0.80
## t5 0.00 0.00 0.00 0.80 1.05
```

```
# Convert the covariance matrix to a correlation matrix
corrmat <- cov2cor(vcv(phy))
# Print the matrix, rounding the numbers to three decimals
round(corrmat,3)
```

```
##       t1    t2    t3    t4    t5
## t1 1.000 0.857 0.476 0.000 0.000
## t2 0.857 1.000 0.476 0.000 0.000
## t3 0.476 0.476 1.000 0.000 0.000
## t4 0.000 0.000 0.000 1.000 0.762
## t5 0.000 0.000 0.000 0.762 1.000
```

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

```
##       t1    t2    t3    t4    t5
## t1 1.000 0.857 0.476 0.000 0.000
## t2 0.857 1.000 0.476 0.000 0.000
## t3 0.476 0.476 1.000 0.000 0.000
## t4 0.000 0.000 0.000 1.000 0.762
## t5 0.000 0.000 0.000 0.762 1.000
```

Now, the diagonal elements equal to 1, indicating that species are perfectly correlated to themselves.

Now we can use this correlation matrix to run a PGLS.

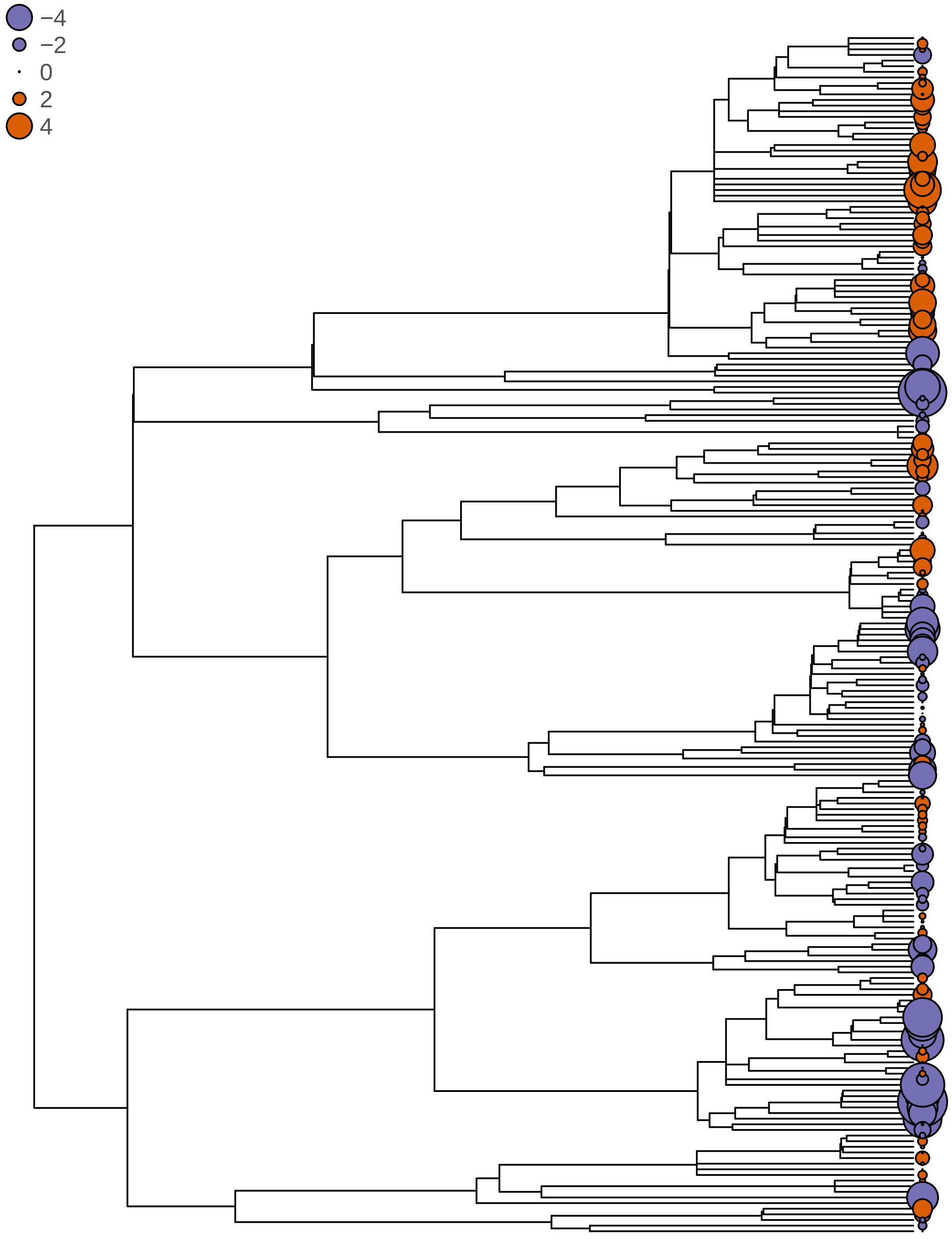
There are several ways to run PGLS in R. For instance, the package `caper` is a very well known package for PGLS. However, we will use the function `gls` here from the `nlme` package, which comes with the base packages in R. This function is robust and has the advantage to be very flexible. Indeed, it allows to easily use more complex models such as logistic or poisson regressions, and mixed effect models.

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)
wbc.gls <- gls(WBC ~ log(AdultBodyMass_g), data=dat)
summary(wbc.gls)
```

```
## Generalized least squares fit by REML
##   Model: WBC ~ log(AdultBodyMass_g)
##   Data: dat
##       AIC      BIC    logLik
##   1027.71 1037.765 -510.8548
##
## Coefficients:
##                          Value Std.Error   t-value p-value
## (Intercept)          10.123992 0.8385685 12.072946  0.0000
## log(AdultBodyMass_g) -0.199236 0.0860393 -2.315641  0.0215
##
##  Correlation:
##                      (Intr)
## log(AdultBodyMass_g) -0.976
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.1850304 -0.6463966 -0.1284923  0.5007144  2.9288004
##
## Residual standard error: 2.646484
## Degrees of freedom: 213 total; 211 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(wbc.gls,"response")`). Let's plot the residuals on the tree:



---

**CHALLENGE**

It looks like there is some strong phylogenetic non-independence in the residuals. Try extracting the residuals and plotting them on the tips of the phylogeny (re-create your own version of the plot above).

---

Let's try a PGLS. 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 $b$ and $a$. 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.

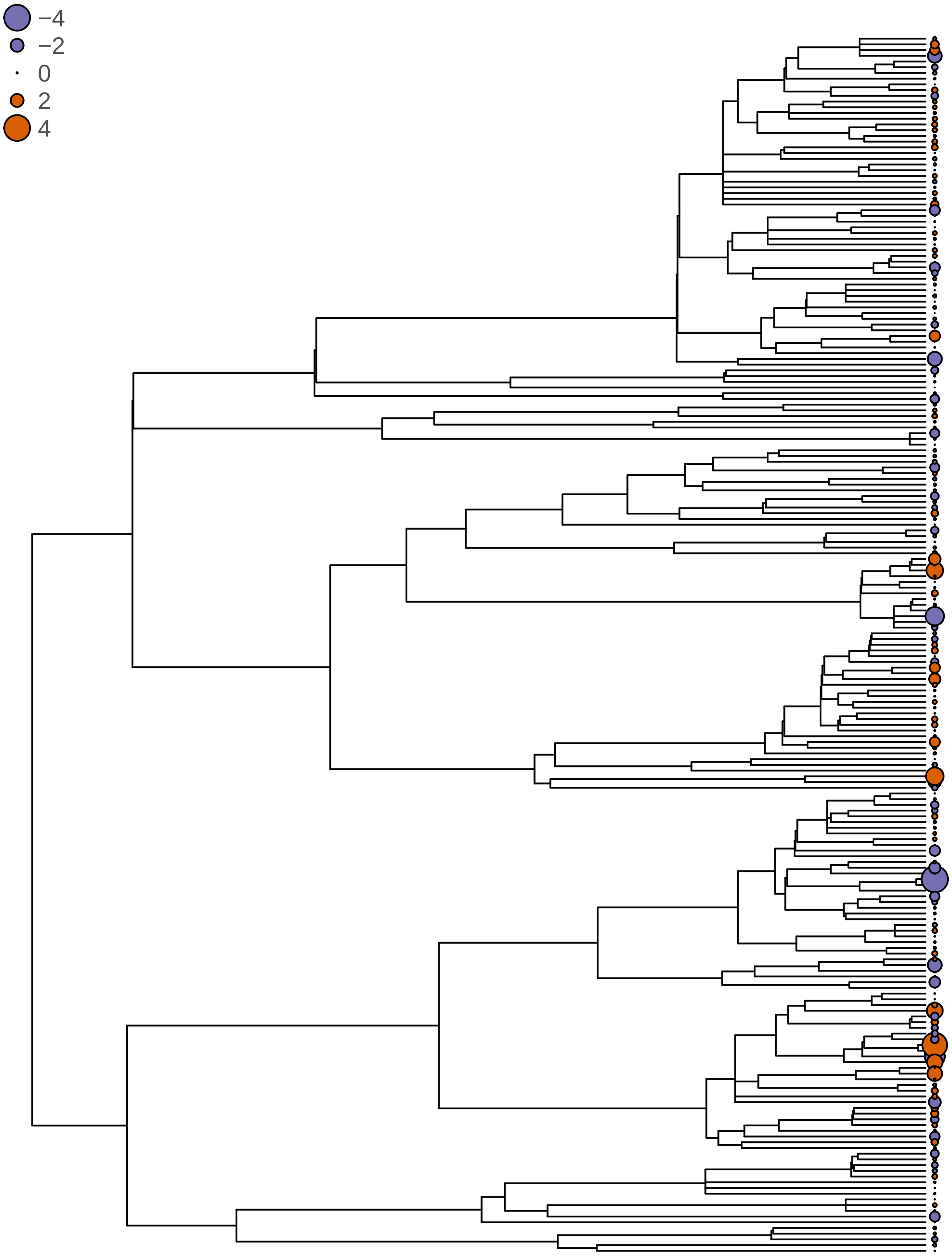
Note that the term `fixed=TRUE` in the corSymm structure indicates that the correlation structure is fixed during the parameter optimization.

```
# Calculate the correlation matrix from the tree
mat <- vcv(tree, corr=TRUE)
# Create the correlation structure for gls
corr.struct <- corSymm(mat[lower.tri(mat)],fixed=TRUE)
# Run the pgls
wbc.pgls1 <- gls(WBC ~ log(AdultBodyMass_g), data = dat, correlation=corr.struct)

# Note, summary(wbc.pgls1) returns the entire correlation matrix. In this example it is quire large wit
summary(wbc.pgls1)$tTable
```

```
##                          Value Std.Error  t-value     p-value
## (Intercept)          3.5801961 2.1075358 1.698759 9.08385e-02
## log(AdultBodyMass_g) 0.5723391 0.1357051 4.217520 3.66542e-05
```

Interestingly, you can see that the coefficient estimate for the slope is now positive (0.572) and larger than with standard regression, and also significant ($p$=0). This is a positive exmple of PGLS. Indeed, the relationship between white blood cell count and body size would have been inferred as negative when obscured by the phylogenetic correlation of the residuals. Once this correlation is accounted for, a significant positive relationship is revealed.

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 using `residuals(wbc.pgls1,type="normalized")`.



Compared to the ordinary least squares regression, 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 using the `corBrownian` structure.

```
# Get the correlation structure
bm.corr <- corBrownian(phy=tree)

# PGLS
wbc.pgls1 <- gls(WBC ~ log(AdultBodyMass_g), data = dat, correlation=bm.corr)
summary(wbc.pgls1)
```

```
## Generalized least squares fit by REML
##   Model: WBC ~ log(AdultBodyMass_g)
##   Data: dat
##        AIC      BIC    logLik
##   878.0487 888.1043 -436.0244
##
## Correlation Structure: corBrownian
##  Formula: ~1
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##                         Value Std.Error  t-value p-value
## (Intercept)          3.580196 2.1075358 1.698759  0.0908
## log(AdultBodyMass_g) 0.572339 0.1357051 4.217520  0.0000
##
##  Correlation:
##                      (Intr)
## log(AdultBodyMass_g) -0.613
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.4298577 -0.6649309 -0.1635332  0.2378296  1.8397768
##
## Residual standard error: 4.526799
## Degrees of freedom: 213 total; 211 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).

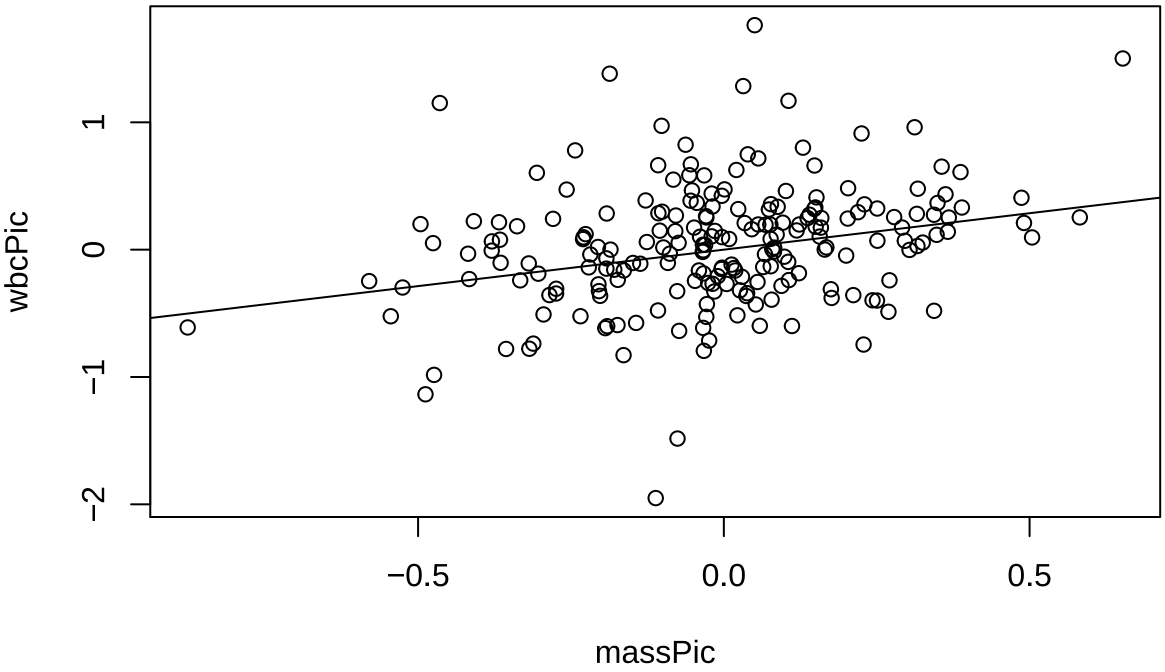
---

### CHALLENGE

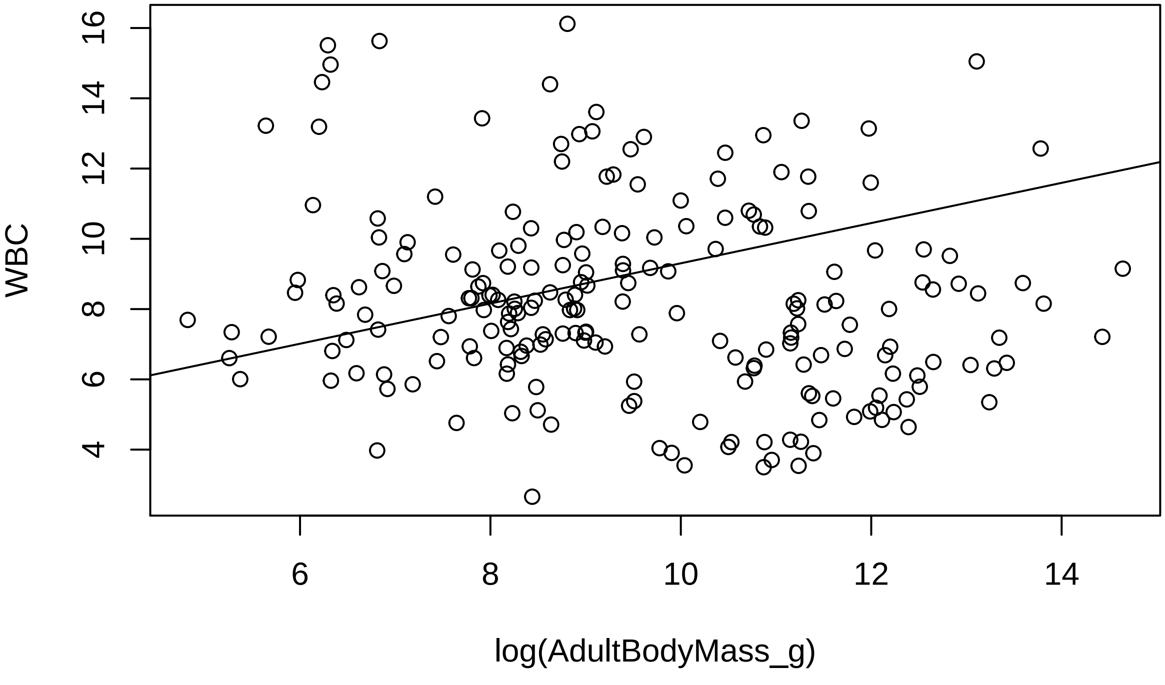
Using the Cooper 2012 data, show that the estimated relationship between body mass and white blood cell count is the same as that returned by PICs when the same model of evolution is used (here the BM model).

Your results should look something like this:

```
##   massPic
## 0.5723391
```



```
##          (Intercept) log(AdultBodyMass_g)
##            3.5801961            0.5723391
```



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

PGLS assumes that the residuals are phylogenetically correlated under a pure BM model. This is very constraining because it means that other sources of errors that are not phylogenetically correlated are not possible. 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 correlation structure that allows to relax this assumption.

### Pagel's correlation structure

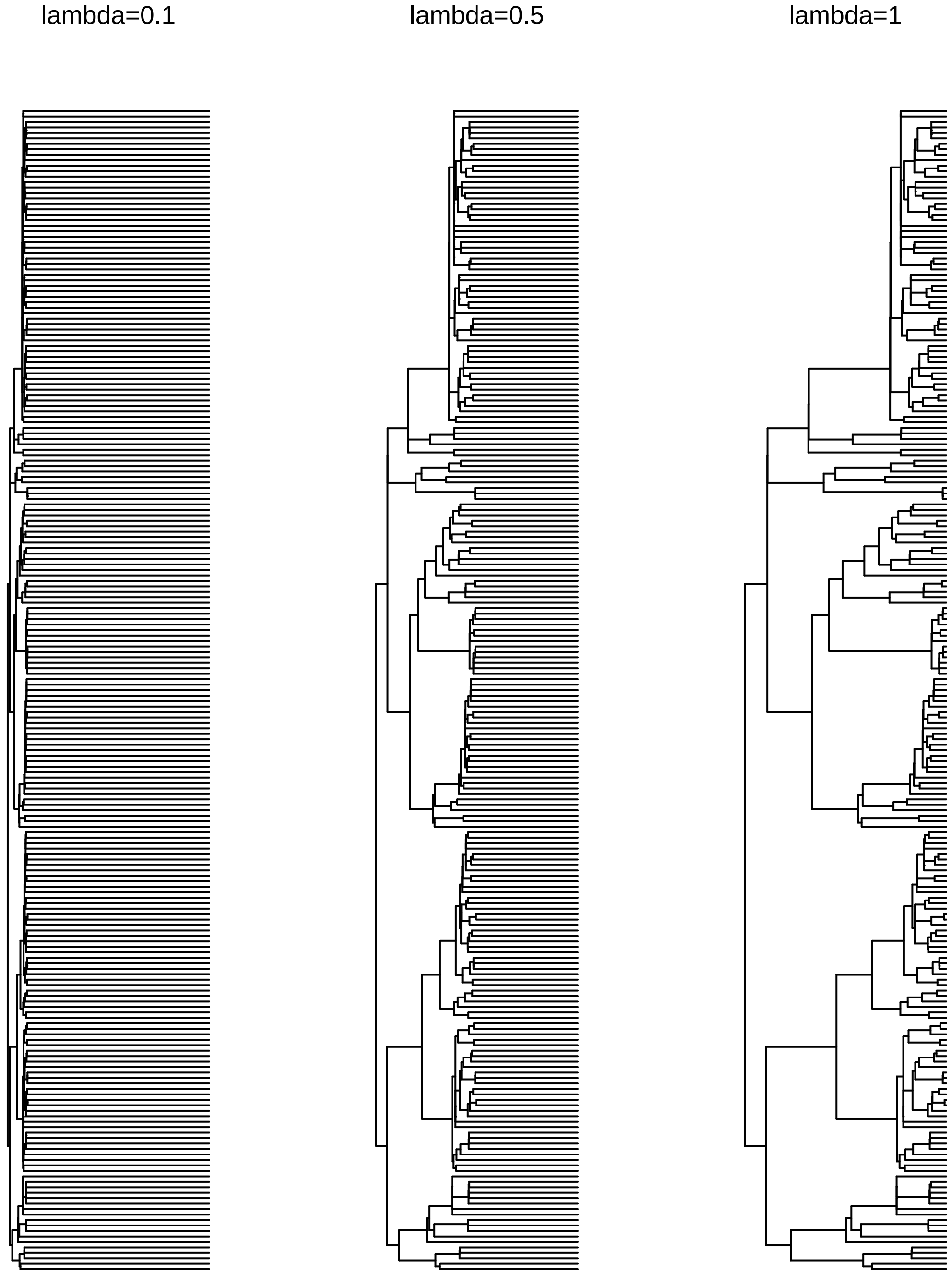
When controling for phylogenetic relationships with phylogenetic generalized least squares, we assume that the residuals are perfectly correlated according to the correlation structure. In practice, it might not be always the case and it is difficult to really know how important it is to control for 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). Pagel's approach to estimate phylogenetic signal is relatively simple. It is based on the idea that under the Brownian Motion model of evolution, the expected covariance matrix between traits is perfectly defined by the phylogenetic tree. Pagel introduced a branch transformation that can downweight the importance of the expected Brownian phylogenetic covariances to fit the observed ones. The $\lambda$ parameter defines this weight. The modified branch length $d_i^*$ for branch $i$ is:

$$d_i^* = \lambda d_i,$$

where $d_i$ is the original length of branch $i$. Essentially, $\lambda$ is a multiplier for the off-diagonal elements of the covariance matrix. In general, $\lambda$ can take a value between 0 and 1. When $\lambda = 1$, it means that the branch lengths are unaffected and the model corresponds to a pure Brownian model. At the opposite, $\lambda = 0$ means that branch lengths will equal 0, resulting in a more star-like phylogeny.

The following figure shows how different lambda values affect the shape of the phylogeny:



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.

**CHALLENGE**

$\lambda$ is essentially a modified BM process that requires one additional parameter to be estimated (when it is not fixed). With fitContinuous(), estimate phylogenetic signal in WBC using Pagel's $\lambda$. Using either AICc or a likelihood ratio test, determine whether this model is favoured over a pure BM model, and a model with lambda fixed at 0.5. *Hint: look at the function* `rescale`

---

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=tree, 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 corPagel
wbc.pgls2 <- gls(WBC ~ log(AdultBodyMass_g), data = dat, correlation=pagel.corr)
summary(wbc.pgls2)
```

```
## Generalized least squares fit by REML
##   Model: WBC ~ log(AdultBodyMass_g)
##   Data: dat
##       AIC      BIC    logLik
##   857.675 871.0824 -424.8375
##
## Correlation Structure: corPagel
##  Formula: ~1
##  Parameter estimate(s):
##    lambda
## 0.9550714
##
## Coefficients:
##                         Value Std.Error  t-value p-value
## (Intercept)          3.518317 1.7794692 1.977172  0.0493
## log(AdultBodyMass_g) 0.576949 0.1258981 4.582669  0.0000
##
##  Correlation:
##                      (Intr)
## log(AdultBodyMass_g) -0.674
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.7711518 -0.8213287 -0.1977877  0.3014702  2.2914759
##
## Residual standard error: 3.648814
## Degrees of freedom: 213 total; 211 residual
```

You can see that gls has estimated the $\lambda$ parameter, which is 0.955 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.

---

**Challenges**

When reliable phylogenies are unavailable, it may be tempting to include higher taxonomic groups to control for phylogenetic non-independence.

Include Order as a co-predictor in the WBC and AdultBodyMass OLS regression, compare this to the OLS model witout Order and determine whether this approach successfully controls for phylogenetic non-independence.

Fit a PGLS model with a Pagel correlation structure for parasite species richness (PSR) predicted by AdultBodyMass_g. Are the residuals as phylogenetically correlated than in the previous regression for WBC? How do you know?

---

## 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 $\lambda$ or Blomberg's $K$ (another common metric), 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 tempting 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 $\lambda$ with the PGLS model. Consequently, if the residuals are not phylogenetically correlated, $\lambda$ will be close to 0 and the model will essentially become OLS. 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 to decide whether it is worth it to use PGLS despite the extra parameters estimated.

### 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.

---

### CHALLENGE

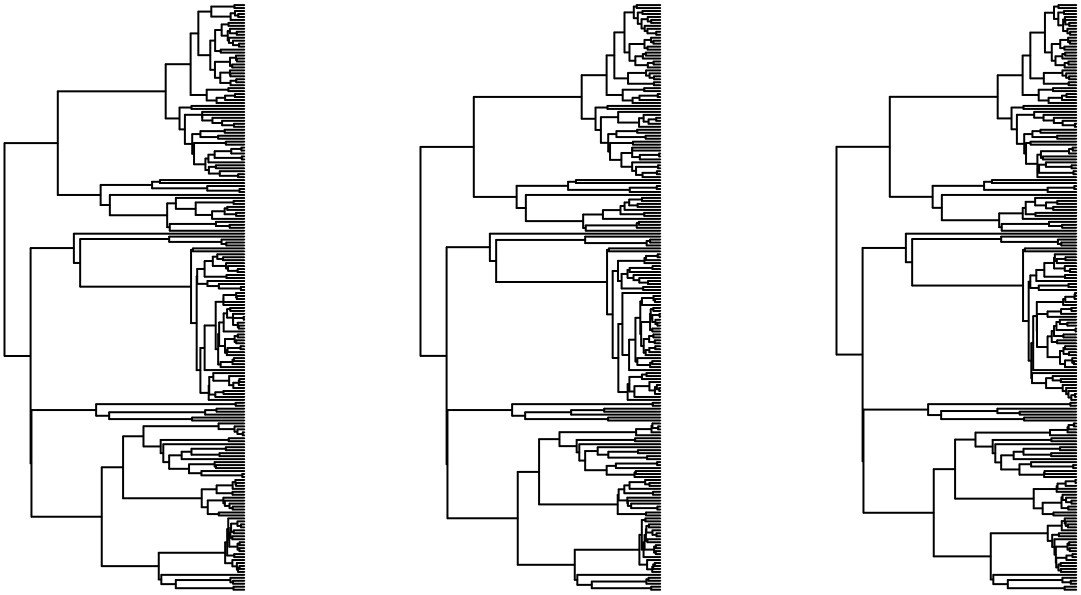
Re-do the above model with a different correlation structure.

---

### Phylogenetic Uncertainty

In the previous examples we used a single phylogeny for all analyses, however it is important to keep in mind that phylogenies represent evolutionary hypotheses, and there is often uncertainty in the topologies and branch lengths of a given phylogeny. Many phylogenies are reconstructed via Bayesian methods, which provides a posterior distribution of trees.

For example, a more recent mammal phylogeny was produced by Faurby et al. 2015. Here is an example of the first three trees for the species in the Cooper data:



Where do the trees differ?

#### CHALLENGE

In the data folder is a nexus file containing 100 trees from the posterior distribution of phylogenies from the Faurby 2015 paper. Fit a Brownian motion to WBC for each tree, save all the results as a list, and extract the estimate of the $sigma^2$ parameter. Plot a histogram of that parameter.

If you are feeling adventurous, try repeating the above PGLS accounting for phylogenetic uncertainty and plot histograms of the model coeffecients.

# BONUS: The Phylogenetic Mixed Model (PMM)

We mentioned previously that it is not a good idea to force all the residuals to be phylogenetically correlated and we saw that using the corPagel correlation structure provides one solution. In this section, we will see another approach that can do this, and much more. It is the phylogenetic mixed model.

### Mixed Models

As the name says, the phylogenetic mixed model is a mixed model. That is, it can be composed of fixed or random effects. Fixed effects are generally the variable of interest in an experiment; their effect on the response variable need to be quantified precisely. In contrast, the random effects in a model are variables that we now can affect the response variable but for which we are not specifically interested in quantifying their effects on the response variable. A good example of random effect is that of blocks in an replicated experiments. We often repeat experiments in blocks to account for spatial effects for instance. In such cases, blocks can be included as a random effect in the model to account for this additional variance that was accounted for. However, we are not interested in this variation per see.

This is very similar to the problem of phylogenetic correlations. Indeed, we know that the phylogenetic relationships between species can affect the relationship between variables and we can describe the correlation structure. However, we are not really interested in the effect of the phylogenetic correlations on the response variable; we just want to account for it in the main statistical test. As you can see, the phylogenetic correlation thus fits very well the definition of a random effect in a mixed model.

---

### The Phylogenetic Mixed Model (PMM)

The phylogenetic mixed model was first proposed by Lynch in 1991. This model was borrowed from quantitative genetic where similar models (generally called the 'animal' model) have been used from quite some time. It made a comeback in 2004 when Houswörth et al. (2004) proposed efficient algorithms to implement it. But it is really following the publication of the MCMCglmm package (Hadfield and Hasagawa, 2010) that Phylogenetic Mixed Model could really be applied widely by biologists.

The phylogenetic mixed model has been described in detail elsewhere (Hadfield et al. 2010, Villemereuil 2014) and this description will be rather brief. In the following, lowercase italic letters represent numbers, lowercase boldface letters vectors and uppercase boldface letters matrices. The phylogenetic mixed model (PMM) has the form:

$$\mathbf{y} = \mu + \beta\mathbf{x} + \mathbf{a} + \mathbf{e},$$

where $\mathbf{y}$ is the response variable, $\mu$ is the intercept, $\mathbf{x}$ is an explanatory variable, $\beta$ the regression coefficient, $\mathbf{a}$ represents the effects due to the phylogenetic structure, and $\mathbf{e}$ the residuals. $\mathbf{x}$ is a fixed effect (there could be more than one), whereas $\mathbf{a}$ is a random effect. The random effect and residuals are assumed to follow normal distributions:

$$\mathbf{a} \sim \mathcal{N}(0, \sigma_a^2\mathbf{A})\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I}).$$

$\sigma_a^2$ is the phylogenetic variance and $\sigma_e^2$ is the residual variance. The matrix $\mathbf{A}$ represents the phylogenetic correlation structure. The identity matrix $\mathbf{I}$ indicates that the residuals are independent and identically distributed. Accordingly, the (co)variance structure ($\mathbf{V}$) of the model is $\mathbf{V} = \sigma_a^2\mathbf{A} + \sigma_e^2\mathbf{I}$.

The major difference with PGLS is the great flexibility. For instance, the residuals are not expected to be all phylogenetically correlated. Indeed, the residuals ($\mathbf{e}$) are explicitely quantified in the model. Moreover, it is possible to add many more random effects. For instance, one could add a random affect that can account for a block design, or for measurement errors (if the response variable is estimated several times). This is simply not possible with PGLS.

Similarly with the PGLS and the estimation of the $\lambda$ parameter with the corPagel structure, it is possible to estimate the pylogenetic signal with the PMM. The idea is to estimate the proportion of the total variance ($\sigma^2 = \sigma_a^2 + \sigma_e^2$) that is due to the genetic structure; this is the heritability ($h^2$) parameter of quantitative geneticists (Housworth et al. 2006). The heritability is the quotient obtained by dividing the phylogenetic variances by the total variance: $h^2 = \sigma_a^2/\sigma^2$. The remaining variance, $1 - h^2 = \sigma_e^2/\sigma^2$, is the non-genetic variance that could be due to the environment or other effects that impact the individuals in a way that is not defined by the genetic correlation structures.

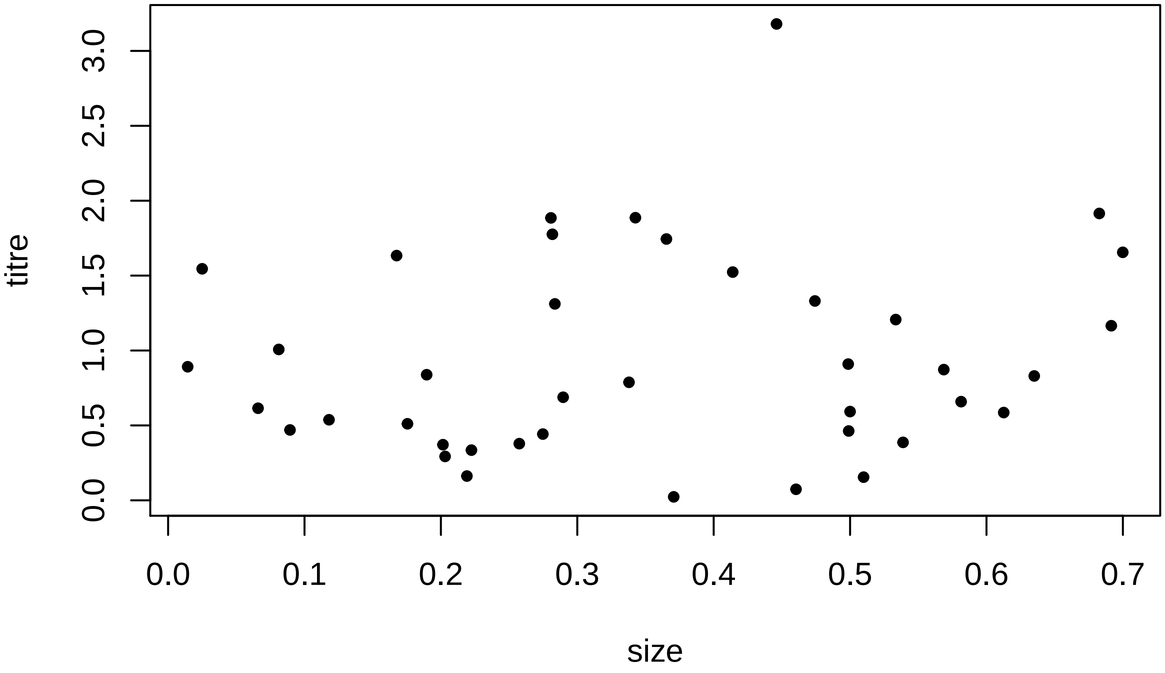
For this reason, the PMM is more interesting than PGLS. Yet, PGLS continues to be widely used. One reason for this is that PMM are still poorly known. The other is that the statistical tools are a bit complex. There are a few ways to fit PGLMMs, including the packages `pez` and `MCMCglmm`, however, we are going to use this opportunity to introduce model fitting in the language `Stan`, using `R` as an interface. If you want to learn more about `Stan` you can check out my intro workshop.

## brms (Bayesian regression models in Stan)

*Note: this section has been adapted from the vignette Estimating Phylogenetic Multilevel Models with brms by* `brms` *developer Paul Bürkner.*

Assume we have measurements of a phenotype, `titre` (say the mean viral titre measured for a given susceptible species), and a `cofactor` variable (say the mean body size).

```
tree <- rcoal(n=40)
size <- abs(fastBM(tree, a=0, mu=0, sig2=0.1))
titre <- abs(1.5 * size + rnorm(length(tree$tip.label), mean = 0, sd=1))
dat <- data.frame(Species = tree$tip.label, titre=titre, size=size)

with(dat, plot(titre~size, pch=20))
```



The `phylo` object contains information on the relationship between species. Using this information, we can construct a covariance matrix of species.

```
corr_phy <- vcv(tree, scaled=TRUE, corr=TRUE)
```

In contrast to **MCMCglmm**, **brms** requires the covariance matrix and not its inverse. Now, we are ready to fit our first phylogenetic multilevel model:

```
require(brms)
```

```
model_simple <- brm(
  titre ~ size + (1|Species), data = dat,
  family = gaussian(), cov_ranef = list(Species = corr_phy),
# ^ this is very similar to lme4 syntax for multilevel models

# OPTIONAL PARAMETERS
  prior = c(
    prior(normal(0, 10), "b"),
    prior(normal(0, 10), "Intercept"),
    prior(student_t(3, 0, 10), "sd"),
    prior(student_t(3, 0, 10), "sigma"))
  , iter=5000, warmup=4000, cores=2,
# #    ^ these are priors passed to brms
    control = list(adapt_delta=0.99, stepsize=0.01, max_treedepth=15)
# #    ^ this is a modification to help convergence for this model

)
```

```
## Compiling the C++ model
```

```
## Start sampling
```

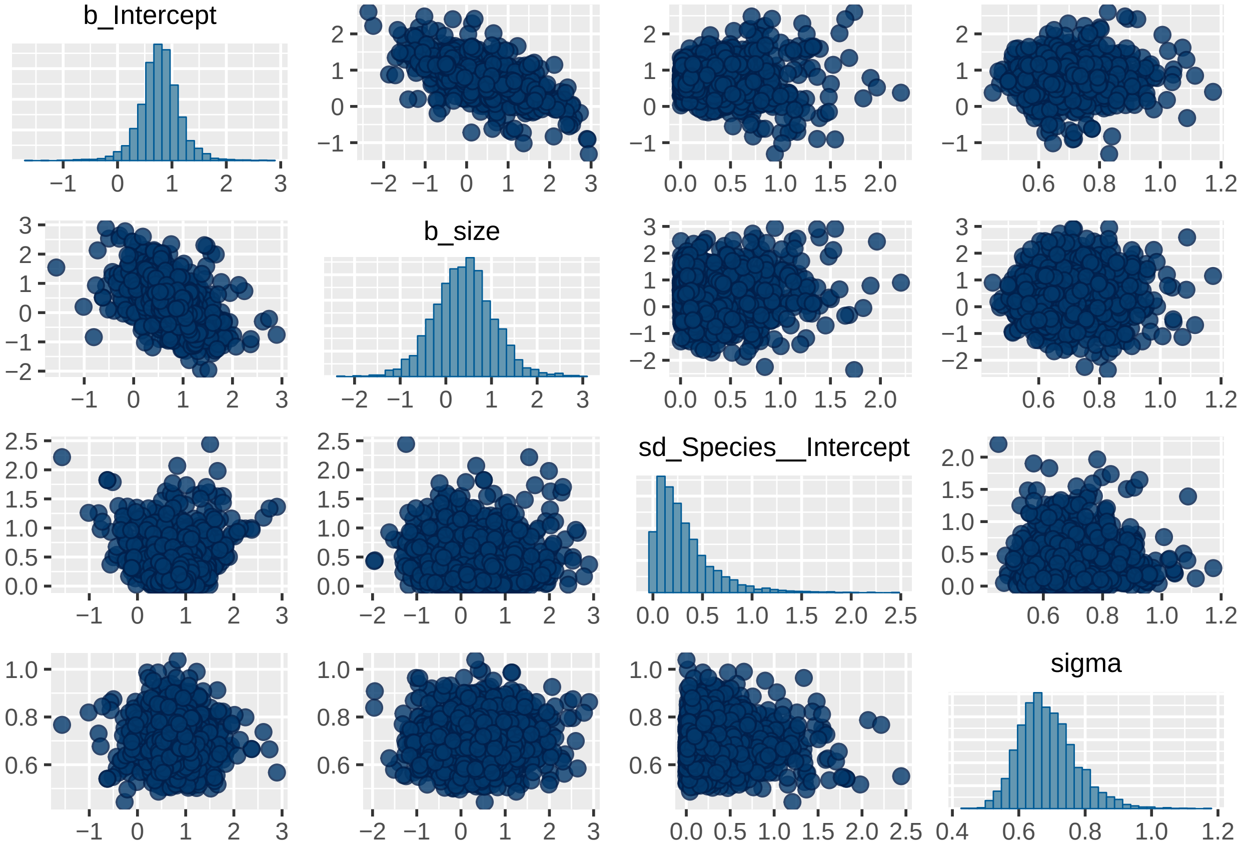
```
## Warning: There were 3 divergent transitions after warmup. Increasing adapt_delta above 0.99 may help
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

With the exception of `cov_ranef = list(phylo = corr_phy)` this is a basic multilevel model with a varying intercept over species (`Species` is an indicator of species in this data set). However, by using the `cov_ranef` argument, we make sure that species are correlated as specified by the covariance matrix `corr_phy`.

Setting priors is not always required as brms uses default priors if no priors are specified. After fitting, the results can be investigated in detail.

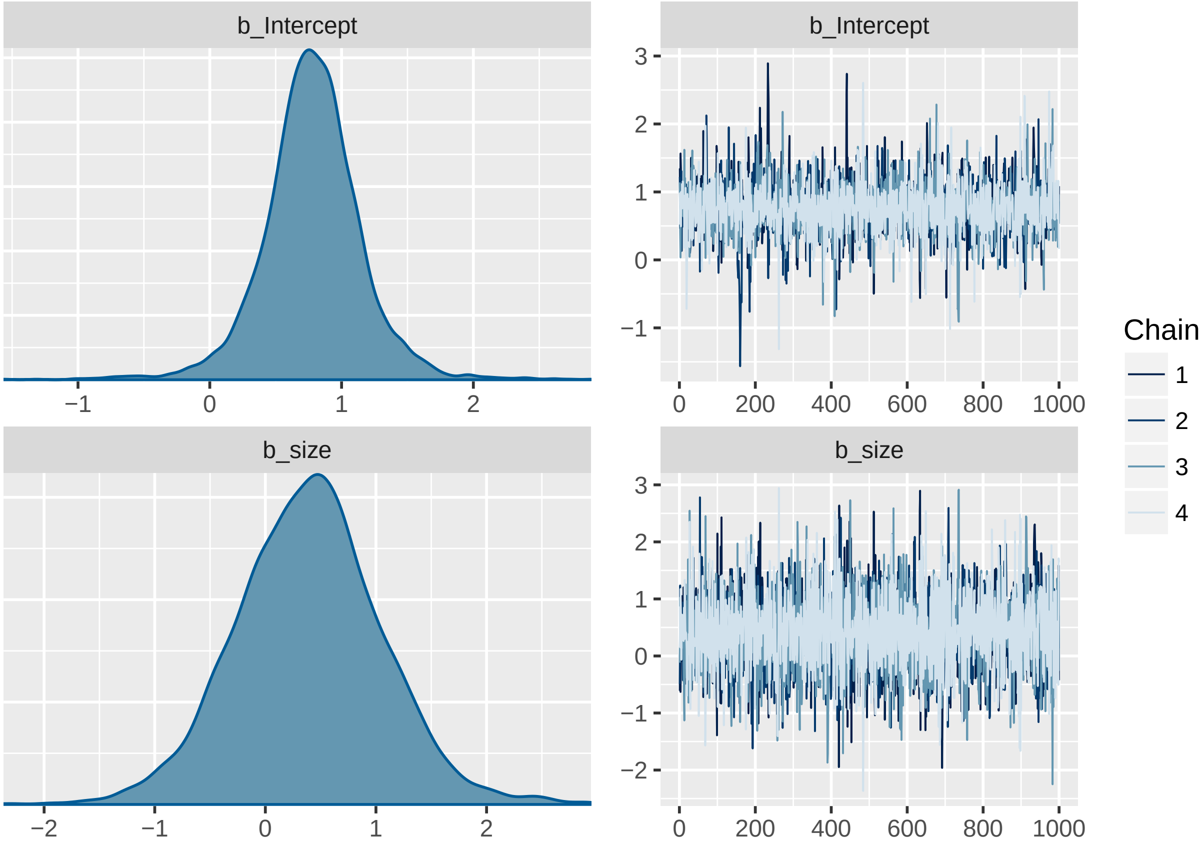
```
pairs(model_simple)
```

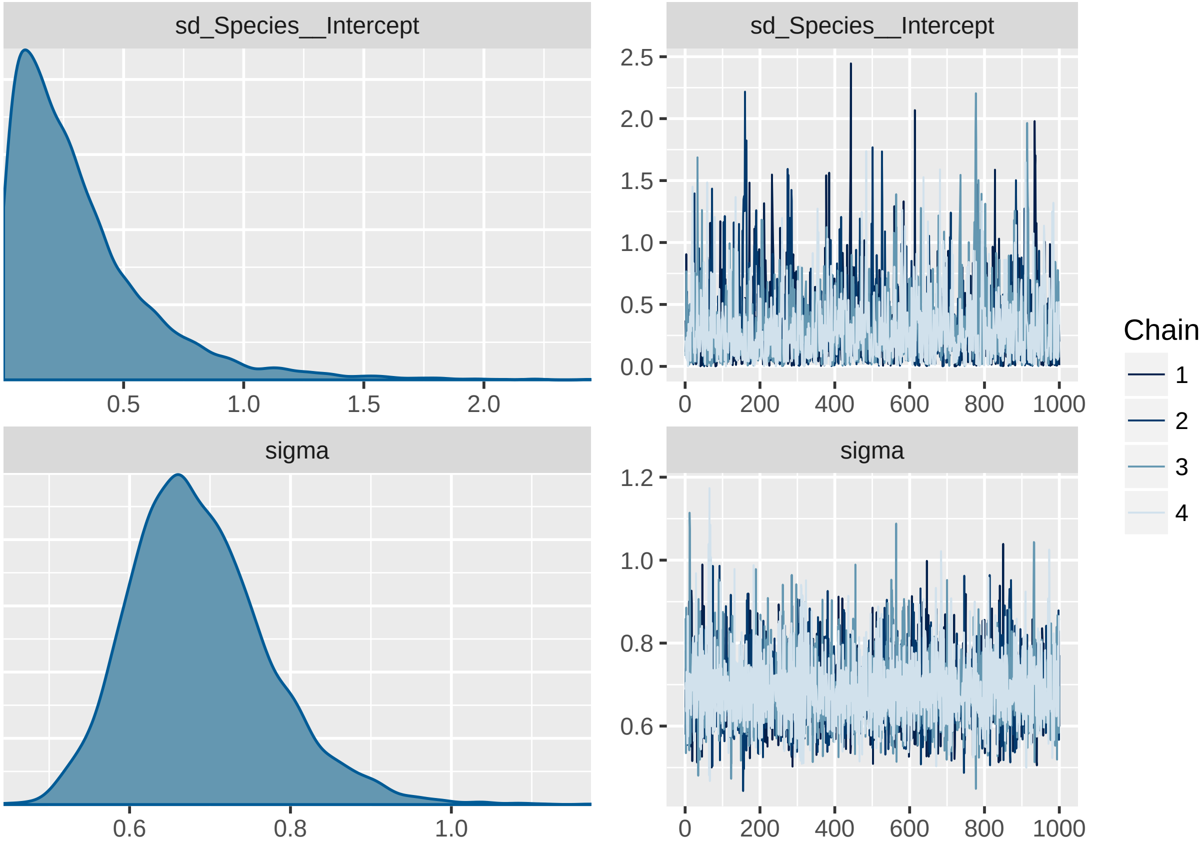


```r
summary(model_simple)
```

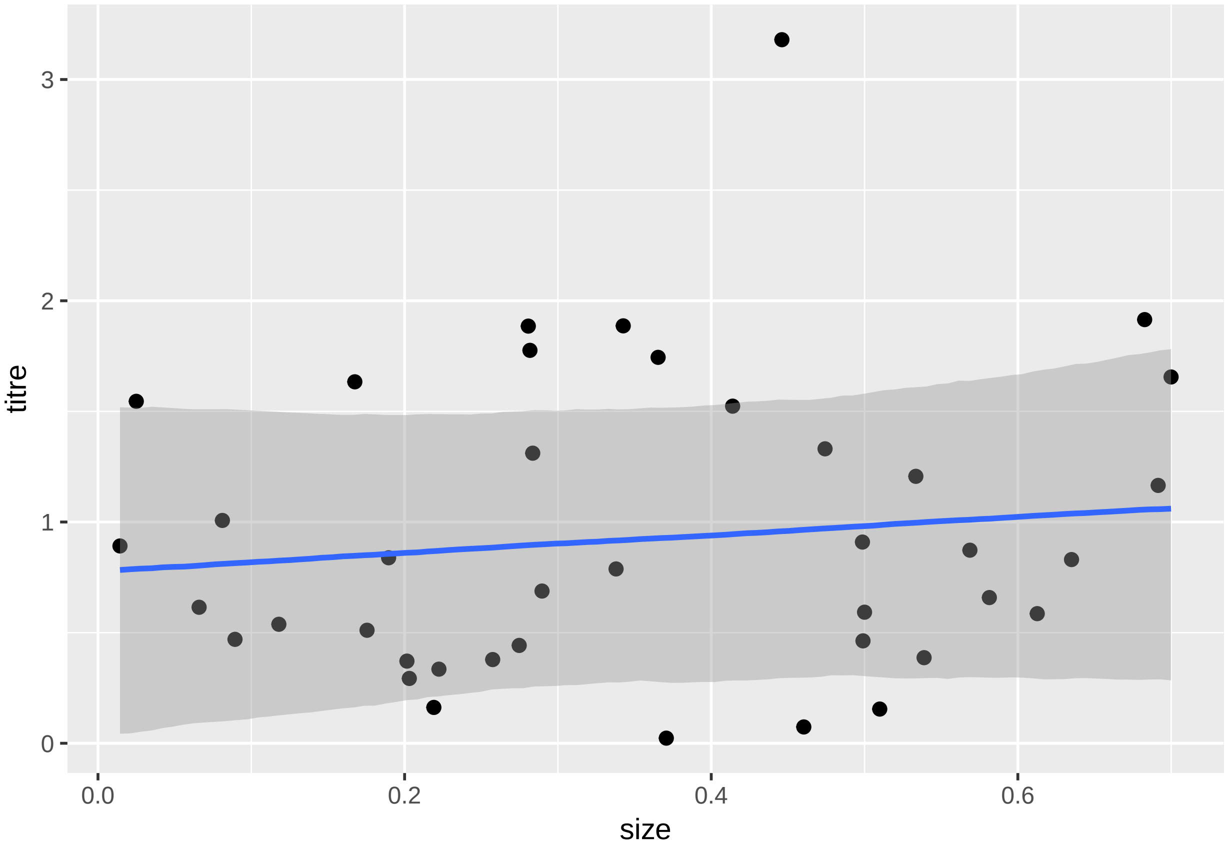
```
## Warning: There were 3 divergent transitions after warmup. Increasing adapt_delta above 0.99 may help
## See http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup

##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: titre ~ size + (1 | Species)
##    Data: dat (Number of observations: 40)
## Samples: 4 chains, each with iter = 5000; warmup = 4000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~Species (Number of levels: 40)
##               Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(Intercept)     0.32      0.29     0.01     1.12       1225 1.00
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## Intercept     0.78      0.38     0.02     1.52       2067 1.00
## size          0.39      0.66    -0.91     1.71       4000 1.00
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sigma     0.69      0.09     0.54     0.89       4000 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
plot(model_simple, N = 2, ask = FALSE)
```





```
plot(marginal_effects(model_simple), points = TRUE)
```

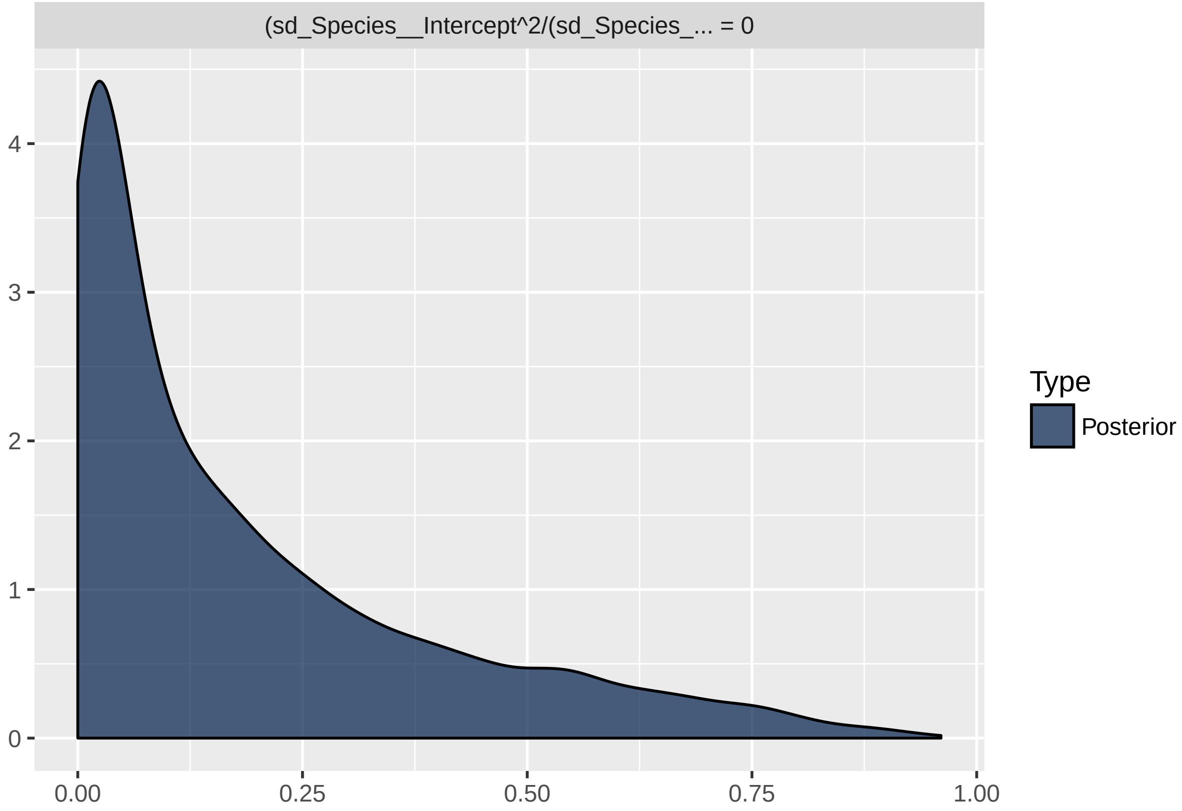


Phylogenetic signal (measured by $\lambda$) can be computed with the `hypothesis` method and is roughly $\lambda = 0.18$ for this example.

```
hyp <- "sd_Species__Intercept^2 / (sd_Species__Intercept^2 + sigma^2) = 0"
hyp <- hypothesis(model_simple, hyp, class = NULL)
hyp
```

```
## Hypothesis Tests for class :
##                 Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio
## 1 (sd_Species__Inte... = 0     0.19      0.21        0     0.73         NA
##   Star
## 1    *
## ---
## '*': The expected value under the hypothesis lies outside the 95%-CI.
```

```
plot(hyp)
```



Note that the phylogenetic signal is just a synonym of the intra-class correlation (ICC) used in the context phylogenetic analysis.

In addition to phylogenetic uncertainty, the PGLS we saw before assumes no intraspecific variation, however, with multilevel models we can have multiple observations per species and allows us fit more complicated phylogenetic models, including phyogenetic meta-analyses.

---

**CHALLENGE**

With Stan, we have a unique measure of convergence problems, which is denoted by "Divergent Transitions". Investigate what this means, and whether or not we can trust this model.

---

If you are interested, you can also see the Stan code auto-generated by `brms` with `stancode` function. The great thing about this is that when you want to make more customizeable models, you can use this as a starting point.

```
stancode(model_simple)
```

```
## // generated with brms 2.3.0
## functions {
## }
## data {
##   int<lower=1> N;  // total number of observations
##   vector[N] Y;  // response variable
##   int<lower=1> K;  // number of population-level effects
##   matrix[N, K] X;  // population-level design matrix
##   // data for group-level effects of ID 1
##   int<lower=1> J_1[N];
##   int<lower=1> N_1;
##   int<lower=1> M_1;
##   // cholesky factor of known covariance matrix
##   matrix[N_1, N_1] Lcov_1;
##   vector[N] Z_1_1;
##   int prior_only;  // should the likelihood be ignored?
## }
## transformed data {
##   int Kc = K - 1;
##   matrix[N, K - 1] Xc;  // centered version of X
##   vector[K - 1] means_X;  // column means of X before centering
##   for (i in 2:K) {
##     means_X[i - 1] = mean(X[, i]);
##     Xc[, i - 1] = X[, i] - means_X[i - 1];
##   }
## }
## parameters {
##   vector[Kc] b;  // population-level effects
##   real temp_Intercept;  // temporary intercept
##   real<lower=0> sigma;  // residual SD
##   vector<lower=0>[M_1] sd_1;  // group-level standard deviations
##   vector[N_1] z_1[M_1];  // unscaled group-level effects
## }
## transformed parameters {
##   // group-level effects
##   vector[N_1] r_1_1 = sd_1[1] * (Lcov_1 * z_1[1]);
## }
## model {
##   vector[N] mu = Xc * b + temp_Intercept;
##   for (n in 1:N) {
##     mu[n] += r_1_1[J_1[n]] * Z_1_1[n];
##   }
##   // priors including all constants
##   target += normal_lpdf(b | 0, 10);
##   target += normal_lpdf(temp_Intercept | 0, 10);
##   target += student_t_lpdf(sigma | 3, 0, 10)
##     - 1 * student_t_lccdf(0 | 3, 0, 10);
##   target += student_t_lpdf(sd_1 | 3, 0, 10)
##     - 1 * student_t_lccdf(0 | 3, 0, 10);
##   target += normal_lpdf(z_1[1] | 0, 1);
##   // likelihood including all constants
##   if (!prior_only) {
##     target += normal_lpdf(Y | mu, sigma);
##   }
## }
## generated quantities {
##   // actual population-level intercept
##   real b_Intercept = temp_Intercept - dot_product(means_X, b);
## }
```