Phylogenetic comparative methods

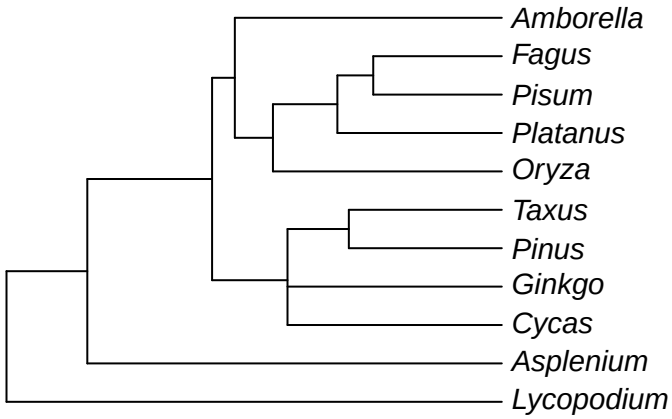
Simon Joly

Fall 2015

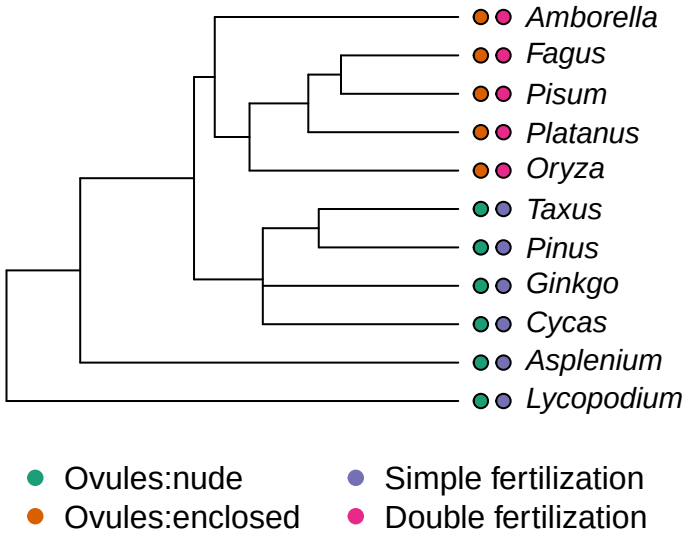
Contents

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 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 their share a more recent common ancestor. In other words, their evolutionary histories are shared over a longer period then with *Asplenium*. As such, they have more chance to have more similar traits (and in fact they do). For instance, take two characters, ovule and fertilization type, within this group.



Ignoring the phylogeny, we might be tempted to see a strong correlation between these two characters. Indeed, the states between the two characters show a perfect correspondance. Using standard statistics, we could do a chi-square test:

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

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  matrix(c(5, 0, 0, 6), ncol = 2)
## X-squared = 7.3364, df = 1, p-value = 0.006757
```

This would suggest that the assotiation is significant. However, we know that the comparisons made are not completely independent. Actually, both characters evolved only once, and this along the same branch. Consequently, these character might be correlated, but it is impossible to test this because these events are not replicated. For these reasons, phylogenetic comparative methods were developped.

The Brownian Motion (BM) model

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

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

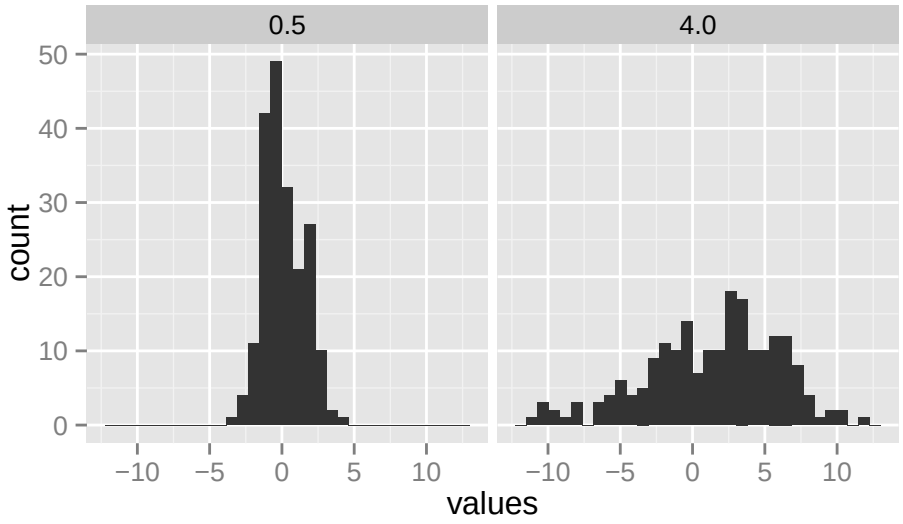
When applied to phylogenies, the Brownian motion model is kind of applied indepenpenty to each branch of the phylogeny. That allows to model the amount of change that occured along a given branch. If the variance of the Brownian motion model is σ^2 per unit of time t, then the net change along a branch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

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

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

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

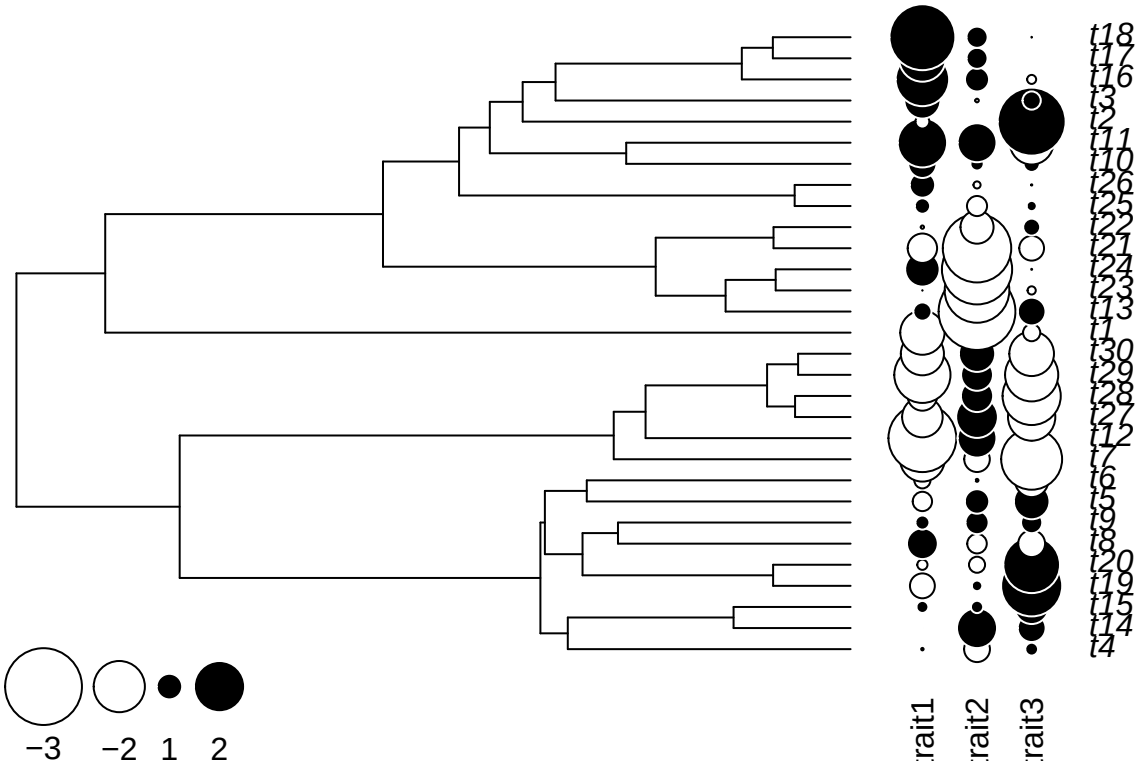
The parameter σ^2 in the model gives the variance, or in other word the speed of evolution. The higher the variance, the faster the character will evolve. Here are two examples of simulated characters on a tree of 200 species with $\sigma^2 = 0.5$ and $\sigma^2 = 4$.



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

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 somtimes be more different than to a distant species. Look at the following figure, that shows three traits simulated under the Brownian motion.

```
## Warning in table.phylo4d(obj, cex.label = 1, cex.symbol = 1, ratio.tree =
## 0.8, : There may not be enough room left to plot data; you may consider
## reducing ratio.tree or cex.label.
```



Phylogenetic independent contrasts (PIC)

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

> Warning. Your tree must be fully resolved and you cannot have missing data in your dataset. Note that you can arbitrarily resolve polytomies using the ape function `multi2di`.

For the PIC examples, we will use data on seed plant functional traits published by Paquette et al in 2015.

```
require(ape)
seedplantstree <- read.nexus("./data/seedplants.tre")
seedplantsdata <- read.csv2("./data/seedplants.csv")
# Remove species for which we don't have complete data
seedplantsdata <- na.omit(seedplantsdata)
# Remove species in the tree that are not in the data matrix
```

```
species.to.exclude <- seedplantstree$tip.label[!(seedplantstree$tip.label %in%
                                                 seedplantsdata$Code)]
seedplantstree <- drop.tip(seedplantstree,species.to.exclude)
rm(species.to.exclude)
```

> It is important to assign names to your character vectors. These will be used to match the names of the tips of the phylogeny. If you don't assign names, then the function `pic` will assume that the characters are in the same order as the tip.label of the phylogeny. But this can lead to errors. In general, it is a good idea to have the data to be in the same order as the tip.labels of the phylo anyway.

```
# 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
```

```
# Name the rows of the data.frame with the species codes used as tree labels
rownames(seedplantsdata) <- seedplantsdata$Code
seedplantsdata <- seedplantsdata[,-1]
# Order the data in the same order as the tip.label of the tree. In the present
# example, this was already the case.
seedplantsdata <- seedplantsdata[seedplantstree$tip.label,]
```

Calculate contrasts

We will now calculate independent contrasts using the `pic` function in `ape`. First, let's create numeric vectors for four traits: wood density (Wd), shade tolerance (Shade), seed mass (Sm), and nitrogen content (N).

```
# Extract trait data into vectors
Wd <- seedplantsdata$Wd
Shade <- seedplantsdata$Shade
Sm <- seedplantsdata$Sm
N <- seedplantsdata$N
# Important: Give names to your vectors
names(Wd) <- names(Shade) <- names(Sm) <- names(N) <- row.names(seedplantsdata)
```

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

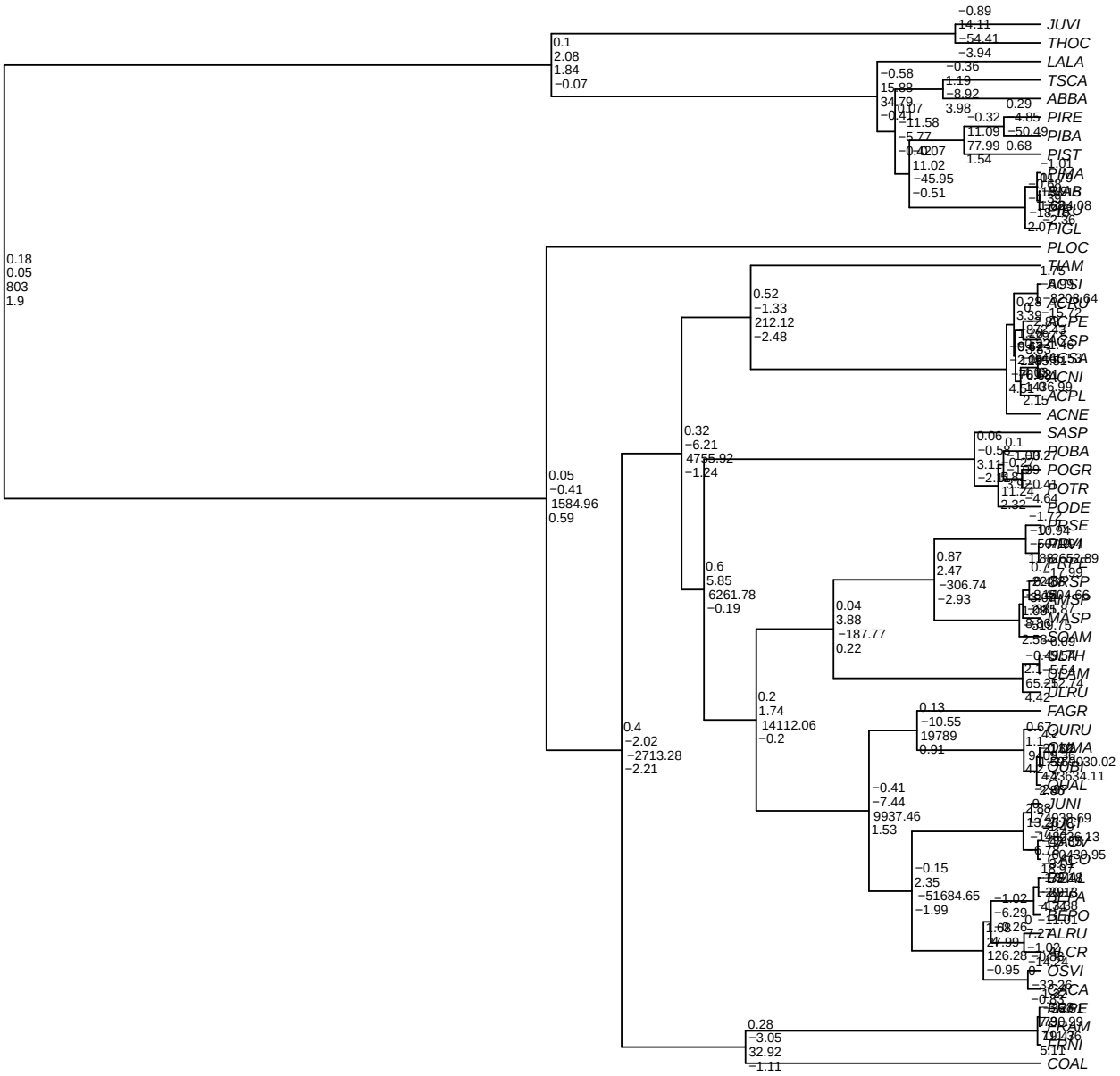
```
# Calculate the contrasts for each trait that have been scaled using the expected
# variances
Wd.contrast <- pic(Wd,seedplantstree,scaled=TRUE)
Shade.contrast <- pic(Shade,seedplantstree,scaled=TRUE)
```

```
Sm.contrast <- pic(Sm,seedplantstree,scaled=TRUE)
N.contrast <- pic(N,seedplantstree,scaled=TRUE)
# If you want to calculate contrasts for several variables, create a matrix with
# your variables in columns, then just use apply:
contrasts.seedplantsdata <- as.data.frame(apply(seedplantsdata[,-1], 2, pic,
                                                seedplantstree))
```

Display contrasts

You can display the contrasts at the nodes of the phylogeny.

```
plot(seedplantstree, label.offset=0.001,cex=0.6)
nodelabels(round(Wd.contrast, 2), adj = c(-0.1, -2), frame="n",cex=0.5)
nodelabels(round(Shade.contrast, 2), adj = c(-0.1, -0.5), frame="n",cex=0.5)
nodelabels(round(Sm.contrast, 2), adj = c(-0.1, 1), frame="n",cex=0.5)
nodelabels(round(N.contrast, 2), adj = c(-0.1, 2.5), frame="n",cex=0.5)
```



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.

Testing for statistical association between traits

The correlation between two traits will be tested. Specifically, we will test two hypotheses. The first one is that wood density affects shade tolerance, and the second that leaf nitrogen content affects seed mass.

Let's first look at a standard regression, that is not controlling for the phylogeny.

```
RegressShade <- lm(Shade ~ Wd, seedplantsdata)
summary.lm(RegressShade)
```

```
##
## 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
```

```
RegressSm <- lm(Sm ~ N, seedplantsdata)
summary.lm(RegressSm)
```

```
##
## Call:
## lm(formula = Sm ~ N, data = seedplantsdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2580.1 -1234.9  -667.1   483.7 13207.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2661.4     1449.7  -1.836   0.0718 .
## N             1759.1      691.3   2.545   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2597 on 55 degrees of freedom
## Multiple R-squared:  0.1053, Adjusted R-squared:  0.08907
## F-statistic: 6.475 on 1 and 55 DF,  p-value: 0.01377
```

You can see that there is no significant effect of wood density on shade tolerance, but that the relation between nitrogen content and seed mass is significant. Now let's look at the same relationships after we corrected for phylogenetic relatedness of species.

```
RegressShade.pic <- lm(Shade.contrast~Wd.contrast -1)
summary.lm(RegressShade.pic)
```

```
##
## Call:
## lm(formula = Shade.contrast ~ Wd.contrast - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -71.943  -4.602   0.781   4.632  21.614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## Wd.contrast    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
```

```
RegressSm.pic <- lm(Sm.contrast~N.contrast -1)
summary.lm(RegressSm.pic)
```

```
##
## Call:
## lm(formula = Sm.contrast ~ N.contrast - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -157012   -1766     -16    1908   71512
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## N.contrast   -480.2      692.9  -0.693    0.491
##
## Residual standard error: 33420 on 55 degrees of freedom
## Multiple R-squared:  0.008657,   Adjusted R-squared:  -0.009368
## F-statistic: 0.4803 on 1 and 55 DF,  p-value: 0.4912
```

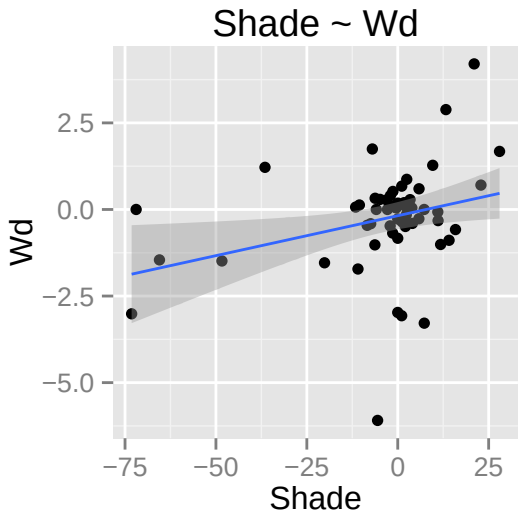
> The -1 in the model specifies that the regression is through the origin (the intercept is set to zero) as recommended by Garland et al., 1992.

When taking phylogenetic information into account, seed mass is not significantly related to nitrogen content anymore. This means that the apparent correlation observed on the raw data was an artefact of their evolutionary histories. The other regression, in contrast, now shows a significant relationship between wood density and shade tolerance. This is a positive example of the application of PIC. Indeed, the application of PIC does not always make the relationships less significant. Sometimes, it helps highlight significant relationships that were obscured by the evolutionary history of species.

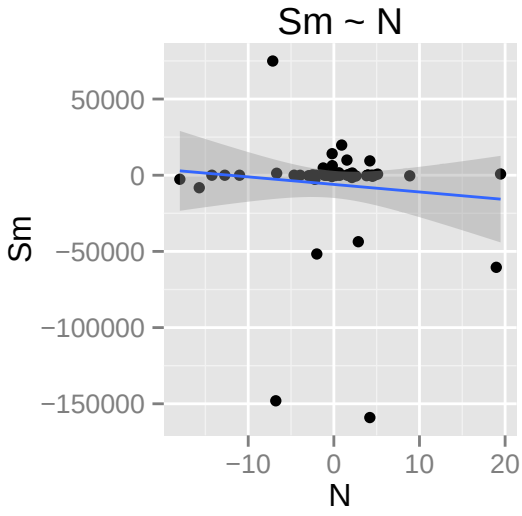
Plot contrasts

It is often useful to plot the contrasts to visualize the regression. This allows to check that there is indeed a linear relationship between the 2 variables:

```
library(ggplot2)
par(mfrow=c(1,2))
qplot(Shade,Wd,data=contrasts.seedplantsdata) + geom_smooth(method="lm") +
  ggtitle("Shade ~ Wd")
```




```
qplot(N,Sm,data=contrasts.seedplantsdata) + geom_smooth(method="lm") + ggtitle("Sm ~ N")
```



Major Axis regression

An alternative to regression when there are no clear dependent variable is major axis regression. Unlike regression where the residuals are only estimated on the y axis, major axis regression is estimated on both x and y axes. This can be done with the package `smatr`. The 'method=2' in the command is for the major axis regression method.

```
library(smatr)
Shade.mar.pic <- slope.test(Shade.contrast,Wd.contrast,test.value=0,
                            intercept=FALSE,method=2)
# Result Shade ~ Wd
data.frame(Slope=Shade.mar.pic$b,p_value=Shade.mar.pic$p)
```

```
##      Slope    p_value
## 1 40.31972 0.01352355
```

```r
Sm.mar.pic <- slope.test(Sm.contrast,N.contrast,test.value=0,intercept=FALSE,method=2)
# Result Sm ~ N
data.frame(Slope=Sm.mar.pic$b, p_value=Sm.mar.pic$p)
```

```
##      Slope   p_value
## 1 -55467.5 0.4951602
```

In these cases, the results are the same as for the standard regression.

Phylogenetic generalized least squares (PGLS)

Phylogenetic generalized least squares are very similar to PIC. The idea is the same, that is to remove the effect of the evolutionary relationships of species when fitting a regression between two variables. Generalized least squares allows to user to specify a covariance structure that characterize the data and which effect should be removed when fitting the regression. 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 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 will see some of these models in detail later in the course, for instance in lecture 5.

In the present case, we will use the Brownian Motion structure, which is also the model behind PIC.

```r
library(nlme)
# Get the correlation structure of the tree
bm.corr <- corBrownian(phy=seedplantstree)
# PGLS: Shade ~ Wd
shade.bm.pgls <- gls(Shade ~ Wd, data = seedplantsdata, correlation = bm.corr)
summary(shade.bm.pgls)
```

```
## 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
```

```
# PGLS: Sm ~ N
sm.bm.pgls <- gls(Sm ~ N, data = seedplantsdata, correlation = bm.corr)
summary(sm.bm.pgls)
```

```
## Generalized least squares fit by REML
##   Model: Sm ~ N
##   Data: seedplantsdata
##        AIC      BIC    logLik
##   1038.066 1044.088 -516.0331
##
## Correlation Structure: corBrownian
##  Formula: ~1
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##                 Value Std.Error    t-value p-value
## (Intercept) 1049.4649  7665.415  0.1369091  0.8916
## N           -480.1715   692.864 -0.6930239  0.4912
##
##  Correlation:
##   (Intr)
## N -0.16
##
## Standardized residuals:
##          Min           Q1          Med           Q3          Max
## -0.043549669 -0.013107618  0.001040479  0.014256968  1.180782954
##
## Residual standard error: 12973.58
## Degrees of freedom: 57 total; 55 residual
```

As you can see, the results are exactly the same as for the PIC results when the same model of evolution is used (here the BM model).

Phylogenetic ANOVA

As a special case of linear models, it is also possible to perform a phylogenetic ANOVA to compare character values between groups while taking into account the phylogenetic relationships among individuals. There are (at least) two main ways of doing a phylogenetic ANOVA. The first one, and often considered the "traditional" one, was proposed by Garland et al. (1993). The idea is to do a classic ANOVA, but then to simulate character datasets on the phylogeny using the Brownian Motion model to obtain a null model of the test statistic (F). The statistical significance is tested by comparing the value obtained with the observed data to the simulated values. This ANOVA approach is implemented in the function `phylANOVA` of the `phytools` package.

The second approach consist in using generalized least squares with phylogenetic correlation of the residuals, as described above, but using a categorical variable as independent variable. This is also an approach that can be used in R for standard ANOVA. For instance, the linear model `lm(Y~X)` would perform an ANOVA if the X variable is categorical.

In one of his blog posts, Liam Revell (2013) has shown that both approaches had correct type I errors, that is they won't reject the null hypothesis when it is true. However, he showed that the PGLS ANOVA approach was much more powerful than the Garland et al. approach. It is thus the method I will show here. To demonstrate how to perform the PGLS ANOVA, we will use an example with simulated data.

```
require(phytools)
#Simulate a random tree
tree<-pbtree(n=200)
#Simulate two correlated characters; they will be store in a matrix with 2 columns
xx<-sim.corrs(tree,matrix(c(1,0.8,0.8,1),nrow=2))
#Create a multistate character (X) from the first character; the second (Y) remains unchanged
X<-cut(xx[,1],breaks=3,labels=c("small","medium","tall"))
Y<-xx[,2]
names(X)<-names(Y)<-rownames(xx)
```

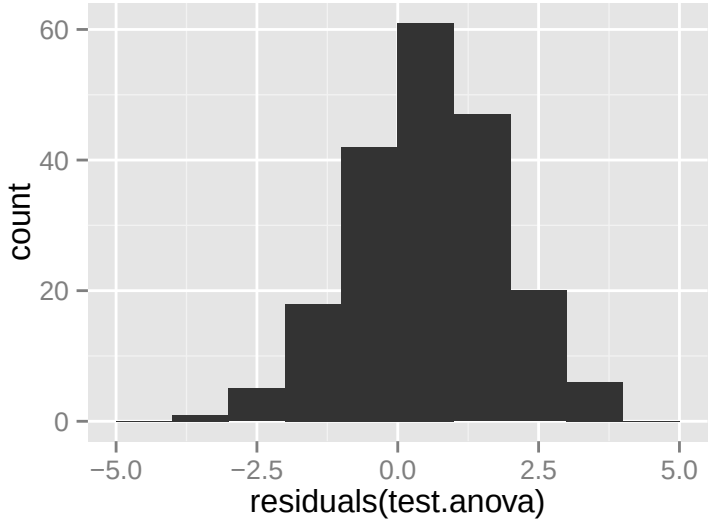
Now, let's test whether the different categories differ according to character Y.

```
# PGLS ANOVA
require(ape)
require(nlme)
tree.corr <- corBrownian(phy=tree)
testdata<-data.frame(Y=Y,X=X)
# The PGLS ANOVA
test.anova <- gls(Y ~ X - 1, data=testdata, correlation = tree.corr)
# The '-1' in the formula is a trick to force R to fit the groups relative to the
# general mean, and not relative to the first category, as it would be with the model
# 'Y~X'. However, the results are exactly identical with the two approaches, apart from
# the group means.
summary(test.anova)
```

```
## Generalized least squares fit by REML
##   Model: Y ~ X - 1
##   Data: testdata
##        AIC      BIC    logLik
##   494.9763 508.1091 -243.4881
##
## Correlation Structure: corBrownian
##  Formula: ~1
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##              Value Std.Error    t-value p-value
## Xsmall  -1.9684584 0.8702686 -2.2618974  0.0248
## Xmedium -0.5142642 0.8605272 -0.5976153  0.5508
## Xtall    0.4656718 0.8708550  0.5347295  0.5934
##
##  Correlation:
##         Xsmall Xmedim
```

```
## Xmedium 0.939
## Xtall   0.922  0.980
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.7030793 -0.1905998  0.2612959  0.6774947  1.9123879
##
## Residual standard error: 1.968924
## Degrees of freedom: 200 total; 197 residual
```

```
# The residuals should be normally distributed
qplot(residuals(test.anova),geom="histogram",binwidth=1)
```




```
# Other plotting alternative:
# hist(residuals(test.anova))
# It is possible to test if the ANOVA model is better than a null model
anova(test.anova)
```

```
## Denom. DF: 197
##   numDF F-value p-value
## X     3 19.2757  <.0001
```

```
# It is also possible to compare directly two different models. Here is a simple model
# with only an intercept:
test.NULL <- gls(Y ~ 1, data=testdata, correlation = tree.corr)
summary(test.NULL)
```

```
## Generalized least squares fit by REML
##   Model: Y ~ 1
##   Data: testdata
##        AIC      BIC    logLik
##   539.0395 545.6261 -267.5197
##
## Correlation Structure: corBrownian
##  Formula: ~1
```

```
##  Parameter estimate(s):
## numeric(0)
##
## Coefficients:
##                   Value Std.Error    t-value p-value
## (Intercept) -0.9248027 0.9608015 -0.9625325   0.337
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.65618657 -0.07037253  0.46329004  0.95035894  1.87858962
##
## Residual standard error: 2.222883
## Degrees of freedom: 200 total; 199 residual
```

```
# The two models can be compared with the 'anova.gls' function
anova(test.anova,test.NULL)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## test.anova     1  4 494.9763 508.1091 -243.4881
## test.NULL      2  2 539.0395 545.6261 -267.5197 1 vs 2 48.06317  <.0001
```

You can see that the ANOVA has a better AIC than the model with only the intercept, which supports this more complex model. The more standard likelihood ratio test is also strongly significant.

Phylogenetic logistic regression

It is also possible to do phylogenetic logistic regression. That is, you can test whether a variable (independent) can affect the outcome of a binary dependent variable.

The package `phylolm` allows you to perform phylogenetic logistic regressions as described in Ives and Garland (2010), using the method of Ho and Ané (2014). To show how it works, let's simulate a tree and three variables, one binary (dependent variable), one quantitative and one categorical.

```
require(phylolm)
set.seed(123456)
# Simulate a tree of 50 species
tre = rtree(50)
# Simulate a continuous trait
conTrait = rTrait(n=1,phy=tre)
# Make a design matrix for the binary trait simulation
X = cbind(rep(1,50),conTrait)
# Simulate a binary trait
binTrait = rbinTrait(n=1,phy=tre, beta=c(-1,0.5), alpha=1 ,X=X)
# Simulate a radom categorical trait
catTrait <- as.factor(sample(c("A","B","C"),size=length(tre$tip.label),replace=TRUE))
# Create data frame
dat = data.frame(binTrait = binTrait, conTrait = conTrait, catTrait = catTrait)
```

Continous variable

Now you can fit the phylogenetic logistic regression. Let first fit a model with the continuous variable as predictor variable.

```
(fit = phyloglm(binTrait ~ conTrait, phy=tre, data=dat))
```

```
## Call:
## phyloglm(formula = binTrait ~ conTrait, data = dat, phy = tre)
##        AIC     logLik Pen.logLik
##      48.54     -21.27     -19.61
##
## Parameter estimate(s) from MPLE:
## alpha: 1.241056
##
## Coefficients:
## (Intercept)    conTrait
##  -0.5056675   1.4126605
```

The logistic regression uses alpha to estimate the level of phylogenetic correlation. As such, the estimate of the alpha parameter by the model inform on whether there is phylogenetic signal in the dependent character. Values > 0 suggest relatively strong phylogenetic signal.

Now let's fit a null model, with only an intercept and no predictor variable.

```
(fit0 = phyloglm(binTrait ~ 1, phy=tre, data=dat))
```

```
## Call:
## phyloglm(formula = binTrait ~ 1, data = dat, phy = tre)
##        AIC     logLik Pen.logLik
##      56.93     -26.46     -25.71
##
## Parameter estimate(s) from MPLE:
## alpha: 1.01695
##
## Coefficients:
## (Intercept)
##  -0.8317861
```

Then you can compare the two model with AIC to see if the fit of the phylogenetic logistic regression on the continuous trait is significantly better than the null model.

```
data.frame(model=c("conTrait","Null model"),
           log_lik=c(logLik(fit)$logLik,logLik(fit0)$logLik),
           df=c(logLik(fit)$df,logLik(fit0)$df),
           AIC=c(AIC(fit),AIC(fit0)))
```

```
##        model   log_lik df      AIC
## 1   conTrait -21.27195  3 48.54391
## 2 Null model -26.46468  2 56.92937
```

You can see that the fit of the model is significantly better than the null model.

Categorical variable

We can also do the same thing with a categorical variable.

```
(fit2 = phyloglm(binTrait ~ catTrait, phy=tre, data=dat))
```

```
## Call:
## phyloglm(formula = binTrait ~ catTrait, data = dat, phy = tre)
##        AIC     logLik Pen.logLik
##      59.07     -25.53     -24.11
##
## Parameter estimate(s) from MPLE:
## alpha: 0.8052537
##
## Coefficients:
## (Intercept)   catTraitB   catTraitC
##  -0.8425245   0.3757810  -0.4956067
```

```
data.frame(model=c("catTrait","Null model"),
           log_lik=c(logLik(fit2)$logLik,logLik(fit0)$logLik),
           df=c(logLik(fit2)$df,logLik(fit0)$df),
           AIC=c(AIC(fit2),AIC(fit0)))
```

```
##        model   log_lik df      AIC
## 1   catTrait -25.53323  4 59.06646
## 2 Null model -26.46468  2 56.92937
```

In this case, the model is worse than the null model. This is expected in this case because the categorical character was a random variable.

Phylogenetic Principal Component Analysis

Principal component analysis (PCA) is a very popular method to represent in a few dimensions the variation of several variables. Again, the representation of the data in an PCA could also partly represents the shared coancestry or organisms. Liam Revell (2009) has proposed a phylogenetic PCA (pPCA) that takes into account the nonindependence of the data due to the shared co-ancestry of the species.

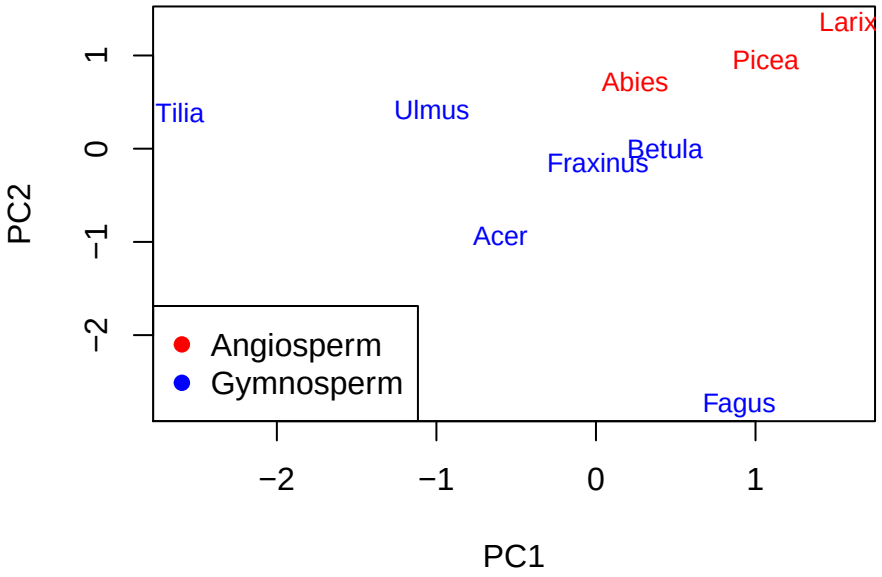
Let's go back to the angiospem data. A PCA of a reduced matrix would look like the following:

```
#Species to keep
keep<-c("ABBA","ACSA","BEAL","FAGR","FRAM","LALA","PIRU","TIAM","ULAM")
spnames<-c("Abies","Acer","Betula","Fagus","Fraxinus","Larix","Picea","Tilia","Ulmus")
category<-as.factor(c("Gymnosperm","Angiosperm","Angiosperm","Angiosperm","Angiosperm",
            "Gymnosperm","Gymnosperm","Angiosperm","Angiosperm"))
exclude<-seedplantstree$tip.label[!(seedplantstree$tip.label %in% keep)]
#reduce the seed plant phylogeny
seedplantstreereduced<-drop.tip(seedplantstree,exclude)
seedplantsdatareduced<-data.frame(seedplantsdata[rownames(seedplantsdata) %in%
                                        seedplantstreereduced$tip.label,-(1:2)])

#PCA on the correlation matrix
seedplant.pca<-prcomp(seedplantsdatareduced,scale=TRUE)
summary(seedplant.pca)
```

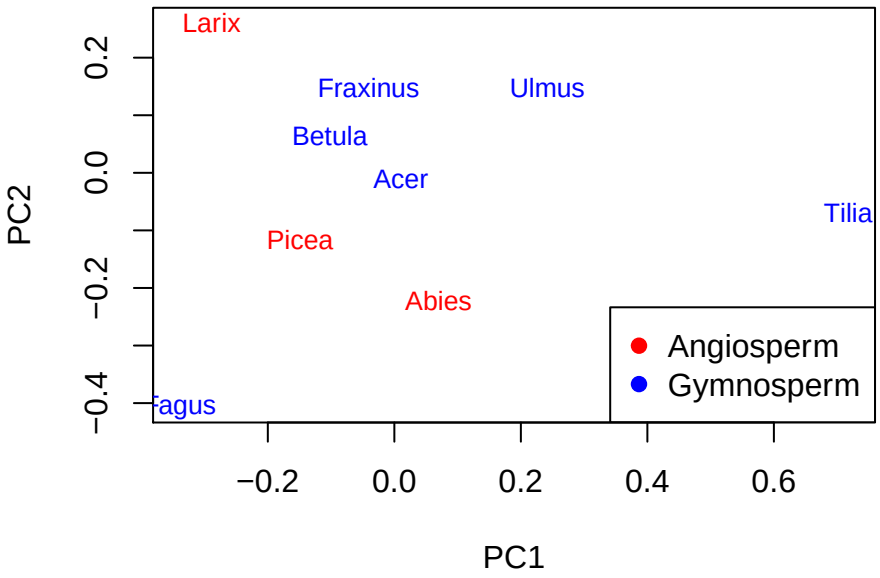
```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.2702 1.2301 1.0868 0.69144 0.46296
## Proportion of Variance 0.3227 0.3026 0.2362 0.09562 0.04287
## Cumulative Proportion  0.3227 0.6253 0.8615 0.95713 1.00000
```

```
colPalette=c("blue","red")
plot(seedplant.pca$x,type="n")
text(seedplant.pca$x,labels=spnames,cex=0.8,col=colPalette[category])
legend("bottomleft",pch=20,legend=levels(category),col=colPalette[category],pt.cex=1.5)
```



Not unexpectedly, you can see that the angiosperms and the gymnosperms occupy different parts of the ordination space. Now, let's compare with a phylogenetic PCA.

```
#pPCA
seedplant.ppca<-phyl.pca(seedplantstreereduced,seedplantsdatareduced,
                         method="BM",mode="corr")
plot(seedplant.ppca$S,type="n")
text(seedplant.ppca$S,labels=spnames,cex=0.8,col=colPalette[category])
legend("bottomright",pch=20,legend=levels(category),col=colPalette[category],pt.cex=1.5)
```



The plots are relatively similar, which means that the phylogeny might not be very important in shaping the traits amongst the species. Yet, there are also some differences. For instance, gymnosperms and angiosperms group mostly together in the standard PCA. In contrast, when the evolutionary history of species are taken into account, we see the affinities of species that are not due to shared ancestry. For instance, *Abies* is farther from *Larix* (another gymnosperm) in the pPCA, and is closer to *Tilia* and *Fagus*, at the opposite side of the plot.

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.

Garland, Jr. T., Harvey, P.H. & Ives, A. R. (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* 41, 18-32.

Ho, L. S. T. and Ané, C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* 63:397-408.

Ives, A. R. and T. Garland, Jr. 2010. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology* 59:9-26.

Purvis, A. & Rambaut, A. (1995) Comparative Analysis by Independent Contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Bioinformatics* 11(3), 247-251.

Garland, T., Jr. and A. R. Ives (2000). Using the past to Predict the Present: Confidence Intervals for Regression Equations in Phylogenetic Comparative Methods. *The American Naturalist*, Vol. 155, No. 3. (Mar., 2000), pp. 346-364.

Revell, L. J. (2009). Size-Correction and Principal Components for Interspecific Comparative Studies. *Evolution* 63: 3258-3268.

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