Exercises for linear mixed effect models, part 2

Timothée Bonnet

March 29, 2019

Contents

1 Random interactions

1.1 Random slopes

* Exercise 1 Random slopes and unbalanced data

Load the dataset `hares.csv`. It contains (fake) measurements of snowshoe hare color (darkness) and their detectability against the background where they live. Measurements were taken in 50 different locations. We want to know whether darkness has an effect on detectability.

1. Fit a simple linear model of detectability on darkness. What is the effect?
2. Add a random intercept to the previous model. Does it change the result quantitatively?
3. Add a random slope. What do you see now?

Answer of exercise 1

```
dat <- read.csv("hares.csv")
summary(lm(detectability ~ 1 + darkness, data = dat) )
summary(lmer(detectability ~ 1 + darkness + (1|location), data = dat) )
summary(lmer(detectability ~ 1 + darkness + (1+darkness|location), data = dat) )
```

The direction of the slope changes when you add a random slope (but not when you add a random intercept).

** Exercise 2 Visualize random slopes

Visualize the effect from the random slope model, that is, plot the relationship detectability over darkness for every location. Add the overall relationship (for instance extracted from a simple linear regression). You can use the functions `ranef()` and `fixef()`. Why did the fixed effect of darkness changed so much when you added the random slope?

Answer of exercise 2

```
mhar <- lmer(detectability ~ 1 + darkness + (1+darkness|location), data = dat)

## Error:  'data' not found, and some variables missing from formula environment

rhar <- ranef(mhar)$location

## Error in ranef(mhar):  object 'mhar' not found

fhar <- fixef(mhar)

## Error in fixef(mhar):  object 'mhar' not found

dark<- seq(from=min(dat$darkness), to=max(dat$darkness), length.out = 100)

## Error in seq(from = min(dat$darkness), to = max(dat$darkness), length.out
= 100):  object 'dat' not found

#first we plot the MEDIAN relationship, that is, the slope to the
# hypothetical location with random effects of zero

plot(x=dark, y=fhar[1]+dark*fhar[2], ylim=c(0,4), type="l",
     lwd=5, lty=2, main="Mean effect (dotted),
     Median effect (dashed) and location slopes")

## Error in plot(x = dark, y = fhar[1] + dark * fhar[2], ylim = c(0, 4), :
object 'dark' not found

#then we plot slopes for each locations
cls <- rainbow(n = nrow(rhar))

## Error in nrow(rhar):  object 'rhar' not found

for(i in 1:nrow(rhar))
{
  lines(x=dark, y=(fhar[1]+rhar[i,1])+#intercept for location i
          dark*(fhar[2]+rhar[i,2]),# slope for location i
    col=cls[i])
}

## Error in nrow(rhar):  object 'rhar' not found

m0har <- lm(detectability ~ 1 + darkness, data = dat)

## Error in is.data.frame(data):  object 'dat' not found

abline(m0har, lwd=5, lty=3)

## Error in abline(m0har, lwd = 5, lty = 3):  object 'm0har' not found

legend(x = "topleft", legend = c("Simple regression", "Random slope"),
       lwd=5, lty = c(3,2))

## Error in strwidth(legend, units = "user", cex = cex, font = text.font):
plot.new has not been called yet
```

The problem with the simple regression comes from the imbalance in the data! One location has a positive relationship, while the 49 others have a negative relationship. The simple regression is also positive. Why does one location have so much weight? Let's look at sample sizes per location:

```
table(dat$location)

## Error in table(dat$location):  object 'dat' not found
```

510 observations come from location 50! That is why it weights so much in the simple regression. In contrast the random slope can tell the mean effect from the mediam slope.

** Exercise 3 Does natural selection vary?

Load the dataset `AllM.txt`. It contains true data from the long term monitoring of a wild animal population. We are interested in quantifying natural selection on Weight. To simplify let's assume natural selection is the slope of `fitnessR` on `Weight`.

1. Fit a linear regression of fitnessR on Weight. Include `Age` as a predictor Is there evidence for selection?
2. Change your model to a mixed model with year as a random intercept. Do you think fitnessR varies a lot among years?
3. Now add a random slope on weight. How much variation is there in selection?
4. Make a graph to visualize selection on different years (the function `ranef()` extract random effects) (you can make the graph for adults, for juveniles, or both).
5. Looking at the estimated variance for the intercept and for Weight, which one looks more important? Is that your impression graphically? Why?
6. Bonus: test for the statistical significance of the variation in selection (you can use `anova()` to compare two models).

Answer of exercise 3

Careful here! The dataset is a text file, not a csv file! You need to load the data using (read.table()) with the argument `header=TRUE`!

```
library(lme4)
allm <- read.table("AllM.txt", header = TRUE)
```

1. Simple linear regression

```r
mlr <- lm(fitnessR ~ 1 + Age + Weight, data = allm)
summary(mlr)

##
## Call:
## lm(formula = fitnessR ~ 1 + Age + Weight, data = allm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.243 -1.258 -0.602  1.058 11.958
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.35759    0.34410    1.04      0.3
## AgeJ        -1.31987    0.15564   -8.48   <2e-16 ***
## Weight       0.06937    0.00808    8.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.31 on 2906 degrees of freedom
##   (114 observations deleted due to missingness)
## Multiple R-squared:  0.227,Adjusted R-squared:  0.227
## F-statistic:  428 on 2 and 2906 DF,  p-value: <2e-16
```

It looks like there is selection for heavier individuals.

2. Random intercept

```
mri <- lmer(fitnessR ~ 1+ Age  + Weight + (1|Year), data = allm)
summary(mri)

## Linear mixed model fit by REML ['lmerMod']
## Formula: fitnessR ~ 1 + Age + Weight + (1 | Year)
##    Data: allm
##
## REML criterion at convergence: 12956
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.161 -0.585 -0.238  0.442  5.062
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Year     (Intercept) 0.595    0.771
##  Residual             4.966    2.228
## Number of obs: 2909, groups:  Year, 8
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.47454    0.43158     1.1
## AgeJ        -1.57284    0.15241   -10.3
## Weight       0.06283    0.00786     8.0
##
## Correlation of Fixed Effects:
##        (Intr) AgeJ
## AgeJ   -0.676
## Weight -0.763  0.830
```

Some years have much higher or much lower fitness on average.

3.Random slope

```
mrs <- lmer(fitnessR ~ 1+ Age  + Weight  + (1+ Weight|Year), data = allm)
summary(mrs)

## Linear mixed model fit by REML ['lmerMod']
## Formula: fitnessR ~ 1 + Age + Weight + (1 + Weight | Year)
##    Data: allm
##
## REML criterion at convergence: 12872
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.478 -0.570 -0.205  0.432  5.174
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Year     (Intercept) 1.8465   1.359
##           Weight      0.0026   0.051    -0.94
##  Residual             4.8037   2.192
## Number of obs: 2909, groups:  Year, 8
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.1515     0.5918    1.95
## AgeJ         -1.5632     0.1510  -10.35
## Weight        0.0444     0.0198    2.24
##
## Correlation of Fixed Effects:
##        (Intr) AgeJ
## AgeJ   -0.490
## Weight -0.932  0.328
```

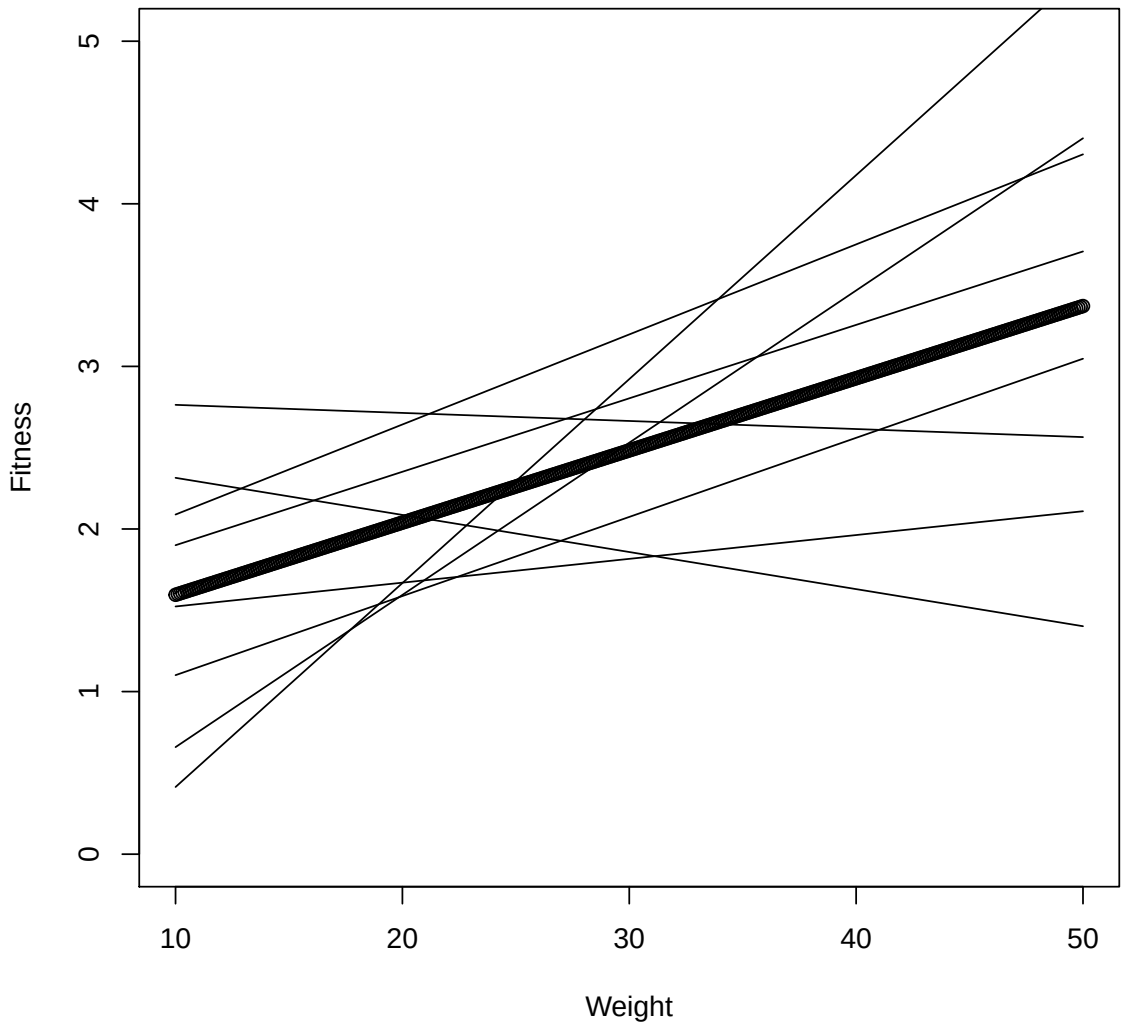
Selection may fluctuate (the variance component for the slope of Weight is not null), but it looks small and it is very correlated to the random intercept. So it is difficult to tell.

4.Visualize

```r
wgt <- seq(10,50, by=0.1)
fixef(mrs)

## (Intercept)        AgeJ      Weight
##     1.15147    -1.56323     0.04439

plot(x=wgt, y=fixef(mrs)[1] + fixef(mrs)[3]* wgt, ylim = c(0,5),
     xlab="Weight", ylab="Fitness")
rde <- ranef(mrs)$Year
for (i in 1:nrow(rde))
{
  lines(x=wgt, y=fixef(mrs)[1]+  rde[i,1] + (fixef(mrs)[3]+rde[i,2])* wgt)
}
```



5. Statistical test

```r
anova(mri,mrs)

## refitting model(s) with ML (instead of REML)

## Data: allm
## Models:
## mri: fitnessR ~ 1 + Age + Weight + (1 | Year)
## mrs: fitnessR ~ 1 + Age + Weight + (1 + Weight | Year)
##     Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## mri  5 12954 12984  -6472    12944
## mrs  7 12877 12918  -6431    12863  81.2      2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can compare the random intercept only model to the random intercept and random slope model to see if adding the latter improves the fit. The p-value is approximate only, but that doesn't matter because it is tiny anyways.

1.2 Random factor interaction

* Exercise 4 Random interaction with a factor

Load the file `interfactor.csv` and fit two random interaction modelw of y as a function of treat, using both the reaction norm and the character state approach. How do the estimate differ? Use the functions `AIC()`, `fitted()` and `resid()` to compare the fit of the two models? What can you conclude?

Answer of exercise 4

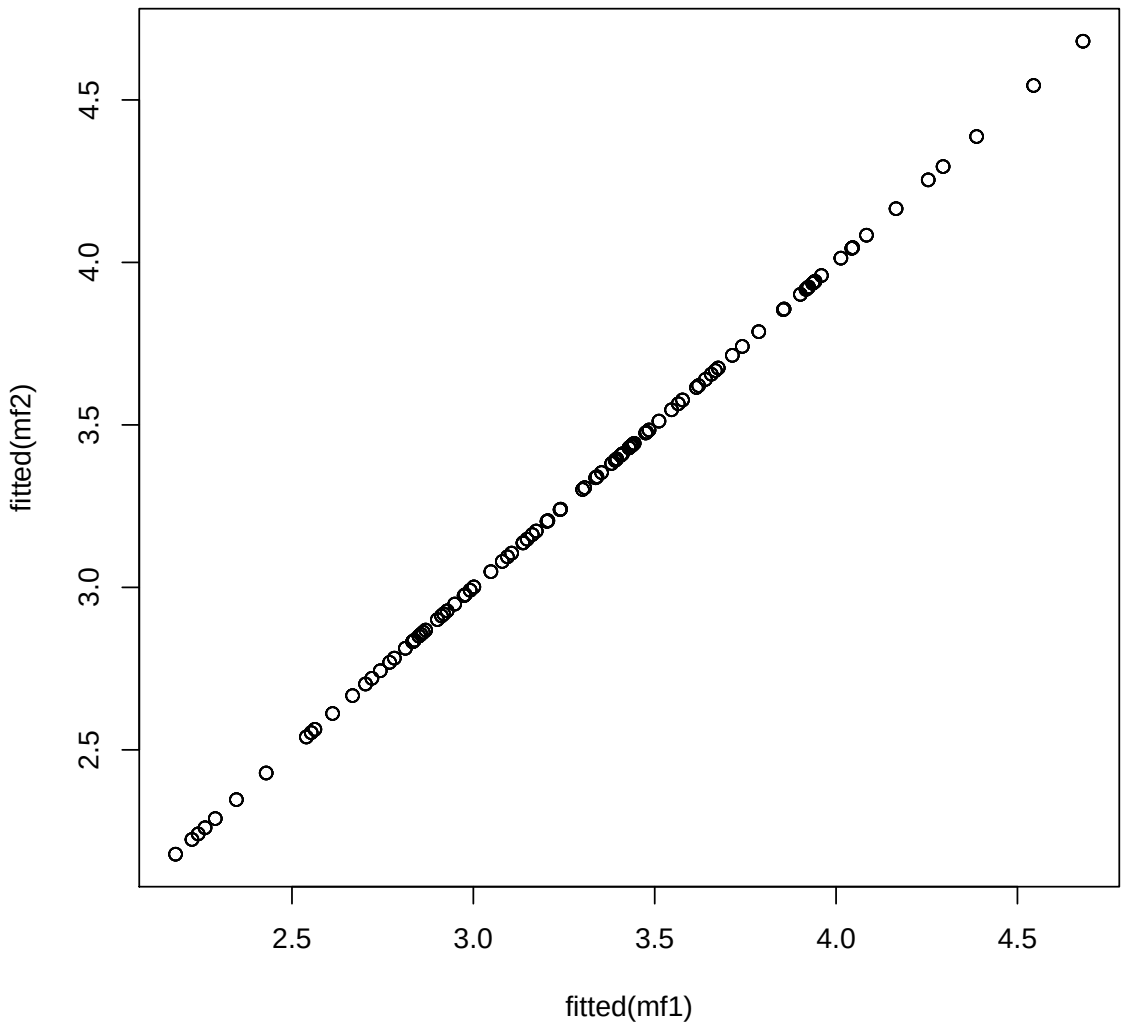
```r
dat <- read.csv("interfactor.csv")
library(lme4)
summary(mf1 <- lmer(y ~ 1 + treat + (1+treat|id), data = dat))

## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + treat + (1 + treat | id)
##    Data: dat
##
## REML criterion at convergence: -2335
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4730 -0.6422 -0.0409  0.6072  2.4375
##
## Random effects:
##  Groups   Name           Variance Std.Dev. Corr
##  id       (Intercept)    0.1740   0.417
##           treattreatment 0.0606   0.246    0.45
##  Residual                0.0001   0.010
## Number of obs: 500, groups:  id, 50
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)      3.0734     0.0590    52.1
## treattreatment   0.4525     0.0348    13.0
##
## Correlation of Fixed Effects:
##             (Intr)
## treattrtmnt 0.451
```

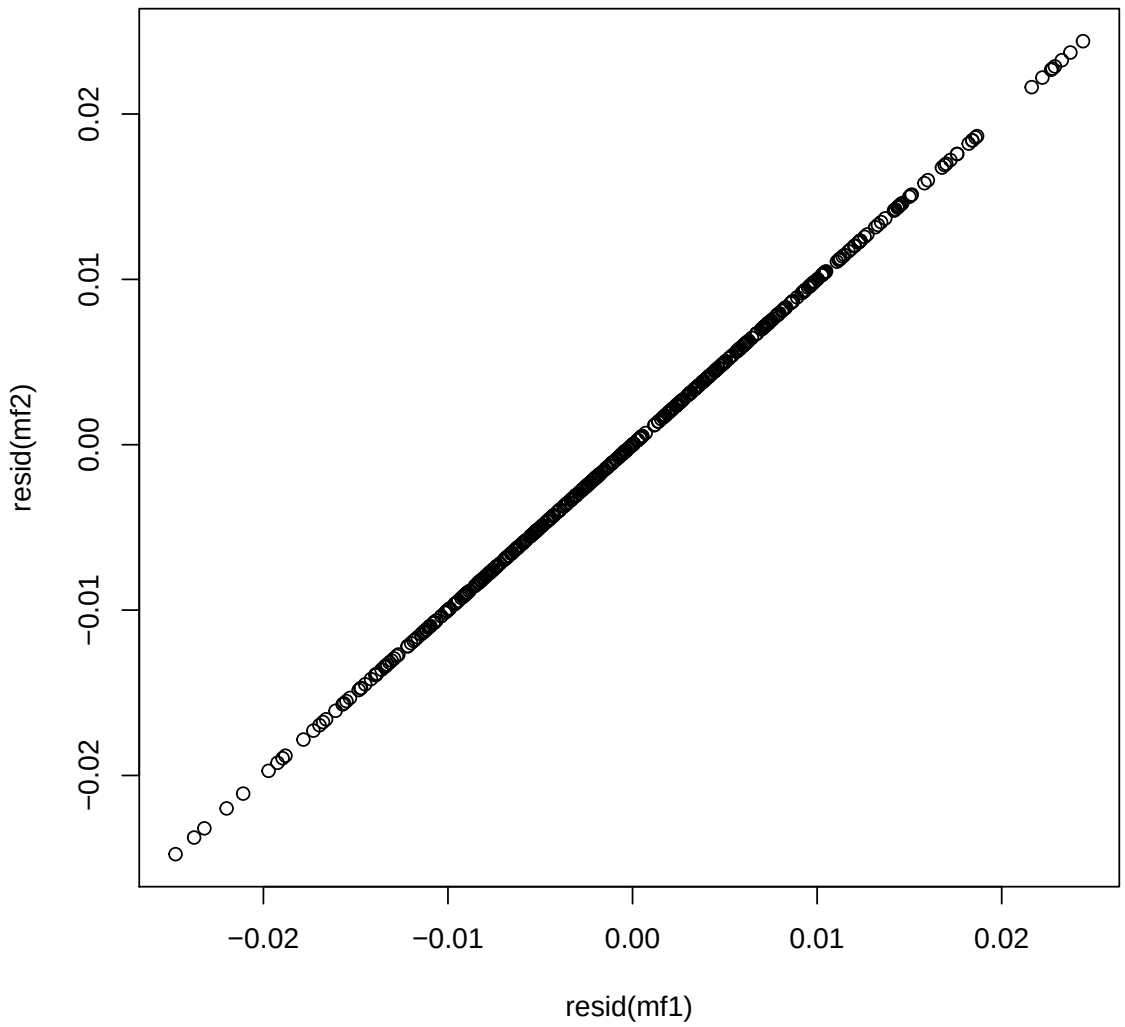
```r
summary(mf2 <- lmer(y ~ 1 + treat + (0+treat|id), data = dat))

## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + treat + (0 + treat | id)
##    Data: dat
##
## REML criterion at convergence: -2335
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4730 -0.6422 -0.0409  0.6072  2.4375
##
## Random effects:
##  Groups   Name           Variance Std.Dev. Corr
##  id       treatcontrol   0.1740   0.417
##           treattreatment 0.3272   0.572    0.92
##  Residual                0.0001   0.010
## Number of obs: 500, groups:  id, 50
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)      3.0734     0.0590    52.1
## treattreatment   0.4525     0.0348    13.0
##
## Correlation of Fixed Effects:
##             (Intr)
## treattrtmnt 0.451

plot(fitted(mf1), fitted(mf2))
```




```
plot(resid(mf1), resid(mf2))
```




```
AIC(mf1)

## [1] -2323

AIC(mf2)

## [1] -2323
```

The two models are equivalent. Actually you can convert one into the other:

```r
vc1 <- VarCorr(mf1)
vm1 <- matrix(vc1$id[1:4], nrow = 2)

vc2 <- VarCorr(mf2)
vm2 <- matrix(vc2$id[1:4], nrow = 2)

trs <- matrix(c(1,1, 0,1), nrow = 2)

trs %*% vm1 %*% t(trs)

##        [,1]   [,2]
## [1,] 0.1740 0.2203
## [2,] 0.2203 0.3272

vm2

##        [,1]   [,2]
## [1,] 0.1740 0.2203
## [2,] 0.2203 0.3272
```

** Exercise 5 Beetles: build a model

Load the dataset "beetles.csv". It contains (fake) data from an (real) experiment on gene-by-environment interactions. The variable of interest is the mass of beetles born in two different environments, from different parents, and in different cages. Assuming that we can measure genetic varition with parent random effects, we wonder if different genomes respond differently to different environments. **Build the model corresponding to this question in lme4.**

(hints: you could start from a lm() of mass modeled by environment, then add random intercepts, and finally a little something more).

** Exercise 6 Beetles: look at the model

What are the variances related to genetic differences? How are they correlated? Does genetic variation explain a lot of the total variation we observe? Try and draw a representation of genetic variation in the two environments.

*** Exercise 7 Beetles: interpret

Interpret model outputs (use raw numbers and / or graphes) to answer the following:
Is there evidence for genetic variation? Do the two environment differ in their effects on beetles?
Is there evidence for genetic variation in the response to the environment?
Does that mean that genomes good at environment 1 are bad at environment 2?

2 Correlated random effects

In all of the above, we have assumed that random effect levels to be perfectly correlated (e.g., observations from the same year) or not at all correlated (e.g., observations from different years). It can be very interesting to allow for intermediate values, in particular for models of spatio-temporal autocorrelation, phylogenetics, quantitative genetics.

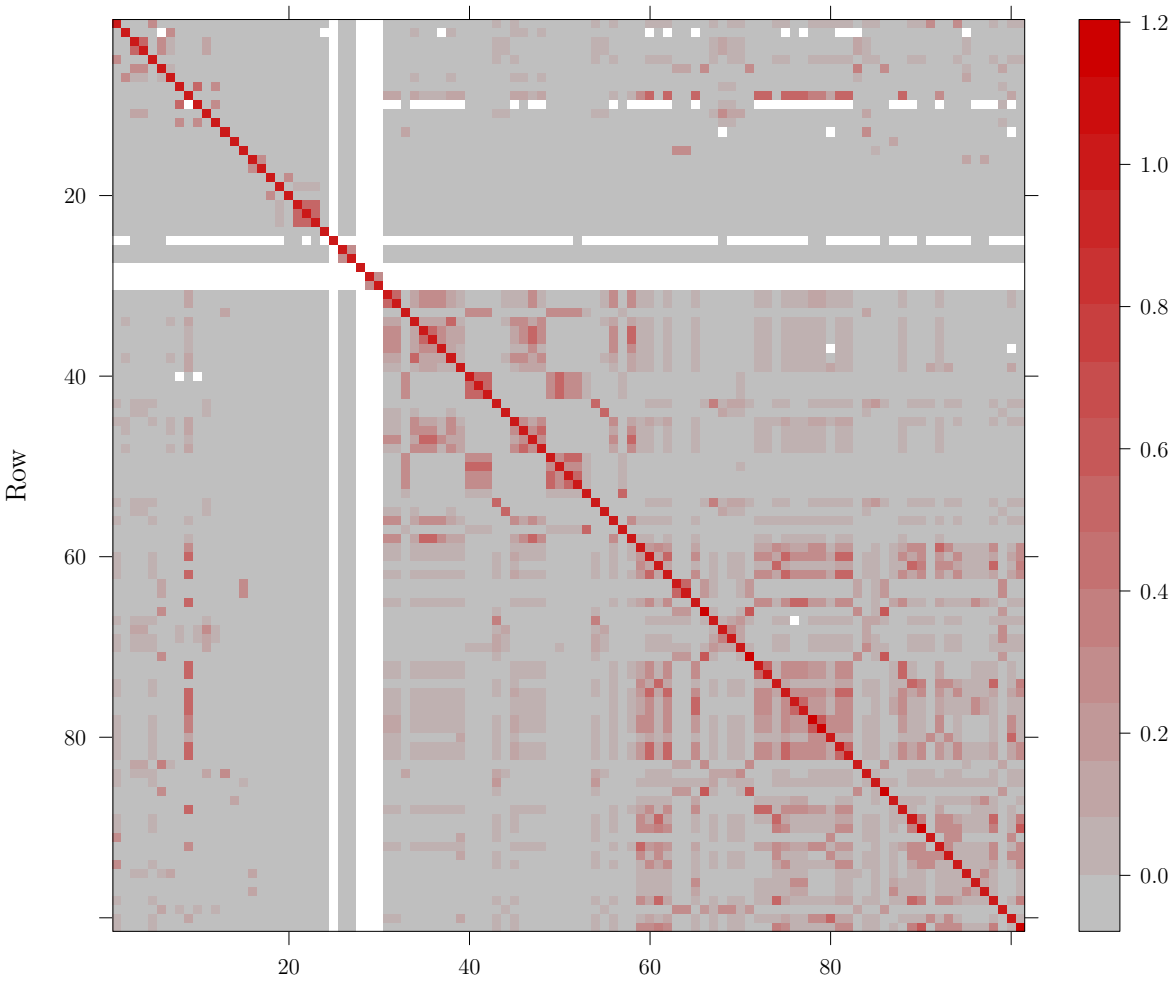
2.1 Quantitative genetics

Genes are transmitted from parents to offspring in a very predictable way: their genes have a correlation of 50%. Therefore it is possible to estimate the genetic variance without any DNA sequencing when a dataset contains parents and offspring. Going back to the long term monitoring of the wild animal population, let's load the population pedigree (ped.txt) and calculate this relatedness matrix:

```
library(MCMCglmm)

## Loading required package:  coda
## Loading required package:  ape

ped <- read.table("ped.txt", header = TRUE)
ainv <- inverseA(ped)$Ainv #the inverse of relatedness matrix
image(solve(ainv)[500:600,500:600], useRaster=T) #the relatedness matrix
```



Dimensions: 101 x 101

Is there genetic variation in weight? Here is a demonstration of fitting a quantitative genetic model.

```
allm <- read.table("AllM.txt", header = TRUE)

mweight <- MCMCglmm(Weight ~ 1 + Age*Sex, random=~Year + id,
                    ginverse = list(id=ainv), data=allm)
summary(mweight)
```

The effect "id" has a large variance attached to it, suggesting the presence of a lot of genetic variation.

* Exercise 8 Is it really genetic?

However that was a bit cheating, because individuals had several observations, so the "genetic" random effect may be just repeated measurements. Let's add another random effect for individual, but not connected to the relatedness matrix (you can use the variable "animal" which is a duplicate of "id").

Answer of exercise 8

```
mweightG <- MCMCglmm(Weight ~ 1 + Age*Sex, random=~Year + id + animal,
                     ginverse = list(animal=ainv), data=allm, nitt = 30000)
summary(mweightG)
```

2.2 Phylogenetic model

In a phylogenetic tree some species have a longer common evolutionary history than others. What those species look like may be influenced by the common evolutionary history. We can model that by considering phylogenetic correlations between two lineage as the time of common evolution relative to their outgroup.

* Exercise 9 Demo of Phylogeny and correlated phenotypes

Load a phylogeny of bird families and some bird phenoypes:

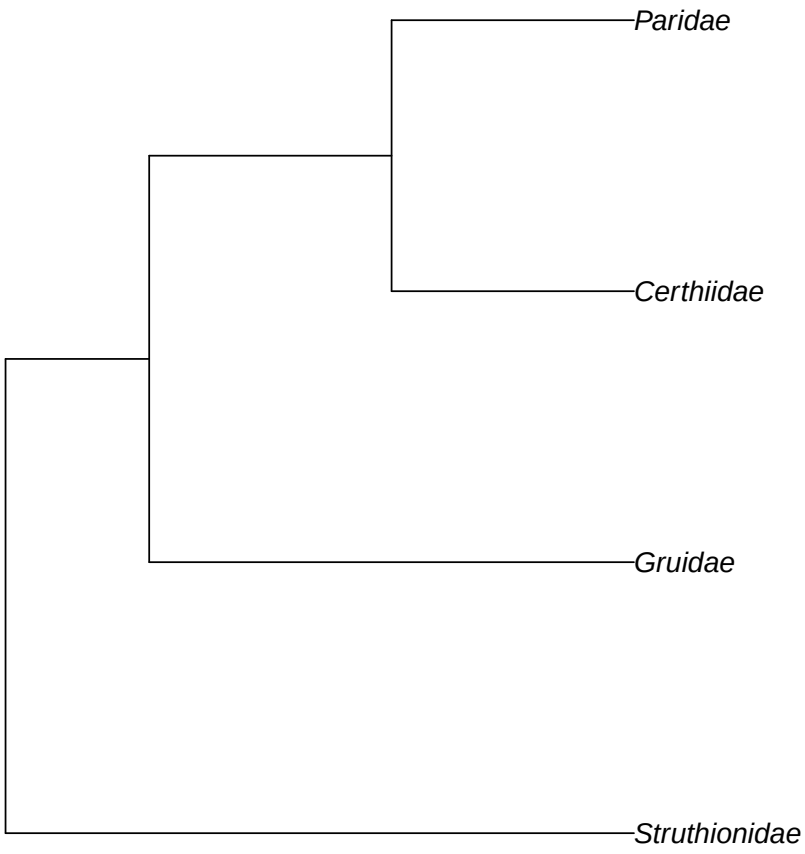
```
load("birdfamilies") #phylogeny. Loading creates the object bird.families
birds <- read.csv("birdpheno.csv") #family phenotypes
```

To start, we subset the phylogeny to a few families

```
bird.families <- makeNodeLabel(bird.families)
some.families <- c("Certhiidae", "Paridae", "Gruidae",
"Struthionidae")
Nphylo <- drop.tip(bird.families, setdiff(bird.families$tip.label,
some.families))
```

So we can easily visualize it:

```
plot(Nphylo)
```



Which families will tend to have more similar phenotypes do you think?

We can calculate the variance covariance matrix of that tree as:

```
library(MCMCglmm)
INphylo <- inverseA(Nphylo)
sA <- as.matrix(solve(INphylo$Ainv))
colnames(sA) <- rownames(sA) <- rownames(INphylo$Ainv)
sA
```

```
##               Node58 Node122 Struthionidae Gruidae Certhiidae Paridae
## Node58        0.2286  0.2286             0  0.2286     0.2286  0.2286
## Node122       0.2286  0.6143             0  0.2286     0.6143  0.6143
## Struthionidae 0.0000  0.0000             1  0.0000     0.0000  0.0000
## Gruidae       0.2286  0.2286             0  1.0000     0.2286  0.2286
## Certhiidae    0.2286  0.6143             0  0.2286     1.0000  0.6143
## Paridae       0.2286  0.6143             0  0.2286     0.6143  1.0000
```

What are the zero?

Coming back to the full dataset, we can fit a phylogenetic model as:

```
INphylofull <- inverseA(bird.families) # Object contains inverse relatedness matrix,
prior0 <- list(G=list(G1=list(V=1, nu=1, alpha.mu=0, alpha.V=100)),
               R=list(V=1, nu=0.002))
m1 <- MCMCglmm(y ~ 1, random = ~id, ginverse = list(id=INphylofull$Ainv),
               data = birds, prior = prior0)
summary(m1)
```

Answer of exercise 9