Appendix S2: Analysis of budburst data using MCMCglmm

Simon Joly and Elizabeth Wolkovich

October 2017

Contents

Preface

This file describe the analysis of the budburst data using Phylogenetic Mixed Models using the MCMCglmm R package.

Data preparation

Load packages

```
# Packages -----
library(ape)
library(MCMCglmm)
library(ggplot2)
library(plyr)
```

Prepare the datasets

Let's load the datasets that we will need to fit the models.

```
# Import the data files
load(file="./data/budbust.Rdata")
```

There are three objects:

1. `thedata` contains the budburst data. It contains the species information (`sp`), the time to budburst (`lday`), the warming treatment (`warm`), the photoperiod treatment (`photo`), the population from where the individual was collected (`site`), and a code identifying the individuals (`ind`).
2. `treephylo` contains the phylogenetic tree
3. `intra.mat` contains the block diagonal matrix of the intraspecific genetic similarities.

Let's have a look at the first lines of the data.

```
head(thedata,n=8)
```

```
##                   sp lday warm photo site          ind
## ACEPEN01_HF1 ACEPEN   48   15    12   HF ACEPEN_MA_01
## ACEPEN01_HF2 ACEPEN   34   20     8   HF ACEPEN_MA_01
## ACEPEN01_HF3 ACEPEN   26   20    12   HF ACEPEN_MA_01
## ACEPEN01_HF4 ACEPEN   68   15     8   HF ACEPEN_MA_01
## ACEPEN02_HF1 ACEPEN   40   15    12   HF ACEPEN_MA_02
## ACEPEN02_HF2 ACEPEN   68   15     8   HF ACEPEN_MA_02
## ACEPEN02_HF3 ACEPEN   NA   20    12   HF ACEPEN_MA_02
## ACEPEN02_HF4 ACEPEN   40   20     8   HF ACEPEN_MA_02
```

Format the data for the MCMCglmm analysis

```
# MCMCglmm data preparation ----

# Single value decomposition of the intraspecific structure
# follow code of Stone et al. 2012
intra.svd <- svd(intra.mat)
intra.svd <- intra.svd$v %*% (t(intra.svd$u) * sqrt(intra.svd$d))
rownames(intra.svd) <- colnames(intra.svd) <- rownames(intra.mat)

# Remove node names from phylogeny
treephylo$node.label <- NULL
# MCMCglmm needs ultrametric trees
is.ultrametric(treephylo)
```

```
## [1] FALSE
```

```
# Make tree ultrametic
treephylo <- chronos(treephylo)
```

```
##
## Setting initial dates...
## Fitting in progress... get a first set of estimates
##          Penalised log-lik = -59.30403
## Optimising rates... dates... -59.30403
##
## Done.
```

```
class(treephylo) <- "phylo"
# check again
is.ultrametric(treephylo)
```

```
## [1] TRUE
```

```
# Rename the column with species names "animal". This is required for MCMCglmm
colnames(thedata)[1] <- "animal"

# Prepare intraspecific correlation matrix. Specifically, the individuals have to
# appear the same number of times as in the dataset and in the same order
matches <- match(thedata[,"ind"],rownames(intra.svd))
# Reorder intra.svd to fit the data
intra.svd <- intra.svd[unique(matches),unique(matches)]
```

```
# Inflate matrix: duplicate rows and comlums for the number of treatments
# present in the dataset.
intra.svd <- matrix(apply(apply(intra.svd,2,function(c)
  rep(c,times=table(matches))),1,function(c)
    rep(c,times=table(matches))),nrow=sum(table(matches)),ncol=sum(table(matches)))

# Finally, remove rows with missing data
thedata <- na.omit(thedata)
intra.svd <- intra.svd[-attr(thedata,"na.action"),-attr(thedata,"na.action")]
```

Data analysis

Model fitting

```
# Model M0 is a simple model with no random factors
priorpr.m0 <- list(R = list(V = 1, nu = 0.002))
M0 <- MCMCglmm(lday ~ warm * photo, data=thedata, scale=TRUE,
               nitt=105000,thin=20,burnin=5000,prior=priorpr.m0)

# Model M1 is a simple phylogenetic model without intraspecific correlation structure
priorpr.m1 <- list(R = list(V = 1, nu = 0.002),
                   G = list(G1 = list(V = 1, nu = 0.002)))
M1 <- MCMCglmm(lday ~ warm * photo,
               random = ~ animal, pedigree = treephylo, data=thedata,
               scale=TRUE,nitt=105000,thin=20,burnin=5000,prior=priorpr.m1)

# Model M2 has only intraspecific structure
M2 <- MCMCglmm(lday ~ warm * photo,
               random =~ idv(intra.svd), data=thedata, scale=TRUE,
               nitt=105000,thin=20,burnin=5000,prior=priorpr.m1)

# Model M3 has both phylogenetic and intraspecific structure
priorpr.m3 <- list(R = list(V = 1, nu = 0.002),
                   G = list(G1 = list(V = 1, nu = 0.002),
                            G2 = list(V = 1, nu = 0.002)))
M3 <- MCMCglmm(lday ~ warm * photo,
               random =~ idv(intra.svd) + animal,
               pedigree = treephylo, data=thedata, scale=TRUE,
               nitt=105000,thin=20,burnin=5000,prior=priorpr.m3)
```

Model comparison

Now that we ran the models, we can compare the different models using the Deviance Information Criterion (DIC; lower values are best).

```
# Compare fit of the models (deviance information criterion)
data.frame(models=c("M0","M1","M2","M3"),
           random.effects=c("NA","inter","intra","inter+intra"),
           DIC=c(M0$DIC,M1$DIC,M2$DIC,M3$DIC))
```

```
##   models random.effects      DIC
```

```
## 1     M0             NA 2425.950
## 2     M1          inter 2134.513
## 3     M2          intra 2104.556
## 4     M3    inter+intra 2097.295
```

It is clear from these results that the M3 model with the inter + intra specific structure is the best.

Run convergence

Before looking at this model more closely, it is important to look at the convergence of the MCMC run. To evaluate this, it is useful to run an independent analysis.

```
# Do another run to check for convergence
M3b <- MCMCglmm(lday ~ warm * photo,
                random =~ idv(intra.svd) + animal,
                pedigree = treephylo, data=thedata, scale=TRUE,
                nitt=105000,thin=20,burnin=5000,prior=priorpr.m3)
```

Once this is done, we can calculate the Potential scale reduction factors for the random effects.

```
# Potential scale reduction factors (PSRF)

# PSRF of fixed effects
gelman.diag(mcmc.list(M3$Sol,M3b$Sol))
```

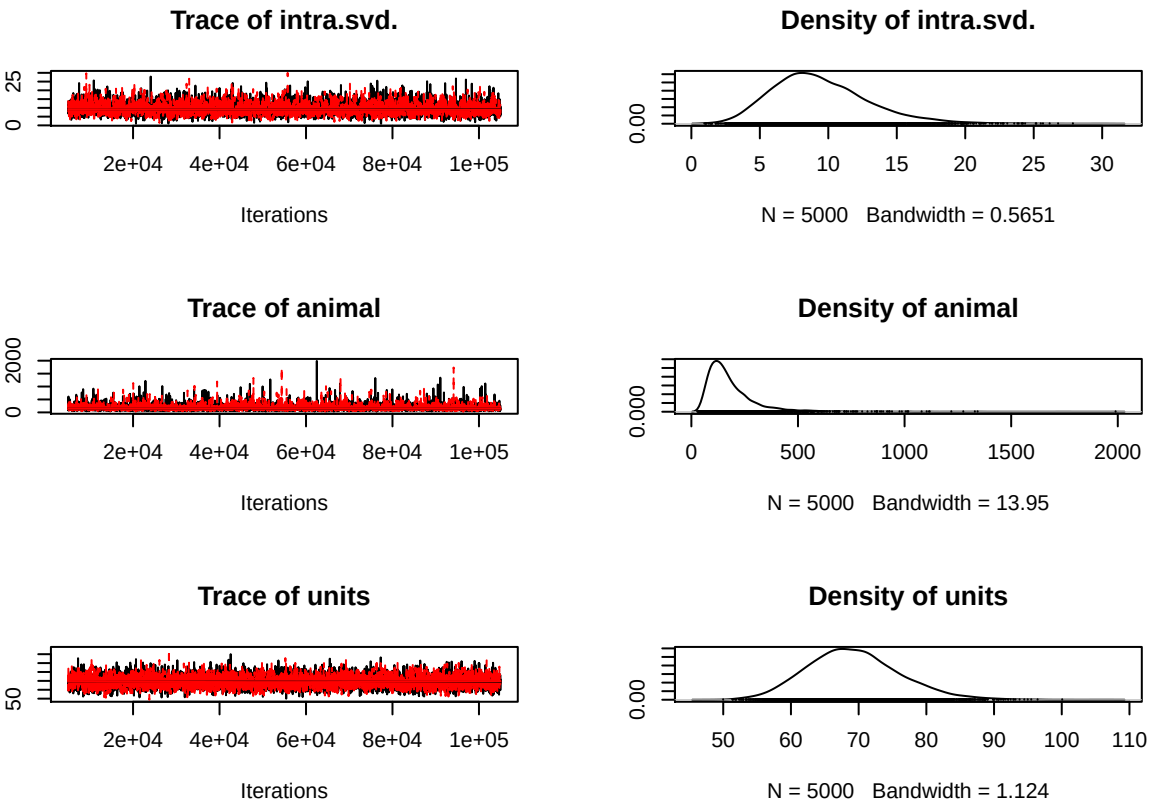
```
## Potential scale reduction factors:
##
##                 Point est. Upper C.I.
## (Intercept)              1       1.00
## warm20                   1       1.00
## photo12                  1       1.01
## warm20:photo12           1       1.00
##
## Multivariate psrf
##
## 1
```

```
# PSRF of random effects
gelman.diag(mcmc.list(M3$VCV,M3b$VCV))
```

```
## Potential scale reduction factors:
##
##            Point est. Upper C.I.
## intra.svd.          1       1.01
## animal              1       1.00
## units               1       1.00
##
## Multivariate psrf
##
## 1
```

You can see that these are very close to 1, suggesting very good convergence. This is also evident when looking at the plot of the values generation per generation as the mixing is very good.

```
# look at MCMC chain sampling
plot(mcmc.list(M3$VCV,M3b$VCV))
```



Model summary

Now that we are convinced that the analyses have converged, we can look at the results summary (here from a single run).

```
# Summary of best model
summary(M3)
```

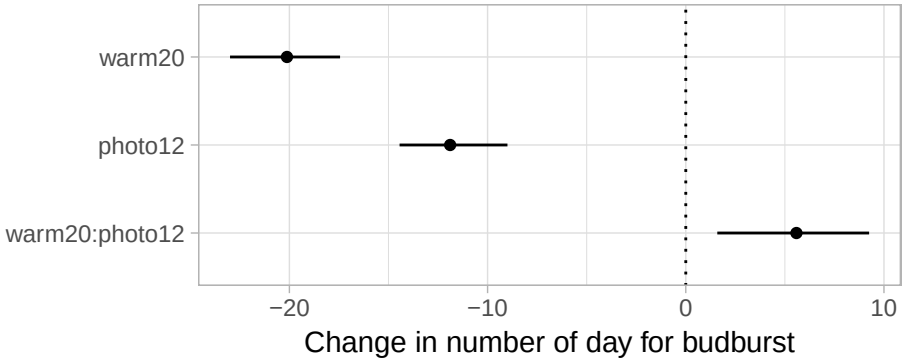
```
##
##  Iterations = 5001:104981
##  Thinning interval  = 20
##  Sample size  = 5000
##
##  DIC: 2097.295
##
##  G-structure:  ~idv(intra.svd)
##
##             post.mean l-95% CI u-95% CI eff.samp
## intra.svd.        9.5      3.4    17.06     2515
##
##                ~animal
##
##        post.mean l-95% CI u-95% CI eff.samp
## animal     183.4    36.34    415.3     5000
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
```

```
## units     69.34    56.81    83.31     3668
##
##  Location effects: lday ~ warm * photo
##
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)        56.085   45.516   67.006     5000 <2e-04 ***
## warm20            -20.124  -22.995  -17.444     5000 <2e-04 ***
## photo12           -11.889  -14.441   -8.998     5000 <2e-04 ***
## warm20:photo12      5.583    1.590    9.250     5000 0.0064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effective sample sizes (corrected for the autocorrelation along the chain) are all relatively good, both for random and fixed effects.

We can also visualize the fixed effects using a figure.

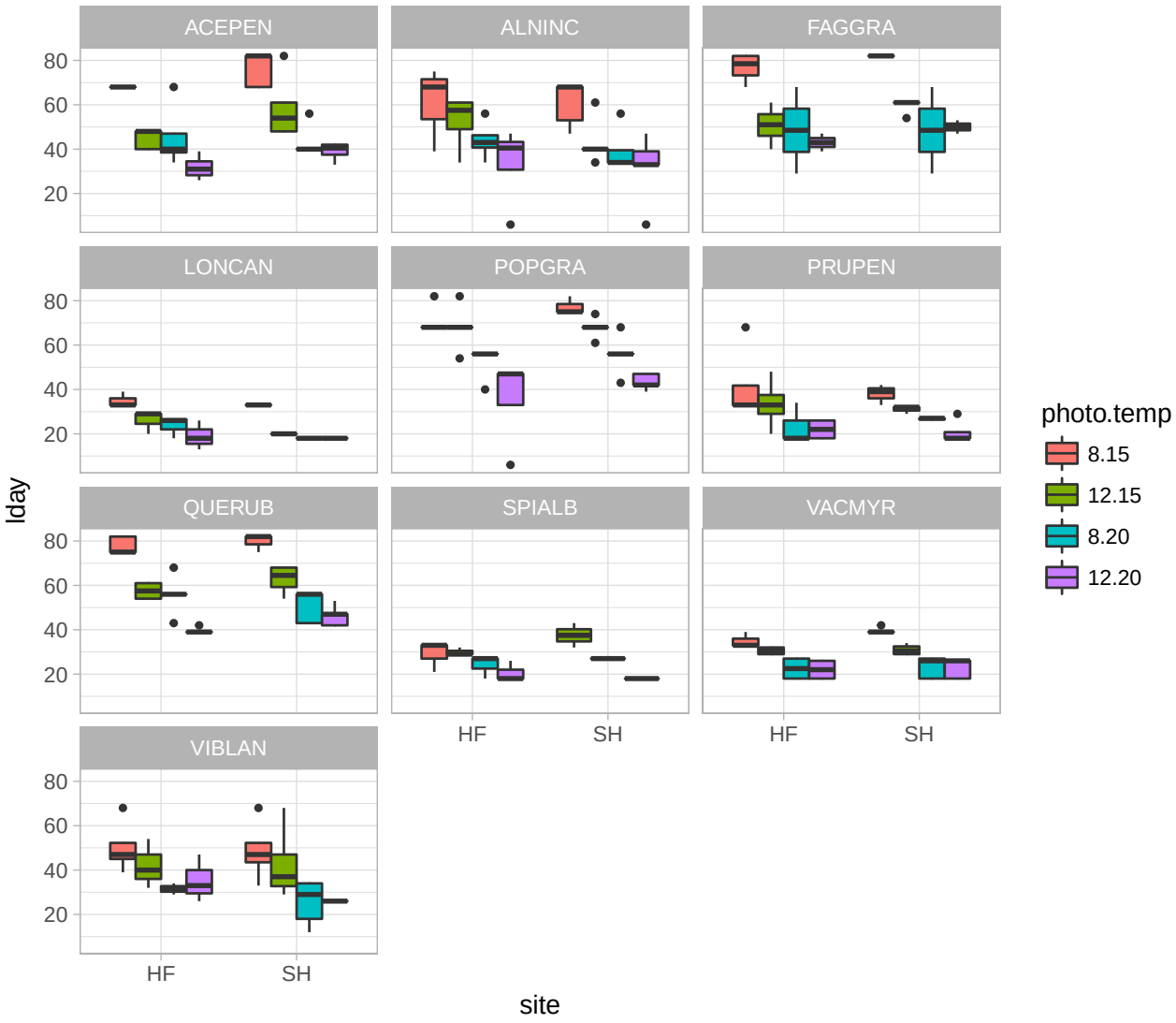
```
#Fixed effects
results.M3 <- as.data.frame(summary(M3)$solutions)
results.M3$effects <- factor(rownames(results.M3),levels=rownames(results.M3))
colnames(results.M3)[2:3] <- c("lowerCI","upperCI")
fixed.p <- ggplot(results.M3[-1,], aes(y=post.mean, x=effects)) +
  geom_point(color="black") +
  geom_linerange(aes(ymin=lowerCI, ymax=upperCI)) +
  geom_hline(yintercept = 0, lty=3) +
  scale_x_discrete(limits = rev(levels(results.M3$effects)[-1])) +
  ylab("Change in number of day for budburst") + xlab("") +
  coord_flip() + theme_light()
fixed.p
```



For the **fixed effects**, the model suggests that temperature and the photoperiod are highly significant. The interaction between the temperature and the photoperiod (`warm:photo`) is also significant.

Let's have a look at the raw data:

```
library(ggplot2)
thedata$photo.temp <- interaction(thedata$photo,thedata$warm)
p <- ggplot(thedata,aes(y=lday,x=site))
p + geom_boxplot(aes(fill=photo.temp),outlier.size = 1) +
  facet_wrap(~animal,ncol=3) + theme_light()
```



It is interesting to see that in general the photoperiod as a stronger effect at 15 C than at 20 C. This explains the significant interaction. This interaction appears to be slightly species specific, however.

Let's now look at the **random effects**. To evaluate variance explained by the random effects, it is useful to look at the relative proportion of the variance explained by each effect and by the residuals (`units`):

```
# Proportion of variance explained by random factors
rand <- M3$VCV/apply(M3$VCV,1,sum)
# Get median values (50%) and 95% quantiles
apply(rand,2,function(c) quantile(c,probs = c(0.025,0.5,0.975)))
```

```
##       intra.svd.    animal     units
## 2.5%  0.01086823 0.4191567 0.1157023
## 50%   0.03793241 0.6562535 0.3013342
## 97.5% 0.09548553 0.8687318 0.5065852
```

```
# Get the mean value
apply(rand,2,mean)
```

```
## intra.svd.     animal      units
##  0.0419605  0.6537490  0.3042905
```

```
# Also get 95% quantiles and mean for Heredity (H^2)
rand <- apply(M3$VCV[,1:2],1,sum)/apply(M3$VCV,1,sum)
quantile(rand,probs = c(0.025,0.5,0.975))
```

```
##      2.5%       50%     97.5%
## 0.4934148 0.6986658 0.8842977
```

```
mean(rand)
```

```
## [1] 0.6957095
```

```
# And the intraspecific variance relative to the phylogenetic variance
rand <- M3$VCV[,1]/apply(M3$VCV[,1:2],1,sum)
quantile(rand,probs = c(0.025,0.5,0.975))
```

```
##       2.5%        50%      97.5%
## 0.01310956 0.05502066 0.17056332
```

```
mean(rand)
```

```
## [1] 0.06454885
```

The phylogenetic component explains the greatest portion of the variance (ca. 65%). The intraspecific component is small (ca. 6.5%), but it is significantly greater than 0 and its addition significantly improved the model fit, as shown above.

It is interesting to compare these results to models where the intraspecific structure is excluded. Here the model with only the phylogenetic genetic structure:

```
summary(M1)
```

```
##
##  Iterations = 5001:104981
##  Thinning interval  = 20
##  Sample size  = 5000
##
##  DIC: 2134.513
##
##  G-structure:  ~animal
##
##        post.mean l-95% CI u-95% CI eff.samp
## animal     198.1    51.52    422.7     5399
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units     87.98    73.47    102.5     5000
##
##  Location effects: lday ~ warm * photo
##
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)        55.992   44.841   66.833     5000 <2e-04 ***
## warm20            -19.982  -23.267  -17.054     4358 <2e-04 ***
## photo12           -11.681  -14.626   -8.568     5000 <2e-04 ***
## warm20:photo12      5.397    1.196    9.709     5000 0.0176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can see that the interaction between temperature and photoperiod is similar in terms of effect size (it is only slightly lower), but it is less significant.

And now the model in which no random effects are taken into account.

```
summary(M0)
```

```
##
##  Iterations = 5001:104981
##  Thinning interval  = 20
##  Sample size  = 5000
##
##  DIC: 2425.95
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units     248.2      206    289.1     5000
##
##  Location effects: lday ~ warm * photo
##
##                post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)       58.019   54.391   61.892     5000 <2e-04 ***
## warm20           -19.627  -24.962  -14.739     5000 <2e-04 ***
## photo12          -10.803  -15.855   -5.585     5000 <2e-04 ***
## warm20:photo12     4.285   -3.187   11.175     5000  0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, the interaction between the temperature and the photoperiod (`warm:photo`) is not significant anymore, and its effect has shrinked!