Community Ecology With GLLVMs

Sam Perrin

Day One

Welcome to the first of two scripts looking at GLLVMs through an ecological lens. Today we'll focus largely on the use of ordinations, while tomorrow will focus a bit more on JSDMs. This script was written to be fully self-containing, and while the lecture slides that go with it add a bit of onfo, you should be able to run everything here independently. There's also an R markdown document that will have the figures produced by the scripts below.

We start by loading all the requisite packages.

```
library(gllvm)
library(dplyr)
library(grDevices)
```

The data we're using for this session was collected largely by the ECOSPAT Group at the University of Lausanne. It looks at co-occurrence patterns in different plants along an elevational gradient in the Swiss Alps. The original dataset can be found at https://doi.org/10.5061/dryad.8mv11 (https://doi.org/10.5061/dryad.8mv11), which also contains links to a fantastic paper by Manuela D'Amen,Heidi Mod, Nicholas Gotelli and Antoine Guisan which uses the data, entitled "Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence". The elevational data we use here comes from SwissTopo.

As a very basic overview, the original data includes presence-absence data for 183 plants over 912 sites, and the following environmental covariates. The dataset included below is a subset, so that this all runs a bit faster during the workshop. Obviously subsetting like this isn't always recommended, but if you'd like to run it with the full dataset that has also been included in the GitHUB.

These are our environmental covariates.

DDEG0 - Days over zero degrees
SOLRAD - Summed annual solar radiation
SLOPE - Slope angle in degrees
MIND - Moisture index
TPI - Topographic position index
ELEVATION - Number of kangaroos present at site (joking it's just elevation)

Each of these envionmental covariates has been standardised to a mean of 0 and SD of 1. This is just to help model convergence.

```
load("WorkshopData.RDA")

Y <- WorkshopData$Y
X <- WorkshopData$X
```

For starters, let's run a basic gllvm using no environmental variables. Note that sd.errors is set to FALSE, since we're not really interested in the effect size of environmental covariates in this particular session.

```
time1 <- Sys.time()
fit_base <- gllvm(Y, num.lv = 2, family = binomial(link="probit"), sd.errors = FALSE)
Sys.time()-time1
```

```
## Time difference of 25.05499 secs
```

Before we look at our latent variables, let's have a look at collinearity. The following code shows any significant collinearity between our environmental covariates.

```
source("http://www.sthda.com/upload/rquery_cormat.r")
colin <- rquery.cormat(X, type = "flatten", graph = FALSE)
```

```
colin$r %>% filter(abs(cor) > 0.5)
```

```
##      row    column   cor       p
## 1 DDEG0      MIND -0.89 6.9e-141
## 2 DDEG0 ELEVATION -0.99  0.0e+00
## 3  MIND ELEVATION  0.90 4.0e-148
```

The high cor values and low p values indicate some serious collinearity between positive degree days, slope, elevation and moisture index.

Let's see if our latent variables correspond to any of these covariates.

We define colours according to the values of covariates. The darker blue indicates a higher value of the relevant covariate.
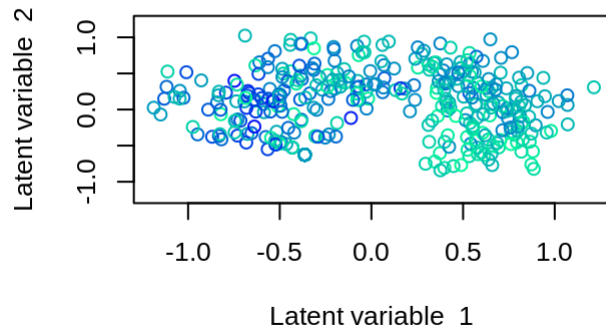
```
par(mfrow=c(2,2))
for (i in 1:length(colnames(X))) {
  covariate <- X[,i]
  rbPal <- colorRampPalette(c('mediumspringgreen', 'blue'))
  Colorsph <- rbPal(20)[as.numeric(cut(covariate, breaks = 20))]
  breaks <- seq(min(covariate), max(covariate), length.out = 30)

  ordiplot(fit_base, main = paste0("Ordination of sites, color: ",colnames(X)[i]),
           symbols = TRUE, s.colors = Colorsph, xlim = c(-1.2,1.2), ylim = (c(-1.2, 1.2
)))
}
```
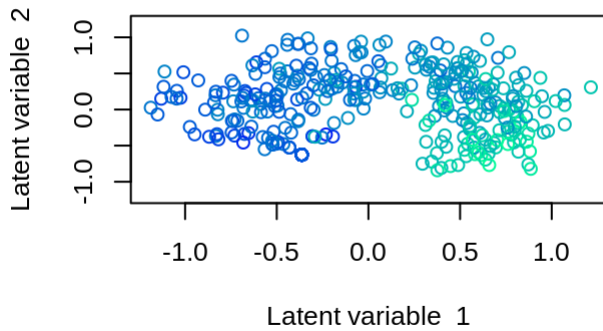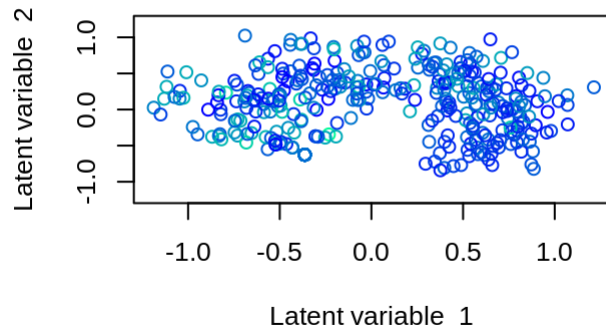
We can see some quite clear gradients related to the four collinear variables we mentioned above. At this point let's take one of the two climate related covariates that could have a direct impact on vegetation (DDEG0) and SLOPE, since MIND is so collinear to DDEG0, and slope might have a more direct impact than elevation on our community.

Let's use these two and use some code from last session to figure out how many latent variables would be appropriate. Basically we're running the same model again and again, but increasing the number of latent variables each time. I'm using the lowest AICc value to determine the model which fits best.

```
fit_list <- list()
for(i in 0:3){
  fit_sub <- gllvm(Y, X, family = binomial(link="probit"), num.lv = i, sd.errors = FALSE,
                formula = ~ SLOPE + MIND, seed = 1234)
  fit_list[[i+1]] <- fit_sub
}
```

Let's have a look at our AICc values.

```
AICc <- sapply(fit_list, function(X) {summary(X)$AICc})
data.frame(AICc, model= paste0("LV-",0:3))
```
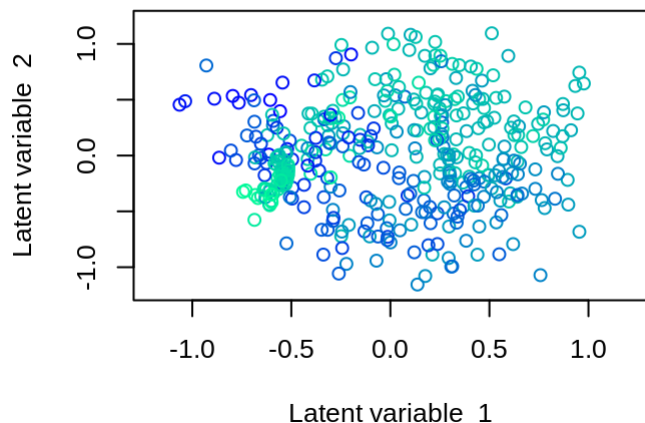
```
##       AICc model
## 1 14687.36  LV-0
## 2 12957.51  LV-1
## 3 12625.36  LV-2
## 4 12970.21  LV-3
```

We can see that the best model here uses 2 latent variables. Let's have a look at how these variables compares to our remaining environmental covariates.

```
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
remaining_covariates <- c("DDEG0","SOLRAD","ELEVATION","TPI")

for(i in 1:length(remaining_covariates)) {
  covariate <- X[,remaining_covariates[i]]
  rbPal <- colorRampPalette(c('mediumspringgreen', 'blue'))
  Colorsph <- rbPal(20)[as.numeric(cut(covariate, breaks = 20))]
  breaks <- seq(min(covariate), max(covariate), length.out = 30)
  ordiplot(fit_list[[3]], main = paste0("Ordination of sites, color: ",remaining_covariates[i]),
         symbols = TRUE, s.colors = Colorsph, xlim = c(-1.2,1.2), ylim = (c(-1.2, 1.2
)))
}
```
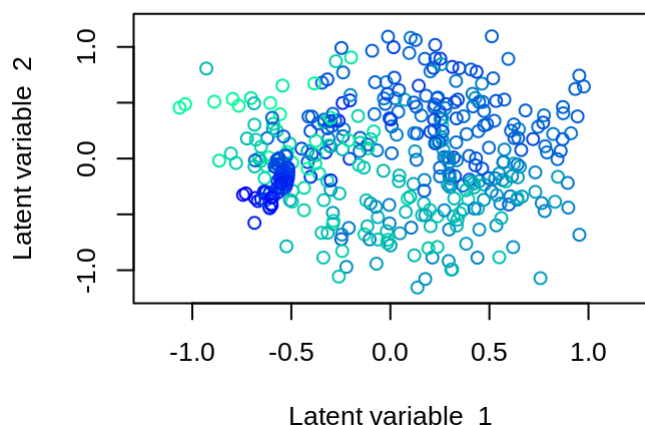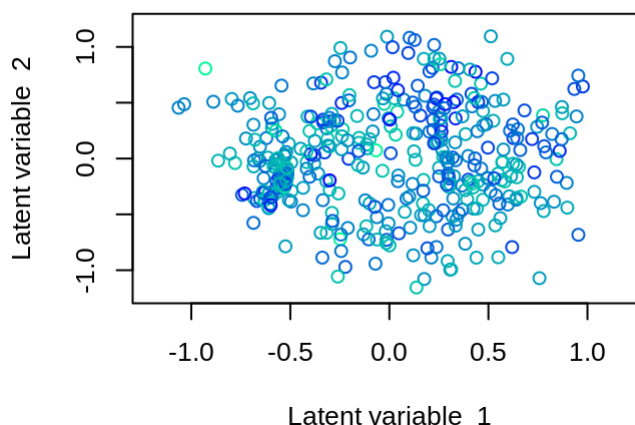
Ordination of sites, color: DDEG0
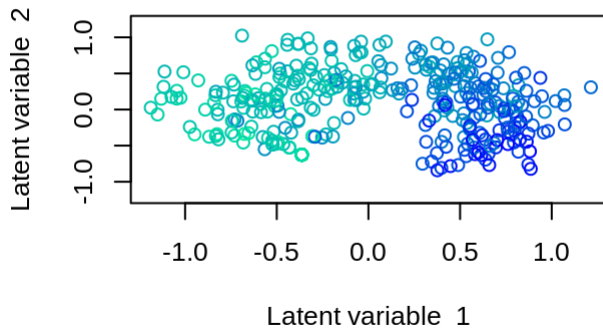
Ordination of sites, color: SOLRAD
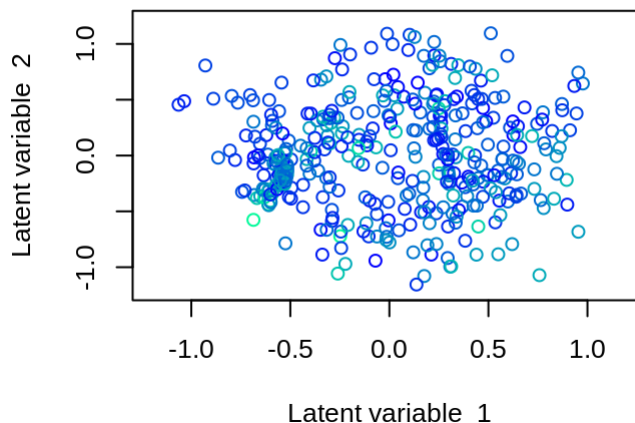
Ordination of sites, color: ELEVATION

Ordination of sites, color: TPI

From this we can see that there is still a bit of variation explained by positive degree days, despite its collinearity with moisture index (and by elevation, but we'll focus on that tomorrow). Let's see what happens when we include degree days over zero.

```
fit_DegreeDays <- gllvm(Y, X, family = binomial(link="probit"), num.lv = 2, sd.errors =
FALSE,
            formula = ~ SLOPE + MIND + DDEG0, seed = 1234)

summary(fit_DegreeDays)$AICc
```

```
## [1] 12743.35
```

We can see the AICc values stay pretty much the same, even rising a bit. But leaving it out means we may attribute variation to our latent variable that is the result of the environment.

▼ EXTENDED QUESTION

Have a look at the coefficient effects using the basic command below, after switching sd.errors to TRUE in your gllvm commands. How do the covariate effects change with the introduction of new variables?

coefplot(fit_Elevation, cex.ylab = 0.5)

Breakout questions

What is a species association?
If we accounted for all possible covariates producing environmental variation, would we still see species associations?

Species Associations

What I've previously done here is group species together based on approximately what elevation their occurrence peaks at. This left us with three groups; montane, subalpine and alpine species. The colour plots below mean we can see each group easily in our ordination plots.

```
colour.groups <- c("red","blue","green")[WorkshopData$elevation_classes]

par(mfrow=c(1,1))
ordiplot.col(fit_base, biplot=TRUE, main = "Ordination of sites: no covariates",
         symbols = TRUE, s.colors = "white", xlim = c(-4,4),ylim=c(-3,3), spp.colors=co
lour.groups)
```
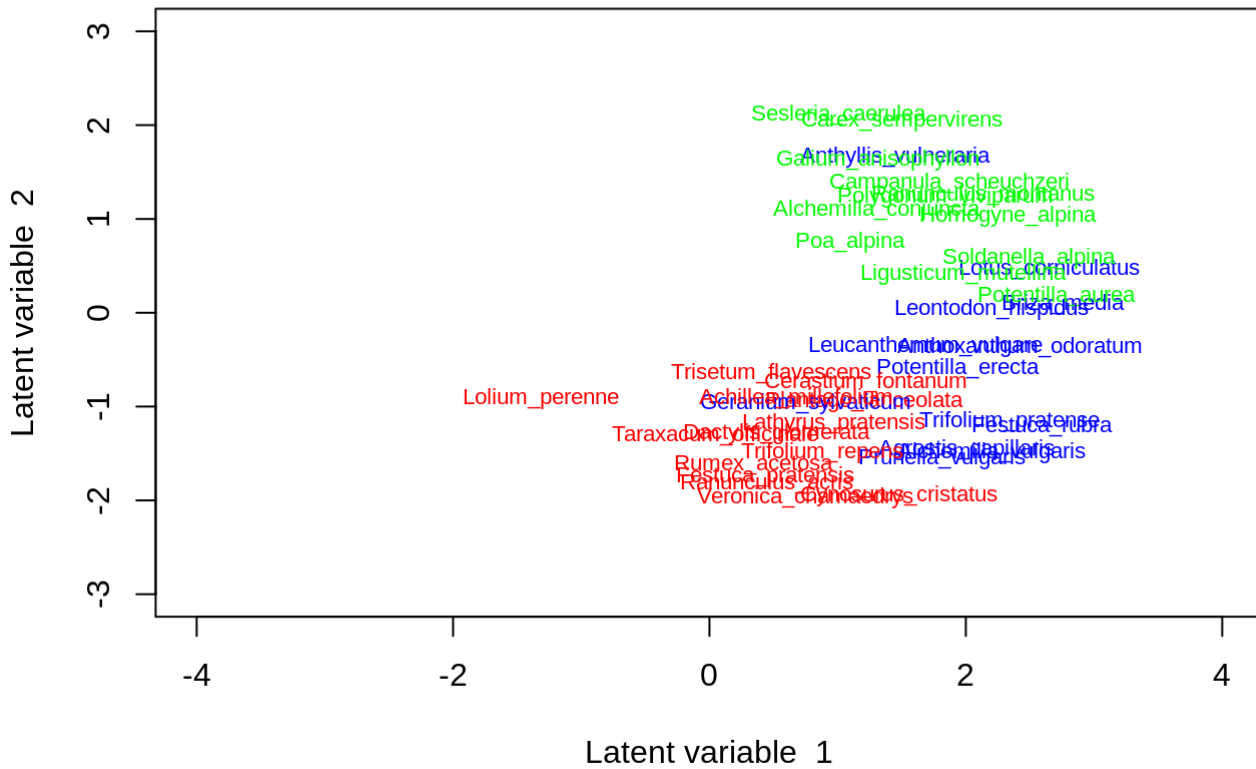
Ordination of sites: no covariates
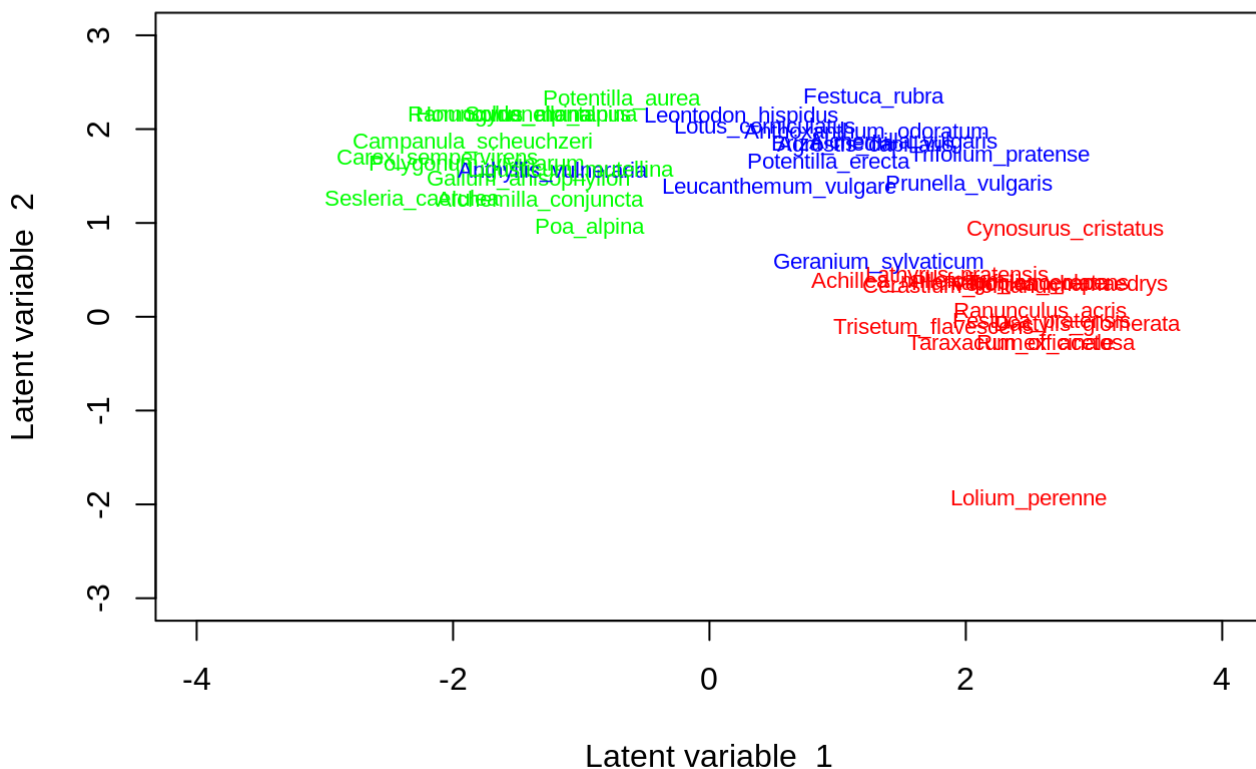
There are some very obvious trends here. Let's see what happens when we introduce MIND and DDEG0.

```
ordiplot.col(fit_list[[3]], biplot=TRUE, main = "Ordination of sites: two covariates",
         symbols = TRUE, s.colors = "white", xlim = c(-4,4),ylim=c(-3,3), spp.colors=co
lour.groups)
```
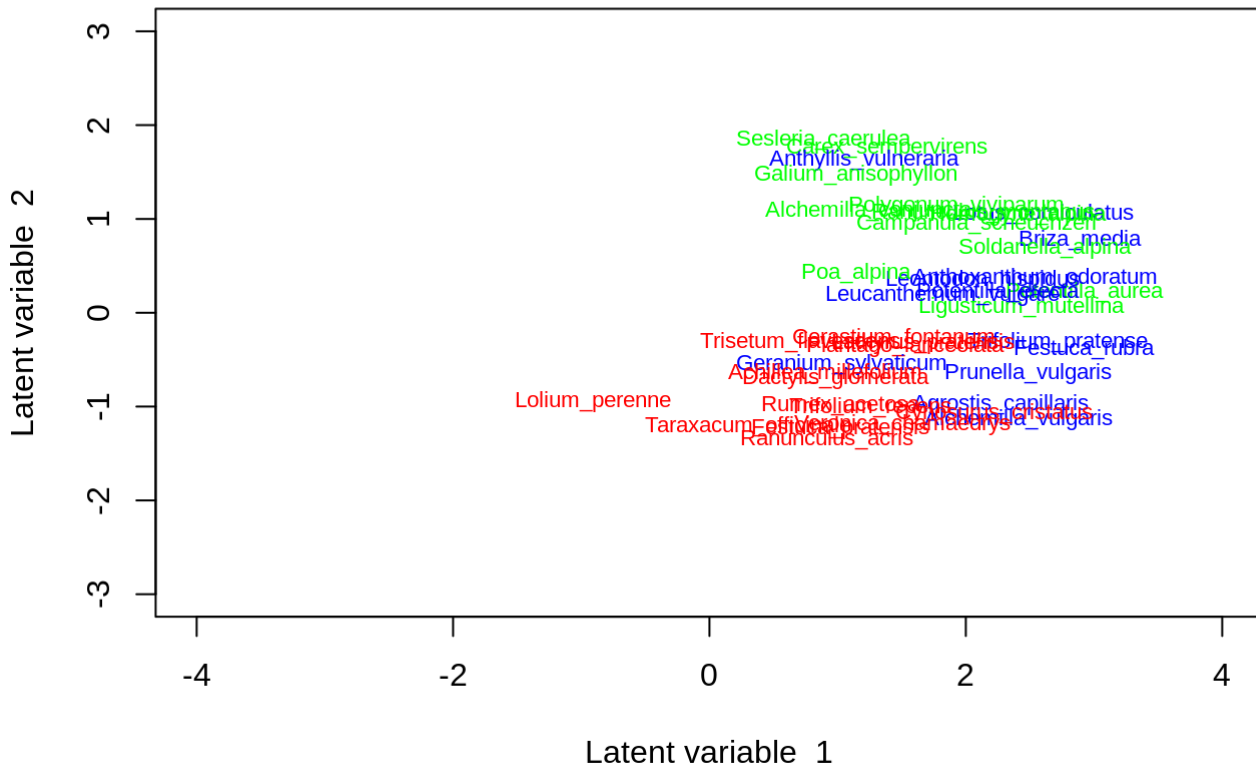
Ordination of sites: two covariates

And now when we introduce degree days as an extra covariate.

```
ordiplot.col(fit_DegreeDays, biplot=TRUE, main = "Ordination of species: three covariat
es",
        symbols = TRUE, s.colors = "white", xlim = c(-4,4),ylim=c(-3,3), spp.colors=co
lour.groups)
```

Ordination of species: three covariates

You can see that the species group together more clearly, as the effect of the latent variable becomes weaker.
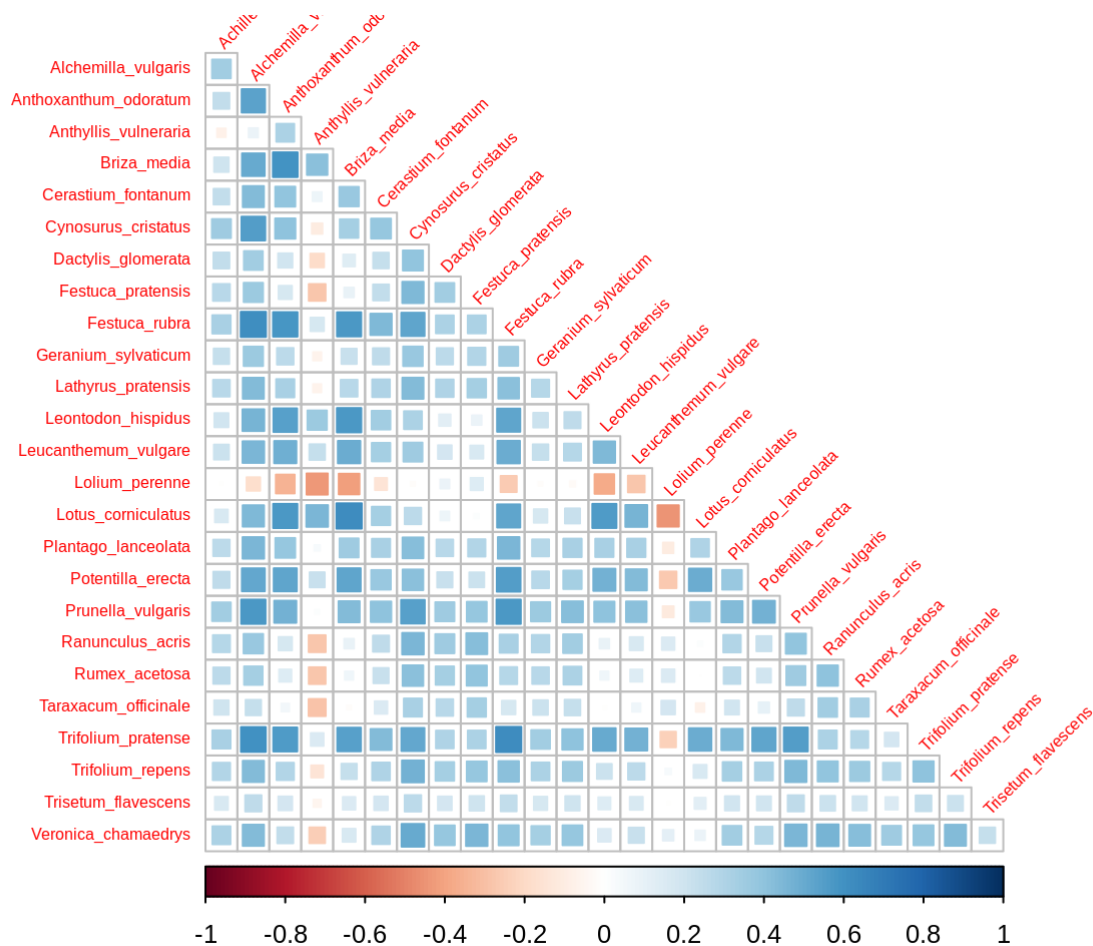
▼ EXTENDED QUESTION

What happens when we incorporate elevation into the equation as well?

Lastly, just for a taste of tomorrow, let's check out a co-occurrence plot.

```
colline_species <- WorkshopData$colline_species

cr1 <- getResidualCor(fit_list[[3]])
corrplot(cr1[colline_species,colline_species], diag = FALSE, type = "lower",
         method = "square", tl.cex = 0.5, tl.srt = 45, tl.col = "red")
```

The blue squares indicate positive associations, which means that the two species are positively associated. Red squares indicate the two species are unlikely to co-occur. Larger, darker squares indicate stronger relationships.

Have any extra questions? Get in touch with us via the conference app! You can also contact me directly via email at sam.perrin@ntnu.no (mailto:sam.perrin@ntnu.no) or on Twitter at @samperrinNTNU.