IllumFUN_Q2a: Which variables to include in path analysis?

Marissa Lee

12/16/2019

Q2. How are differences in fungal composition across the landscape explained by environmental variables?

Table of contents

0. Load data and pre-process ASV matrix

See IllumFUN_Q1.Rmd

A. Determine which environmental variables to include in path analysis

1. Select initial continuous variables
2. Remove variables that are highly correlated (>0.8)
3. Transform to normal all environmental variables
4. Again, check for correlated variables
5. Variable selection w/ LASSO

Load packages, functions, paths

Load custom functions

A. Determine which environmental variables to include in path analysis

1. Select initial continuous variables

Include stand.age as a continuous predictor?

Yes – just need to include a conservative estimate for LWR-BHO's stand age. It is over 10 yrs old, but unclear how old, so fill in as 11 yrs for now.

Calculate max basal area. Use the max basal width and length to calculate ellipse area

```r
# # A = pi * .5(width) * .5(length)
sam %>%
  mutate(basal.area.m2 = pi * (0.5* max.basallength.m) * (0.5* max.basalwidth.m)) -> sam
```

Use bulk density(g/cm3) in NCDA&CS soil report to convert to element conc (mg/dm3) into (ug/g soil)

```
mg.dm3_to_ug.g <- function(x.mg.dm3, bulk){
  # convert from dm3 to cm3
  x.mg.cm3 <- x.mg.dm3 / 1000
  # convert from cm3 to g soil with bulk soil (g/cm3)
  x.mg.g <- x.mg.cm3 / bulk
  # convert from mg/g to ug/g
  x.ug.g <- x.mg.g * 1000
  return(x.ug.g)
}
```

Select initial subset of continuous variables

- *Exclude silt* since sand + clay + silt = 100
- *Exclude nh4 and no3* since nh4 + no3 = TIN
- *Exclude BS, Ac, and CEC* since this is represented by ph and texture

Update phyloseq objects with stand.age and basal area decisions

Print the Site characteristics table

Print the correlation matrix

2. Remove variables that are highly correlated (>0.8)

Bivariate correlations

```
##              var1         var2       cor
## 1              Mg           Ca 0.9469382
## 2              Ca           Mg 0.9469382
## 3          perc.N       perc.C 0.9345720
## 4          perc.C       perc.N 0.9345720
## 5       perc.sand          W.V 0.8694023
## 6             W.V    perc.sand 0.8694023
## 7          perc.N watercontent 0.8515803
## 8    watercontent       perc.N 0.8515803
## 9          perc.C          SOM 0.8327327
## 10            SOM       perc.C 0.8327327
## 11       samp.lon        MAT.C 0.8220813
## 12          MAT.C     samp.lon 0.8220813
## 13         perc.C          mbc 0.8178698
## 14            mbc       perc.C 0.8178698
## 15            doc            S 0.8013372
## 16              S          doc 0.8013372
```

```
##             var1         var2        cor
## 1              S          W.V -0.8518902
## 2            W.V            S -0.8518902
## 3 watercontent          W.V -0.8992811
## 4            W.V watercontent -0.8992811
```

Decide on which correlated variables to exclude

- Soil %C and %N (r = 0.93): *Exclude perc.N* because some mixed-tree plots have distinctly greater %C, not %N
- Soil Mg and Ca (r = 0.91): *Exclude both.* Both are mobile in the form of cations and so are they highly correlated with ph/CEC/texture.

Decide on more variables to exclude

- Soil perc.sand and bulk density (W.V) (r = 0.87): *Exclude bulk density*
- Soil perc.C and SOM (r = 0.83), soil perc.C and mbc (r = 0.82): *Exclude SOM and mbc*

Examine more variables…

Remove: perc.N, Ca, W.V, SOM, mbc

Now there are *20* continuous environmental variables

3. Transform to normal all environmental variables

Transform all predictor variables to normally-distributed since this is required for SEM

Climate

Soil resources

Texture

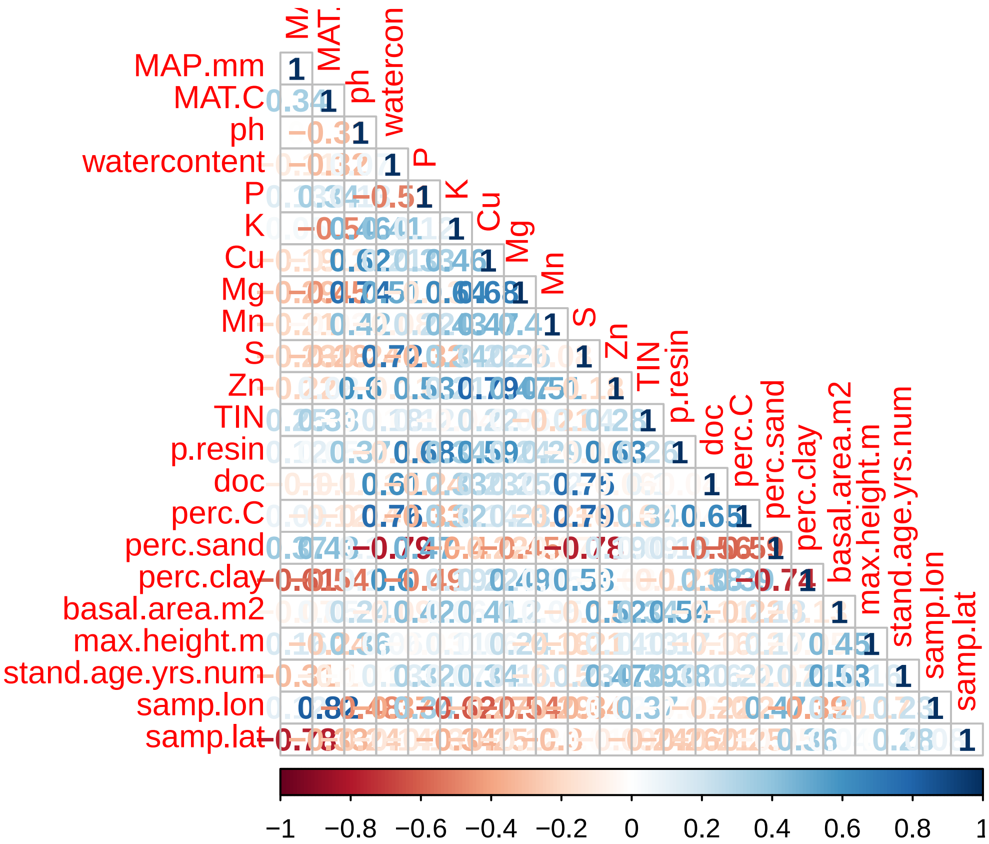
Plant size and stand age

Lat and lon

Save transformed environmental variables

Examine the correlation of the transformed variables

```
## Joining, by = "SiteSamp"
```

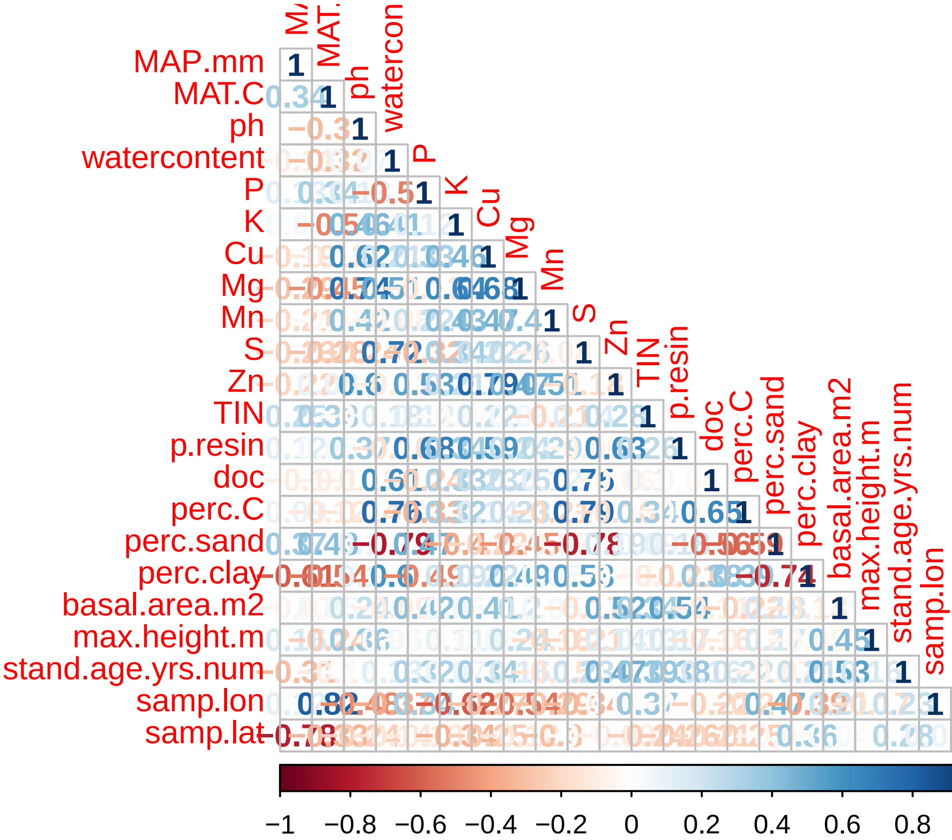


Zn and Cu are highly correlated (0.79); remove Zn. perc.sand and watercontent are highly correlated (-.79); remove watercontent. S and perc.C are highly correlated (0.79); remove S.

lon and MAT are highly correlated (r = 0.82). lat and MAP are highly correlated (r = -0.78). Don't remove lat or lon

Remove highly correlated variables

Need to equalize variances; otherwise get this message from lavaan "Warning message: In lav_data_full(data = data, group = group, cluster = cluster, :lavaan WARNING: some observed variances are (at least) a factor 1000 times larger than others; use varTable(fit) to investigate"



Re-examine the correlation of the variables

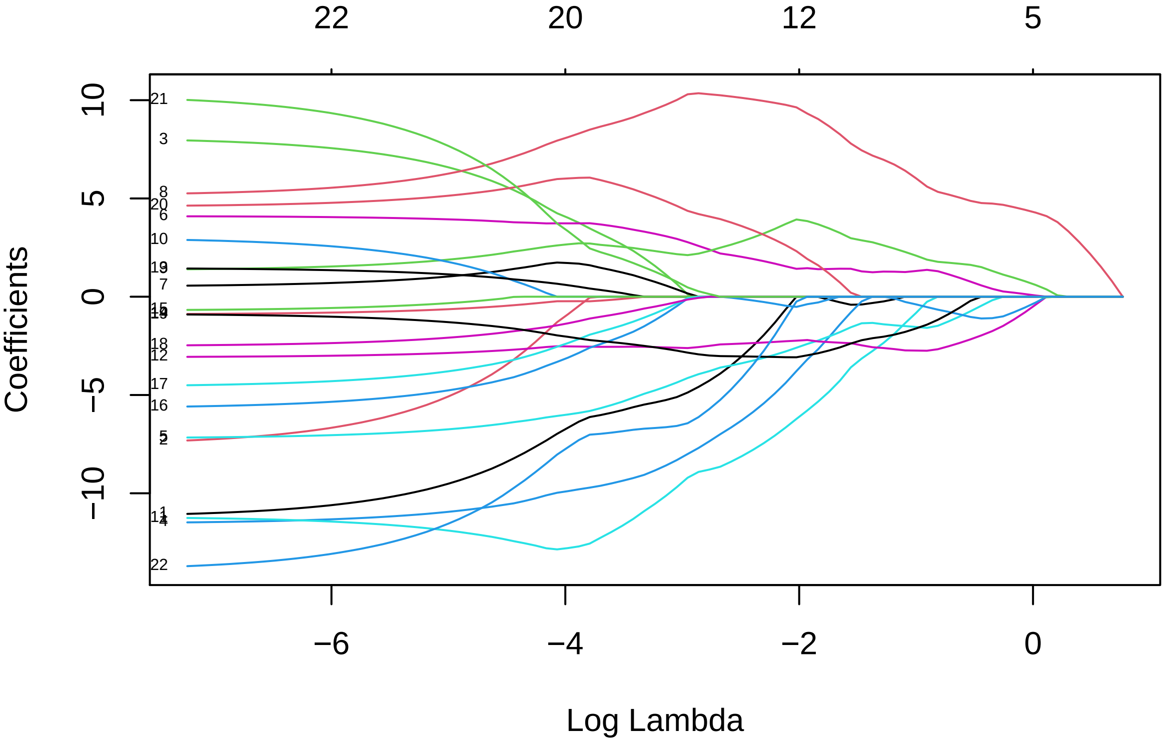
4. Variable selection with LASSO

Leaf

```
# load phylo obj
ps <- readRDS(file = file.path(merged_path, "phyloseq_samps_env_trimTreeASVs_leaf.RData"))

# calc dpcoa
#dpcoa <- DPCoA(ps, correction = cailliez, scannf = FALSE)
dpcoa <- readRDS("output/illumina/Q0/dpcoa_leaf.RData")
df.dpcoa <- data.frame(sample.name.match = row.names(dpcoa$li), dpcoa$li, row.names = NULL)
# # load normalized environmental variables
mat.t <- read.csv(file = file.path(out_path, "normTransformed_contvars_trim_scaled.csv"), row.names = 1

# make dataframe
sam <- data.frame(sample_data(ps))
sam %>%
  select(sample.name.match, Site, SiteSamp) -> sam
sam %>%
  left_join(df.dpcoa) %>%
  left_join(mat.t) %>%
  select(-c(sample.name.match, Site, SiteSamp, Axis2)) -> data
```

```
## Joining, by = "sample.name.match"

## Joining, by = c("Site", "SiteSamp")
```

```
data1 <-as.matrix(data)
#data1<- data1[,!colnames(data1) %in% c("samp.lon","samp.lat")]

# fit LASSO model with range of lambda
require(glmnet)
fit <- glmnet(x = data1[,-1], y = data1[,1], family = "gaussian")
plot(fit, xvar = "lambda", label = T)
```



```
# do cv to find appropriate lambda
cvfit = cv.glmnet(x = data1[,-1], y = data1[,1], family = "gaussian")
#plot(cvfit)
# # extract suggested variables
extract.lambda_uni(cvfit, s = "lambda.1se")
```

```
## [1] "P"              "K"              "Mg"             "Mn"             "TIN"
## [6] "perc.sand"      "max.height.m"
```

```
extract.lambda_uni(cvfit, s = "lambda.min")
```

```
##  [1] "MAP.mm"              "watercontent"      "P"
##  [4] "K"                   "Mg"                "Mn"
##  [7] "Zn"                  "TIN"               "perc.sand"
## [10] "max.height.m"        "stand.age.yrs.num" "samp.lat"
```

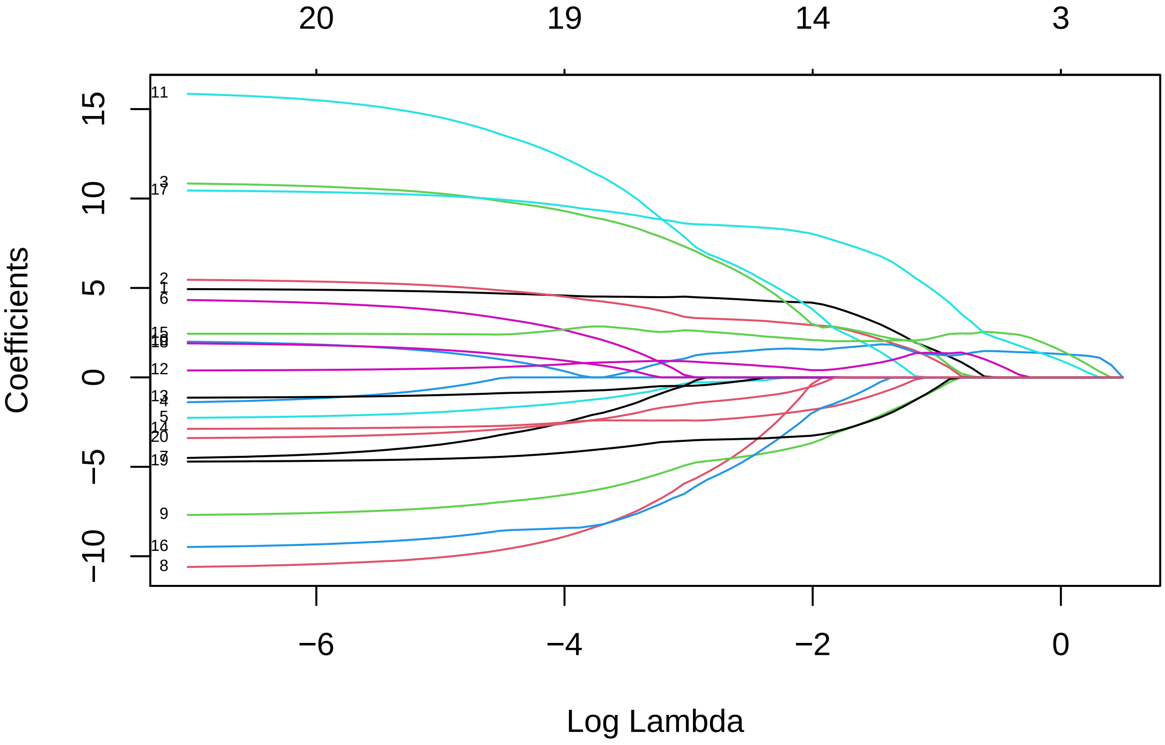
```
#
# # save plots
# pdf(file = file.path(out_path, "leaf_dpcoa_glmnet.pdf"), width = 4, height = 6)
# par(mfrow = c(2,1))
# plot(cvfit)
# plot(fit, xvar = "lambda", label = T)
# dev.off()

# save data
# mat <- data.frame(sam[,c("sample.name.match","Site","SiteSamp")], data1)
# vars <- extract.lambda_uni(cvfit, s = "lambda.min")
# vars
# mat %>%
#   select(sample.name.match, Site, SiteSamp, Axis1, vars) -> mat.vars
# mat.vars
# write.csv(mat.vars, file = file.path(out_path, "leaf_dpcoa_SEMdata.csv"))
```

Note that suggested variables may differ from the manuscript due to sampling stochasticity in the glmnet functions

Root

```
## Joining, by = "sample.name.match"
```

```
## Joining, by = c("Site", "SiteSamp")
```



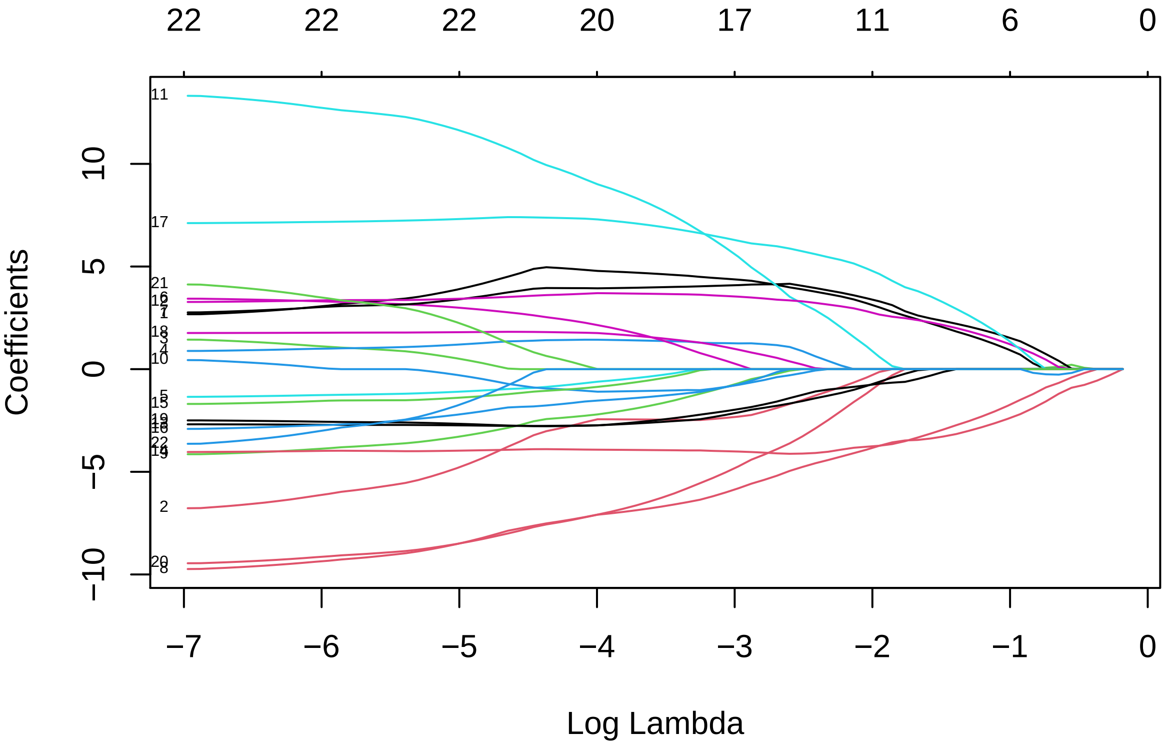
```
## NULL
```

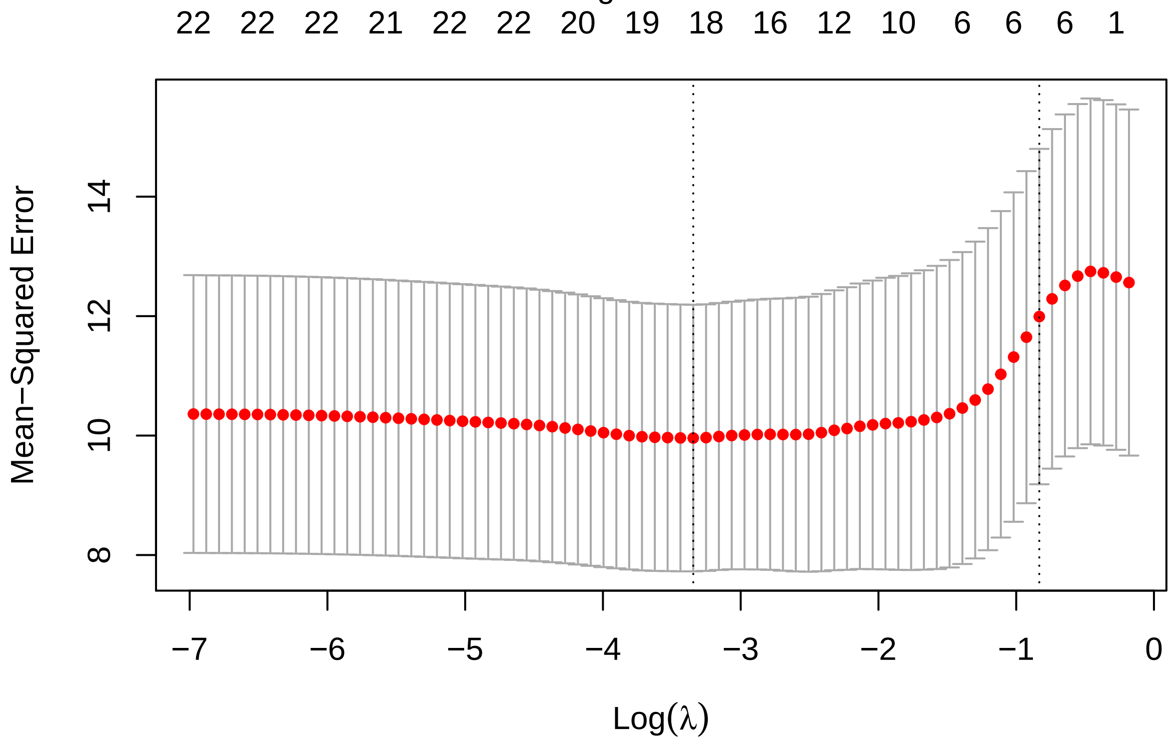
```
##  [1] "MAP.mm"       "MAT.C"        "ph"           "watercontent" "Mn"
##  [6] "Zn"           "TIN"          "doc"          "perc.C"       "perc.sand"
## [11] "perc.clay"    "max.height.m"
```

Soil – Flip the DPCoA axis to help with interpretation

```
## Joining, by = "sample.name.match"
```

```
## Joining, by = c("Site", "SiteSamp")
```





```
## [1] "MAP.mm"             "Cu"                 "TIN"
## [4] "doc"                "perc.clay"          "stand.age.yrs.num"
## [7] "samp.lat"
```

```
##  [1] "MAP.mm"             "MAT.C"              "watercontent"
##  [4] "P"                  "K"                  "Cu"
##  [7] "Mg"                 "Mn"                 "S"
## [10] "Zn"                 "TIN"                "p.resin"
## [13] "doc"                "perc.C"             "perc.sand"
## [16] "perc.clay"          "basal.area.m2"      "max.height.m"
## [19] "stand.age.yrs.num"
```