

multivariate models solutions

Dan McGlinn

2025-02-13

```
# load code dependancies  
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.6-4
```

```
source('./scripts/utility_functions.R')  
data(dune)  
data(dune.env)  
?dune  
# there are few nomial variables in the dataset that make modeling a  
# bit of a pain let's convert those to numeric vectors or  
# to plain un-ranked factors so that they are easier to work with and interpret.  
dune.env$Moisture <- as.numeric(dune.env$Moisture)  
dune.env$Manure <- as.numeric(dune.env$Manure)  
dune.env$Management <- factor(dune.env$Management, ordered = FALSE)  
dune.env$Use <- factor(dune.env$Use, ordered = FALSE)  
# clean up names so they are easier to work with
```

1. Conduct an indirect ordination on the dune plant community. Specifically, visually examine a NMDS plot using the bray-curtis distance metric. Below is some code to help you develop a potential plot that emphasizes the role of the environmental variable “Moisture”. Describe how you interpret the graphic. What is the goal of creating such a plot? Does this analysis suggest any interesting findings with respect to the dune vegetation?

```
dune_mds <- metaMDS(dune, trace = 0)  
# correlate envir. variables to ordination space  
# for a post-hoc examination  
dune_fit <- envfit(dune_mds, dune.env)  
dune_fit
```

```
##  
## ***VECTORS  
##  
##           NMDS1    NMDS2    r2 Pr(>r)  
## A1          0.96474  0.26321 0.3649 0.018 *
```

```

## Moisture  0.98272  0.18510 0.7296  0.001 ***
## Manure    -0.19582 -0.98064 0.6932  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
##
## ***FACTORS:
##
## Centroids:
##           NMDS1  NMDS2
## ManagementBF -0.4534 -0.0102
## ManagementHF -0.2636 -0.1282
## ManagementNM  0.2958  0.5790
## ManagementSF  0.1506 -0.4670
## UseHayfield   -0.1568  0.3248
## UseHaypastu   -0.0412 -0.3370
## UsePasture     0.2854  0.0844
##
## Goodness of fit:
##           r2 Pr(>r)
## Management 0.4134  0.003 **
## Use         0.1871  0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999

```

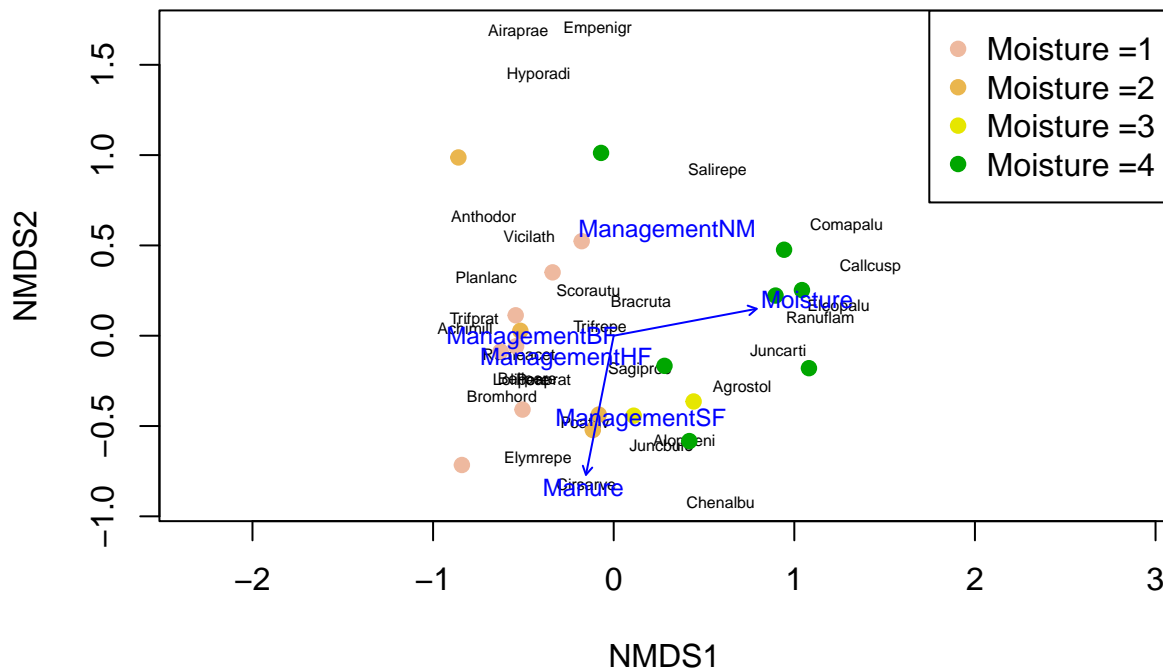
The output from the function `envfit` suggests that the categorical variables `Moisture` and `Management` are the two most important variables for explaining the site placements in ordination space. In other words, these variables correlate the strongest to the primary axes of variation in species composition identified by the analysis.

Let's examine our ordination plot to see if this seems reasonable.

```

# plot results
plot(dune_mds, type='n')
text(dune_mds, 'sp', cex=.5)
# generate vector of colors
mois_lvs <- sort(unique(dune.env$Moisture))
color_vect <- rev(terrain.colors(length(mois_lvs) + 1))[-1]
points(dune_mds, 'sites', pch=19,
       col=color_vect[dune.env$Moist])
# add enviornmental variables for which their p value was less than 0.01
plot(dune_fit, p.max = 0.01, cex=.75, add = TRUE)
legend('topright', paste("Moisture =", mois_lvs, sep=''),
       col=color_vect, pch=19)

```



In the above plot, the points are colored by their Moisture level. It appears that sites of different moisture levels are arrayed along the first ordination axis (i.e., x-axis) which by definition is the dominant gradient in species composition. The management levels are arrayed vertically along the second axis indicating that they are associated with an secondary orthogonal (i.e., independent) axis of variation in the plant community.

Notice too that the graphic is a bit cluttered and messy. Cleaning up these plots takes time and care. One simple fix would be to rename the variables **Moisture** and **Management** so that their labels are shorter. Alternatively, if the effect of **Moisture** is largely linear as it appears in the ordination then it could be treated as a single numeric vector rather than a factor with mutiple centroids.

2. Carry out a direct ordination using CCA in order to test any potential hypotheses that you developed after examining the MDS plot. Specifically, carry out a test of the entire model (i.e., including all constrained axes) and also carry out tests at the scale of individual explanatory variables you included in your model if you included more than one variable. Plot your results.

```
# a very terse way to specify to use all enviornmental variable is use the "."
# notation, but I don't recommend this because it is not the clearest way to
# indicate what variables are in the model for a human reader
dune_cca <- cca(dune ~ . , data=dune.env)
# alternatively I prefer this specification
dune_cca <- cca(dune ~ A1 + Moisture + Management + Use + Manure,
               data=dune.env)
# examine output
dune_cca
```

```
## Call: cca(formula = dune ~ A1 + Moisture + Management + Use + Manure,
```

```
## data = dune.env)
##
##              Inertia Proportion Rank
## Total          2.1153      1.0000
## Constrained    1.2128      0.5733    8
## Unconstrained  0.9025      0.4267   11
## Inertia is scaled Chi-square
##
## Eigenvalues for constrained axes:
##   CCA1   CCA2   CCA3   CCA4   CCA5   CCA6   CCA7   CCA8
## 0.4601 0.2954 0.1579 0.1337 0.0659 0.0406 0.0341 0.0251
##
## Eigenvalues for unconstrained axes:
##   CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8    CA9    CA10
## 0.29542 0.13548 0.10574 0.10085 0.07773 0.05656 0.05106 0.03564 0.02477 0.01288
##   CA11
## 0.00635
```

The output from the CCA model including all the environmental variables indicates the model explains $100 * 1.50 / 2.12 = 71\%$ of the variance in species composition. However, we included a lot of variables in our model so it is probably a good idea to compute the adjusted R-squared statistic as well.

```
dune_cca_r2 <- RsquareAdj(dune_cca, nperm=2000)
dune_cca_r2[2]
```

```
## $adj.r.squared
## [1] 0.2645998
```

After 2000 permutations you can see that the adjusted r-squared stabilized around 0.26 which is quite a bit smaller than the raw R^2 . This indicates that the model was overfit to the data because it had many spurious explanatory variables. Let's examine if the model and particular variables that are statistically significant.

```
# test for model significance
anova(dune_cca)
```

```
## Permutation test for cca under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = dune ~ A1 + Moisture + Management + Use + Manure, data = dune.env)
##           Df ChiSquare      F Pr(>F)
## Model      8   1.21278 1.8478 0.001 ***
## Residual  11   0.90248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# test for marginal effects of each variable
anova(dune_cca, by='margin')
```

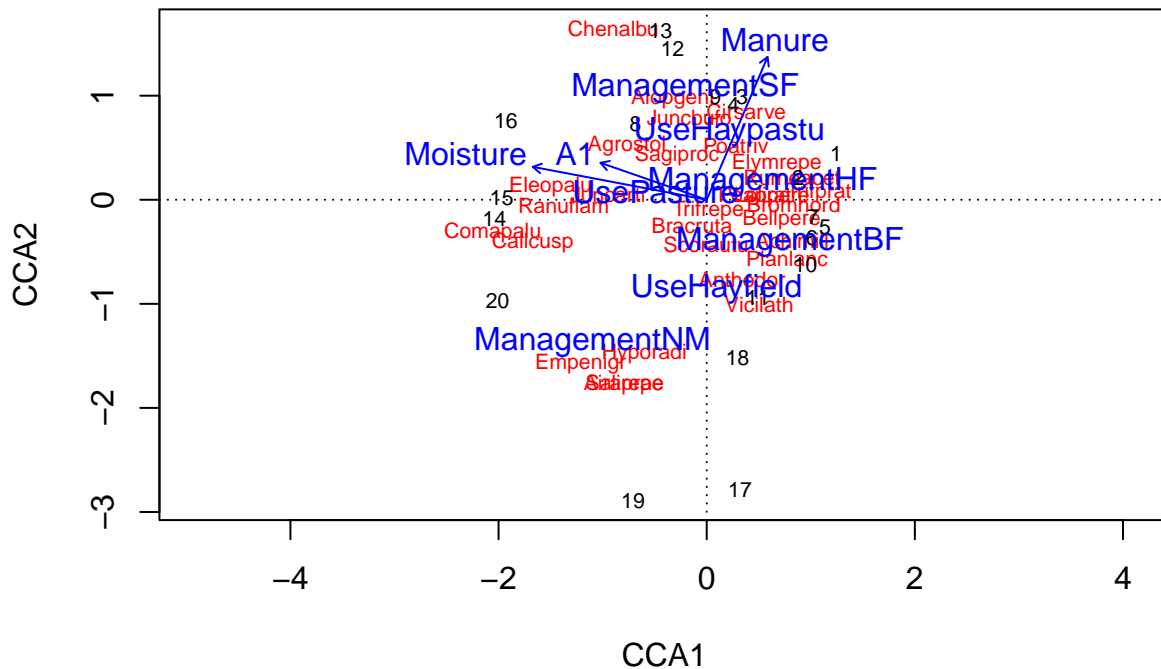
```
## Permutation test for cca under reduced model
## Marginal effects of terms
```

```
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = dune ~ A1 + Moisture + Management + Use + Manure, data = dune.env)
##           Df ChiSquare      F Pr(>F)
## A1          1   0.10858 1.3235 0.200
## Moisture     1   0.17380 2.1183 0.018 *
## Management   3   0.31036 1.2610 0.161
## Use          2   0.13323 0.8120 0.690
## Manure        1   0.09070 1.1055 0.312
## Residual    11   0.90248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

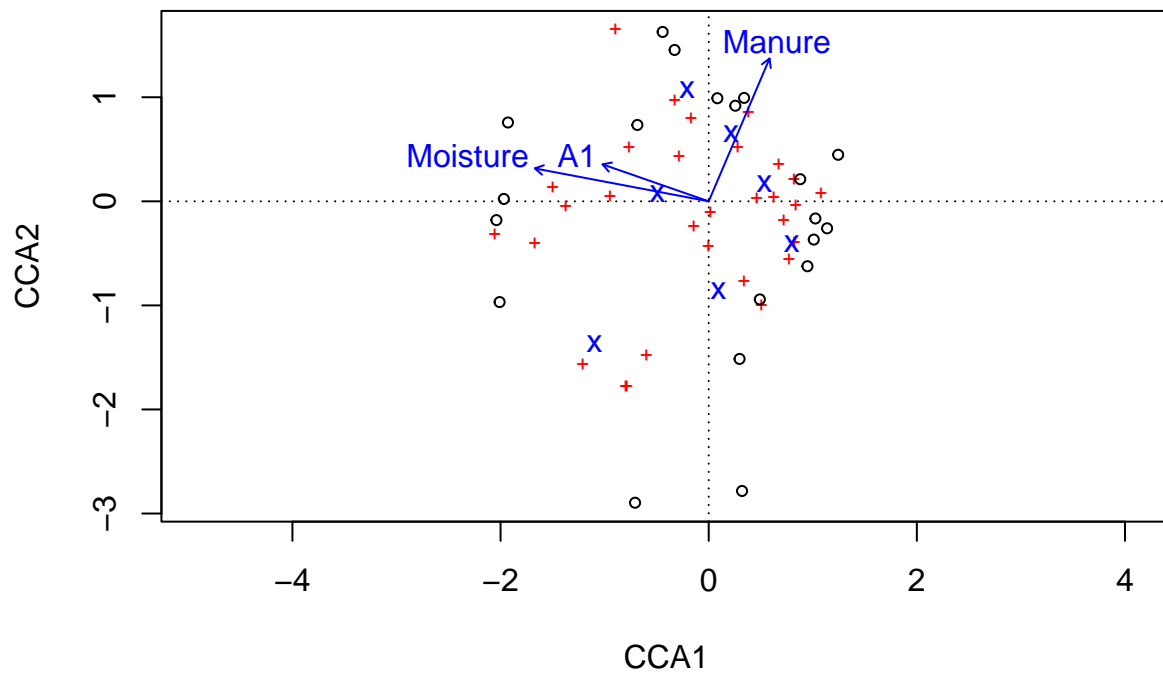
The first permutation-based ANOVA indicates that the model does explain more variance than random, but the effect size is pretty weak as indicated by the small F-statistic. The second ANOVA which examined the marginal effects of each model term indicates that when the variables are forced to compete against one another that no single variable is statistically significant. The variable “A1” which is the thickness of the soil A1 horizon is the most important variable followed by the Moisture and Manure variables as indicated by their F-statistics.

Let’s plot our results now:

```
plot(dune_cca)
```

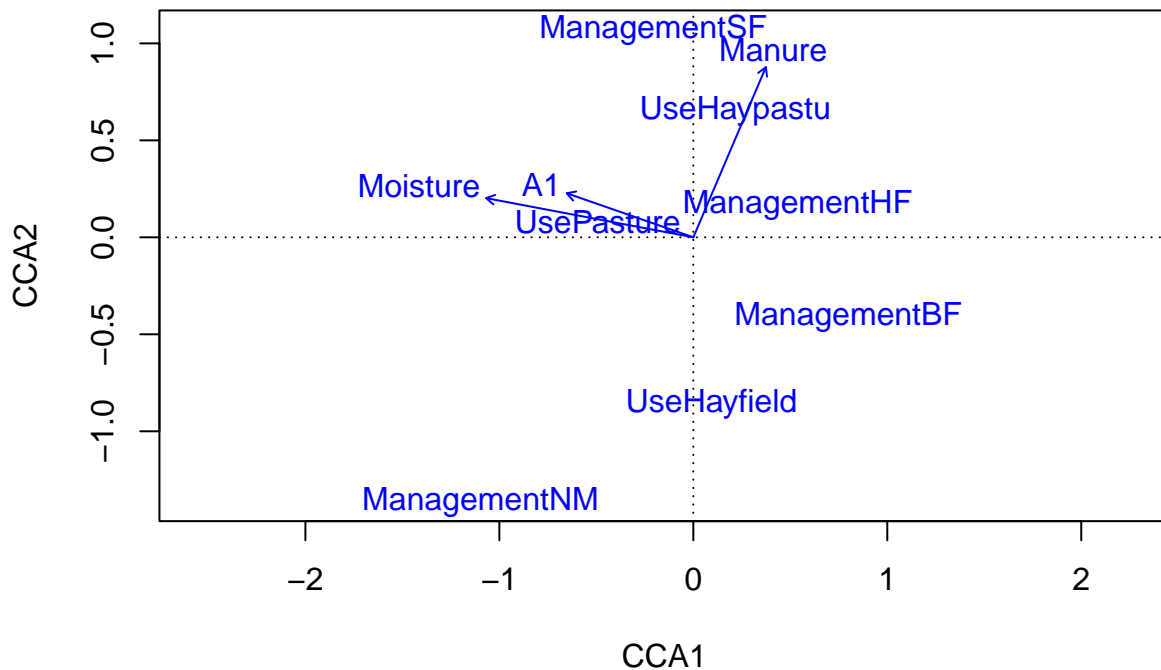


```
# bit of a mess, here's a cleaner option to de-congest the plot
ordiplot(dune_cca)
```



```
# this is cleaner but the blue 'X' are environmental variables that need labels
# if we want to focus in on the environmental variables we can use
plot(dune_cca, display='cn')
```

```
## Warning in max(nrow(g$spe), nrow(g$sit), nrow(g$con), nrow(g$def)): no
## non-missing arguments to max; returning -Inf
```



The third plot which uses the `display = 'cn'` argument to display the centroids of the categorical variables indicates that the first axis is primarily being loaded on by the the moisture variable (Moisture) and the thickness of soil (A1).

The second axis is more strongly related to differences in management and use as the three classes of management and three classes of use are spread broadly along this axis.

Based upon the large difference between the R^2 and R^2 adjusted values, the output of the marginal variables tests, and the graphical results of the CCA indicate that many of the variables included in the model do not have strong explanatory power.

In an effort to improve model interpretation and to decrease the chance of overfitting the community data, let's examine a simpler model with only the moisture and management variables. A1 appeared to have a similar effect as the moisture component (notice both arrows share a similar angle) so we'll ignore it here.

```
dune_cca_MM <- cca(dune ~ Moisture + Management, data=dune.env)
dune_cca_MM_r2 <- RsquareAdj(dune_cca_MM, 2000)
dune_cca_MM_r2
```

```
## $r.squared
## [1] 0.4048847
##
## $adj.r.squared
## [1] 0.2487795
```

```
anova(dune_cca_MM)
```

```
## Permutation test for cca under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = dune ~ Moisture + Management, data = dune.env)
##      Df ChiSquare      F Pr(>F)
## Model    4   0.85644 2.5513 0.001 ***
## Residual 15   1.25883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(dune_cca_MM, by='margin')
```

```
## Permutation test for cca under reduced model
## Marginal effects of terms
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = dune ~ Moisture + Management, data = dune.env)
##      Df ChiSquare      F Pr(>F)
## Moisture    1   0.25260 3.0099 0.002 **
## Management   3   0.44563 1.7700 0.019 *
## Residual    15   1.25883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

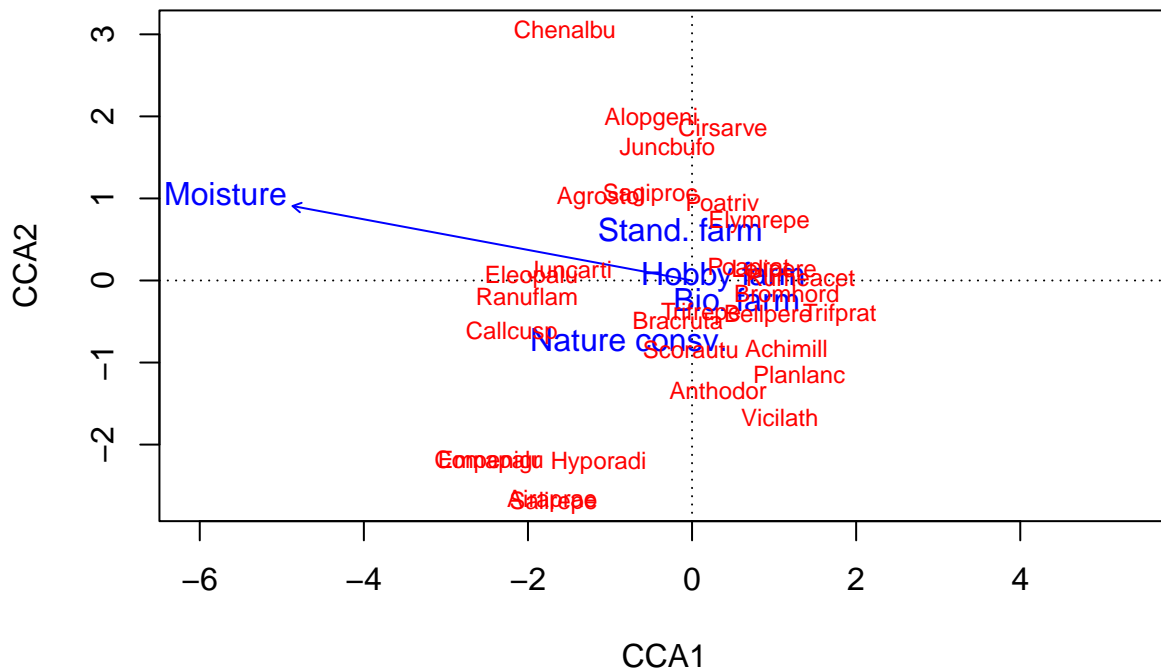
```
anova(dune_cca_MM, dune_cca)
```

```
## Permutation tests for cca under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model 1: dune ~ Moisture + Management
## Model 2: dune ~ A1 + Moisture + Management + Use + Manure
##   ResDf ResChiSquare Df ChiSquare      F Pr(>F)
## 1     15      1.25883
## 2     11      0.90248 4   0.35635 1.0858 0.34
```

The above analysis indicates that this reduced model model is less overfit (note it actually has a higher adjusted R^2 because the spurious variables were dropped). Both moisture and management are relevant variables (i.e., statistically significant), but their effects are not mind blowing.

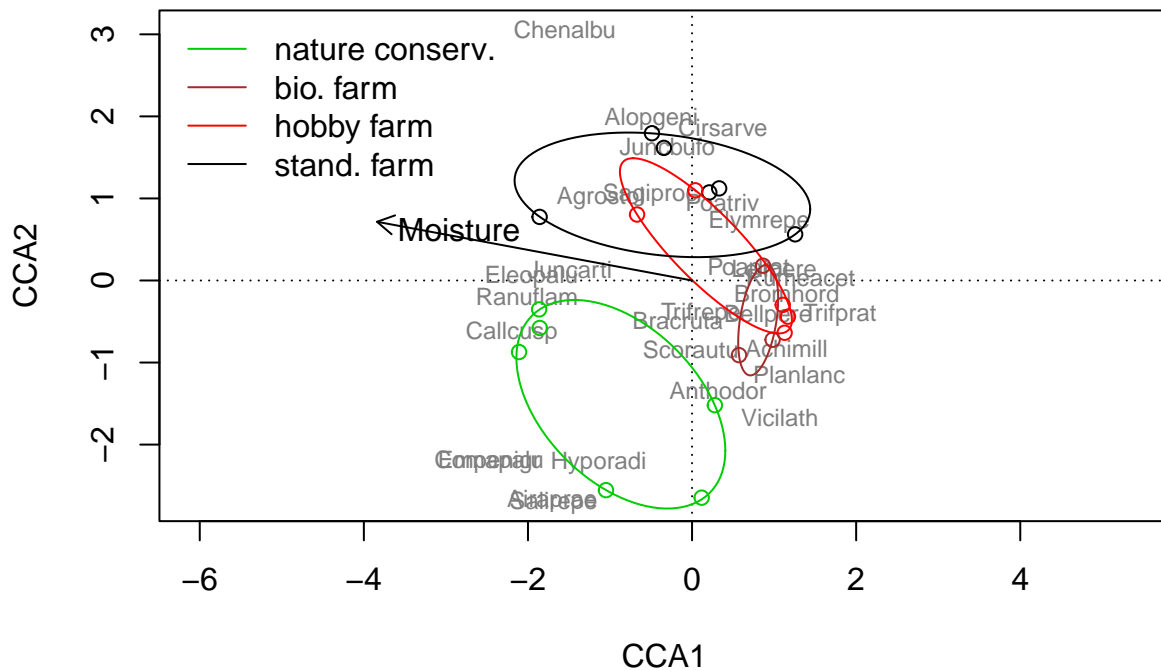
Note that it is difficult to visually compare the strength of a continuous variable such as A1 with the categorical variables because the categorical variable is displayed as a centroid rather than an arrow. The ANOVA analysis is useful for parsing out the relative importance when different classes of variables are involved.

```
plot(dune_cca_MM, type='n', scaling = 'sites')
text(dune_cca_MM, 'cn', col = 'blue', scaling = 'sites',
     labels = c('Bio. farm', 'Hobby farm', 'Nature consv.', 'Stand. farm', 'Moisture'))
text(dune_cca_MM, 'species', cex = 0.75, col = 'red', scaling = 'sites')
```

```
#points(dune_cca, 'sites')
#ordiellipse(dune_cca, Management, col=1:4, lwd=3)

# trying to get it prettier
col_vecs <- c('brown','red', 'green3', 'black')
plot(dune_cca_MM, type='n', scaling = 'sites')
# put species on
text(dune_cca_MM, 'species', cex = 0.75, col = 'grey50', scaling = 'sites')
# put sites on
points(dune_cca_MM, 'wa', pch = 1, col = col_vecs[dune.env$Management])
# place ellipse around sites
ordiellipse(dune_cca_MM, dune.env$Management,
            col=c('brown','red', 'green3', 'black'), label = FALSE, kind="ehull")
# still need to add arrow for just Moisture variable
bip <- scores(dune_cca_MM, display = 'bp', scaling = 'sites')
bip_scl <- (bip * ordiArrowMul(bip, fill = .8))[1, ]
arrows(0, 0, bip_scl[1], bip_scl[2], length = 0.1)
text(bip_scl[1] + 1, bip_scl[2] - 0.1, labels = 'Moisture')
#points(dune_cca_lin, 'sites', pch = 1, cex=0.75)
legend("topleft",
      c('nature conserv.', 'bio. farm', 'hobby farm', 'stand. farm'),
      col = c('green3','brown', 'red', 'black'), lty = 1, bty='n')
```



3. Do your two analyses agree with one another or complement one another or do these two analyses seem to be suggesting different take home messages? Which analysis do you find to be more useful?

This is a more subjective question but my personal take in this case is that the analyses are somewhat complementary. The NMDS is nice because you are ensuring that you are examining the dominant axes of variation not just the axes that the environment can explain (as in CCA). However, the direct ordination provides us a clear measures of variable importance and variance explained which are intuitive an easier to communicate. With the NMDS bringing in the environment is always a posthoc approach. One point to note is that the distance measures used in these two analyses are different. In the NMDS analysis the distance measure was the bray-curtis distance measure but in the CCA analysis chi-squared distances are computed. In the case of CCA the chi-squared distances have a clear ecological meaning, species are assumed to have an optima along an environmental gradient and decrease in prevalence away from the optima. In other words species are expected to have unimodal responses to the environment which is consistent with traditional ecological niche theory. The bray-curtis distance measure has no such theoretical foundation and its interpretation is thus much more vague.