multivariate models solutions

Dan McGlinn

3/24/2016

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

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

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

```
## This is vegan 2.4-5
```

```
source("./scripts/utility_functions.R")
data(dune)
data(dune.env)
`?`(dune)
```

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)
```

```
## Run 0 stress 0.1192678
## Run 1 stress 0.1192679
## ... Procrustes: rmse 6.226909e-05  max resid 0.0001827537
## ... Similar to previous best
## Run 2 stress 0.1886532
## Run 3 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 0.02027853  max resid 0.06499737
## Run 4 stress 0.1812961
## Run 5 stress 0.1192686
## Run 6 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 0.0001192118  max resid 0.000384896
## ... Similar to previous best
## Run 7 stress 0.1808913
## Run 8 stress 0.1192678
## Run 9 stress 0.1192678
## Run 10 stress 0.1192679
## Run 11 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 4.57106e-05  max resid 0.0001439983
## ... Similar to previous best
## Run 12 stress 0.1192681
## Run 13 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 1.278406e-05  max resid 4.108067e-05
## ... Similar to previous best
## Run 14 stress 0.2035424
```

```
## Run 15 stress 0.1183186
## ... Procrustes: rmse 9.20184e-06  max resid 1.933613e-05
## ... Similar to previous best
## Run 16 stress 0.1192686
## Run 17 stress 0.1812934
## Run 18 stress 0.1183186
## ... Procrustes: rmse 2.721447e-05  max resid 6.737372e-05
## ... Similar to previous best
## Run 19 stress 0.1808914
## Run 20 stress 0.3726987
## *** Solution reached
```

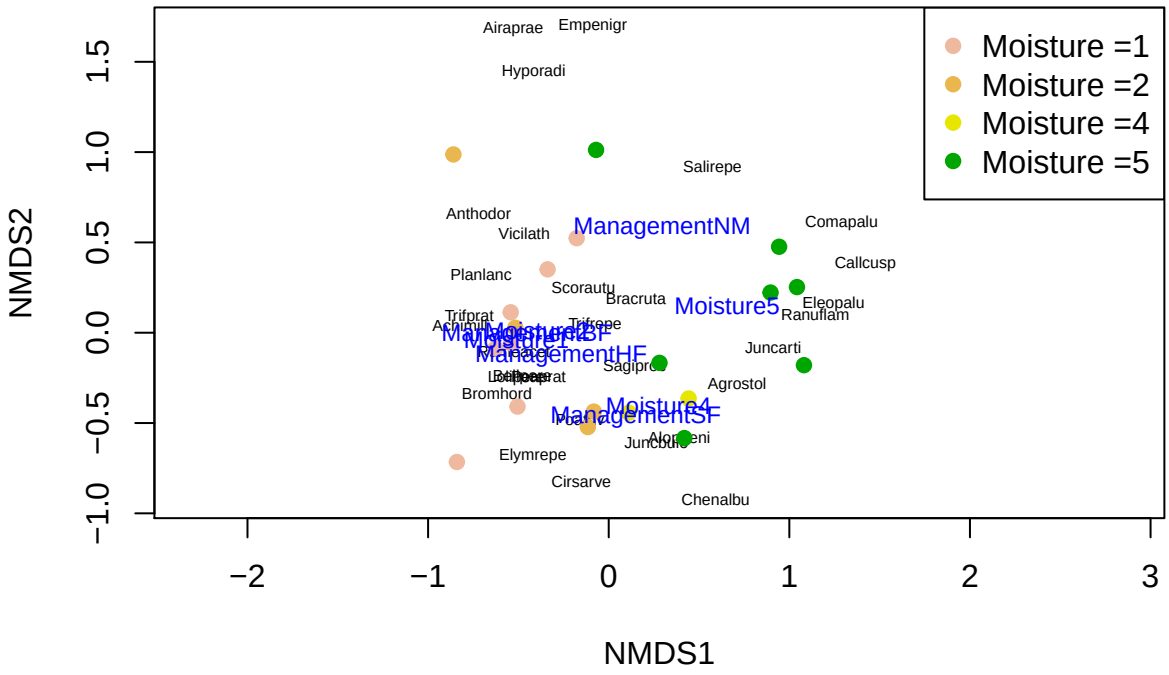
```
# fit enviornmental variables to ordination space
dune_fit = envfit(dune_mds, dune.env)
dune_fit
```

```
##
## ***VECTORS
##
##      NMDS1   NMDS2     r2 Pr(>r)
## A1 0.96474 0.26321 0.3649  0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
##
## ***FACTORS:
##
## Centroids:
##                NMDS1   NMDS2
## Moisture1    -0.5101 -0.0403
## Moisture2    -0.3938  0.0139
## Moisture4     0.2765 -0.4033
## Moisture5     0.6561  0.1476
## 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
## Manure0       0.2958  0.5790
## Manure1      -0.2482 -0.0215
## Manure2      -0.3079 -0.1866
## Manure3       0.3101 -0.2470
## Manure4      -0.3463 -0.5582
##
## Goodness of fit:
##                r2 Pr(>r)
## Moisture   0.5014  0.001 ***
## Management 0.4134  0.007 **
## Use        0.1871  0.110
## Manure     0.4247  0.018 *
## ---
## 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 "Moisture" and "Management" are the two most important variables for explaning the site placements in ordination space. In other words these variables coorelate the strongest to the primary axes of variation in species composition identified by the analysis. Let's examine our oridination plot to see if this seems reasonable.

```
plot(dune_mds, type = "n")
text(dune_mds, "sp", cex = 0.5)
# generate vector of colors
mois_lvs = sort(unique(dune.env$Moisture))
color_vect = rev(terrain.colors(length(mois_lvs) + 1))[-1]
points(dune_mds, "sites", pch = 19, col = color_vect[dune.env$Moisture])
# add enviornmental variables for which their p value was less than 0.01
plot(dune_fit, p.max = 0.01, cex = 0.75)
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 orthogonal axis of variation in the plant community.

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 tearse 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 perfer this specification
dune_cca = cca(dune ~ A1 + Moisture + Management + Use + Manure, data = dune.env)
```

```r
# 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    1.5032     0.7106   12
## Unconstrained  0.6121     0.2894    7
## Inertia is mean squared contingency coefficient
## Some constraints were aliased because they were collinear (redundant)
##
## Eigenvalues for constrained axes:
##   CCA1   CCA2   CCA3   CCA4   CCA5   CCA6   CCA7   CCA8   CCA9  CCA10
## 0.4671 0.3410 0.1761 0.1532 0.0953 0.0703 0.0589 0.0499 0.0318 0.0260
##  CCA11  CCA12
## 0.0228 0.0108
##
## Eigenvalues for unconstrained axes:
##     CA1     CA2     CA3     CA4     CA5     CA6     CA7
## 0.27237 0.10876 0.08975 0.06305 0.03489 0.02529 0.01798
```

The output from the CCA model including all the enviornmental 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.

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

```
## $adj.r.squared
## [1] 0.2308469
```

After 3000 permutations you can see that the adjusted r-squared stablized around 0.23 which is quite a bit smaller than the raw r2. 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 are statistically significant.

```r
# 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    12    1.5032 1.4325  0.023 *
## Residual  7    0.6121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 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   0.11070 1.2660  0.248
## Moisture    3   0.31587 1.2041  0.195
## Management  2   0.15882 0.9081  0.562
## Use         2   0.13010 0.7439  0.760
## Manure      3   0.25490 0.9717  0.464
## Residual    7   0.61210
```

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.

We could examine to see if single variable models yeild statistically significant effects but I don't see much point in this excercise. We know that the model does capture some of the variance in species composition but that the variables in general are fairly weak. We could test the specific hypothesis that was suggested by the NMDS which was that moisture and management captured orthogonal axes of variation in vegetation.

```
dune_cca_MM = cca(dune ~ Moisture +  Management, data=dune.env)
dune_cca_MM_r2 = RsquareAdj(dune_cca_MM, 1000)
dune_cca_MM_r2
```

```
## $r.squared
## [1] 0.4738772
##
## $adj.r.squared
## [1] 0.2427248
```

```
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     6    1.0024 1.9515  0.001 ***
## Residual 13    1.1129
## ---
## 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    3   0.39854 1.5518  0.033 *
## Management  3   0.37407 1.4565  0.048 *
```

```
## Residual   13   1.11289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

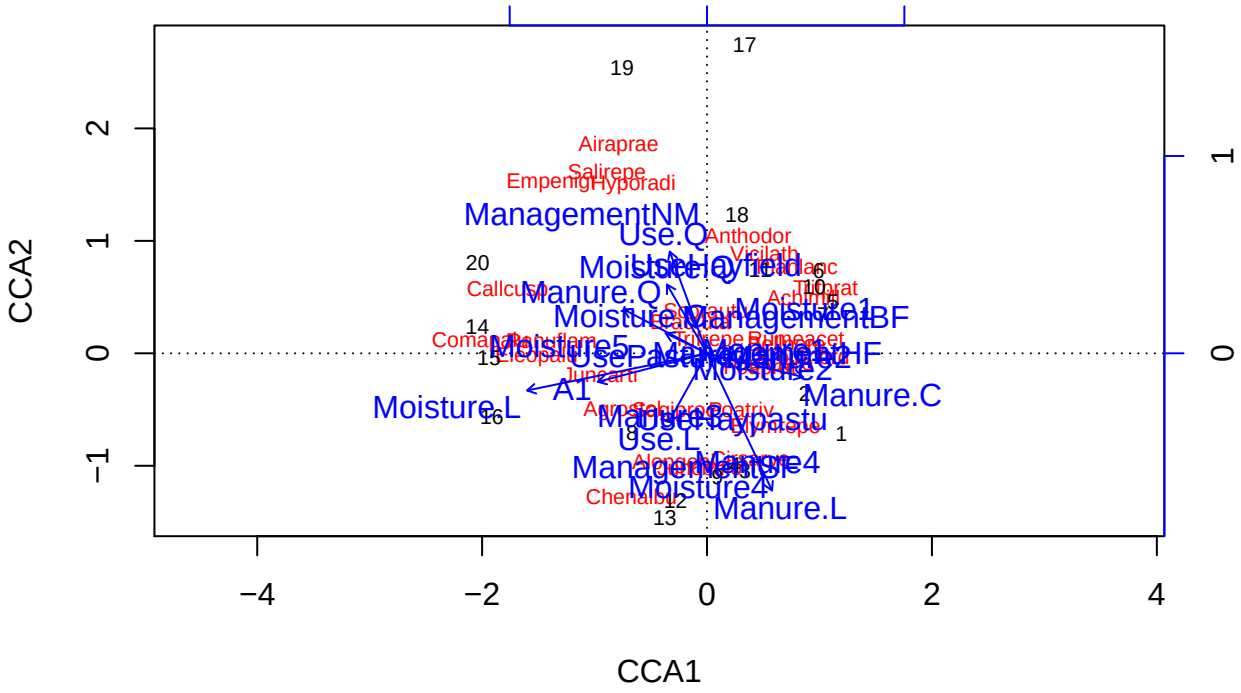
```
anova(dune_cca_MM, dune_cca)
```

```
## 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    13       1.1129
## 2     7       0.6121  6   0.50079 0.9545  0.532
```

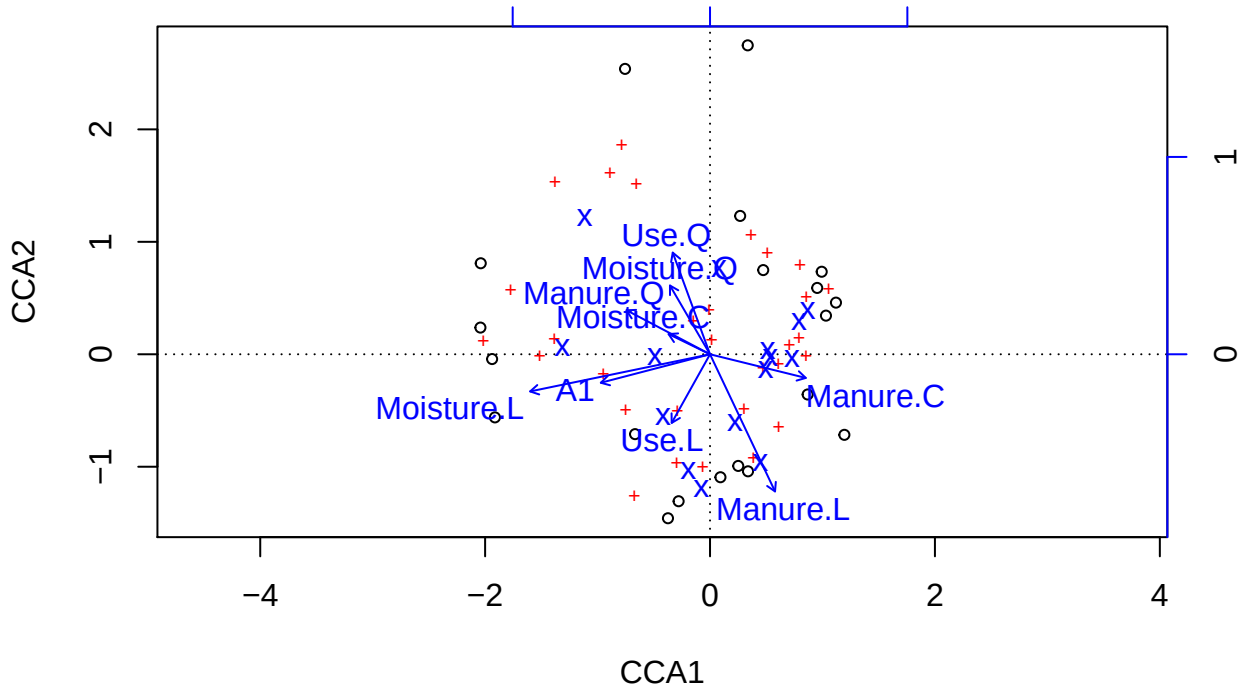
The above analysis indicates that this reduced model model is less overfit (note it actually has a higher adjusted r2 because the spurious variables were dropped). Both moisture and management are relevant variables (i.e., statistically significant), but their effects are not mind blowing.

Let's plot our results now:

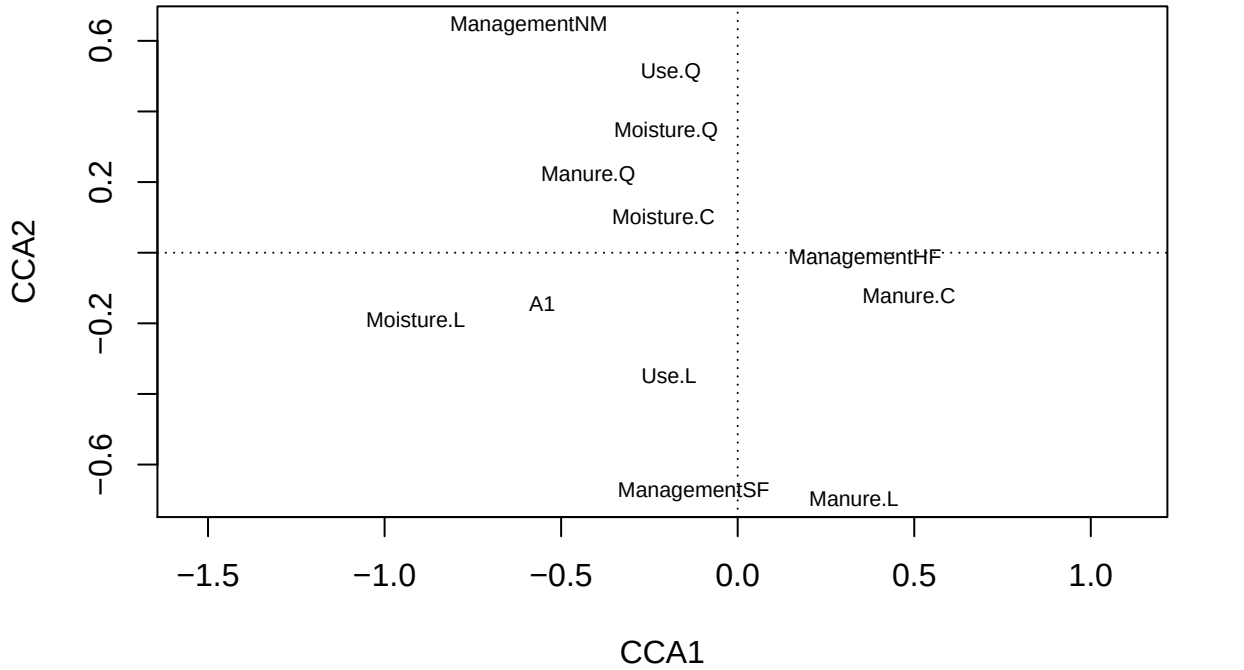
```
plot(dune_cca)
```




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




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



The second plot indicates that the first axis is primarily being loaded on by the linear component of the moisture variable and the cubic component of the Manure variable. The second axis is more strongly related to differences in management as the three classes of management are spread broadly along this axis.

Note that it is difficult to visually compare the strength of a continous variable such as A1 with the categorical variables because the relative distance of the label seems to be scaled different. The ANOVA analysis is useful for parsing out the relative importances when different classes of variables are involved.

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 enviornment can explain (as in CCA). However, the direct ordination provides us a clear measures of variable importance and variance explained which are intutitive an easier to communicate. With the NMDS bringing in the enviornment 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 enviornmental gradient and decrease in prevalance away from the optima. In other words species are expected to have unimodal responses to the enviornment which is consistent with traditional ecological niche theory. The bray-curtis distance measure has no such theorectical foundation and its interpretation is thus much more vauge.