Canonical Correspondence Analysis

KEY

April 10, 2020

Variation of bird species composition at the San Dieguito River Park estuary

This data set is interesting both because it is from a local source, and because we can learn a method for the proper way to incorporate circular variables, such as month, into your analysis.

To begin, import the data:

```
library(readxl)
data.frame(read_excel("sdrp_waterbirds.xlsx")) -> birds
```

Make a list of the species (all columns are species except the first four):

```
species <- colnames(birds)[-c(1:2)]
```

Load the vegan library:

```
library(vegan)
```

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

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

```
## This is vegan 2.5-5
```

Run a correspondence analysis using the cca() function with only the bird data (no "environmental" matrix yet):

```
cca(birds[species]) -> birds.ca
birds.ca
```

```
## Call: cca(X = birds[species])
##
##               Inertia Rank
## Total           2.075
## Unconstrained   2.075   71
## Inertia is scaled Chi-square
##
## Eigenvalues for unconstrained axes:
##    CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8
## 0.4624 0.3783 0.2829 0.1393 0.1082 0.0778 0.0685 0.0541
## (Showing 8 of 71 unconstrained eigenvalues)
```

Get a report of variation associated with each axis, and the running total across axes:

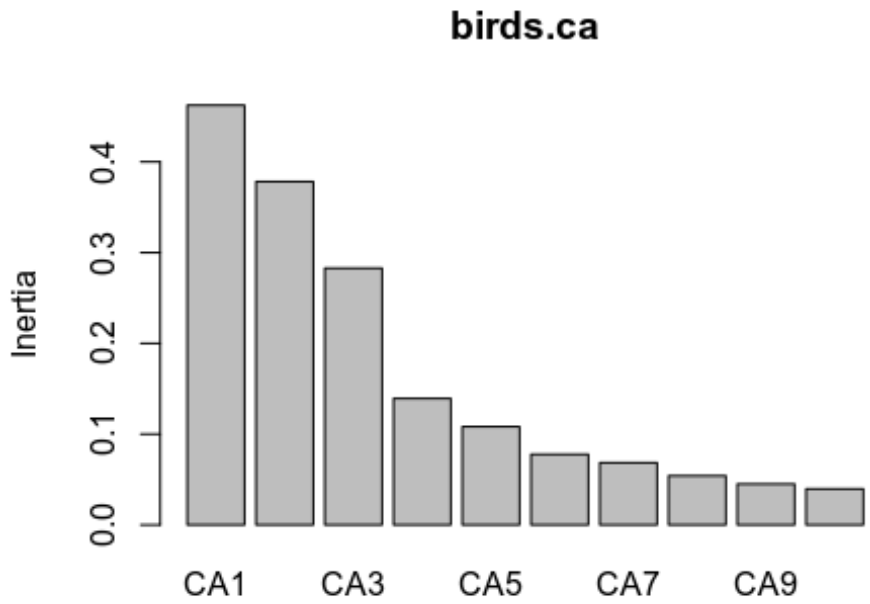
```
summary(birds.ca)$cont$importance[,1:8]
```

```
##                             CA1       CA2       CA3        CA4        CA5
## Eigenvalue            0.4624209 0.3783135 0.2828878 0.13929106 0.10822183
## Proportion Explained  0.2228561 0.1823220 0.1363331 0.06712904 0.05215573
## Cumulative Proportion 0.2228561 0.4051781 0.5415112 0.60864028 0.66079601
##                             CA6        CA7        CA8
## Eigenvalue            0.07775887 0.06847001 0.05413810
## Proportion Explained  0.03747461 0.03299800 0.02609097
## Cumulative Proportion 0.69827063 0.73126863 0.75735960
```

Question: what proportion of variance is explained by the first two CA axes?

22.2% for CA1, 18.2% for CA2, for a total of 40.5% Get a screeplot of the eigenvalues:

```
screeplot(birds.ca)
```



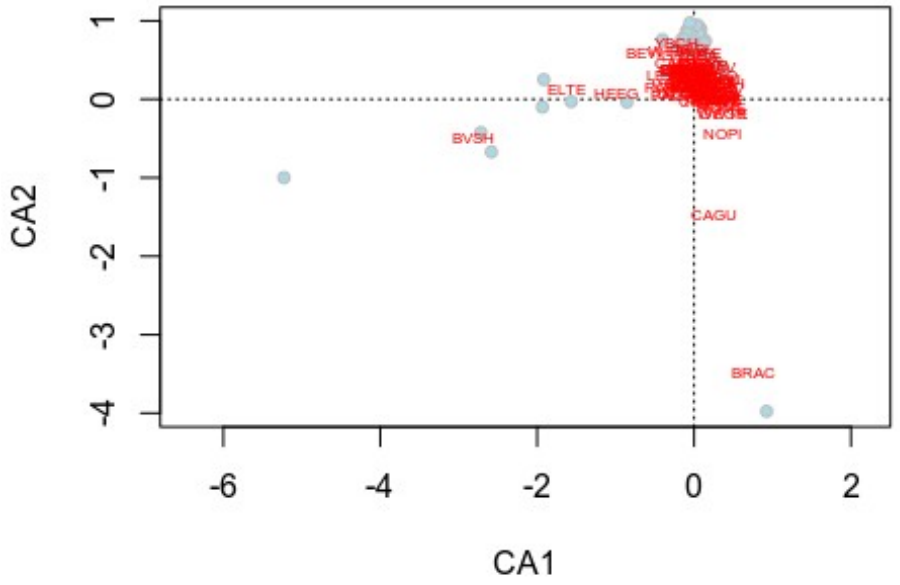
Question: is there a noticeable dropoff in the size of the eigenvalues across these first eight CC axes? Between which axes?

Yes, after the third one it drops off.

Construct the biplot:

```
plot(birds.ca, type = "n")
points(birds.ca, display = "sites", cex = 0.8, pch = 21, col = "gray", bg =
```

```
"lightblue")
text(birds.ca, display = "spec", cex = 0.5, col = "red")
```



Question: according to the biplot, which species are most strongly associated with CA1? Which species would you expect to see in large numbers in the points at CA1 values of 2 or more?

BVSH and BRAC. BRAC would be common.

Question: which species are most strongly associated with CA2? Which species would you expect to see in large numbers in points with CA2 scores of -3 or less?

BRAC and YBCH are strongly associated with CA2. Low values of CA2 are associated with BRAC.

Calculate the CA 1 scores:

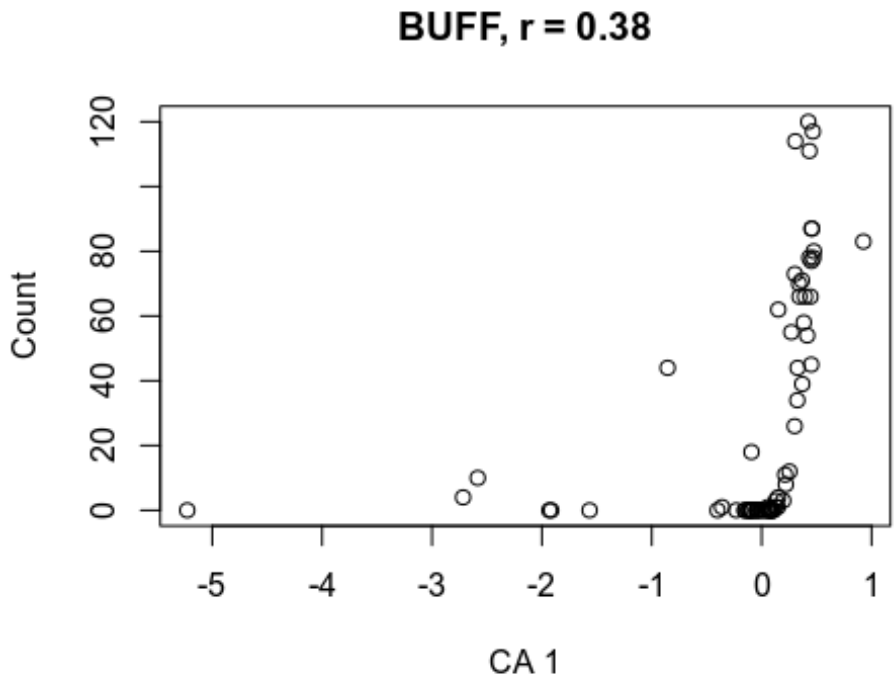
```
scores(birds.ca, display = "sites", choices = c(1)) -> birds.ca1
```

Correlate the bird counts with these CA 1 scores:

```
cor(birds[species], birds.ca1) -> birds.ca1.cor
```

Plot buffleheads, as an example of a strong positive correlation:

```
plot(birds$BUFF ~ birds.ca1, xlab = "CA 1", ylab = "Count", main = "BUFF, r =
0.38")
```

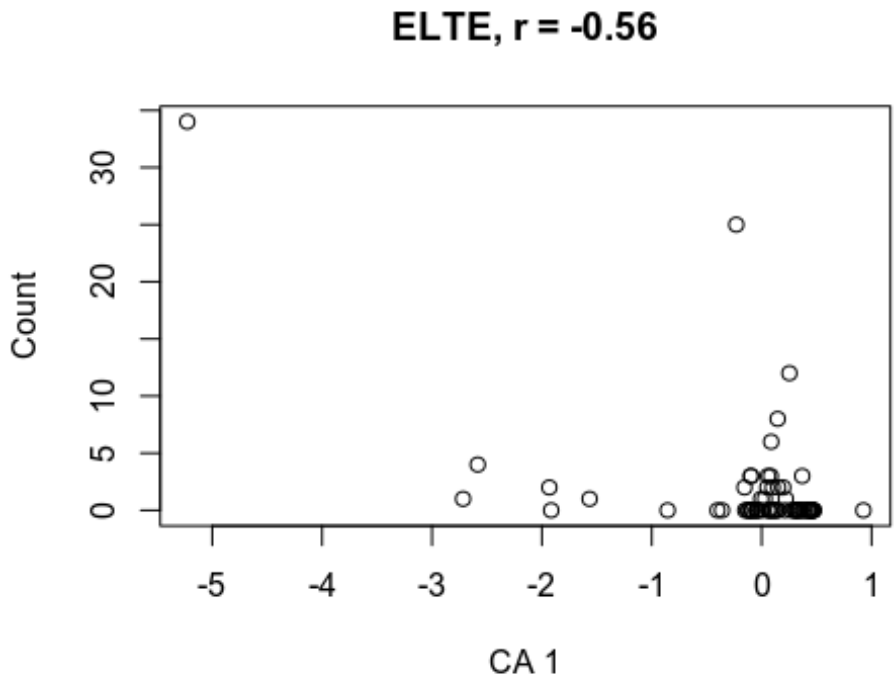


Question: Is the relationship between BUFF and CA1 linear? Does the correlation coefficient accurately represent the relationship between BUFF counts and CA1 scores?

No, it's curved, so the correlation coefficient is under-estimating the strength of the relationship.

Now plot elegant terns, as an example of a strong negative correlation:

```
plot(birds$ELTE ~ birds.ca1, xlab = "CA 1", ylab = "Count", main = "ELTE, r =
-0.56")
```

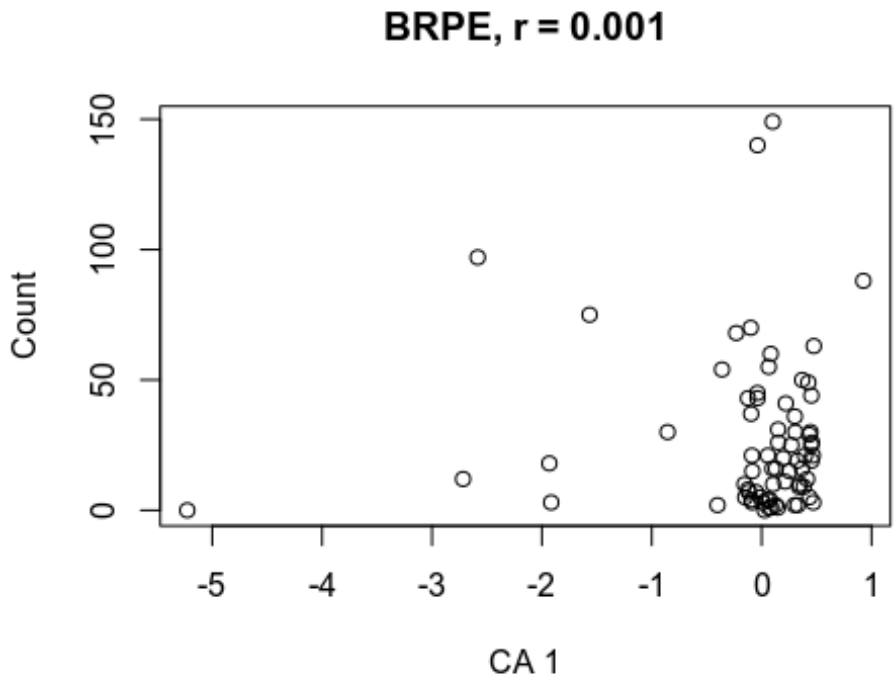


Question: can you see why there is a negative correlation between ELTE and CA1? Is the relationship linear?

The negative relationship is due to one really large count at the left side of the graph. The relationship is not really linear, there are lots of points with low counts clustered at 0 and a few scattered points away from 0.

Plot brown pelicans as an example of a correlation near 0:

```
plot(birds$BRPE ~ birds.ca1, xlab = "CA 1", ylab = "Count", main = "BRPE, r =
0.001")
```



Question: there are two reasons why you might get a low correlation between counts for a species and an axis: species that are restricted to a narrow range in the middle of the axis, or species that are found in roughly equal numbers along the entire axis. Which is true here for brown pelicans?

The species is pretty randomly distributed - there are both high and low counts all along CA1.

Canonical correspondence analysis

We will use month and year as our "environmental matrix", but we want month to be represented by springness and winterness to avoid having December and January at opposite ends of the scale.

First, we need an ordered factor for month:

```
birds$month <- ordered(birds$month, levels =
c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))
```

Convert the months to radians on a circle:

```
2 * pi * as.numeric(birds$month)/12 -> months.rad
```

Calculate springness and add it to the birds data set:

```
birds$springness <- sin(months.rad)
```

Calculate winterness and add it to the birds data set:

```
birds$winterness <- cos(months.rad)
```

Now run the CCA, relating the birds to year, springness, and winterness:

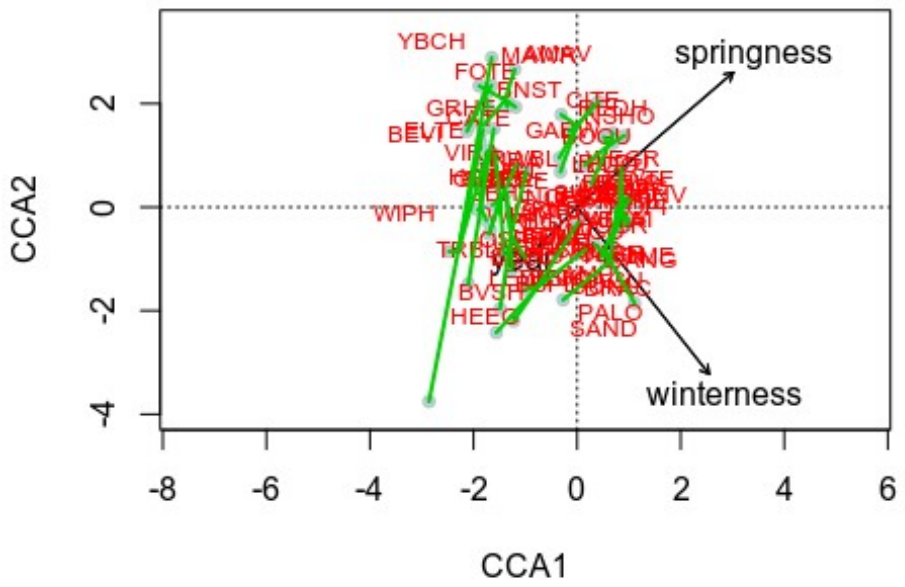
```
cca(birds[species] ~ springness + winterness + year, data = birds) ->
birds.cca

birds.cca

## Call: cca(formula = birds[species] ~ springness + winterness +
## year, data = birds)
##
##               Inertia Proportion Rank
## Total          2.0750     1.0000
## Constrained    0.5156     0.2485    3
## Unconstrained  1.5594     0.7515   68
## Inertia is scaled Chi-square
##
## Eigenvalues for constrained axes:
##    CCA1    CCA2    CCA3
## 0.30535 0.12428 0.08595
##
## Eigenvalues for unconstrained axes:
##    CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8
## 0.4018 0.3190 0.1171 0.0752 0.0698 0.0502 0.0458 0.0409
## (Showing 8 of 68 unconstrained eigenvalues)
```

Question: the eigenvalue for CCA1 is 0.30925. Does this indicate that it explain 30.9% of the variation in bird species, or does it explain 30.9% of the 24.5% that is explained across all of the CCA axes?

Bad question, dropped. The sum of the eigenvalues are the constrained inertia, so 0.305 is 0.305/0.5156 = 0.59, or 59% of the total explained variation. The total explained is 24.8%, so CCA1 explains 59% of 24.8%, or 14.7% of the inertia.

Make a triplot, with site scores, species scores, and vectors indicating the predictors:

```
plot(birds.cca, type = "n", xlim = c(-6,4), ylim = c(-4,3.5))

points(birds.cca, display = "sites", cex = 0.8, pch = 21, col = "gray", bg =
"lightblue")

points(birds.cca, display = "cn")

text(birds.cca, display = "cn")

text(birds.cca, display = "spec", cex = 0.75, col = "red", scaling = 1)

ordispider(birds.cca, birds$month, col = 3, lwd = 2)
```



Question: which season has the least variation from year to year in species composition? Hint: look for the spiders with the shortest legs, and see wheich quadrant of the graph they're in.

Spring has the least variation in species composition, because the spiders have the shortest legs in teh spring.

Question: which species have the highest relative frequencies in each season?

In spring the highest relative frequencies are for species like REDH and NSHO. For winter it is SAND and PALO. In summer it is YBCH, and in the fall it is HEEG and BVSH.

Get a randomization test for the CCA to see if the amount of variation in bird community composition that is explained by season and time is statistically significant:

```
anova(birds.cca, by = "term")
```

```
## Permutation test for cca under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = birds[species] ~ springness + winterness + year, data
= birds)
##            Df ChiSquare      F Pr(>F)
## springness  1   0.21710 9.4672  0.001 ***
## winterness  1   0.20587 8.9773  0.001 ***
## year        1   0.09261 4.0386  0.001 ***
## Residual   68   1.55939
```

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

Question: which of the predictors are statistically significant predictors of species composition?

All three of the predictors are statistically significant.

To calculate loadings, first extract the site scores:

```
scores(birds.cca, choices = c(1,2,3), display = "sites") -> birds.cca.scores
```

Then correlate the bird counts by the site scores:

```
 cor(birds[species],birds.cca.scores) -> birds.loadings
```

Question: which species are most strongly positively correlated with CCA1? Which are strongly negatively correlated with CCA1?

Strong positive correlations from BUFF and EAGR, strong negative correlations from GRHE and GREG.

Correlate the site scores with the predictor variables:

```
cor(birds[,c("springness","winterness","year")],birds.cca.scores)
```

```
##                  CCA1       CCA2        CCA3
## springness  0.6757928  0.3624347 -0.01303199
## winterness  0.6553143 -0.6502531 -0.17085610
## year       -0.1174535 -0.2829198 -0.63191989
```

Question: which axis is most strongly associated with change over time in species composition (year)?

The third CCA axis has a correlation of -0.63 with year.

To better understand the change over time, correlate the species that have the highest loadings on CCA3 with year:

```
cor(birds[,c("BRAC","CAGU","LBCU","SEPL")], birds$year)
```

```
##            [,1]
## BRAC -0.1743265
## CAGU -0.2292200
## LBCU  0.5533789
## SEPL  0.3294975
```

Question: based on these correlations, how are the four species changing over time (which are increasing, which are decreasing)?

BRAC and CAGU are declining over time, and LBCU and SEPL are increasing.

To make sure that we're justified in using year as a linear predictor in the model, fit a cca with year as a linear predictor, followed by year as a factor:

```r
cca(birds[species] ~ year + as.factor(year) + winterness + springness, data =
birds) -> birds.yr.fact.cca

anova(birds.yr.fact.cca, by = "terms")
```

```
## Permutation test for cca under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = birds[species] ~ year + as.factor(year) + winterness
+ springness, data = birds)
##                 Df ChiSquare       F Pr(>F)
## year             1   0.09414  4.2603  0.001 ***
## as.factor(year)  4   0.15126  1.7113  0.007 **
## winterness       1   0.18987  8.5931  0.001 ***
## springness       1   0.22555 10.2074  0.001 ***
## Residual        64   1.41416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: is year as a factor significant, after the linear change over time is accounted for? What does this mean - is there variation from year to year that isn't a directional trend over time?

Yes, year as a facotr is significant, so there is additional variation from year to year that isn't accounted for by the linear trend modeled by year.