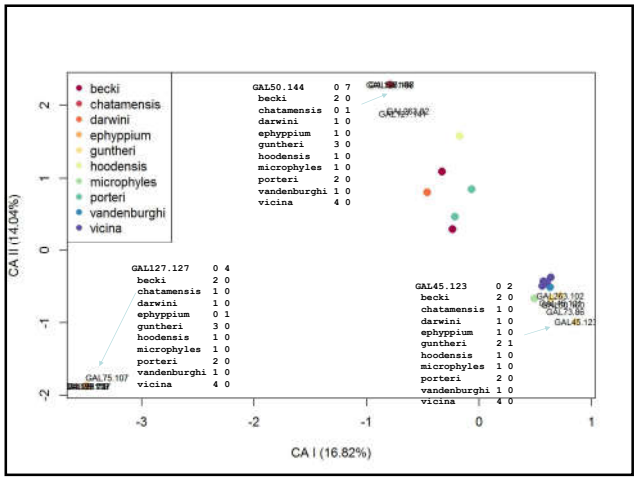
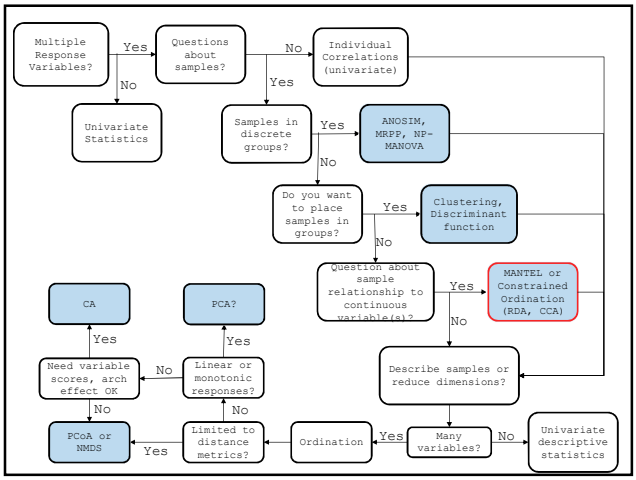
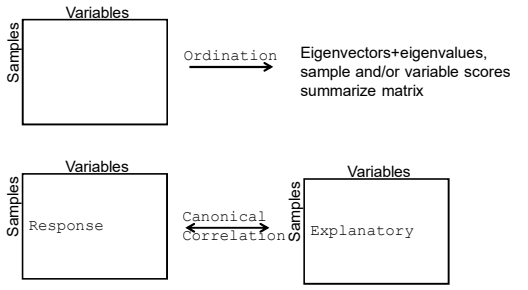


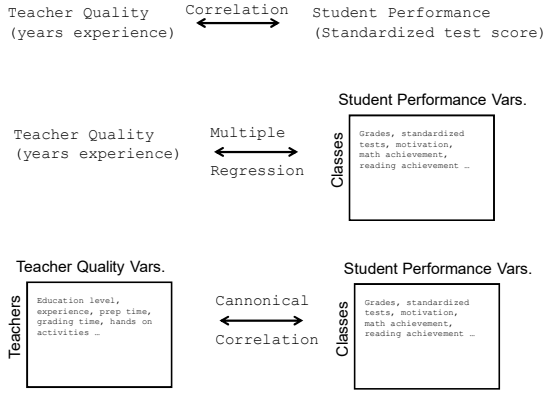


Canonical Analyses

- Analysis aimed at identifying the relationship between two multivariate datasets.



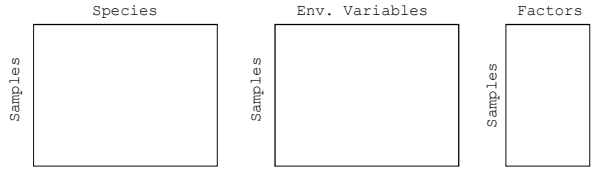
Cannonical Correlation



Canonical Correspondence Analysis (CCA)

Ter Braak, C. J. F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

- Canonical – "in simplest or standard form"
- Good choice if you have clear and strong *a priori* hypotheses regarding gradients and you are not interested in general structure of the data.
- Poor choice for exploratory analysis or "fishing expedition"



CCA

- Cannonical correlation of site scores on environmental variables (this is the **constraining** part)
- CA ordination of fitted values from correlation (this is the **oridination** part)
- Results in:
  - Site scores
  - Site constraints
  - Species scores
- Same assumptions and issues are involved with CA. Rare species are over emphasized, arch effect.

Use CCA with caution

- CCA uses multiple regression, has all of those associated assumptions and potential issues.
  - Multicollinearity
  - Outliers or errors in environmental data
  - Linear relationships
- Not well suited to exploratory analyses.
- Only use when you have a good idea of what environmental variables are structuring a gradient.

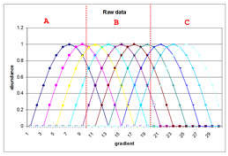
Use CCA with caution

- The rational for doing ordination is to summarize multidimensional data.
- The environmental data are used to **constrain** and change the ordination.
- An alternative that is more exploratory is to simply examine correlations between environmental variables and axes scores for NMDS or other ordination.

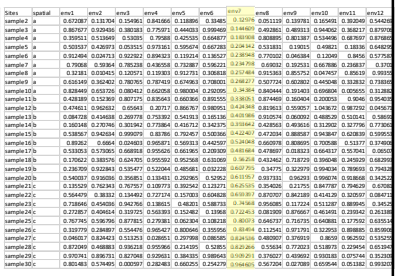
Running CCA in R

- Function cca in vegan package (see also cca function in ade4 package)
- Code
  - `base_cca<-cca(community)`
  - `base_cca<-cca(community ~ .,environmental)`
- Models must be specified, example above includes all environmental variables in the model – usually a bad idea.
- Model with variables specified:
  - `base_cca<-cca(community ~ DO + pH + substrate, environmental)`
- Models can include interaction terms but interpretation will be difficult.
  - `base_cca<-cca(community ~ DO * pH * substrate, environmental)`

Sample Dataset



Environmental Data Matrix



11 random variables, 1 correlated with gradient (env7)
1 spatial variable (spatial)

Variables env2 and env10 highly correlated with each other
All other variables random

CCA Example

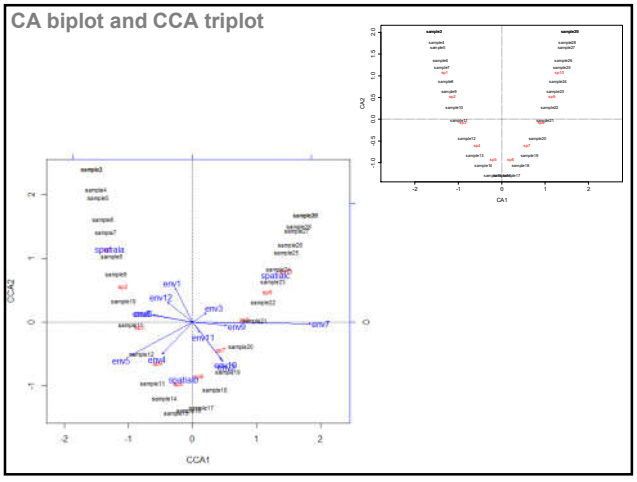
```
Call:
cca(X = community)

Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total           1.780          1
Unconstrained   1.780          1
```

Standard CA of our sample dataset.

```
Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total           1.7795      1.000
Constrained     1.4308      0.804
Unconstrained   0.3488      0.196

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                         CCA1   CCA2   CCA3   CCA4   CCA5   CCA6…
Eigenvalue             0.831 0.474 0.0756 0.0337 0.01079 0.00286…
Proportion Explained   0.467 0.266 0.0425 0.0189 0.00606 0.00161…
Cumulative Proportion  0.467 0.733 0.7757 0.7946 0.80068 0.80229…

Accumulated constrained eigenvalues
Importance of components:
                         CCA1   CCA2   CCA3   CCA4   CCA5   CCA6   CCA7   CCA8   CCA9
Eigenvalue             0.831 0.474 0.0756 0.0337 0.01079 0.00286 0.00163 0.000833 0.000589
Proportion Explained   0.581 0.331 0.0528 0.0236 0.00754 0.00200 0.00114 0.000580 0.000410
Cumulative Proportion  0.581 0.912 0.9648 0.9883 0.99587 0.99787 0.99901 0.999590 1.000000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
```

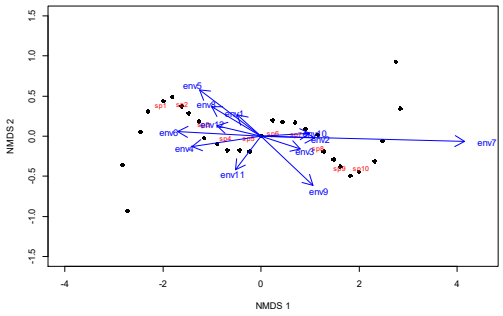
CCA using all environmental variables.

CA biplot and CCA triplot



For comparison…NMDS with weighted averages for env variables

There is no constraining here. This is simply plotting environmental variables in species-space.



CCA Example

Species scores

```
        CCA1     CCA2     CCA3      CCA4     CCA5      CCA6
sp1  -1.3105  1.13548 -0.47923 -0.107058  0.08418  0.009805
sp2  -1.0867  0.55649  0.24995  0.171998 -0.17493 -0.008235
…
```

Site scores (weighted averages of species scores)

```
             CCA1     CCA2     CCA3      CCA4     CCA5     CCA6
sample2  -1.57764  2.39518 -6.3383  -3.17773  7.80222  3.42777
…
```

Site constraints (linear combinations of constraining variables)

```
             CCA1     CCA2     CCA3      CCA4     CCA5     CCA6
sample2  -1.61526  1.8103 -4.7058 -1.00532  2.73522  2.63292
sample3  -1.83546  2.0029 -2.7875 -3.03380  2.21514  0.95788
```

Biplot scores for constraining variables

```
              CCA1      CCA2      CCA3      CCA4      CCA5         CCA6
spatialb -0.13370 -0.91039 -0.14483 -0.01543 -0.07510  1.020e-01
spatialc  0.80013  0.44393  0.22107 -0.04865  0.26046  1.858e-02
env1     -0.14776  0.28676 -0.02818 -0.30078 -0.20745  3.078e-01
…
```

Centroids for factor constraints

```
             CCA1     CCA2     CCA3      CCA4     CCA5     CCA6
spatiala -1.2845  1.1361 -0.1305  0.114251 -0.35088 -0.25748
spatialb -0.1343 -0.9116 -0.1325 -0.002708 -0.07323  0.09816
```

CCA vif scores

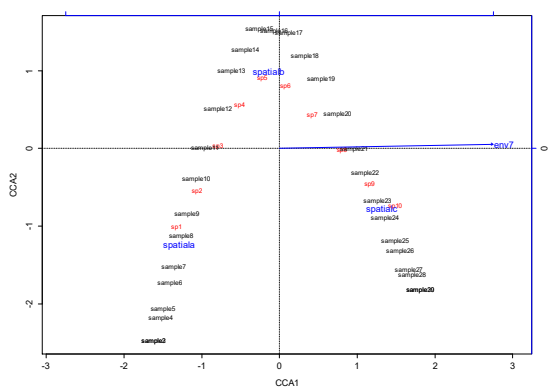
- VIF – variance inflation factor
- Measure of covariance among variables in constraining (environmental) matrix
- Recall that first step of analysis is cannoncial correlation of environmental and species matrix.
- Model contains variables with very high (>5-10) VIF scores…indicating redundancy in some variables (env2 and env10).

```
> vif.cca(base_cca)
  spatialb  spatialc      env1      env2      env3      env4      env5      env6      env7      env8      env9
  8.903666 19.974361  3.008475 332.023395  1.757367  3.599143  2.427301  2.187329  9.626137  3.031498  2.572406
     env10     env11     env12
350.060286  2.176076  2.399144
```

Reduced CCA model

```
Call:
cca(formula = community ~ spatial + env7, data = envdata)

Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total          1.7795     1.0000
Constrained    1.2399     0.6967
Unconstrained  0.5396     0.3033

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                          CCA1   CCA2   CCA3     CA1     CA2     CA3     CA4     CA5     CA6     CA7     CA8     CA9
Eigenvalue              0.8204 0.4049 0.0146 0.2735 0.13900 0.07275 0.02382 0.01248 0.009799 0.004215 0.00275 0.001338
Proportion Explained    0.4610 0.2275 0.0082 0.1537 0.07811 0.04088 0.01339 0.00701 0.005510 0.002370 0.00155 0.000750
Cumulative Proportion   0.4610 0.6885 0.6967 0.8504 0.92855 0.96943 0.98282 0.98983 0.995330 0.997700 0.99925 1.000000

Accumulated constrained eigenvalues
Importance of components:
                          CCA1   CCA2    CCA3
Eigenvalue              0.8204 0.4049 0.01460
Proportion Explained    0.6616 0.3266 0.01177
Cumulative Proportion   0.6616 0.9882 1.00000
```

Recall, all 13 environmental variables explained 80.4%, just 2 explain 70%



CCA model selection

- How do you know which variables to put in the model?
- You should have a hypothesis to test.
- Exploratory approach is usually flawed. Mathematical properties of multiple regression:
  - As you add environmental variables, the variance accounted for by the environmental variables will go up.
  - Even if environmental variables are **random numbers**, variance explained will go up.
  - If the number of environmental variables $\geq$ the number of samples then 100% of variance will be explained (even if they are **random numbers**)
- **Goal** – most explanatory power with the least number of variables
- **Recall** – you are using CCA because "important gradients are known and measured". If this is true, do not throw additional variables in the analysis to "see how it works".

CCA model selection approaches

- Specify your own model based on *a priori* hypothesis
  - Test for and eliminate redundant variables
- Use stepwise procedures to select the best model
  - Function ordistep

```
fullmodel_cca<-cca(community ~ .,environmental)
smallmodel_cca<-cca(community ~ 1, environmental)
fit_model <- ordistep(smallmodel_cca, scope=formula(fullmodel_cca))
```

  - Iterative procedure, works with model P-values
    - Forward, backward or both directions
    - Can select the sensitivity for adding or dropping variables (pin and pout options)
    - Maximum number of permutations
    - Maximum number or steps
  - Alternative is ordiR2step which works with model r^2 and not P

CCA model selection

```
Start: community ~ 1

          Df    AIC      F N.Perm Pr(>F)
+ spatial  2 22.297 21.5441    199 0.0050 **
+ env7     1 30.811 22.9271    199 0.0050 **
+ env5     1 43.932  4.7571    199 0.0100 **
+ env2     1 46.724  1.8425    299 0.1167
+ env10    1 46.883  1.6844    199 0.1650
etc…

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Step: community ~ spatial

          Df    AIC      F N.Perm Pr(>F)
- spatial  2 46.638 21.544     99   0.01 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

         Df    AIC      F N.Perm Pr(>F)
+ env7    1 18.035 6.0248    199   0.01 **
+ env6    1 22.388 1.7011     99   0.17
+ env5    1 22.849 1.2791     99   0.27
+ env8    1 23.053 1.0954     99   0.36
etc…

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Step: community ~ spatial + env7
```

Ordistep and ordiR2step select different final models

ordistep above, ordiR2 below

```
cca(formula = community ~ spatial + env7, data = envdata)

Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total          1.7795     1.0000
Constrained    1.2399     0.6967
Unconstrained  0.5396     0.3033

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                         CCA1   CCA2   CCA3    CA1     CA2…
Eigenvalue            0.8204 0.4049 0.0146 0.2735 0.13900
Proportion Explained  0.4610 0.2275 0.0082 0.1537 0.07811
Cumulative Proportion 0.4610 0.6885 0.6967 0.8504 0.92855

cca(formula = community ~ spatial, data = envdata)

Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total          1.7795     1.0000
Constrained    1.1099     0.6237
Unconstrained  0.6697     0.3763

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                         CCA1   CCA2    CA1     CA2    CA3
Eigenvalue            0.7052 0.4047 0.3030 0.14730 0.1100
Proportion Explained  0.3963 0.2274 0.1703 0.08277 0.0618
Cumulative Proportion 0.3963 0.6237 0.7939 0.87669 0.9385
```

CCA model selection summary

- 1. Use only variables specified in *a priori* hypotheses.
- 2. If you don't have an *a priori* hypothesis (you're probably doing this wrong):
  - Eliminate all variables that are correlated
  - Run stepwise procedure using R^2 as criteria
  - Examine VIF scores of selected model. If VIF scores are high, eliminate redundant variables and start over.

CCA output and interpretation

- Correlation matrix from constraining (environmental) matrix
- Total, constrained and unconstrained inertia (species variance). Eigenvalues proportional to variance explained for each axis.
- **Species scores** - weighted averages of sample scores
- **Two sets of sites scores** – one a weighted average of species scores and one from the multiple regression with environmental variables.
- **Environmental variable scores** – derived from cannonical correlates: correlations between the environmental variables and CCA axes, weighted by the eigenvalues of those axes.
- Proportion of constrained variation explained by each axis. These numbers will be very high but not necessarily meaningful.

Monte Carlos Hypothesis Tests

- Unlike unconstrained ordinations, we have a very specific hypothesis to test. Is there a relationship between the constraining matrix (environmental variables) and the response matrix (species)?
- Significance tested through various permutation tests
- Test the significance of the overall ordination:
  - Community data permuted
  - Pseudo-F calculated as ratio of constrained to unconstrained variance accounted for
  - Null – environmental variables not related, zero constrained
  - P = proportion of random communities producing more than the observed constrained variation.
- Test significance of each axis
  - Similar, test significance of each axis separately
- Test significance of constraining (environmental) variables
  - Variables tested sequentially, order in the model will effect results.

CCA Options

- Detrending (DCCA), but recall problems with detrending
- Multiple regression options
  - Model can include interactions (difficult interpretation)
  - A third matrix to be partialed out (partial CCA)
  - Forward, backward selection
- Number of permutations in Monte Carlo procedures
- Data transformations – same considerations as with CA

Assignment

- Sample data and script
  - Community and environmental matrix.
  - CA, CCA, RDA and model selection
- Chapter 6: CCA, RDA, discriminant function and canonical correlation
- Reading: Titeux, N. et al. 2004. Multivariate analysis of a fine-scale breeding bird atlas using a geographic information system and partial canonical correspondence analysis: environmental and spatial effects. Journal of Biogeography 31: 1841-1856.

Assignment

- Community data from 40 samples (56 species) in black creek
  - Eliminate rare species (less than 3 occurrences)
  - Standardize using the Hellinger method in function decostand (square root of proportions)
- Environmental data from 40 samples (11 variables)
  - Depth
  - CV depth
  - Substrate (mean size on a scale)
  - Mid-water column current velocity
  - CV mid-water velocity
  - Surface velocity
  - CV surface velocity
  - % cover
  - % vegetation
  - Bank stability
  - CV bank stability

Assignment

- Perform CCA (use stepwise procedure, ordistep or ordiR2step)
- Is there a relationship between community structure and environmental variables?
  - Your synthesis needs to:
    - Quantify the relationship between matrices (constrained, unconstrained and individual axis variance explained)
    - Explain how you got to your final model (stepwise procedure or other approach)
    - Report your check of correlation among environmental variables
    - Test for significance of the CCA model, including a formal reporting of the statistics
    - Include a triplot for the final model