SUMMARY:

- MULTIVARIATE DATA & DISTANCE
- REGRESSION BIPLOT
- CORRESPONDENCE ANALYSIS BIPLOTS
- CONSTRAINED BIPLOTS
  (e.g. Canonical Correspondence Analysis, CCA)

The general topic of this course is the interpretation and quantification of relationships between sets of variables.

I use mostly ecological data sets from marine biology, but have two medical examples at the end.
Regression of a response onto predictors

\[ d = 6.14 - 1.39 \text{Pollution} + 0.148 \text{Depth} \]

Dynamic transition from CA to CCA
Rotation of 3-dimensional solution space where prediction takes place; test samples are classified to the closest centroid of the training set.

Five biological and four environmental variables \((n=30\) sites\))

<table>
<thead>
<tr>
<th>Samples</th>
<th>SPECIES COUNTS</th>
<th>ENVIRONMENTAL PREDICTORS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a  b  c  d  e</td>
<td>Depth  Pollution  Temperature  Sediment</td>
</tr>
<tr>
<td>s1</td>
<td>0  2  9  14  2</td>
<td>72  4.8  3.5  S</td>
</tr>
<tr>
<td>s2</td>
<td>26  4  13  11  0</td>
<td>75  2.8  2.5  C</td>
</tr>
<tr>
<td>s3</td>
<td>0  10  9  8  0</td>
<td>59  5.4  2.7  C</td>
</tr>
<tr>
<td>s4</td>
<td>0  0  15  3  0</td>
<td>64  8.2  2.9  S</td>
</tr>
<tr>
<td>s5</td>
<td>13  5  3  10  7</td>
<td>61  3.9  3.1  C</td>
</tr>
<tr>
<td>s6</td>
<td>31  21  13  16  5</td>
<td>94  2.6  3.5  G</td>
</tr>
<tr>
<td>s7</td>
<td>9  6  0  11  2</td>
<td>53  4.6  2.9  S</td>
</tr>
<tr>
<td>s8</td>
<td>2  0  0  0  1</td>
<td>61  5.1  3.3  C</td>
</tr>
<tr>
<td>s9</td>
<td>17  7  10  14  6</td>
<td>68  3.9  3.4  C</td>
</tr>
<tr>
<td>s10</td>
<td>0  5  26  9  0</td>
<td>69  10.0  3.0  S</td>
</tr>
<tr>
<td>s11</td>
<td>0  8  8  6  7</td>
<td>57  6.5  3.3  C</td>
</tr>
<tr>
<td>s12</td>
<td>14  11  13  15  0</td>
<td>84  3.8  3.1  S</td>
</tr>
<tr>
<td>s13</td>
<td>0  0  19  0  6</td>
<td>53  9.4  3.0  S</td>
</tr>
<tr>
<td>s14</td>
<td>13  0  0  9  0</td>
<td>83  4.7  2.5  C</td>
</tr>
<tr>
<td>s15</td>
<td>4  0  10  12  0</td>
<td>100  6.7  2.8  C</td>
</tr>
<tr>
<td>s16</td>
<td>42  20  0  3  6</td>
<td>84  2.8  3.0  G</td>
</tr>
<tr>
<td>s17</td>
<td>4  0  0  0  0</td>
<td>96  6.4  3.1  C</td>
</tr>
<tr>
<td>s18</td>
<td>21  15  33  20  0</td>
<td>74  4.4  2.8  G</td>
</tr>
<tr>
<td>s19</td>
<td>2  5  12  16  3</td>
<td>79  3.1  3.6  S</td>
</tr>
<tr>
<td>s20</td>
<td>0  10  14  9  0</td>
<td>73  5.6  3.0  S</td>
</tr>
<tr>
<td>s21</td>
<td>8  0  0  0  4</td>
<td>59  4.3  3.4  C</td>
</tr>
<tr>
<td>s22</td>
<td>35  10  0  9  17</td>
<td>54  1.9  2.8  S</td>
</tr>
<tr>
<td>s23</td>
<td>6  7  1  17  10</td>
<td>95  2.4  2.9  G</td>
</tr>
<tr>
<td>s24</td>
<td>18  12  20  7  0</td>
<td>64  4.3  3.0  C</td>
</tr>
<tr>
<td>s25</td>
<td>32  26  0  23  0</td>
<td>97  2.0  3.0  G</td>
</tr>
<tr>
<td>s26</td>
<td>32  21  0  10  2</td>
<td>78  2.5  3.4  S</td>
</tr>
<tr>
<td>s27</td>
<td>24  17  0  25  6</td>
<td>85  2.1  3.0  G</td>
</tr>
<tr>
<td>s28</td>
<td>16  3  12  20  2</td>
<td>92  3.4  3.3  G</td>
</tr>
<tr>
<td>s29</td>
<td>11  0  7  8  0</td>
<td>51  6.0  3.0  S</td>
</tr>
<tr>
<td>s30</td>
<td>24  37  5  18  1</td>
<td>99  1.9  2.9  G</td>
</tr>
</tbody>
</table>

Biological variables are abundance counts, three environmental variables are measured on a continuous scale, sediment is categorical:

= C (clay), S (sandy) or G (gravel)
• **Histograms of continuous variables**: Bars are shown connected. Depth distribution looks uniform; Pollution and Temperature look skew.

• **Bar-chart of categorical variable**: Bars are shown separated for three categories of Sediment. Distribution approximately uniform.

**Five species**

• **Histograms of count variables**: all the distribution are ‘skew to the right’
Scatterplots: continuous variables

Showing smoothed relationship between the variables ("scatterplot smoother").

The correlations are shown opposite their scatterplot. Assuming normality, the critical value at 5% level (two-sided test) is $\pm 0.361$

Pearson's $r = -0.396$
($P = 0.030$)

Permutation test on $r$
($P = 0.027$)

Spearman $\rho = -0.423$
($P = 0.020$)

Permutation test on $\rho$
($P = 0.022$)

function `pairs` in R

Scatterplots: count variables
Scatterplots: count vs continuous

Showing coefficients of determination (measure of explained variance, usually expressed as a percentage) - in this simple linear regression case $R^2$ is the square of $r$, the correlation coefficient.

The sign of $r$ can be deduced by the slope of the regression line.

Often the count variable is transformed, for example by a power transformation.

Box-and-whisker plots: continuous vs categorical
Box-and-whisker plots: count vs categorical

$F = 5.53$

$F = 9.74$

$F = 0.54$

$F = 9.61$

$F = 0.24$

Showing $F$ statistics for the analysis of variance (ANOVA), which tests the difference in group means. In this case the critical point of the test at the 1% level is 5.49.

Standardized Euclidean distance

\[ d? \]

$x^* = \frac{x - \text{mean}(x)}{\text{sd}(x)} = \frac{x - 4.52}{2.14}$

$y^* = \frac{y - \text{mean}(y)}{\text{sd}(y)} = \frac{y - 74.4}{15.6}$
(Weighted) Euclidean distance

distance between rows (sites) \(i\) and \(i'\)

\[
d(i, i') = \sqrt{(x_i^* - x_{i'}^*)^2 + (y_i^* - y_{i'}^*)^2}
\]

\[
= \sqrt{\left(\frac{x_i - \bar{x}}{s_x} - \frac{x_{i'} - \bar{x}}{s_x}\right)^2 + \left(\frac{y_i - \bar{y}}{s_y} - \frac{y_{i'} - \bar{y}}{s_y}\right)^2}
\]

\[
= \sqrt{\frac{1}{s_x^2}(x_i - x_{i'})^2 + \frac{1}{s_y^2}(y_i - y_{i'})^2}
\]

For more than two variables, just add similar terms for the other variables.

Inter-site distances

for the three continuous variables depth, pollution and temperature

\[
d(i, i') = \sqrt{\frac{1}{s_x^2}(x_i - x_{i'})^2 + \frac{1}{s_y^2}(y_i - y_{i'})^2 + \frac{1}{s_z^2}(z_i - z_{i'})^2}
\]
The chi-square distance is an alternative measure of dissimilarity that is applicable to frequencies and which is the basis of correspondence analysis (CA).

First convert the rows of frequencies into profiles, that is the frequencies relative to their totals. The row of column sums is also expressed as proportions, called the average profile.

### Chi-square distance

#### Between the First Two Sites

Calculate the chi-square distance between the first two sites as follows:

\[
\begin{align*}
\chi^2 &= \sum_{ij} (o_{ij} - e_{ij})^2 / e_{ij} \\
&= [(s2 - s1) - (s2 - s1)]^2 / (s2 - s1) \\
&= (0.481 - 0.481)^2 / (0.481) \\
&= 0.000
\end{align*}
\]

Therefore

\[
\chi^2 = 0
\]

In general, for profiles \(x_{ij}\) (sample \(i\), species \(j\)) with average profile \(\bar{x}_j\) (for species \(j\)), the chi-square distance between samples is:

\[
d(i,i') = \sqrt{\sum_j \left( \frac{x_{ij} - \bar{x}_j}{\bar{x}_j} \right)^2}
\]

standardization using average
Dealing with heterogeneous data

Gower’s general coefficient of (dis)similarity: standardize each variable and multiply all the columns corresponding to dummy variables by $1/\sqrt{2} = 0.7071$, a factor which compensates for their 0/1 coding:

Continuous variables

<table>
<thead>
<tr>
<th>Station</th>
<th>Depth</th>
<th>Temperature</th>
<th>Salinity</th>
<th>Tarehola</th>
<th>Skognes</th>
<th>Njosken</th>
<th>Storura</th>
<th>Clay</th>
<th>Silt</th>
<th>Sand</th>
<th>Gravel</th>
<th>Stone</th>
</tr>
</thead>
<tbody>
<tr>
<td>s3</td>
<td>30</td>
<td>3.15</td>
<td>33.52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>s8</td>
<td>29</td>
<td>3.15</td>
<td>33.52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s25</td>
<td>30</td>
<td>3.00</td>
<td>33.45</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>s84</td>
<td>66</td>
<td>3.22</td>
<td>33.48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

mean: 58.15 3.086 33.50 0.242 0.273 0.242 0.242 0.606 0.152 0.364 0.182 0.061
s.d.: 32.45 0.100 0.076 0.435 0.452 0.435 0.435 0.496 0.364 0.489 0.392 0.242

compute Euclidean distance between stations

Clustering of sites

based on chi-square distances calculated on species abundances
Simple linear regression

species $d$ versus pollution ($y$)

$d = 19.10 - 1.815y$ ($R^2 = 0.34$)
Multiple linear regression

\[ d = 6.14 - 1.389y + 0.148x \]

Regression model is a (hyper)plane

Regression of a response onto predictors

\[ d = 6.14 - 1.39 \text{Pollution} + 0.148 \text{Depth} \]
Multiple linear regression, variables standardized

\[ d^* = -0.446y^* + 0.347x^* \]

Explanatory variables \( x \) and \( y \) and response variable \( d \) standardized

Response & predictors all standardized

\[ d.z = -0.446 \text{Pollution.z} + 0.347 \text{Depth.z} \]
Line of steepest ascent and contours

<table>
<thead>
<tr>
<th>d.z</th>
<th>Depth.z</th>
<th>Pollution.z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.47</td>
<td>-0.16</td>
<td>0.13</td>
</tr>
<tr>
<td>0.02</td>
<td>0.04</td>
<td>-0.80</td>
</tr>
<tr>
<td>-0.44</td>
<td>-0.99</td>
<td>0.41</td>
</tr>
<tr>
<td>-1.19</td>
<td>-0.67</td>
<td>1.72</td>
</tr>
<tr>
<td>-0.14</td>
<td>-0.86</td>
<td>-0.29</td>
</tr>
<tr>
<td>0.77</td>
<td>1.25</td>
<td>-0.90</td>
</tr>
<tr>
<td>0.02</td>
<td>-1.37</td>
<td>0.04</td>
</tr>
<tr>
<td>-1.64</td>
<td>-0.86</td>
<td>0.27</td>
</tr>
<tr>
<td>0.47</td>
<td>-0.41</td>
<td>-0.29</td>
</tr>
<tr>
<td>-0.29</td>
<td>-0.35</td>
<td>2.56</td>
</tr>
<tr>
<td>-0.74</td>
<td>-1.12</td>
<td>0.93</td>
</tr>
<tr>
<td>0.62</td>
<td>0.61</td>
<td>-0.33</td>
</tr>
<tr>
<td>-1.64</td>
<td>-1.37</td>
<td>2.28</td>
</tr>
<tr>
<td>-0.29</td>
<td>0.55</td>
<td>0.09</td>
</tr>
<tr>
<td>0.17</td>
<td>1.64</td>
<td>1.02</td>
</tr>
<tr>
<td>-1.19</td>
<td>0.61</td>
<td>-0.80</td>
</tr>
<tr>
<td>-1.64</td>
<td>1.38</td>
<td>0.88</td>
</tr>
<tr>
<td>1.37</td>
<td>-0.03</td>
<td>-0.05</td>
</tr>
<tr>
<td>0.77</td>
<td>0.29</td>
<td>-0.66</td>
</tr>
<tr>
<td>-0.29</td>
<td>-0.09</td>
<td>0.51</td>
</tr>
<tr>
<td>-1.04</td>
<td>-0.99</td>
<td>-0.10</td>
</tr>
<tr>
<td>-0.29</td>
<td>-1.31</td>
<td>-1.22</td>
</tr>
<tr>
<td>0.92</td>
<td>1.32</td>
<td>-0.99</td>
</tr>
<tr>
<td>-0.59</td>
<td>-0.67</td>
<td>-0.10</td>
</tr>
<tr>
<td>1.82</td>
<td>1.45</td>
<td>-1.18</td>
</tr>
<tr>
<td>-0.14</td>
<td>0.23</td>
<td>-0.94</td>
</tr>
<tr>
<td>2.12</td>
<td>0.68</td>
<td>-1.13</td>
</tr>
<tr>
<td>1.37</td>
<td>1.12</td>
<td>-0.52</td>
</tr>
<tr>
<td>-0.44</td>
<td>-1.50</td>
<td>0.69</td>
</tr>
<tr>
<td>1.07</td>
<td>1.57</td>
<td>-1.22</td>
</tr>
</tbody>
</table>

Scatterplot of standardized values
Everything on the “ground”

\[ d.z = -0.446 \text{Pollution}.z + 0.347 \text{Depth}.z \]

variance explained \((R^2)\): 44.2%

Regressions of all five responses

variance explained:
\[ a: 52.9\% \]
\[ b: 39.1\% \]
\[ c: 21.8\% \]
\[ d: 44.2\% \]
\[ e: 23.5\% \]

overall:
36.3%

significance:
\[ P \quad D \]
\[ a: ** \]
\[ b: ** \]
\[ c: * \]
\[ d: ** \]
\[ e: * \]
Generalized linear model biplots

- Usual linear regression model:
  Mean $\mu$ of species $d$: $\mu = c + ax + by$
  $d$ normally distributed about this (varying) mean

- Poisson regression model:
  Log of mean $\mu$ of species $d$: $\log(\mu) = c^* + a^*x + b^*y$
  $d$ Poisson distributed …

- Logistic regression model:
  Logit of probability $p$ of species $d$: $\log\left(\frac{p}{1-p}\right) = c^° + a^°x + b^°y$
  Presence of $d$ binomially distributed …

  $\logit(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177 y^* - 0.137 x^*$

---

Logistic regression biplots

```
<table>
<thead>
<tr>
<th>SITE</th>
<th>SPECIES COUNTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO.</td>
<td>a</td>
</tr>
<tr>
<td>s1</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>26</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
</tr>
<tr>
<td>s4</td>
<td>0</td>
</tr>
<tr>
<td>s5</td>
<td>13</td>
</tr>
<tr>
<td>s6</td>
<td>31</td>
</tr>
<tr>
<td>s7</td>
<td>9</td>
</tr>
<tr>
<td>s8</td>
<td>2</td>
</tr>
<tr>
<td>s9</td>
<td>17</td>
</tr>
<tr>
<td>s10</td>
<td>0</td>
</tr>
<tr>
<td>s11</td>
<td>0</td>
</tr>
<tr>
<td>s12</td>
<td>14</td>
</tr>
<tr>
<td>s13</td>
<td>0</td>
</tr>
</tbody>
</table>
```

transform to presence/absence (0/1) data

```
0 1 1 1 1
1 1 1 1 0
0 1 1 1 0
0 0 1 1 0
1 1 1 1 1
1 1 1 1 1
1 1 0 1 1
1 0 0 0 1
1 1 1 1 1
0 1 1 1 0
0 1 1 1 1
1 1 1 1 0
0 0 1 0 1
```

$logit(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177 y^* - 0.137 x^*$
What happens for three predictors?

- Each regression model can be represented as a point vector in three-dimensional space.
- Reconstruct the data from projections of cases onto variable directions, but only as well as measured by $R^2$; in this example the increase in explained variance from two-dimensional to three-dimensional (adding temperature as an explanatory variable) is from 36.3% to 37.1%, hence temperature is explaining very little extra variance.
Dimensional Transmogrifier =
Singular Value Decomposition (SVD)

- The SVD is a matrix decomposition of the (possibly centred and possibly normalized) $Y$:
  \[ Y = U D_\alpha V^T \]
  where $U$ and $V$ are matrices with orthonormal columns: $U^T U = V^T V = I$ and $D_\alpha$ is a diagonal matrix of singular values $\alpha_1 \geq \alpha_2 \geq \cdots > 0$

Standard coordinates of columns: $V$
Principal coordinates of rows: $F = UD_\alpha$

So:
\[ Y = UD_\alpha V^T = F V^T = \]
Principal coordinates
Standard coordinates

first two columns of $V$
first two columns of $F$
Each species is regressed onto the dimensions

Do a (weighted) regression of the variable on the two dimensions, then use the regression coefficients as coordinates.
Adding explanatory variables to the map

Do a (weighted) regression of the variable on the two dimensions, then use the regression coefficients as coordinates.

Explanatory variables not specifically “optimized” in the display (this is called “indirect gradient analysis”)

Variance explained of supplementary variables:

- Polln: 69.5%
- Depth: 30.4%
- Temp: 2.1%
Canonical correspondence analysis (CCA)

CCA analyses the counts, finds their main dimensions and then we can relate these dimensions to the environmental variables (e.g., as supplementary points) – they might or might not be related to the environmental variables.

CCA does a more specific analysis: the dimensions are forced to be linear combinations of the environmental variables – we call this a constrained, or restricted, version of CA.

Thus the CCA axes are not the best axes of the species data (these would be found by CA). They are the best axes in the space correlated directly with the environmental predictors.

Recall: Indirect Gradient Analysis

Variance explained of supplementary variables:

Polln: 69.5%
Depth: 30.4%
Temp: 2.1%
CCA with 3 continuous explanatory variables

Variance explained of constraining variables in 2-d solution:
- Polln: 99.4%
- Depth: 91.3%
- Temp: 3.7%

Using R function `cca` in `vegan` package

```
plot(cca(abcde,xyz), display=c("lc","bp","sp"))
```
### CCA inertia decomposition for $abcde/pdt$ data

**Recall CA results**

Explained inertia:
- Axis 1: 0.288 (53.0%)
- Axis 2: 0.120 (22.2%)
- In 2-d: 0.408 (75.2%)

Total inertia: 0.5436

**inertia in full space**

- 0.5436 (100%)

**inertia in restricted space**

- 0.2399 (44.1%)

**inertia outside restricted space**

- 0.3038 (55.9%)

(we could also decompose this inertia NOT explained by environmental variables, along principal axes)

**inertia explained by 2-d solution**

- 0.2352 (98.0%)

**inertia not explained by 2-d solution**

- 0.0047 (2.0%)

---

### Data set “benthos”

A marine ecological data set with 92 species counted at 13 locations: 11 polluted locations (S..) and 2 unpolluted reference locations (R..)

<table>
<thead>
<tr>
<th>Species</th>
<th>S4</th>
<th>S8</th>
<th>S9</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
<th>S15</th>
<th>S18</th>
<th>S19</th>
<th>S23</th>
<th>S24</th>
<th>R40</th>
<th>R42</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Myri_ocul</em></td>
<td>193</td>
<td>79</td>
<td>150</td>
<td>72</td>
<td>141</td>
<td>302</td>
<td>114</td>
<td>136</td>
<td>267</td>
<td>114</td>
<td>136</td>
<td>267</td>
<td>114</td>
</tr>
<tr>
<td><em>Chae_seto</em></td>
<td>34</td>
<td>4</td>
<td>247</td>
<td>19</td>
<td>52</td>
<td>250</td>
<td>331</td>
<td>12</td>
<td>125</td>
<td>37</td>
<td>12</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td><em>Amph_falc</em></td>
<td>49</td>
<td>58</td>
<td>66</td>
<td>47</td>
<td>78</td>
<td>92</td>
<td>113</td>
<td>38</td>
<td>96</td>
<td>76</td>
<td>37</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><em>Myse_bide</em></td>
<td>30</td>
<td>11</td>
<td>36</td>
<td>65</td>
<td>35</td>
<td>37</td>
<td>24</td>
<td>3</td>
<td>20</td>
<td>156</td>
<td>12</td>
<td>58</td>
<td>43</td>
</tr>
<tr>
<td><em>Goni_macu</em></td>
<td>35</td>
<td>39</td>
<td>41</td>
<td>37</td>
<td>32</td>
<td>45</td>
<td>41</td>
<td>41</td>
<td>31</td>
<td>29</td>
<td>64</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td><em>Amph_fili</em></td>
<td>19</td>
<td>39</td>
<td>11</td>
<td>38</td>
<td>18</td>
<td>20</td>
<td>11</td>
<td>22</td>
<td>30</td>
<td>40</td>
<td>3</td>
<td>55</td>
<td>65</td>
</tr>
<tr>
<td><em>Timo_ovat</em></td>
<td>1</td>
<td>78</td>
<td>5</td>
<td>95</td>
<td>20</td>
<td>0</td>
<td>7</td>
<td>55</td>
<td>50</td>
<td>44</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><em>Falc_cros</em></td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Epit_trev</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Ophi_flex</em></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Eucl_sp_</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Scal_infl</em></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Eumi_ocke</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>Modi_modi</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

A marine ecological data set with 92 species counted at 13 locations: 11 polluted locations (S..) and 2 unpolluted reference locations (R..)
Asymmetric CA biplot of ‘benthos’

explained inertia = 57.5%

CA contribution biplot of ‘benthos’
(each species point is multiplied by the square root of its mass)

explained inertia = 57.5%
Transition to contribution biplot

Data set “benthos” again

(Greenacre, Correspondence Analysis in Practice, 2007: chaps10&24)

### 13 sampling stations (last 2 references)

<table>
<thead>
<tr>
<th>Species</th>
<th>S4</th>
<th>S8</th>
<th>S9</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
<th>S15</th>
<th>S18</th>
<th>S19</th>
<th>S23</th>
<th>S24</th>
<th>R40</th>
<th>R42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myri_ocul</td>
<td>193</td>
<td>79</td>
<td>150</td>
<td>72</td>
<td>141</td>
<td>302</td>
<td>114</td>
<td>136</td>
<td>267</td>
<td>271</td>
<td>992</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Chae_seto</td>
<td>34</td>
<td>4</td>
<td>247</td>
<td>19</td>
<td>52</td>
<td>250</td>
<td>331</td>
<td>12</td>
<td>125</td>
<td>37</td>
<td>12</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Amph_falc</td>
<td>49</td>
<td>58</td>
<td>66</td>
<td>47</td>
<td>78</td>
<td>92</td>
<td>113</td>
<td>38</td>
<td>96</td>
<td>76</td>
<td>37</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Myse_bide</td>
<td>30</td>
<td>11</td>
<td>36</td>
<td>65</td>
<td>35</td>
<td>37</td>
<td>21</td>
<td>3</td>
<td>20</td>
<td>156</td>
<td>12</td>
<td>58</td>
<td>43</td>
</tr>
<tr>
<td>Gonl_macu</td>
<td>35</td>
<td>39</td>
<td>41</td>
<td>37</td>
<td>32</td>
<td>45</td>
<td>41</td>
<td>41</td>
<td>31</td>
<td>29</td>
<td>64</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td>Amph_fil</td>
<td>19</td>
<td>39</td>
<td>11</td>
<td>38</td>
<td>18</td>
<td>20</td>
<td>11</td>
<td>22</td>
<td>30</td>
<td>40</td>
<td>3</td>
<td>55</td>
<td>65</td>
</tr>
<tr>
<td>Timo_ovat</td>
<td>1</td>
<td>78</td>
<td>5</td>
<td>95</td>
<td>20</td>
<td>0</td>
<td>7</td>
<td>55</td>
<td>50</td>
<td>44</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Sthe_lim</td>
<td>22</td>
<td>33</td>
<td>29</td>
<td>26</td>
<td>16</td>
<td>19</td>
<td>27</td>
<td>21</td>
<td>22</td>
<td>22</td>
<td>12</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>Chae_niti</td>
<td>25</td>
<td>21</td>
<td>35</td>
<td>14</td>
<td>16</td>
<td>17</td>
<td>9</td>
<td>11</td>
<td>41</td>
<td>36</td>
<td>11</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Tric_sp</td>
<td>9</td>
<td>26</td>
<td>5</td>
<td>30</td>
<td>35</td>
<td>2</td>
<td>11</td>
<td>13</td>
<td>5</td>
<td>63</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Expl. vars

<table>
<thead>
<tr>
<th>Element</th>
<th>S4</th>
<th>S8</th>
<th>S9</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
<th>S15</th>
<th>S18</th>
<th>S19</th>
<th>S23</th>
<th>S24</th>
<th>R40</th>
<th>R42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ba</td>
<td>1656</td>
<td>1373</td>
<td>3680</td>
<td>2094</td>
<td>2813</td>
<td>4493</td>
<td>6466</td>
<td>1661</td>
<td>3580</td>
<td>2247</td>
<td>2034</td>
<td>40</td>
<td>85</td>
</tr>
<tr>
<td>Fe</td>
<td>2022</td>
<td>2398</td>
<td>2985</td>
<td>2535</td>
<td>2612</td>
<td>2515</td>
<td>3421</td>
<td>2381</td>
<td>3452</td>
<td>3457</td>
<td>2311</td>
<td>1804</td>
<td>1815</td>
</tr>
<tr>
<td>PELITE</td>
<td>2.9</td>
<td>14.9</td>
<td>3.8</td>
<td>5.3</td>
<td>4.2</td>
<td>9.1</td>
<td>5.3</td>
<td>4.1</td>
<td>7.4</td>
<td>3.1</td>
<td>6.5</td>
<td>2.5</td>
<td>2</td>
</tr>
</tbody>
</table>

(transformed to logarithms for CCA)
Regular CA: explanatory variables added

Top (>1%) contributing species shown (account for 75% of the solution)

Axes 1 and 2 have regressions on log(Ba), log(Fe) and log(PE) with $R^2$ of 0.494 and 0.319.

Three explanatory variables have $R^2$ with two axes of 0.648 (Ba), 0.326 (Fe), 0.126 (PE).

Canonical CA: dimensions restricted to be linear combinations of explanatory variables

Top (>1%) contributing species shown (account for 83% of the solution)

Axes equal to linear combinations of log(Ba), log(Fe) and log(PE) - $R^2$ of 1 and 1.

Three explanatory variables have $R^2$ with two axes of 0.998 (Ba), 0.886 (Fe), 0.529 (PE).
CCA inertia decomposition of “benthos”

inertia in full space
0.783 (100%)

inertia in restricted space
0.280 (35.8%)

inertia outside restricted space
0.503 (64.2%)

we could also decompose along principal axes this inertia which is uncorrelated with the explanatory variables

inertia explained by (e.g., 2-d) solution 0.251
(89.7% in restricted space)
(32.1% in full space)

inertia not explained by (e.g., 2-d) solution 0.029 (10.3%)

Dynamic transition from CA to CCA

CA is based on the SVD of:

$S = D_r^{-rac{1}{2}}(P - rc^T)D_c^{-rac{1}{2}} = U\Gamma V^T$

CCA involves (weighted) projection onto the explanatory variables in $X$:

$Q = D_r^{-rac{1}{2}}X(X^TX)^{-1}X^TD_r^{-rac{1}{2}}$

So in CCA the SVD is computed of:

$S^\perp = QS$

Parametrize the “projection” matrix:

$Q^* = Q^*(\delta) = \delta I + (1 - \delta)Q$

Generate a sequence of analyses by letting $\delta$ vary from 1 to 0, each time “projecting” using $Q$ (function of $\delta$):

$S^* = Q^*S$
Sampling area in Barents Sea: 1992-2004

Fish community structure in Barents Sea


- 1360 stations on a grid in the period 1992–2004
- 82 fish species
- environmental variables depth, bottom temperature, geographical region, year
- all environmental variables coded as categorical: depth, temperature and region all coded as fuzzy variables

Fuzzy Sets and Systems.

A medical example
The Framingham data set

Large cohort study – we use here a subset of the Framingham data provided with the online edX course from Harvard School of Public Health: “Health in Numbers: Quantitative Methods in Clinical Research” (www.edx.org). N=4012

Look at interrelationships of some of the principal quantitative measures: systolic and diastolic blood pressure, heart rate, glucose and cholesterol level,..., and the association with gender and age.

Code all quantitative variables fuzzily, 5 categories in this case.
CCA map of constrained part of the data

Total variance (inertia) = 1.884
Part that is explained by sex–age interaction = 0.080 (4.3%)
Most of this 4.3% explained in above CCA map

CCA map of constrained part of the data
Contribution Biplot
An example from Hastie et al. (2009)
*The Elements of Statistical Learning. 2nd Edition.*

2308 genes

63 tumor samples

BL (Burkitt lymphoma)
EWS (Ewing’s sarcoma)
NB (neuroblastoma)
RMS (rhabdomyo-sarcoma)

20 tumor samples and 5 unknowns

Objective: to identify a small subset of genes that is determinant in classifying the type of tumour

Guided tour through n-dimensional space
Guided tour through n-dimensional space

63-d space

- tumor
- tumor average ("centroid")
Guided tour through n-dimensional space

How do we do the tour?

Let weight $\beta$ vary smoothly from 1 to 0 (typically 1, 0.99, 0.98, ..., 0.01, 0) – each time we get a different view as we move into the multidimensional space. $\beta$ is a tuning parameter.
Training data

Test data

Voyage through 63-dimensional space, looking for best predictor of tumors in test data

Quality of representation of tumor centroids almost perfect (98.7%)

Lowest misclassification achieved (10%)

This frame is at approximate 0.4 : 0.6 distribution of weight between tumors and tumor centroids ($\beta = 0.6$).
Peeling away genes with least predictive power, seeing effect on prediction of test data.

46 genes (indicated by dots in this joint plot)

No errors of prediction in 3-dimensional space

Rotation of 3-dimensional solution space where prediction takes place; test samples are classified to the closest centroid of the training set.