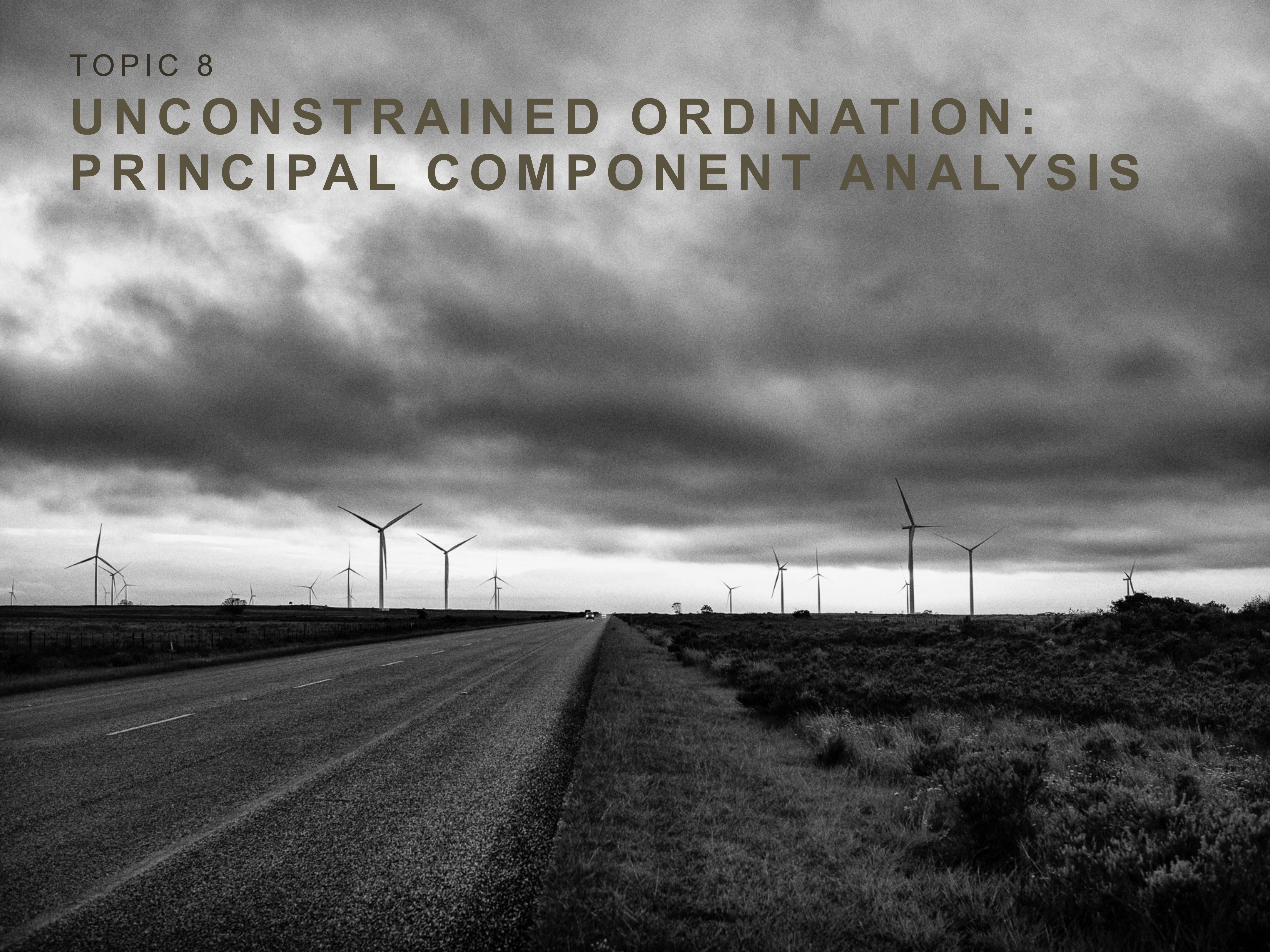
TOPIC 8

UNCONSTRAINED ORDINATION: PRINCIPAL COMPONENT ANALYSIS



Principal Component Analysis (PCA)

- see https://sites.google.com/site/mb3gustame/indirect-gradient-analysis/pca and https://www.davidzeleny.net/anadat-r/doku.php/en:pca
- in ecology, PCA is done by the eigen-decomposition of an distance matrix or association matrix
- this yields a scaling and rigid rotation of axes in that the positions of points relative to one another (Euclidean distances) are maintained during rotation
- the scaled and rotated axes are referred to as principle components (the PC axes)
- the higher the degree of correlation between env. variables (or associations between species) is, the more strongly 'focused' the data cloud is, and therefore the more 'successful' the PCA is in terms of being able to represent many variables by only a few new variables (i.e. a few PC axes)
- since the association matrix is a correlation matrix, the sum of the eigenvalues along the diagonal equals the number of 'species' (here env. vars.)
- it maximises the 'variance explained'

https://youtu.be/_UVHneBUBW0
https://youtu.be/FgakZw6K1QQ

Principal Component Analysis (PCA)

```
> env
# A tibble: 29 x 11
     dfs   alt   slo   flo    pH   har   pho   nit   amm   oxy   bod
   <dbl> <int> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
 1   0.3   934  48    0.84   7.9    45  0.01  0.2   0     12.2   2.7
 2   2.2   932   3    1      8      40  0.02  0.2   0.1   10.3   1.9
 3  10.2   914   3.7  1.8    8.3    52  0.05  0.22  0.05  10.5   3.5
 4  18.5   854   3.2  2.53   8      72  0.1   0.21  0     11     1.3
 5  21.5   849   2.3  2.64   8.1    84  0.38  0.52  0.2    8     6.2
 6  32.4   846   3.2  2.86   7.9    60  0.2   0.15  0     10.2   5.3
 7  36.8   841   6.6  4      8.1    88  0.07  0.15  0     11.1   2.2
 8  70.5   752   1.2  4.8    8      90  0.3   0.82  0.12   7.2   5.2
 9  99     617   9.9 10      7.7    82  0.06  0.75  0.01  10     4.3
10 123.    483   4.1 19.9    8.1    96  0.3   1.6   0     11.5   2.7
# ... with 19 more rows
```

```
# PCA based on a correlation matrix
# Argument scale = TRUE calls for a standardization of the variables and it
# creates a correlation matrix; this is necessary because the variables each
# have a different measurement scale
env.pca <- rda(env, scale = TRUE)
```

...intermediate correlation matrix produced by `scale = TRUE`...

```
> cor(env)
            dfs         alt        slo         flo          pH         har        pho         nit        amm        oxy        bod
dfs  1.00000000 -0.93837894 -0.3947527  0.94742121  0.01604625  0.73277932  0.4729141  0.73809569  0.4076923 -0.5700685  0.4346159
alt -0.93837894  1.00000000  0.4571298 -0.86289080 -0.05035674 -0.78551547 -0.4371091 -0.75260515 -0.3811344  0.4248230 -0.3825222
slo -0.39475266  0.45712978  1.0000000 -0.35761430 -0.22222041 -0.52669175 -0.1953638 -0.31433990 -0.1746442  0.3076363 -0.1738556
flo  0.94742121 -0.86289080 -0.3576143  1.00000000  0.03312637  0.73662526  0.3786878  0.59315083  0.2925252 -0.4210945  0.2951891
pH   0.01604625 -0.05035674 -0.2222204  0.03312637  1.00000000  0.08451511 -0.0794025 -0.04046292 -0.1220018  0.1923980 -0.1617056
har  0.73277932 -0.78551547 -0.5266918  0.73662526  0.08451511  1.00000000  0.3731861  0.53495392  0.2961360 -0.3736374  0.3369747
pho  0.47291414 -0.43710911 -0.1953638  0.37868776 -0.07940250  0.37318607  1.0000000  0.80093149  0.9699345 -0.7575015  0.9091698
nit  0.73809569 -0.75260515 -0.3143399  0.59315083 -0.04046292  0.53495392  0.8009315  1.00000000  0.8022323 -0.6867146  0.6832300
amm  0.40769227 -0.38113442 -0.1746442  0.29252515 -0.12200180  0.29613600  0.9699345  0.80223230  1.0000000 -0.7462700  0.9028247
oxy -0.57006855  0.42482297  0.3076363 -0.42109447  0.19239804 -0.37363740 -0.7575015 -0.68671462 -0.7462700  1.0000000 -0.8398165
bod  0.43461586 -0.38252216 -0.1738556  0.29518914 -0.16170564  0.33697473  0.9091698  0.68322998  0.9028247 -0.8398165  1.0000000
```

Principal Component Analysis (PCA)

```
> # What is the total (unconstrained) inertia? Remember, since the
> # association matrix is a correlation matrix, the sum of the eigenvalues
> # along the diagonal equals the number os 'species' (here env. vars.)
> sum(env.pca$CA$eig)
[1] 11
```

```
> env.pca
Call: rda(X = env, scale = TRUE)

              Inertia Rank
Total              11
Unconstrained      11   11
Inertia is correlations

Eigenvalues for unconstrained axes:
  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11
6.098 2.167 1.038 0.704 0.352 0.319 0.165 0.112 0.023 0.017 0.006
```

continue…

```
> # The inertia associated with the first PC axis is
> env.pca$CA$eig[1]
     PC1
6.097948
```

Principal Component Analysis (PCA)

```
> summary(env.pca) # Default scaling 2

Call:
rda(X = env, scale = TRUE)

Partitioning of correlations:
              Inertia Proportion
Total              11          1
Unconstrained      11          1

Eigenvalues, and their contribution to the correlations

Importance of components:
                         PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8      PC9    PC10      PC11
Eigenvalue            6.0979 2.1672 1.03761 0.70353 0.35174 0.31912 0.16453 0.11173 0.023134 0.01737 0.0060521
Proportion Explained  0.5544 0.1970 0.09433 0.06396 0.03198 0.02901 0.01496 0.01016 0.002103 0.00158 0.0005502
Cumulative Proportion 0.5544 0.7514 0.84571 0.90967 0.94164 0.97065 0.98561 0.99577 0.997870 0.99945 1.0000000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
* General scaling constant of scores:  4.189264
```

continue…

- eigenvalues are the amount of variation (inertia) explained by the new variables (PC axes), *i.e.* (6.0979/11)=0.5544 … if using a correlation matrix (the default)!
- percentage and cumulative percentage
- the first principal component explains 55.44% of the variation
- the second adds an additional 19.70%
- the total amount of explanation offered by PC1 + PC2 is 75.14%
- all of the axes (here 11) account for 100% of the variation

Principal Component Analysis (PCA)

- **species scores:** the loadings (a.k.a. scores or *scaled eigenvectors*) indicate the 'strength of contribution' of the original variables to the new variables, the Principal Components (PC1, PC2, etc.)
- they indicate how much each of the original variables contribute to PC1, PC2, etc.
- larger (more +ve) and smaller (more -ve) values indicate a greater contribution (albeit in opposite directions)
  - here, PC1 is made up of uneven contributions from most of the original variables
  - largest value is nitrate (1.1432) and smallest is oxygen (-1.0089)… these contribute most towards the differences between sites… places with the most nitrate will have the least dissolved oxygen (which makes ecological sense too)
  - pH and slope are least important
- given the strength of this principal component (it explains 55.44% of the inertia), one might hypothesise that its constituent variables influence many aspects of the community, but the vars oxygen and nitrate are most influential

…continue

Species scores

	PC1	PC2	PC3	PC4	PC5	PC6
dfs	1.0842	0.5150	-0.25749	-0.16168	0.21132	-0.09485
alt	-1.0437	-0.5945	0.17984	0.12282	0.12464	0.14022
slo	-0.5752	-0.5103	-0.55499	-0.80204	0.02764	0.20077
flo	0.9577	0.6412	-0.30654	-0.19427	0.18401	0.03068
pH	-0.0586	0.4820	1.03444	-0.51378	0.14421	0.05821
har	0.9072	0.6181	-0.02280	0.15767	-0.27865	0.50738
pho	1.0460	-0.6093	0.18734	-0.11866	-0.15113	0.04888
nit	1.1432	-0.1290	0.01203	-0.18471	-0.21270	-0.34907
amm	0.9954	-0.6989	0.18597	-0.08277	-0.19234	-0.04979
oxy	-1.0089	0.4578	-0.00918	-0.23450	-0.50552	-0.05764
bod	0.9899	-0.6836	0.11962	0.03646	0.08542	0.21993

continue…

Principal Component Analysis (PCA)

…continue

Site scores (weighted sums of species scores)

	PC1	PC2	PC3	PC4	PC5	PC6
sit1	-1.41243	-1.47560	-1.74593	-2.95533	0.23051	0.49227
sit2	-1.04173	-0.81761	0.34075	0.54364	0.92835	-1.76876
sit3	-0.94881	-0.48823	1.36059	-0.21768	1.05289	-0.69640
sit4	-0.88070	-0.29459	0.21014	0.66428	-0.23934	-0.06427
sit5	-0.42588	-0.66503	0.77631	0.78777	0.62942	1.17850
sit6	-0.77730	-0.74514	-0.06764	0.90839	0.46945	-0.32923
sit7	-0.78155	-0.09448	0.39335	0.23079	-0.45431	1.17306
sit8	-0.28732	-0.47352	0.29471	1.13215	0.69812	1.05344
sit9	-0.49324	-0.44884	-1.31854	0.78932	-0.38574	0.41597
sit10	-0.28009	0.43091	0.12225	-0.11790	-1.07206	0.45807
sit11	-0.44849	0.33200	-0.53096	0.60345	-0.96682	0.11691
sit12	-0.38850	0.68558	0.10462	0.08107	-1.10978	0.84504
sit13	-0.24996	0.74160	0.88642	-0.46709	-0.96946	0.74682
sit14	-0.31329	0.93929	1.93010	-1.27078	0.06283	0.14773
sit15	-0.14329	0.31112	-0.21270	0.24363	-0.61744	-0.52902
sit16	0.08956	0.29836	-0.18601	0.23428	-0.73164	-0.44261
sit17	0.05649	0.34914	-0.22049	0.14198	-0.76039	-0.60408
sit18	0.04513	0.40790	0.12272	-0.20091	-0.49665	-0.87755
sit19	0.16126	0.36126	-0.28796	-0.05345	-0.79294	-1.36200
sit20	0.16079	0.32644	-0.74873	0.40912	0.17568	-0.90766
sit21	0.14178	0.53551	-0.08106	-0.07021	0.58856	-0.24654
sit22	1.37614	-1.19047	0.74766	-0.35075	-0.22921	0.75808
sit23	0.98260	-0.51434	0.01123	0.40978	1.01197	0.84836
sit24	2.18633	-2.04860	0.35017	-0.29583	-1.26009	-0.39052
sit25	0.88257	-0.11921	-0.64734	0.34001	0.85793	-0.14280
sit26	0.63983	0.39438	-0.15997	-0.30089	1.09889	-0.66497
sit27	0.75833	0.80559	0.51015	-0.96863	0.42032	-0.74305
sit28	0.65324	1.09406	-1.68227	0.37796	0.43707	0.65309
sit29	0.73849	1.36252	-0.27161	-0.62819	1.42387	0.88211

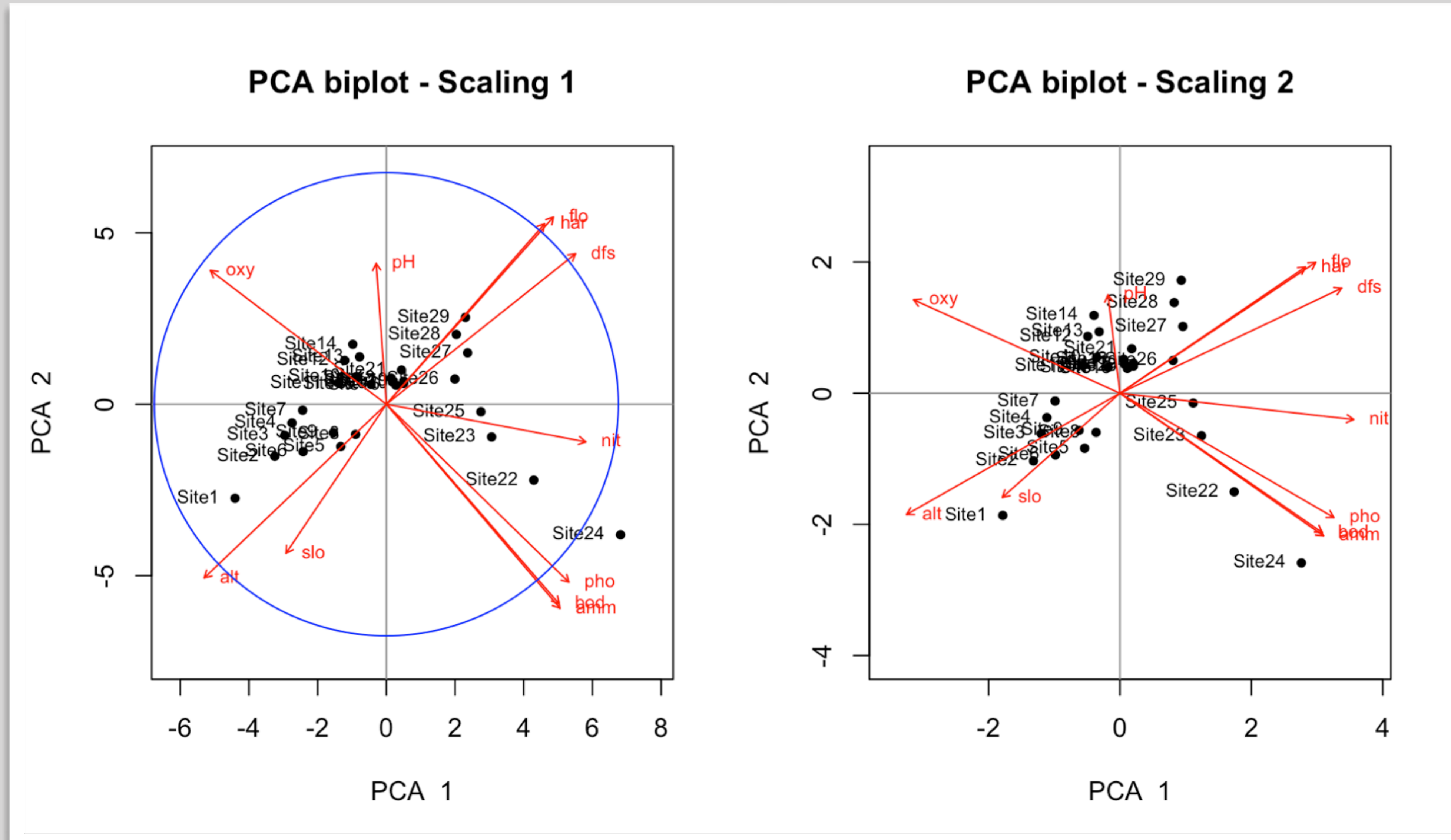
- **site scores:** the (scaled) coordinates of the objects (sites)

Principal Component Analysis (PCA)

```
# Two PCA biplots: scaling 1 and scaling 2
# ****************************************

# Plots using biplot.rda (top row)
# dev.new(width = 12, height = 6, title = "PCA biplots - environmental variables - biplot.rda")
# Plots using cleanplot.pca - NEW VERSION OF THIS FUNCTION which changed syntax (bottom row)
# A rectangular graphic window is needed for the two plots
# dev.new(width = 12, height = 6, title = "PCA biplots - environmental variables - cleanplot.pca")
par(mfrow = c(2, 2))
biplot(env.pca, scaling = 2, choices = c(1, 2), main = "PCA - scaling 1")
biplot(env.pca, choices = c(1, 2), main = "PCA - scaling 2")  # Default scaling 2
cleanplot.pca(env.pca, scaling = 1, mar.percent = 0.08)
cleanplot.pca(env.pca, scaling = 2, mar.percent = 0.04)
```

Principal Component Analysis (PCA)



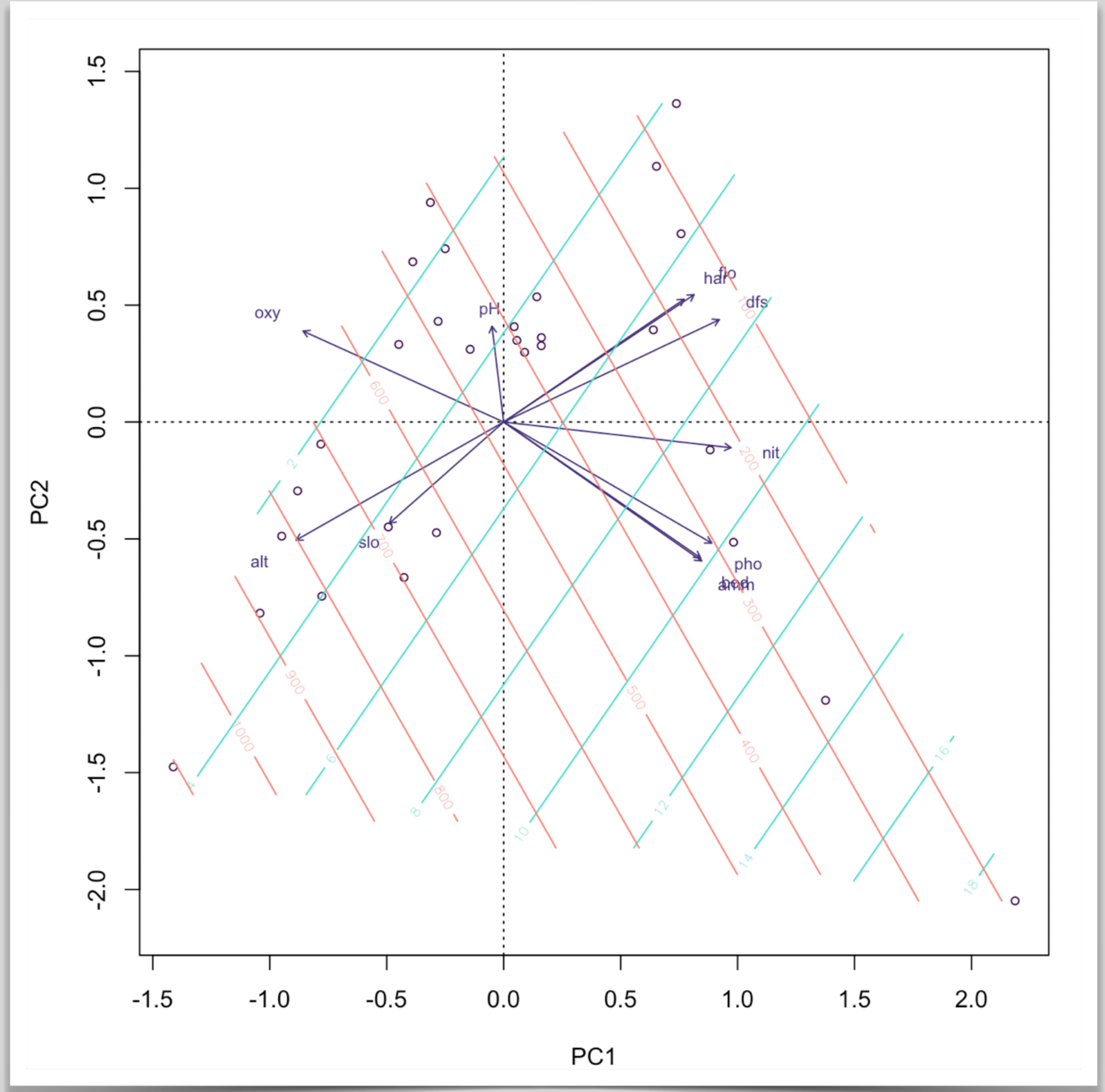
Principal Component Analysis (PCA)

```r
# Ordisurf of the PCA
# ****************************************

par(mfrow = c(1, 1))
palette(viridis(8))
biplot(env.pca, type = c("text","points"))
tmp <- ordisurf(env.pca ~ bod, env, add = TRUE, col = "turquoise", knots = 1)
tmp <- ordisurf(env.pca ~ alt, env, add = TRUE, col = "salmon", knots = 1)
```

Principal Component Analysis (PCA)

- the contours form a linear trend surface, *i.e.* they are perpendicular to the arrow
- **this is the main problem of PCA**, as community data are *nonlinear*
- in general, **PCA should *not* be used for community data**
- PCA is useful for linear data
- can be used to derive new combined variables for other analyses when PCA explains a large proportion of variance
- often displays a ‘horseshoe’ effect, indicating a non-linear response



Read more at http://www.flutterbys.com.au/stats/tut/tut14.2.htm

Self study of PCA

Replicate the analysis shown above on the

1. The Doubs River data

 * DoubsEnv.csv … the environmental data

2. The Seaweed Data

 * ordiStat_58_v2.0.RData … the environmental data

Notes:

- consult the book Numerical Ecology with R thoroughly
- you will need the **vegan** package
- the file cleanplot.pca.R is provided as it contains a function that you will use
- background reading to Doubs River data in https://www.davidzeleny.net/anadat-r/doku.php/en:data:doubs
- background reading to Seaweed data in Smit et al. (2017)
- submit the annotated R script next week Monday, 17 August 2020, at 23:55; make sure to include an explanation of the findings for both the Doubs River and Seaweed environmental data

FAQ

Question 1:

- One of the graphs are in a circle and the other does not. Why is this? What does the circle represent?
- What does the circle of equilibrium mean?

Question 2:

- Will the PCA plots always only have the two axis?
- To calculate the proportion of variance, do you only use the values of the which ever axis are shown (e.g. PC1 or PC2)?

Question 3:

- How do we compare variables between different sites exactly?

Question 4:

- What does the length of the arrows mean, why are some of them short and others longer?

Question 5:

- Why do some of the principal components produce negative values in their eigenvalues?

Question 6:

- Why should you standardise variables first before running PCA?

Question 7:

- Should you leave some variables unstandardised if you want certain variables to have more weighting than others?

Question 8:

- Can PCAs be used for things like GIS or molecular work involving large Datasets? In order to summarise climate data for ecological niche modeling.?

Question 9:

- I do not understanding why and how PCAs are relevant for certain datasets and not for others.

Question 10:

- Explain what the distance between objects mean?

Question 11:

- How do you know if the variables are positively or negatively correlated with one another by looking at the graph?