Introduction to Multivariate Analysis in R

George Perry

Te Whare Wānanga o Tāmaki Makaura | University of Auckland

16th September, 2019

http://spatialecol.com/learning/durham-mva/

pwd: ordin4tion

Thanks to Olivia Burge for some of the materials these slides build on

Structure

- The nature of multivariate data
- Clustering, classification and ordination
  - agglomerative classification, *k*-means, *k*-medioids
  - unconstrained ordination: PCA and nMDS
  - *post-hoc* analysis of dissimilarity: visualisation and distance tests
  - constrained ordination: distance-based redundancy analysis
  - indicator (species) anaysis
- Model-based approaches (as opposed to distance-based)

The focus is on the **how** in R, rather than the theoretical underpinnings of the methods

Getting up and running

You need R installed (current version is 3.6.1), RStudio (makes life easier) and the following: tidyverse, vegan, ggvegan, ggbiplot, pvclust, NbClust, plus dependencies

```
install.packages('tidyverse')
install.packages('factoextra')
install.packages('vegan')
install.packages('pvclust')
install.packages('dendextend')
install.packages('mgcv')
install.packages('mvabund')
install.packages('devtools')
devtools::install_github('gavinsimpson/ggvegan')
devtools::install_github('vqv/ggbiplot')
```

Getting up and running

We'll use two datasets to explore some multivariate methods:

1. data from Lee et al. 2017 and Lee et al. 2018, comprising a set of ecological and environmental data used to investigate the effects of land-use change and the invasive fish *Gambusia affinis* on stream ecosystems

2. community composition data from the Grampians National Park (Aust.) with information on plots with different burn ages

Why multivariate analysis?

Environmental (including ecological) sampling: many variables at many sites

Objectives

- condense data to look for trends and differences
- generate questions and hypotheses
- model-based inference

What and why?

The nature of multivariate data

Data from multi-site monitoring programs (e.g. water quality, ecological communities) are multivariate and matrix-like (sites × variable / species or *vv*)

Sites/samples [objects] →

Species [variables] ↓

	99-1-1	99-2-1	99-3-1	99-3-2	99-4-1	99-5-1	96-1-1	96-1-2	96-1-3	96-2-1	96-2-2
Acacia mitchellii	0.2	0	0	0	0	0	0	0	0	0	0
Acacia myrtifolia	0	0	0	0	0	0	0.2	0	0.2	0	0
Acacia pycnantha	0	0	0	0	0.2	0.2	0	0	0.2	0	0.2
Acacia ulicifolia	0	0	0	0	0	0	0	0	0	0	0
Acacia verniciflua	0	0	1	1	0	0	0	0	0	0	0
Agrostis sp	0	0	0	0	0	0	0.2	0	0	0	0
Aira sp	0	0	1	0	0	0	0	0	0.2	0	0
Amphipogon strictus	0	0	0	0	0	0	0.2	0	0	0	0
Anthropodium stricta	0	0	0	0	0	0.2	0	0	0	0	0
Asteraceae sp	0	0	0	0	0.2	2	0	0	0	0	0
Astroloma conostephoides	0.2	0	1	0	0	0	2	0.2	0	0.2	0.2
Astroloma humifusum	0	0	0	0	0	0	0	0	1	0	0
Astroloma pinifolium	0	1	0	0	0	0	0.2	0	0.2	0	0.2
Austrodanthonia sp	0	0	0	0	0	0	0	0	0	0	0
Baeckea crassifolia	0	0	0	0	0	0	0	0.2	0	0.2	0
Banksia marginata	0	0	0	0	0	0	0.2	0	0	0.2	0

The nature of multivariate data

Multivariate data have some characteristics that pose some challenges:

- sparse: the majority of entries consists of zeros
- many factors influence sample (local species) composition
- but, the number of important factors is few
- a small number of factors (can) explain the majority of the variation

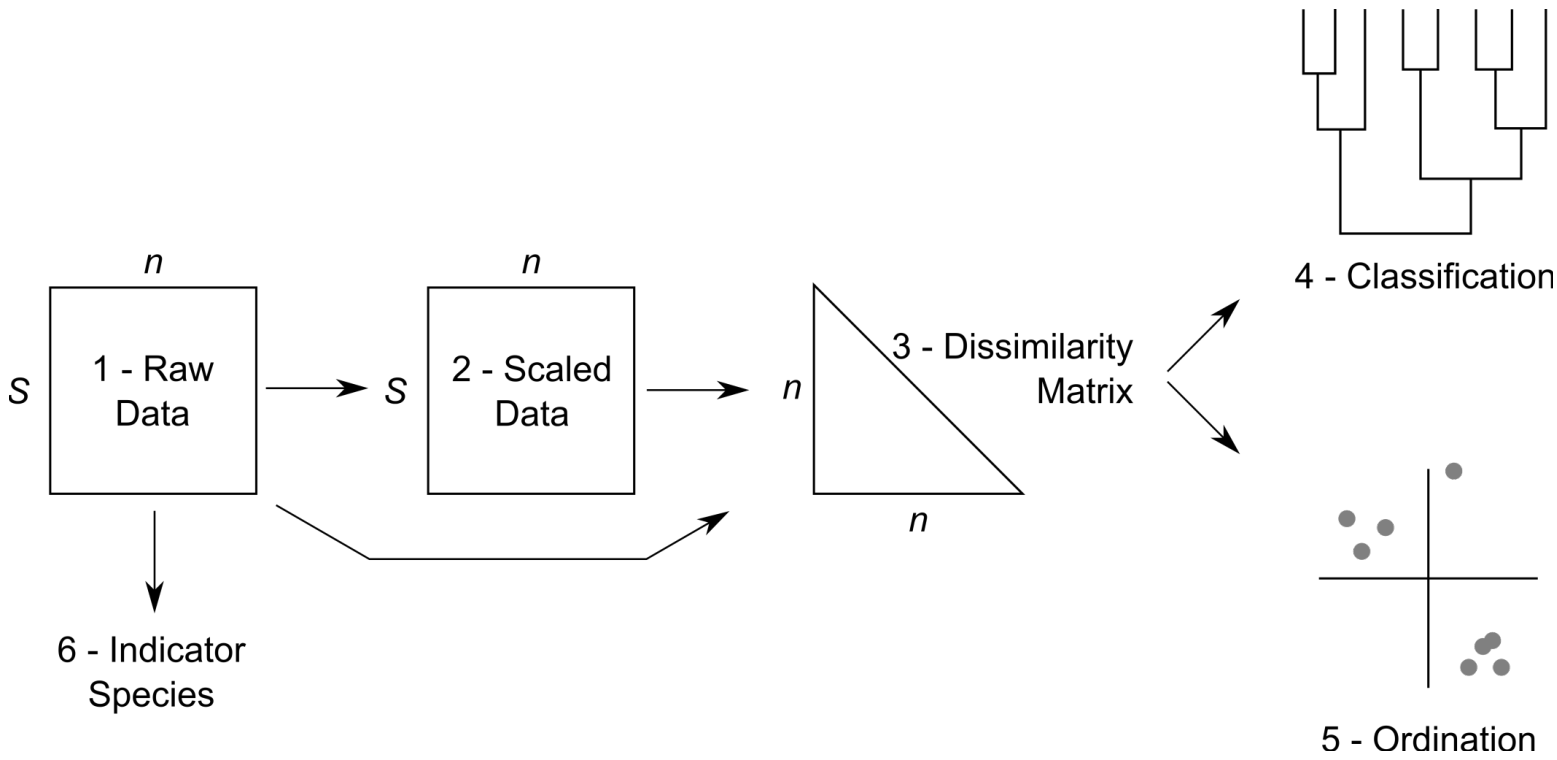
The nature of multivariate data

Multivariate data have some characteristics that pose some challenges:

- noise: even under ideal conditions, replicate samples vary substantially from each other: natural variation, human error, etc.
- redundant information: species share distributions

But it's this redundancy that allows us to make sense of multivariate data

The analysis workflow



After: Field et al. (1982)

Standardisation and transformation

Standardisation weights data so that differences are relative not absolute

- crucial if data on different scales are used
- are differences in total abundance between samples of biological relevance?

Transformation changes the entire matrix (often post-relativisation)

- should the analysis reflect common or rare taxa?

Standardisation and transformation

```r
# Set the data up and select variables of interest
  physical <- read.csv('./data/physicochemical.csv', header = T, row.names = 1)
  physical.lee17 <- select(physical, SRP, Ammonia, Velocity, ProportionMacrophyte,
                           Temperature, Conductivity, Pasture, Scrub, Forest,
                           BankGradient, Elevation, DistanceToSea, CatchmentSize)

  inverts <- read.csv(file = './data/inverts.csv', header = T, row.names = 1)

# Scale using either base::scale or vegan::decostand
  physical.sc <- scale(physical.lee17)
  physical.sc <- decostand(physical.lee17, method = "standardize")  # Vegan
```

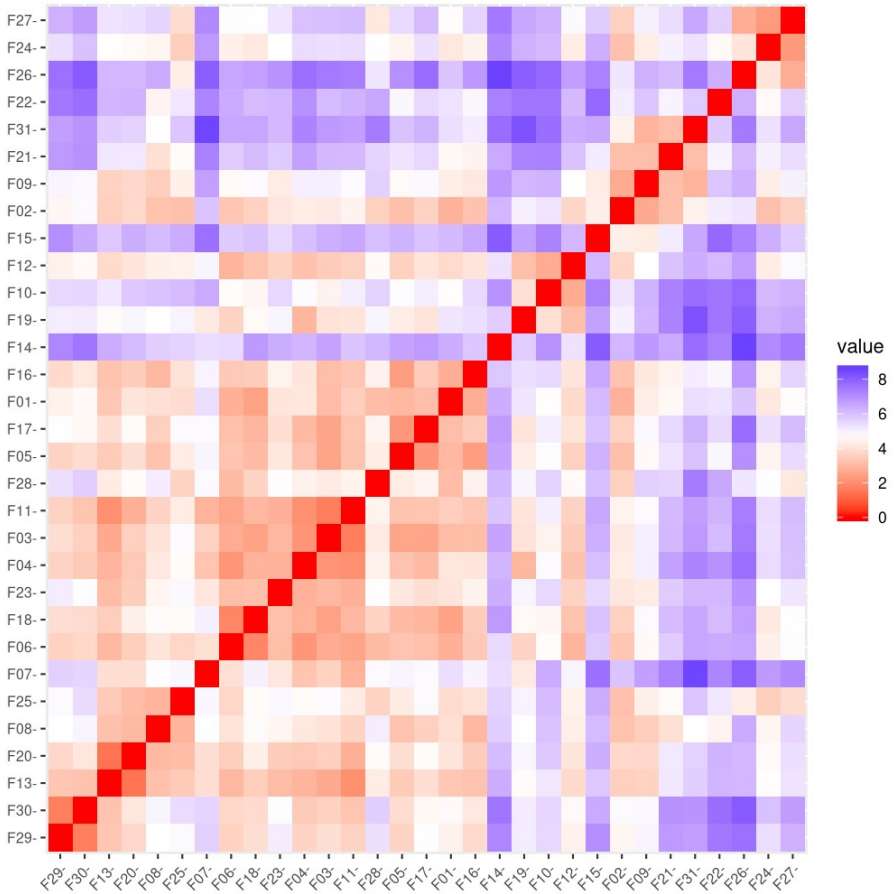
Calculating 'distance'

- Defining the relationship between each pair of samples (or species) is the starting point for distance-based multivariate methods
  - there are many ways to compute 'distance' depending on the nature of the data

- The end-product will be a triangular distance or (dis)similarity matrix

```
library(vegan)
physical.dist <- dist(physical.sc)   # In base R - default distance is Euclidean
inverts.bc <- vegdist(inverts, method = "bray")  # In vegan
```

```r
factoextra::fviz_dist(physical.dist, lab_size = 8)
```



Clustering and classification

"The art of putting things in groups"

Classification

- Classification hierarchically groups samples of increasing similarity to allow:
  1. all samples to be compared simultaneously
  2. explanatory factors to be compared between units

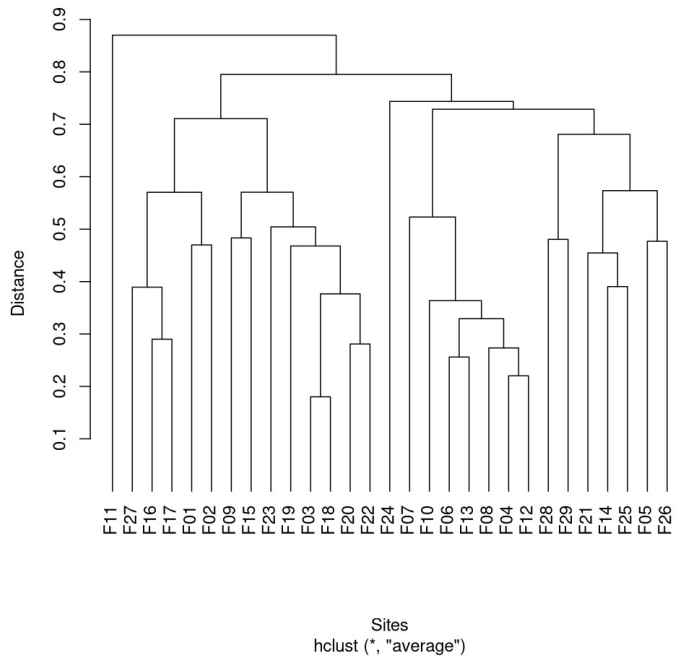
- Agglomerative clustering is a frequently used method of classification
  - group the individual objects until they form a single large group
  - need to decide how to do this grouping (the linkage method)

Hierarchical agglomerative clustering

1. Build matrix of (dis)similarities between objects $[D_{ij}]$
2. First cluster is formed between the objects that are most similar
3. Similarities between this cluster and all other objects are then recalculated
4. Second cluster is formed between cluster 1 and the object most similar to cluster 1
5. Repeat until **all** objects are linked in clusters

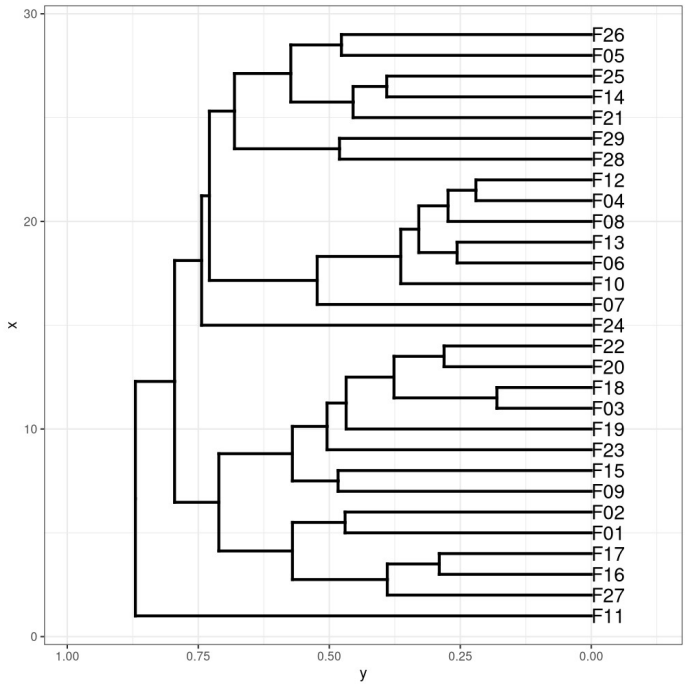
Hierarchical agglomerative clustering: `hclust()`

```
invert.hc <- hclust(inverts.bc, method = "average")  # stats library
plot(invert.hc, hang = -1, main = "", xlab = 'Sites', ylab = 'Distance')
```



Improve visualisation with dendextend

```
library(dendextend)
invert.hc.gg <- as.ggdend(as.dendrogram(invert.hc))
ggplot(invert.hc.gg, horiz = TRUE) + theme_bw()
## Warning: Removed 57 rows containing missing values (geom_point).
```



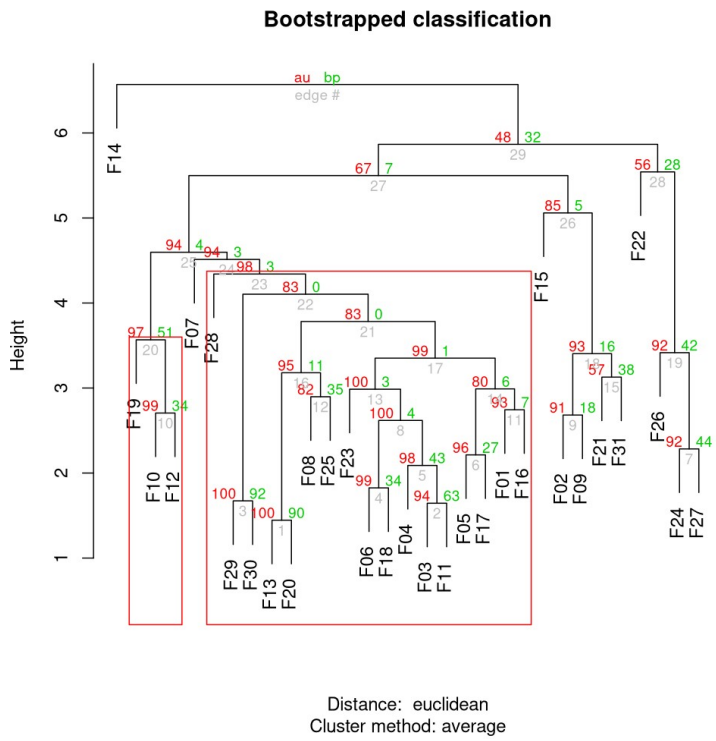
Bootstrapping to assess the classification

- `pvclust` lets you assess the support for each cluster using a bootstrapping approach:

```
# this analysis is very computationally expensive (control with nboot)
physical.pc <- pvclust(t(physical.sc), method.dist="euclidean",
                       method.hclust="average", nboot=1000, quiet = TRUE)  # <<
```

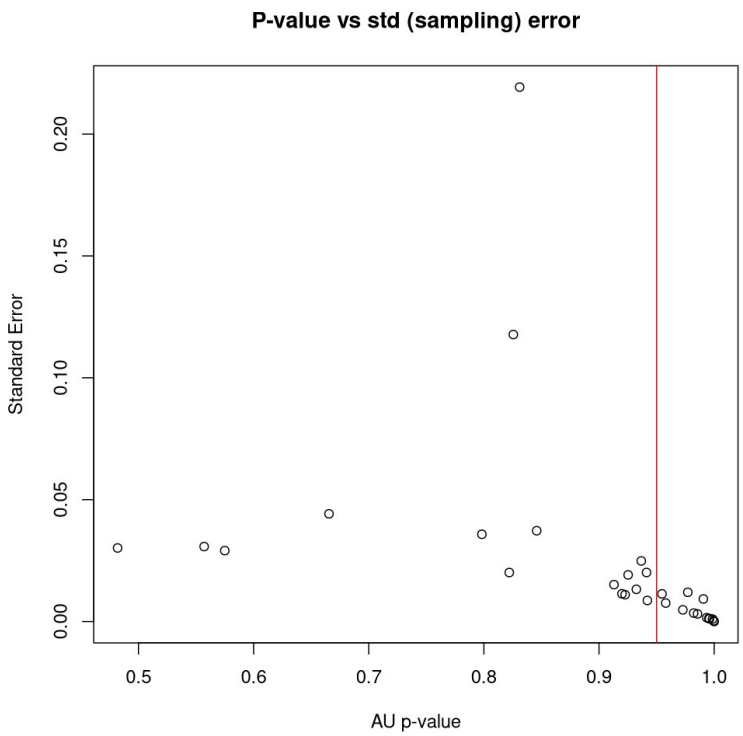
Bootstrapping to assess a classification

```
plot(physical.pc, main = "Bootstrapped classification")
pvrect(physical.pc, alpha = 0.95)
```



Bootstrapping to assess a classification

```
seplot(physical.pc, main = 'P-value vs std (sampling) error')
abline(v = 0.95, col = 'red')
```

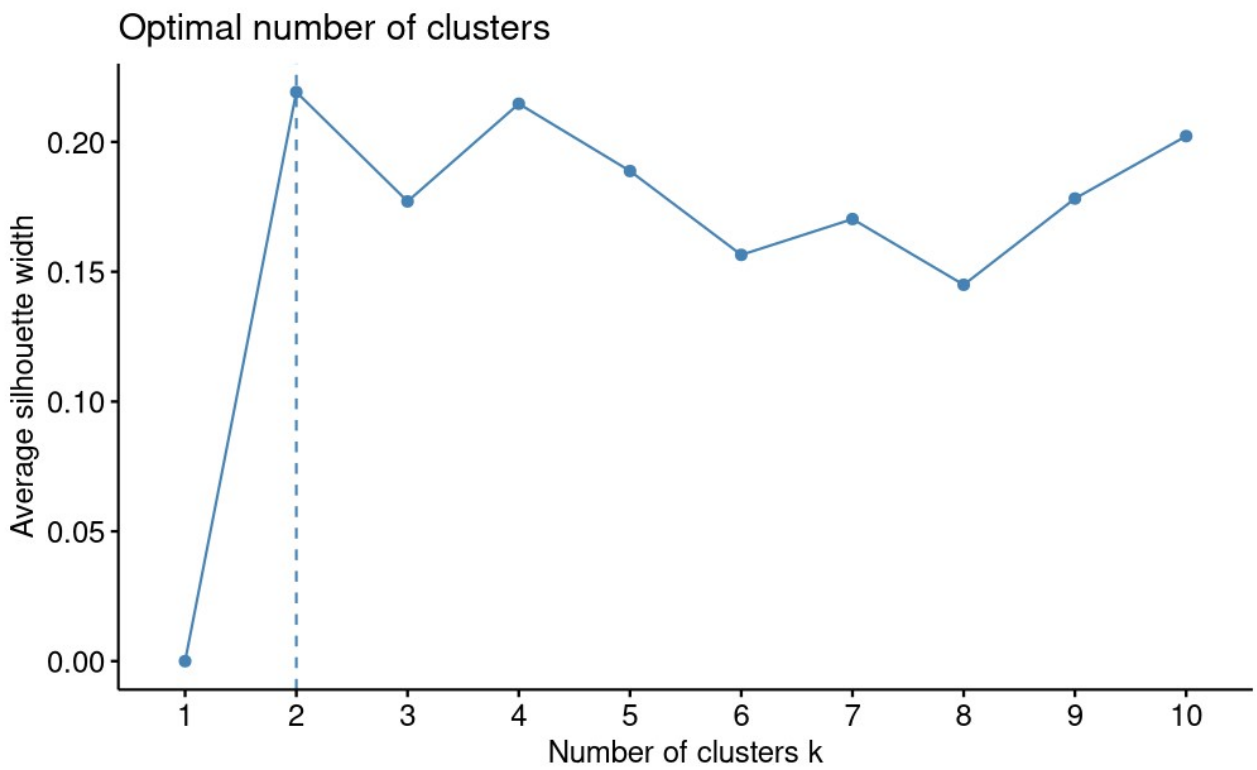


Partitioning: *k*-means

- *k*-means clustering is a popular partitioning method: seeks to minimise the pairwise squared distance within clusters
- You specify the number of clusters to extract, but there are graphical and other approaches to help find the 'best' *k*

Finding the optimal *k*

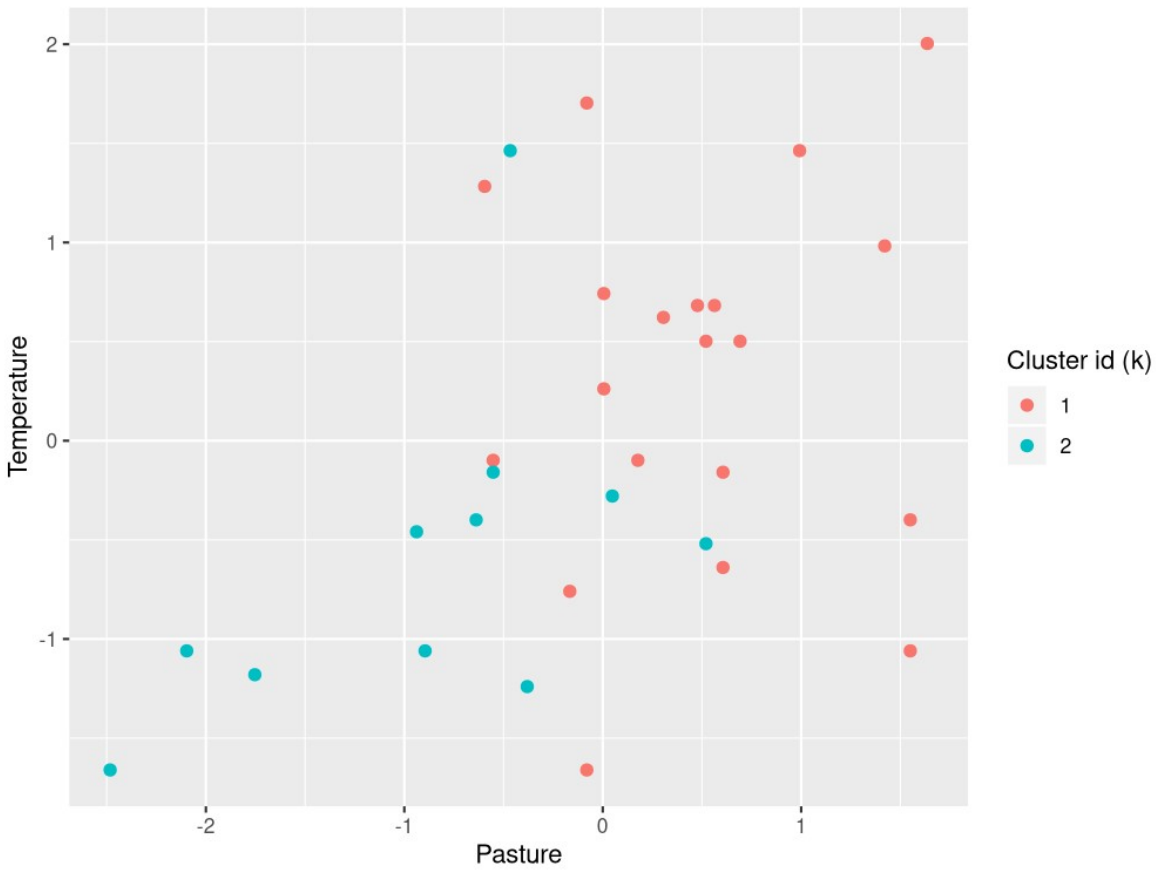
```
# Various options also available in NbClust package
fviz_nbclust(physical.sc, kmeans, method='silhouette')
```



k-means on the Lee et al. physical data

```
# Do the kmeans analysis
site.km <- kmeans(x = physical.sc, centers = 2)           # centres = groups (K)
physical.sc.df <- data.frame(physical.sc)
physical.sc.df$km.cluster <- as.factor(site.km$cluster)   # add cluster id as a colu

# Plot it via ggplot...
site.km.plot <- ggplot(data = physical.sc.df) +
  geom_point(aes(x = Pasture, y = Temperature, col = km.cluster), size = 2) +
  labs(colour = 'Cluster id (k)') +
  coord_equal()
```

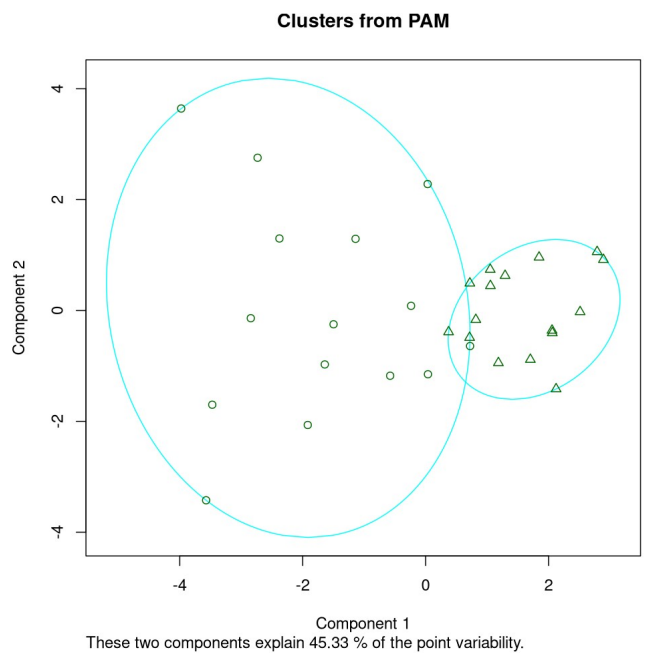

Temperature
2
1
0
-1
-2
-1
0
1
Pasture
Cluster id (k)
1
2

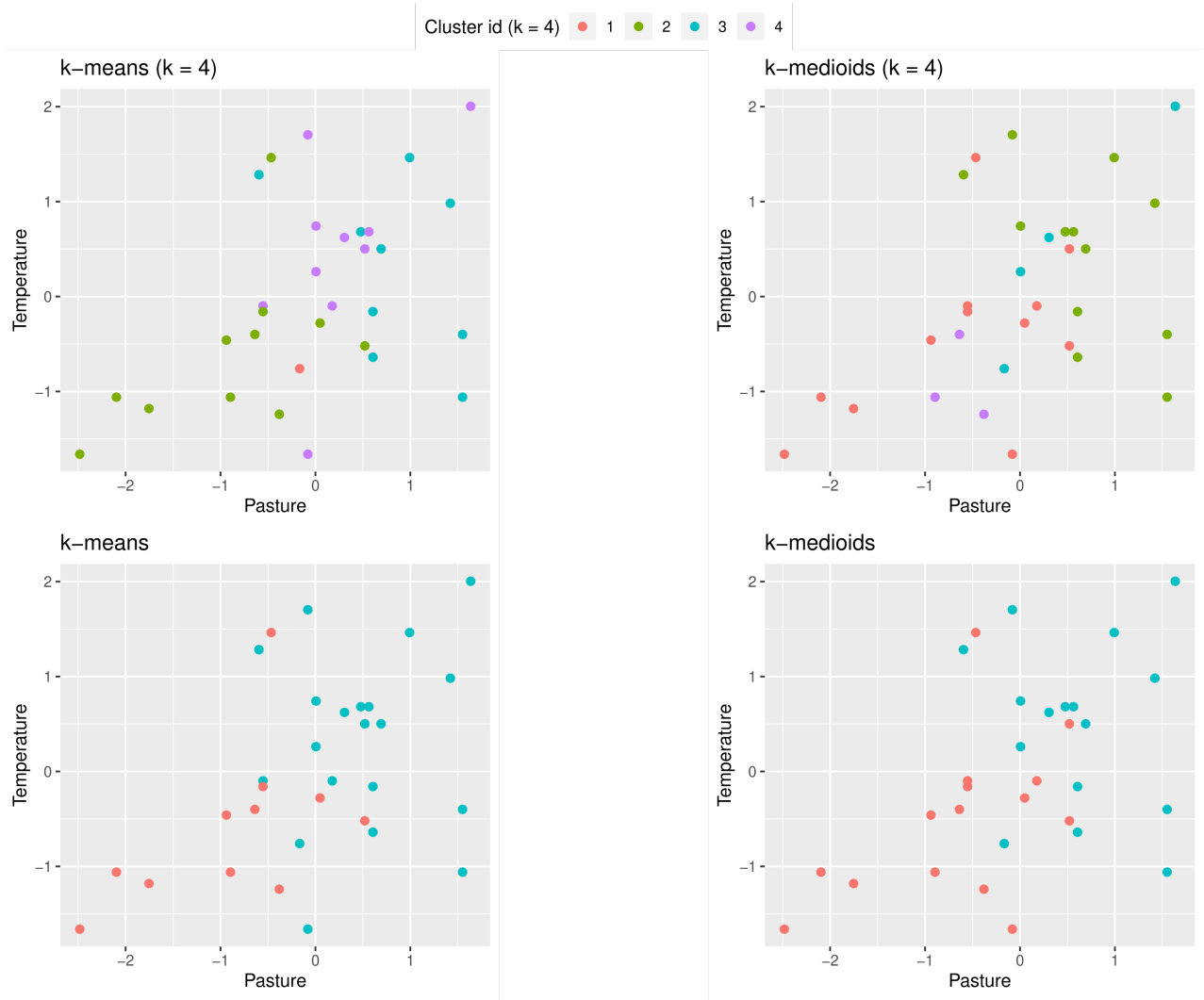
A more robust option: *k*-medioids (`pam()`)

- A robust version of *k*-means based on mediods can be calculated via `cluster::pam` instead of `kmeans`
  - minimise the summed distance between points in cluster and a point designated as the center of that cluster

A more robust option: *k*-medioids (pam( ))

```
library(cluster)
site.pam <- pam(x = physical.sc, k = 2)
clusplot(site.pam, main = 'Clusters from PAM')
```




Cluster id (k = 4)
1
2
3
4
k−means (k = 4)
k−medioids (k = 4)
k−means
k−medioids
Temperature
Pasture
2
1
0
−1
−2

Other more computationally intensive approaches

- Affinity propagation clustering: `apcluster`
- Density-based spatial clustering of applications with noise: `dbscan`
- Model-based clustering: `mclust`
- Dynamic time-warping: `dtw`

- ...

- See the CRAN Task View: Cluster Analysis & Finite Mixture Models

Ordination

"Ordination primarily endeavors to represent sample and species relationships as faithfully as possible in a low-dimensional space" - Gauch (1982)

Why ordination?

- Data and pattern simplification, outlier detection
- Variable decomposition or data reduction: need a latent variable for path analysis or regression
- Variable selection for experimental studies
- Interpretation and understanding
- Generating testable hypotheses

Ordination: Constrained or Unconstrained

- Unconstrained ordinations seek to maximise explained variation in community data: PCA, nMDS, etc.
- Constrained ordinations attempt to maximise explained variation according to constraints, with axes synthetic combinations of predictors: RDA, CCA, etc.

Unconstrained - PCA

- Principal components analysis (PCA) reorganises variables into a new set of components ('axes') equal to the number of original variables
  - typically environmental data for variable reduction

- The components:
  - are independent (orthogonal)
  - decrease in the amount of variance from the originals
  - only a subset retained some for further study (dimensional reduction)
  - suited to environmental data, **not** community data

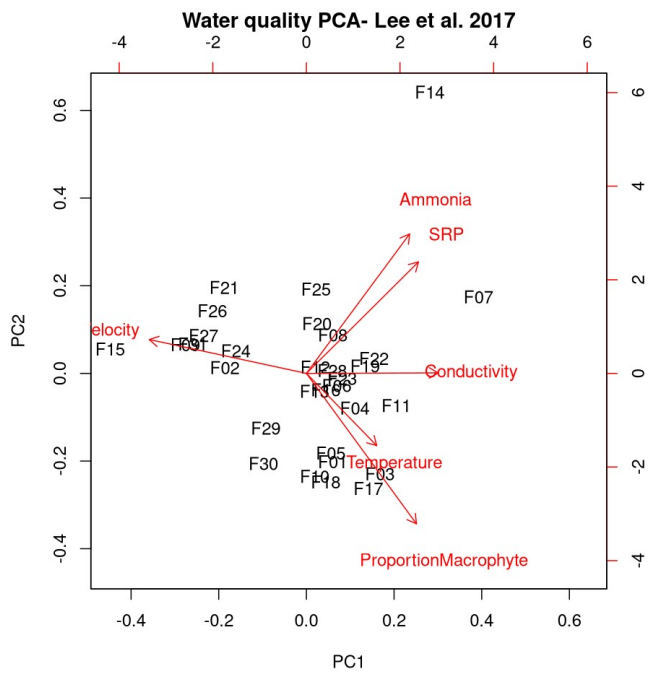
Unconstrained - PCA

- Reproduce Fig. 4 from Lee et al. 2017 (PCA of site conditions)

```
library(tidyverse)
f4a.dat <- select(physical.lee17, SRP, Ammonia, Velocity, ProportionMacrophyte,
                  Temperature, Conductivity)
f4b.dat <- select(physical.lee17, Pasture, Scrub, Forest, BankGradient,
                  Elevation, DistanceToSea, CatchmentSize)
```

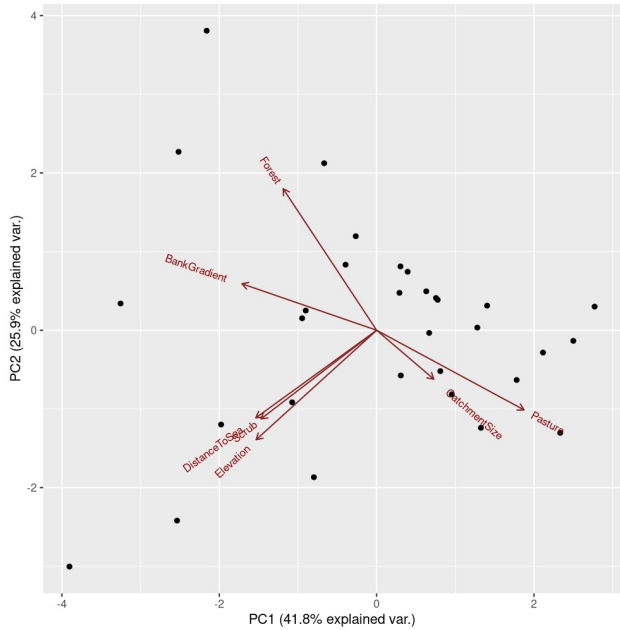
Unconstrained - PCA

```
# prcomp preferred to princomp (?princomp)
chem.pca <- prcomp(f4a.dat, scale = TRUE)  # scale if data on different scales
lu.pca <- prcomp(f4b.dat, scale = TRUE)
biplot(chem.pca, main = "Water quality PCA- Lee et al. 2017")
```



Unconstrained - PCA

```
library(ggbiplot)
ggbiplot(lu.pca, obs.scale = 1, var.scale = 1) +
  scale_color_discrete(name = '') +
  theme(legend.direction = 'horizontal', legend.position = 'top')
```



Loadings and scores

The loadings are effectively the weights for each original variable when calculating the principal component

```
summary(chem.pca)
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6
## Standard deviation     1.3763 1.1834 1.0720 0.8254 0.74806 0.56140
## Proportion of Variance 0.3157 0.2334 0.1915 0.1136 0.09327 0.05253
## Cumulative Proportion  0.3157 0.5491 0.7407 0.8542 0.94747 1.00000
```

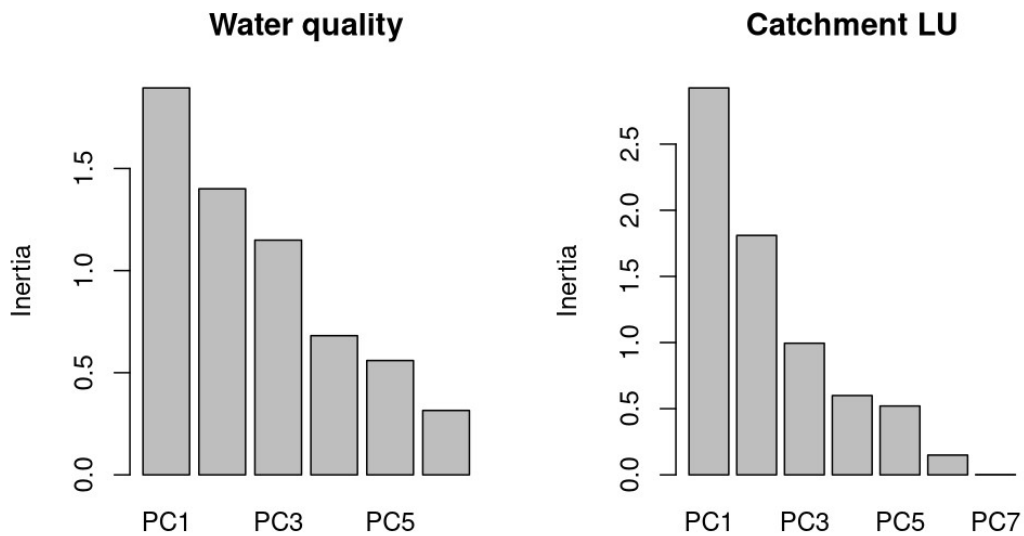
Loadings and scores

The scores are the original data in a rotated coordinate system

```
scores(chem.pca)   # vegan::scores (for all ordinations)
##            PC1         PC2         PC3          PC4         PC5
## F01  0.4604310 -1.32910787 -0.24559273 -0.911614079 -0.56071508
## F02 -1.4277889  0.08695647 -0.18598098  0.362080136 -0.32307972
## F03  1.2822211 -1.51089841  0.32006116  0.052959495  0.70032802
## F04  0.8433508 -0.52033170  1.02683403  0.289654174  0.10429497
## F05  0.4259347 -1.19787574 -1.34462942  1.128862943  0.05260333
## F06  0.5285894 -0.18345649  0.69867795 -0.750724860 -0.96435301
## F07  3.0063146  1.15029128  1.16007381  0.157237565  1.22796611
## F08  0.4609773  0.57812925 -1.01600753  0.942454484  0.37092928
## F09 -2.1183892  0.43001925  0.24800138  0.006815663  0.24875303
## F10  0.1436229 -1.54731091  0.36258845 -0.754223048 -0.69405069
## F11  1.5709335 -0.48418247  0.54005885 -0.309937436  0.46948532
```

Screeplots for factor importance

```
par(mfrow = c(1,2))
screeplot(chem.pca, main = "Water quality")
screeplot(lu.pca, main = "Catchment LU")
```



Unconstrained - PCA summary

- R packages::commands are:
  - `base::prcomp`, `base::princomp`, `base::biplot`
  - `vegan::bstick`, `vegan::screeplot`
  - various other packages implement various factor analyses

- Some recent extensions: sparse PCA (`elasticnet`), non-linear PCA (`pcaMethods`), outlier-resistant (weighted) PCA (`rospca`)

Unconstrained - non-metric multidimensional scaling

- nMDS is a widely used method of unconstrained ordination (by ecologists, at least)
  - Bray-Curtis distance measure is widely recommended where zeros are common and we want to extract underlying gradients
  - robust unconstrained ordination method for ecological species data (Minchin 1987)
  - attempts to best represent the position of samples (communities) in multidimensional space
  - computationally intensive randomisation

Unconstrained - nMDS

1. Maximises correlation between distance measure and distance in ordination space

2. Points are moved to minimise stress (mismatch between the two distances)

3. Assume that (dis)similarity is monotonically related to ecological distance

4. Stress: how well do distances in ordination space reflect ecological dissimilarities

    - ideally, stress < 0.1 (but this is rare)
    - if stress > 0.3 then the plot is non-interpretable
    - reduce stress by using a higher-dimension solution

Unconstrained - nMDS

- Basic vegan command is `metaMDS` - **many** options!
  - other R options: `cmdscale`, `MASS:isoMDS`, ...

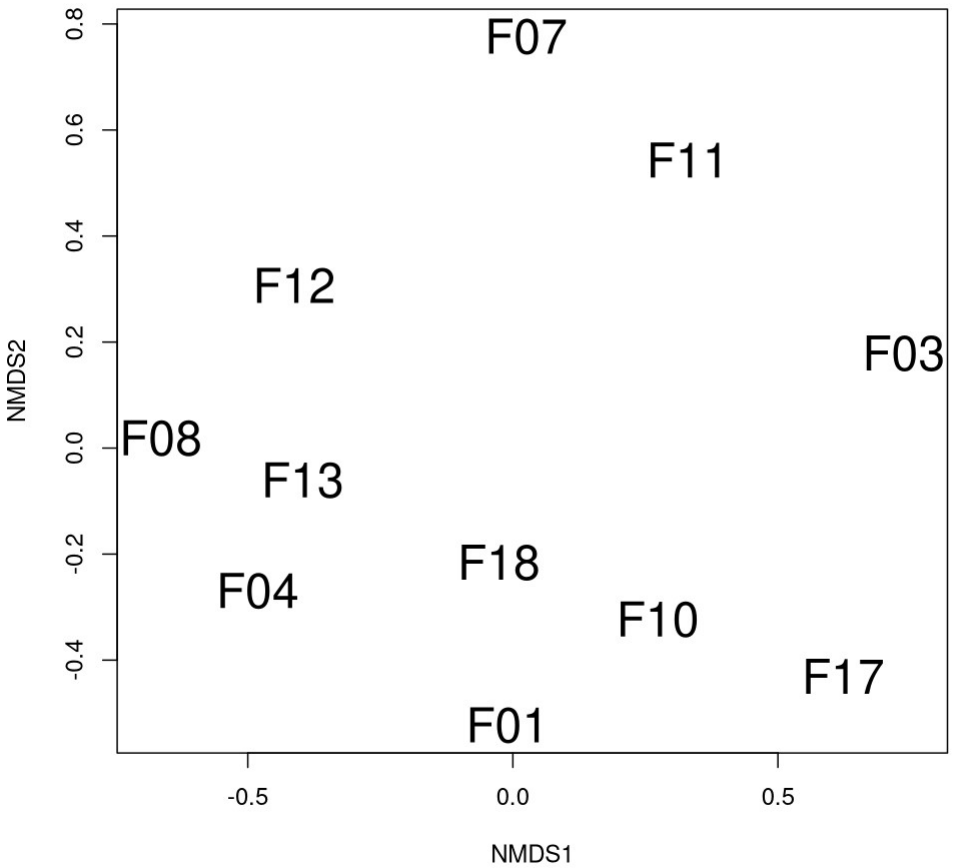
```
# Set up data as per Lee et al. 2018
sites18 <- c('F01', 'F03', 'F04', 'F07', 'F08', 'F10', 'F11', 'F12', 'F13', 'F17',
inverts.lee18 <- inverts[rownames(inverts) %in% sites18, ]
physical.lee18 <- physical.lee17[rownames(physical.lee17) %in% sites18, ]
```

Unconstrained - nMDS

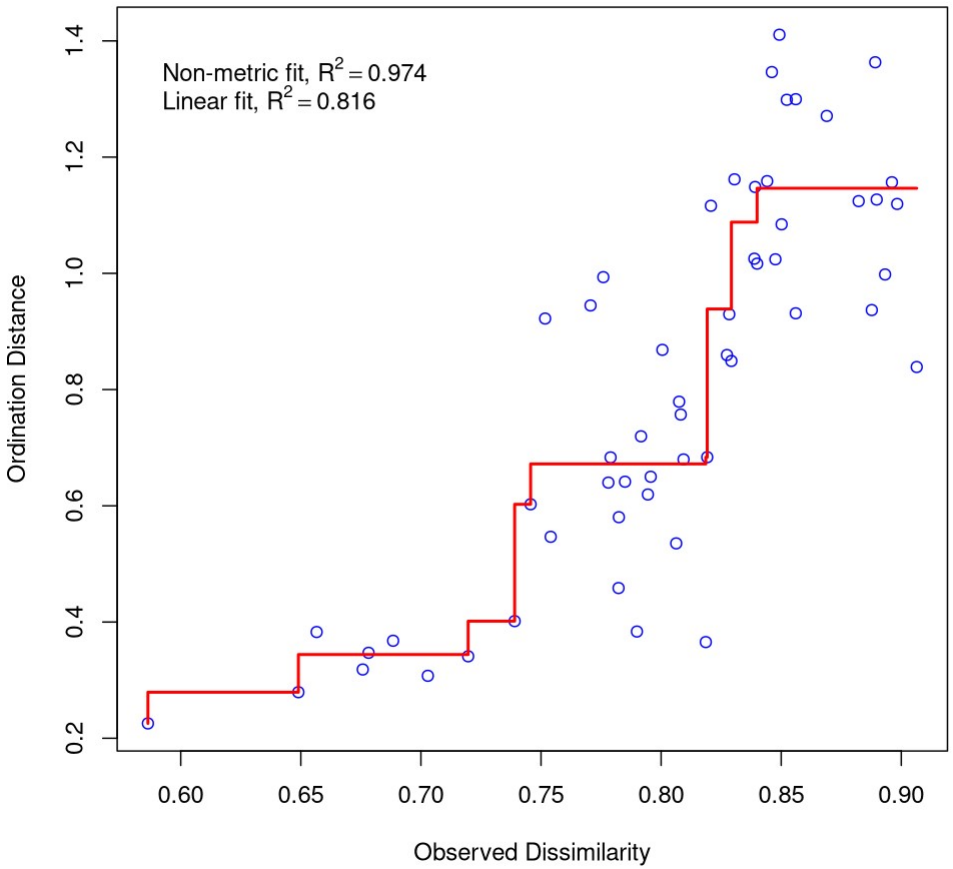
Now conduct the nMDS:

```
# Here we use Jaccard's distance to follow Lee; default is BC
invert.mds <- metaMDS(inverts.lee18, distance = "jaccard",
                      autotransform = TRUE, wascores = TRUE)
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.1605532
## Run 1 stress 0.1820904
## Run 2 stress 0.2071441
## Run 3 stress 0.1662082
## Run 4 stress 0.2152496
## Run 5 stress 0.2234302
## Run 6 stress 0.1662071
## Run 7 stress 0.2506111
```

```r
plot(invert.mds, type = 't', display = 'sites', cex = 2)
```

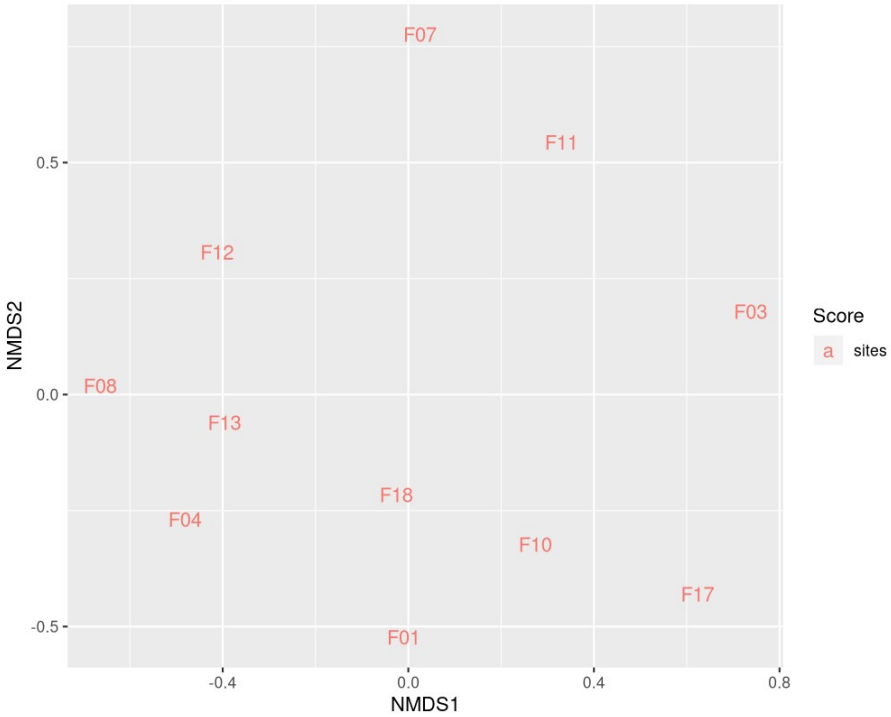


```
invert.mds$stress
## [1] 0.1605531
stressplot(invert.mds)
```



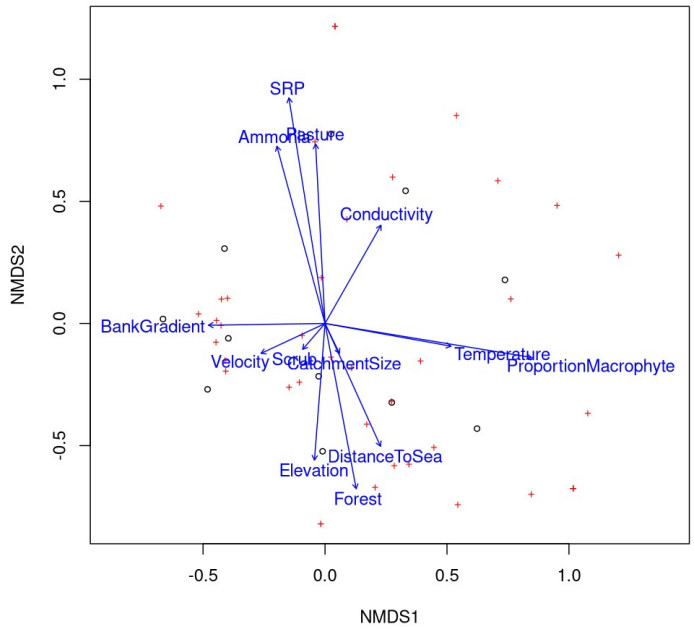
Use the ggvegan library

```
autoplot(invert.mds, geom = "text", layers = 'sites', size = 2)
```



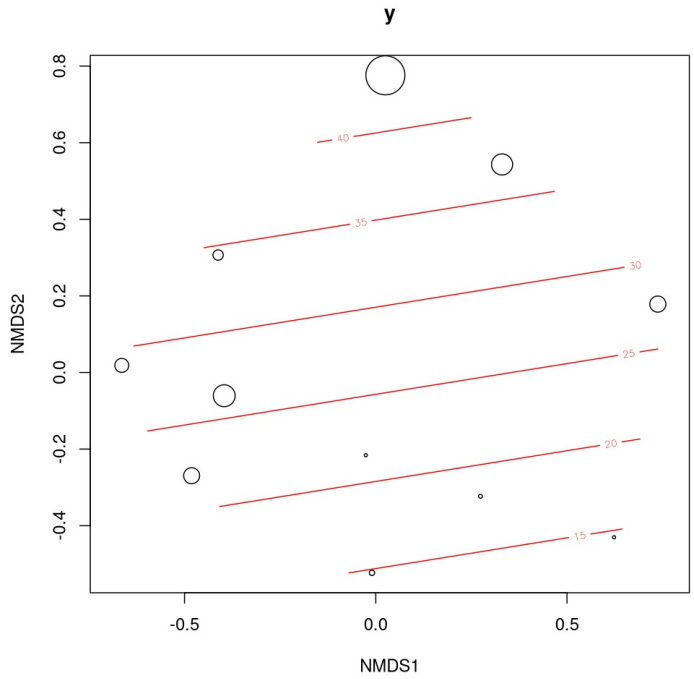
Vector-fitting

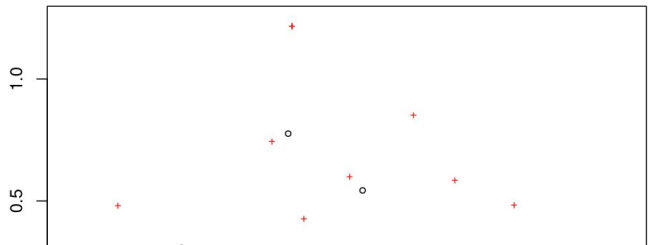
```
phys.fit <- envfit(invert.mds, physical.lee18, perm = 9999, p.max = 0.10)
plot(invert.mds)
plot(phys.fit)
```



Surface-fitting (ordisurf)

```
phys.surf <- ordisurf(invert.mds ~ SRP, physical.lee18, bubble = 5)
plot(invert.mds, main = '')
plot(phys.surf)
```





Post-hoc assessment of dissimilarity matrices

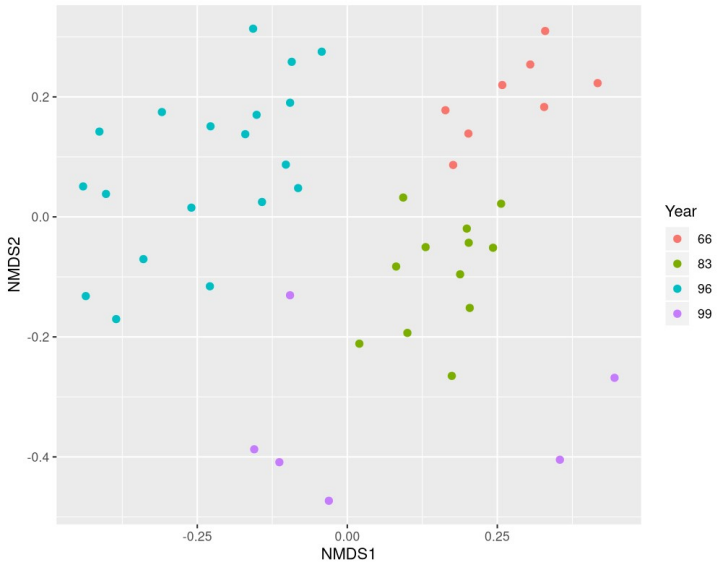
- Load Grampians vegetation data set:

```
grampians <- read.csv(file = './data/gramps99.csv', header = T, row.names = 1)
gr.dist <- vegdist(grampians)
yr <- as.factor(substr(rownames(grampians),1,2))  # fire year

# Do the nMDS analysis
gramps.mds <- metaMDS(gr.dist, wascores = FALSE, autotransform = FALSE, trace = 0)
# trace = 0 supresses messages about stress
gramps.mds$stress
## [1] 0.2009278
```

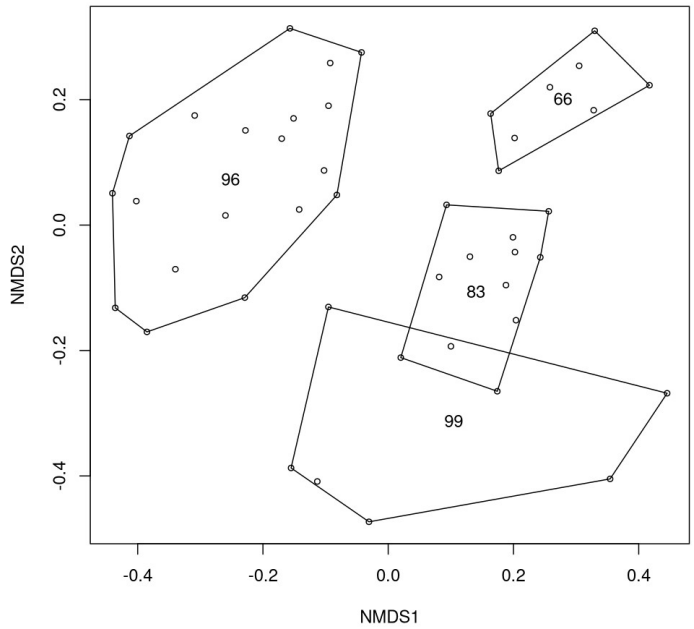
Post-hoc assessment of dissimilarity matrices

```
grp.mds <- data.frame(scores(gramps.mds), yr  = yr)
ggplot(data = (grp.mds)) +
  geom_point(aes(x = NMDS1, y = NMDS2, col = factor(yr)), size = 2) +
  labs(colour='Year') +
  coord_equal()
```



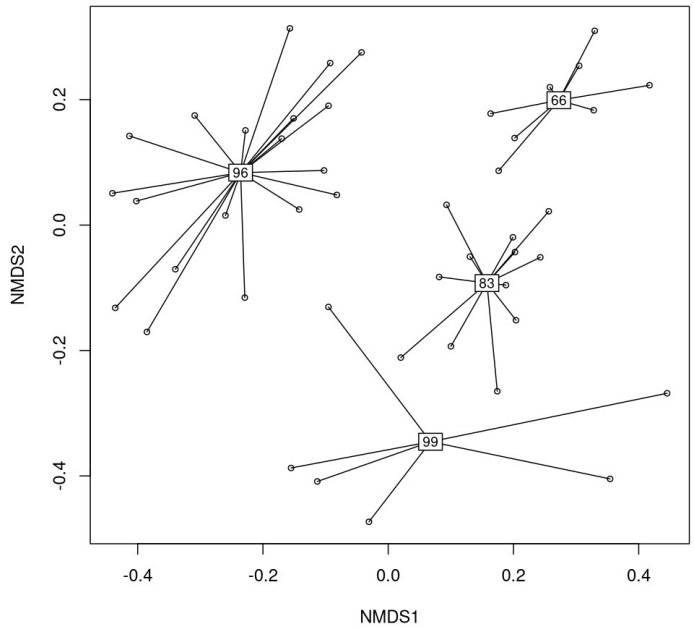
Post-hoc visualisation of dissimilarity matrices

```
plot(gramps.mds)
## species scores not available
grp.hull <- ordihull(ord = gramps.mds, groups = yr, label = TRUE)
```



Post-hoc visualisation of dissimilarity matrices

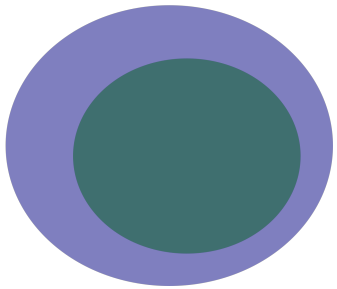
```
plot(gramps.mds)
## species scores not available
grp.spider <- ordispider(ord = gramps.mds, groups = yr, label = TRUE)
```



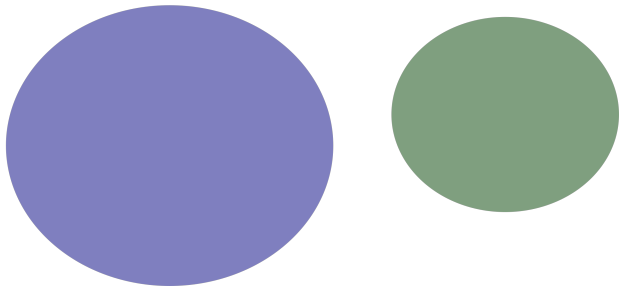
Post-hoc tests on dissimilarity matrices - ANOSIM

- Analysis of similarities (`vegan::anosim`) is based on the difference of mean ranks between vs. within groups
  - tests between multiple groups using any type of dissimilarity
  - conceptually allied with nMDS

ANOSIM likely **unsignificant**



ANOSIM likely **significant**



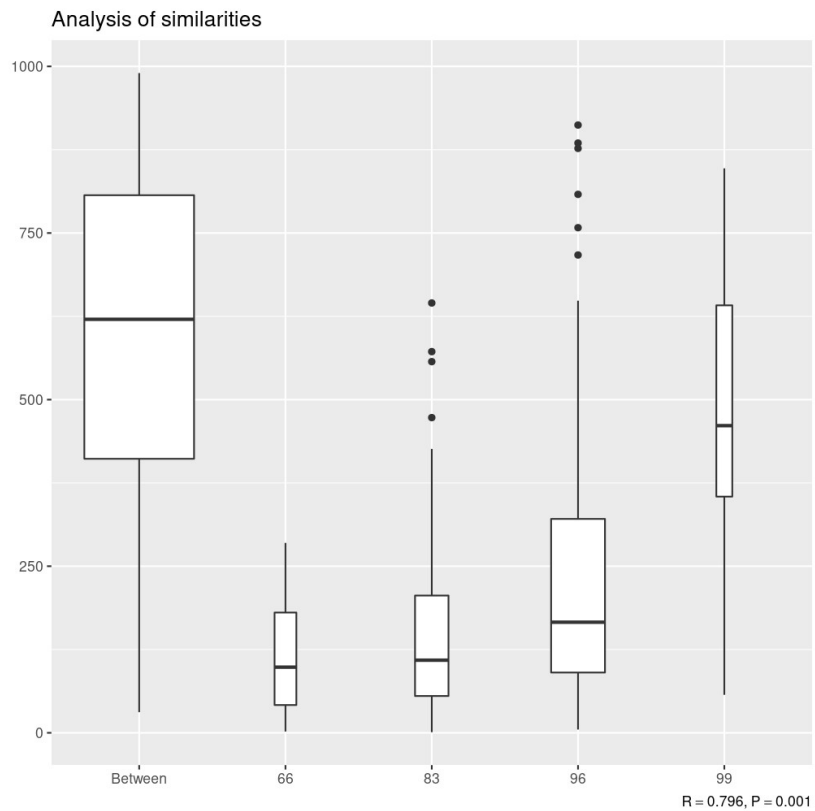
- Assumes ranked dissimilarities within groups have roughly equal median and range

Post-hoc tests on dissimilarity matrices - ANOSIM

```
gramps.ano <- anosim(x = grampians, grouping = yr)
gramps.ano
##
## Call:
## anosim(x = grampians, grouping = yr)
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.7957
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
```

Post-hoc tests on dissimilarity matrices - ANOSIM

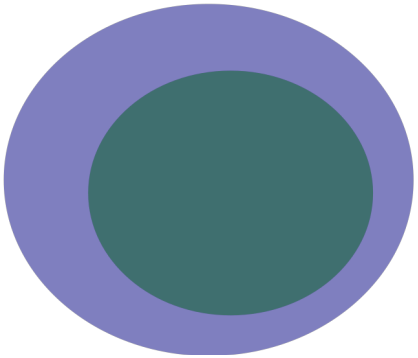
```
autoplot(gramps.ano, notch = FALSE)   # or just base plot...
```



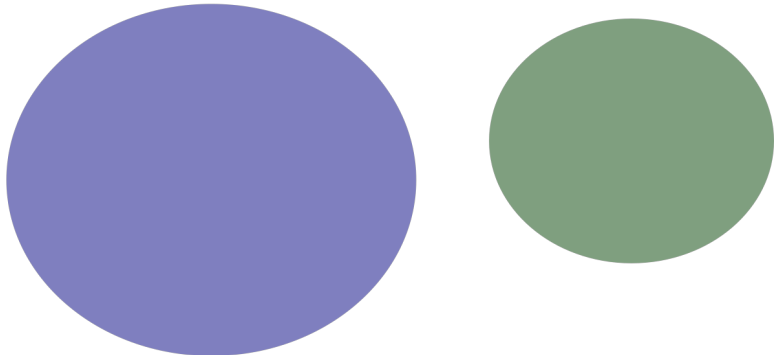
Post-hoc tests on dissimilarity matrices - PERMANOVA

- Adonis (`vegan::adonis2`) is permutational multivariate analysis of variance (PERMANOVA) using distance matrices
- Considered the best test (to date) for examining group means and variance (i.e., community turn-over ß diversity); see Anderson & Walsh (2013)

ADONIS likely **unsignificant**



ADONIS likely **significant**



Post-hoc tests on dissimilarity matrices - PERMANOVA

Read the 'by' argument carefully – the test is sequential (can also specify as marginal)

```
gramps.ad <- adonis2(gr.dist ~ yr)
gramps.ad
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gr.dist ~ yr)
##          Df SumOfSqs      R2      F Pr(>F)
## yr        3   4.0201 0.36447 7.8378  0.001 ***
## Residual 41   7.0098 0.63553
## Total    44  11.0300 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

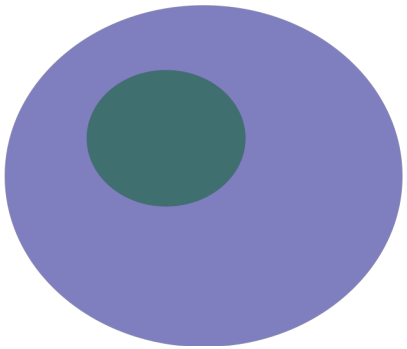
Post-hoc tests on dissimilarity matrices - PERMANOVA

- Also ... the method may confound location and dispersion effects
  - significant differences may arise from different within-group variation (dispersion) instead of different mean values of the groups; see Warton et al. 2012

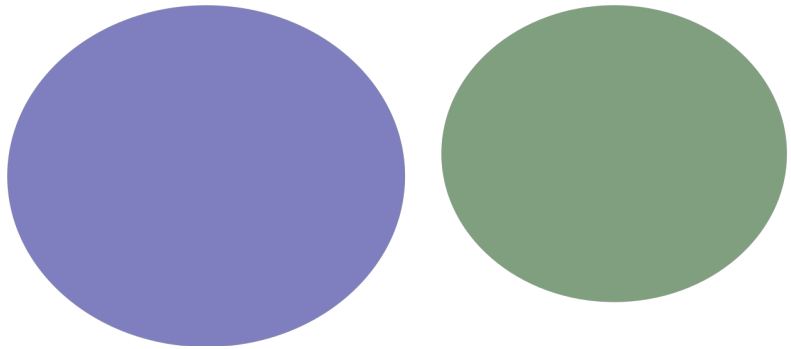
Post-hoc tests on dissimilarity matrices - MHV

- Multivariate homogeneity of variance complements PERMANOVA (`vegan::betadisper`)
- Test developed to assist in determining patterns of community homogenisation

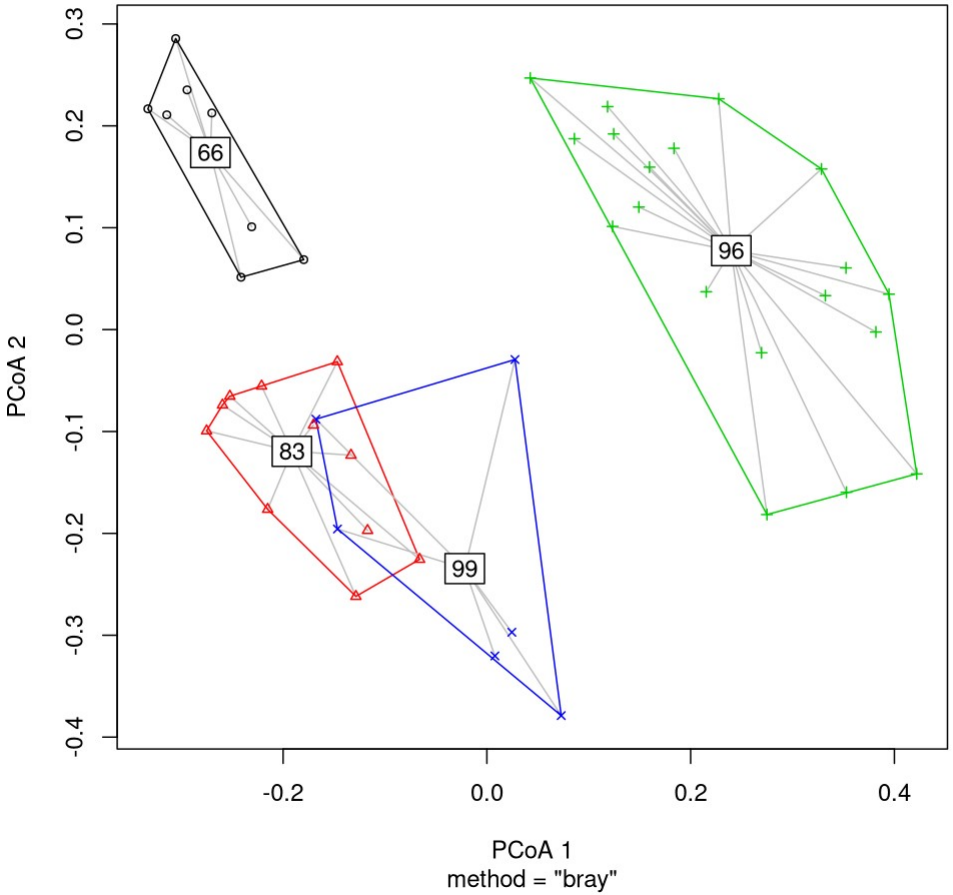
BETADISPER likely **significant**



BETADISPER likely **unsignificant**



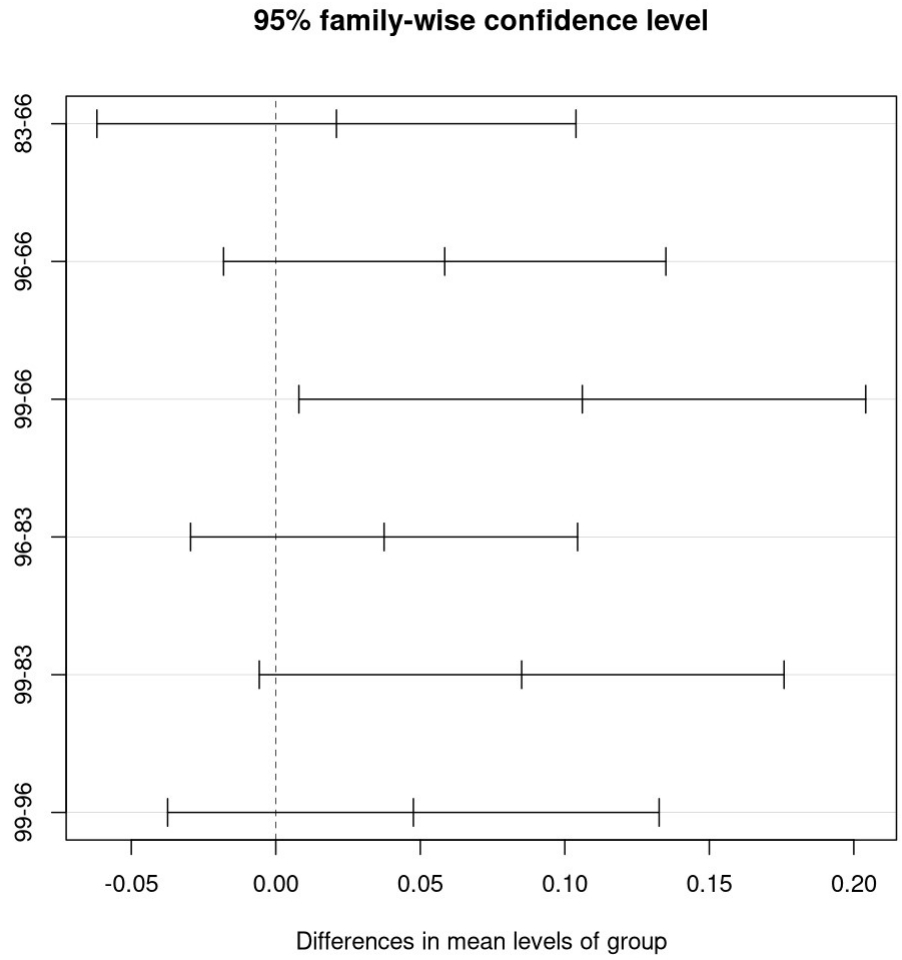
```r
gramps.mod <- betadisper(gr.dist, yr)
plot(gramps.mod, main = "")
```




```r
anova(gramps.mod)
## Analysis of Variance Table
##
## Response: Distances
##           Df   Sum Sq   Mean Sq F value  Pr(>F)
```

```r
#Tukeys HSD test on the permuted model
mod.HSD <- TukeyHSD(gramps.mod)
mod.HSD
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = distances ~ group, data = df)
##
## $group
##             diff          lwr       upr     p adj
## 83-66 0.02098385 -0.061897567 0.1038653 0.9048722
## 96-66 0.05845129 -0.018079792 0.1349824 0.1885431
## 99-66 0.10605706  0.007990451 0.2041237 0.0295927
## 96-83 0.03746744 -0.029488817 0.1044237 0.4477522
## 99-83 0.08507322 -0.005718824 0.1758653 0.0733841
## 99-96 0.04760577 -0.037428765 0.1326403 0.4473454
```

```
plot(mod.HSD)
```


95% family-wise confidence level
83-66
96-66
99-66
96-83
99-83
99-96
-0.05
0.00
0.05
0.10
0.15
0.20
Differences in mean levels of group

Post hoc tests summary

- Visualise groups using `vegan::ordihull`, `vegan::ordiellipse`, etc.
- Test for differences between groups in location in ordination space are `vegan::anomsim`, `vegan:mrpp`, `vegan::adonis`
- Test for differences in homogeneity between groups: `vegan::betadisper`
- Ongoing debate re power and balance for some of these tests (see Anderson & Walsh, 2013): variables are correlated, typically with a strong mean-variance relationship

Constrained ordination

- Constrained ordinations seek to maximise explained variation according to constraints
  - canonical correspondence analysis (CCA; `vegan::cca`)
  - redundancy analysis (`vegan::rda` ← a common and accepted choice)
  - distance-based redundancy analysis (`vegan::capscale`), a newer method that allows the use of varying measures of dissimilarity
  - principal coordinates of neighbourhood matrices (PCNM; `vegan::pcnm`) - ordination carried out on coordinates, then constrained, and uses truncated distances. Allows broad to fine scale spatial characterisation

Indicator (species) analysis

- May want to know which objects(species) account for any differences between groups
- `vegan::simper` identifies species that separate groups based on **Bray-Curtis distance**
  - widely used, but potentially confounds mean and variance
- `indicspecies` package includes metrics that identify 'faithful' objects (`indicspecies::multipatt`)

Model-based solutions

- Model-based approaches (GLM for multivariate data): 'what is the effect?' not just 'is there an effect'?

- Tries to address two problems with 'classical' distance-based analyses:
  - power is often low
  - problems with correlated mean-variance relationships

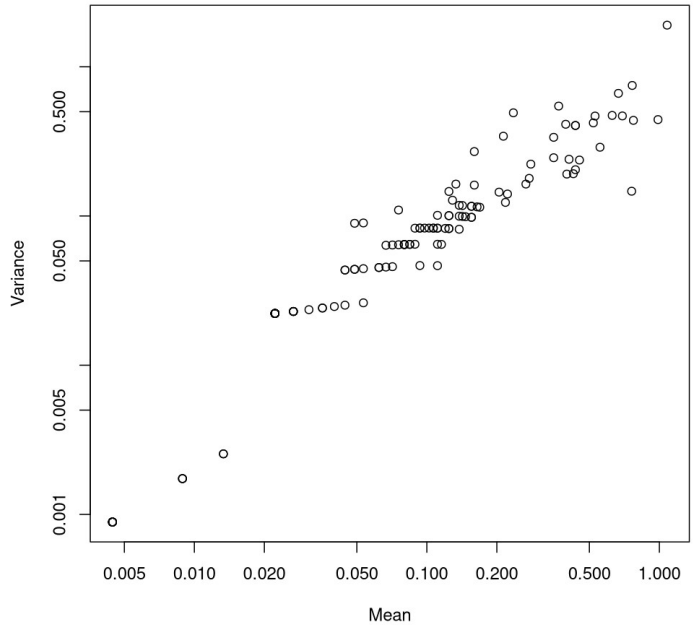
Model-based solutions

- R package mvabund (Wang et al. 2012), with key commands : `manyglm`, `summary`, `anova`

Mvabund

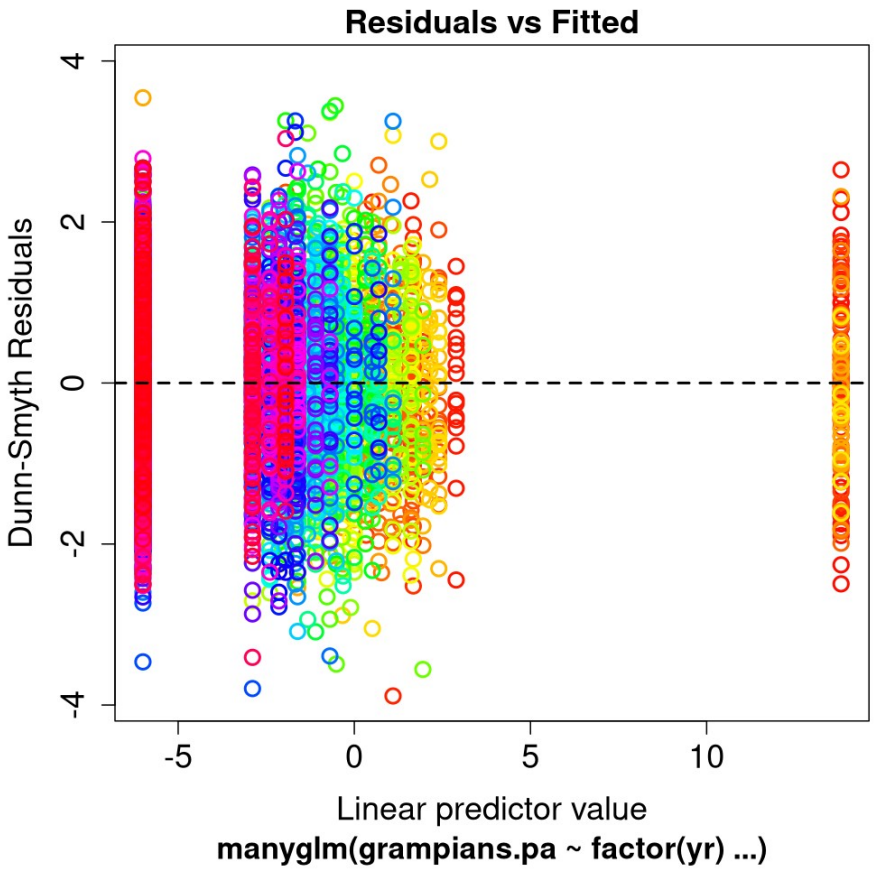
Mean-variance relationships

```r
library(mvabund)
meanvar.plot(grampians, xlab = 'Mean', ylab = 'Variance')
```




```r
# meanvar.plot(mvabund(grampians) ~ yr)
```

```r
# Convert from cover to pres-abs and use a binomial model (Poisson for count)
grampians.pa <- decostand(grampians, method = "pa")
grampians.pa <- mvabund(grampians.pa)
gramps.mod <- manyglm(grampians.pa ~ factor(yr), family="binomial")
plot(gramps.mod)
```



ANOVA to asess the effect of year

```
anova(gramps.mod)
## Time elapsed: 0 hr 0 min 28 sec
## Analysis of Deviance Table
##
## Model: manyglm(formula = grampians.pa ~ factor(yr), family = "binomial")
##
## Multivariate test:
##             Res.Df Df.diff  Dev Pr(>Dev)
## (Intercept)     44
## factor(yr)      41       3 1104    0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Arguments:
##  Test statistics calculated assuming uncorrelated response (for faster computati
##  P-value calculated using 999 resampling iterations via PIT-trap resampling (to
```

ANOVA to asess the effect of year per species

```
anova(gramps.mod, p.uni="adjusted")
## Time elapsed: 0 hr 0 min 32 sec
## Analysis of Deviance Table
##
## Model: manyglm(formula = grampians.pa ~ factor(yr), family = "binomial")
##
## Multivariate test:
##             Res.Df Df.diff  Dev Pr(>Dev)
## (Intercept)     44
## factor(yr)      41       3 1104    0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Other options

- boral: **B**ayesian **O**rdination and **R**egression **A**na**L**ysis (see Hui (2016))
  - effectively a Bayesian extension to mvabund (independent response GLMs, pure latent variable model, correlated response model [explanatory + latent])
  - R package is `boral`

- coral: **C**lustering and **O**rdination **R**egression **A**na***L**ysis (see Hui(2017))
  - simultaneous classification and ordination approach
  - script code (in R) available as SM to the Hui (2017) paper

Very, very, brief synopsis

- Carefully examine the analyses you are using and look at the literature for best practice (it's fast moving!)
- Be critical and analyse with intent. Having effective means of analysing large data sets can lead to:
  - improper study design
  - data-mining without sensible hypotheses, and then
  - inflation of type-I or false positives (*p*-hacking)

Relationships are correlative, not causal!

Useful references

- Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**: 32–46.
- Anderson, M.J. (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**: 245–253.
- Anderson, M.J. & Walsh, D.C.I. (2013) PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs* **83**: 557–574.
- ter Braak, C. J. F., & Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecology* **216**: 683–696.

Useful references

- Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18**: 117–143.

- Legendre, P. & Legendre, L. (1998) *Numerical Ecology* (2nd Ed.). Elsevier, Amsterdam.

- Wang , Y., Naumann , U., Wright, S.T. & Warton, D.I. (2012) mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* **3**: 471–474.

- Warton , D.I., Foster, S.D., De'ath , G., Stoklosa , J. & Dunstan, P.K. (2015) Model-based thinking for community ecology. *Plant Ecology* **216**: 669–682.