# Introduction to multivariate methods: a conceptual overview from an ecologists perspective

Mats Lindegarth

Department of Marine Sciences

The Swedish Institute for the Marine Environment
(Havsmiljöinstitutet)

# Background

- Benthic ecologist with an interest for statistics (examples will be biased towards benthic biodiversity)
- Research and teaching at Tjärnö since 1990
- Not statistician! But research in Sydney under Professor AJ Underwood (logic, statistics, experimental design) and on-going collaboration with trained statisticians.
- Experience of national and regional marine policies since 2000 through independent evaluations of monitoring and 50% of my time spent at Havsmiljöinstitutet (consultations, reference groups etc.)
- Research focussed on the interface between research and management. Principles for monitoring and sampling design generally relevant!

3

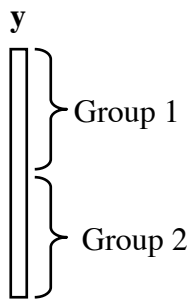
## Aims of introduction

After these lectures you will be able to…

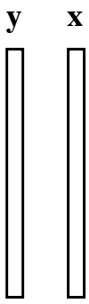
1. Understand fundamental terms and concepts of multivariate analyses
2. Understand the different purposes of multivariate analyses
3. Interpret selected types of multivariate analyses

## Contents

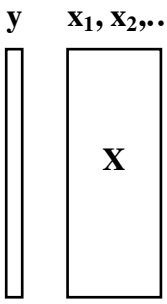
## ’Univariate’




•One response variable
•One factor
= ANOVA

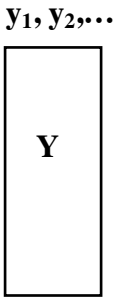



•One response variable
•One predictor variable
= regression

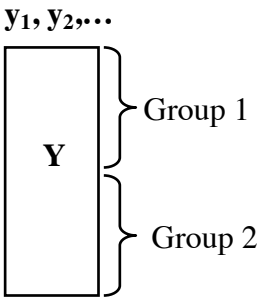



•One response variable
•Several predictor variables
= multiple regression

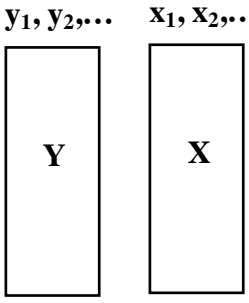
## ’Multivariate’




•Several response variables
= ?




•Several response variables
•One factor
= ?




•Several response variables
•Several predictor variables
= ?

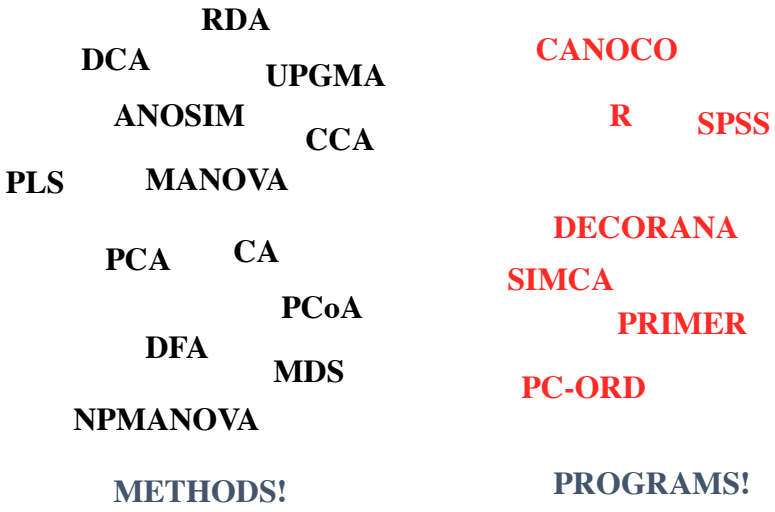
## Where do we encounter multivariate data?

- Number of individuals of different species in a core sample
- Frequencies of different alleles in an individual or a population
- Morphometric data (length, weight, etc) from inidvidual snails
- Concentrations of different nutrients in a sample of water
- and many more…

## Multi- or univariate analysis?

- Multi-…when the question / hypothesis is about patterns of many variables in combination (no variable is more important than the others).
- Uni-…when the question / hypothesis is about patterns of an individual variable

## ”Multivariate methods”!




## The main purposes of multivariate methods

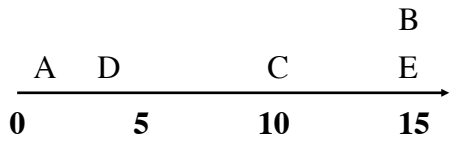
1. Arrange objects (or variables) in relation to each other (’ordination’, ’scaling’)
2. Classify object into groups (classification, clustring, prediction)
3. Test hypotheses about differences among objects or relationsships between response- and predictor variables.

# 1. Ordination

- ≈"arrange objects according to (dis)similarity"
- Illustrate patterns which emerge only when all variabels are taken into account simultaneously
- Reduce the number of variables (dimensions)
- Unconstrained vs constrained (supervised vs unsupervised)

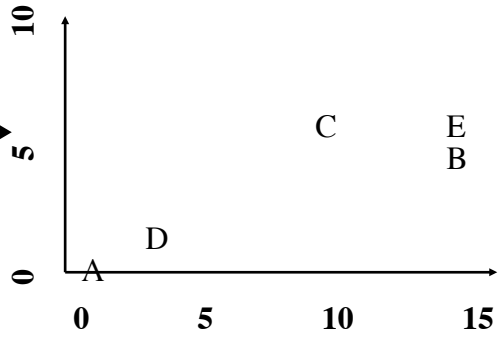
# 1. Ordination
# - one variable (dimension)

$$\begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = \begin{bmatrix} 1 \\ 15 \\ 10 \\ 7 \\ 15 \end{bmatrix}$$



## 1. Ordination
## - two dimensions

$$\begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 15 & 5 \\ 10 & 6 \\ 3 & 2 \\ 15 & 6 \end{bmatrix}$$

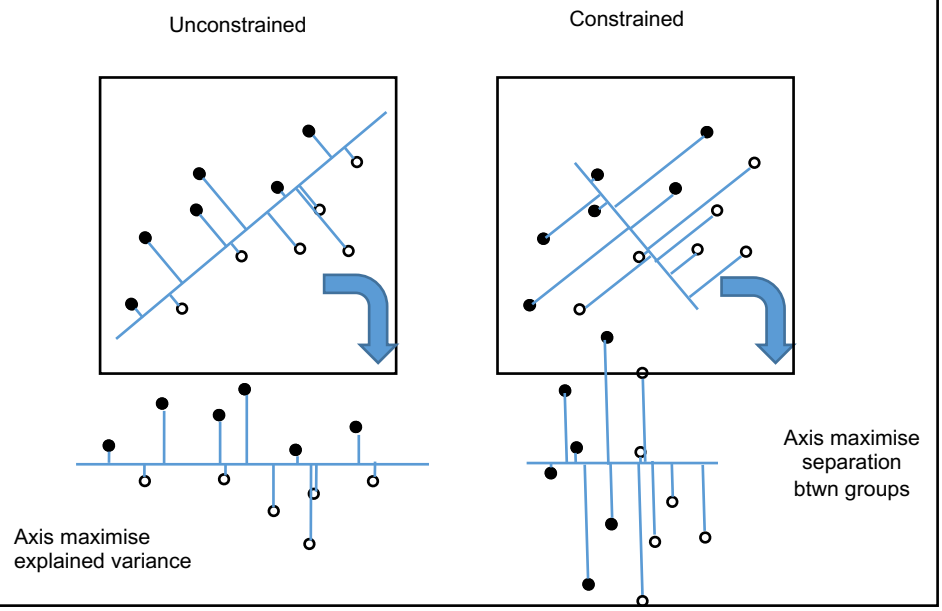



## 1. Ordination
## - > 2-3 dimensions - 'in reduced space'

$$\begin{bmatrix} A \\ B \\ C \\ D \\ E \end{bmatrix} = \begin{bmatrix} 1 & 0 & 10 & 20 & 1 \\ 15 & 5 & 20 & 1 & 10 \\ 10 & 6 & 7 & 8 & 20 \\ 3 & 2 & 1 & 1 & 10 \\ 15 & 6 & 6 & 15 & 15 \end{bmatrix}$$

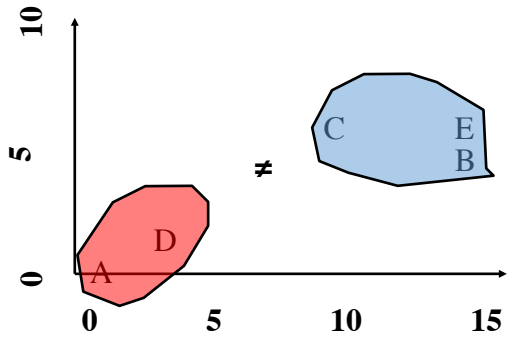
Methods based on:
- (dis)similarity
- eigenanalysis

# 1. Ordination

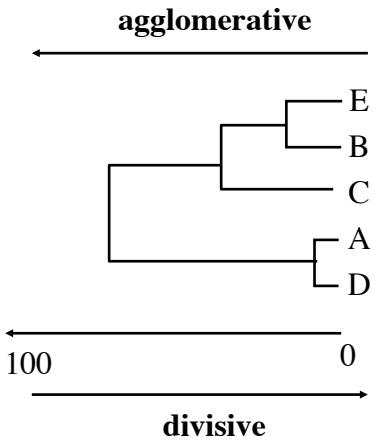



# 2. Classification

≈ ”partition objects into classes”



## 2. Classification
## -cluster analysis




## 3. Hypothesis testing

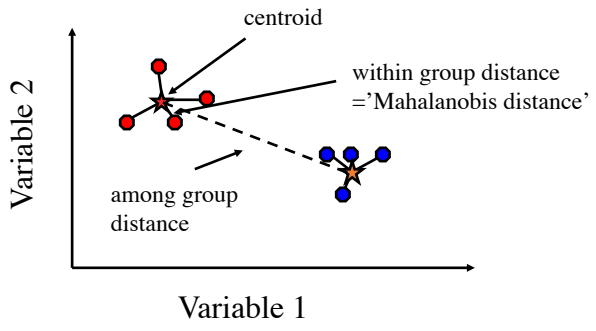
- To test whether there are…

1. Differences among groups of objects
2. Relationships between predictor- and response variables
3. NOTE!! Hypotheses are defined *a priori*.

## Multivariate Analysis of Variance (MANOVA)

- Many different test statistics (Hotelling-Lawley trace,Roy's largest root, Wilks' likelihood ratio criterion, Pillai-Bartlett's trace..)
- All of these are based on eigenanalysis of the var/cov-matrix
- Sensitive to heterogeneity of variances etc.
- Development of randomisation tests (NP-MANOVA)

## 3. Hypothesis testing



## Summary - intro

- There are three main purposes with multivariate methods!!!
- There are an endless number of methods and variations of these methods!
- The choice of method is determinened by purpose, types of data, tradition within the field, etc…

## Terms and concepts

- Standardisation / Transformation
- (Dis)similarity matrix (=distance matrix)
- Variance- / covariancematrix
- Eigenanalysis

# Standardisation (z-transformation)

$$z_i = \frac{x_i - \bar{x}}{s}$$

| | Var 1 | z 1 | | Var 2 | z 2 |
|---|---|---|---|---|---|
| | 4 | -0.69 | | 40 | -0.69 |
| | 8 | 1.04 | | 80 | 1.04 |
| | 3 | -1.13 | | 30 | -1.13 |
| | 5 | -0.26 | | 50 | -0.26 |
| | 8 | 1.04 | | 80 | 1.04 |
| | | | | | |
| mean | 5.6 | 0.0 | | 56 | 0.0 |
| std | 2.3 | 1.0 | | 23.0 | 1.0 |

To give all variables equal weight!

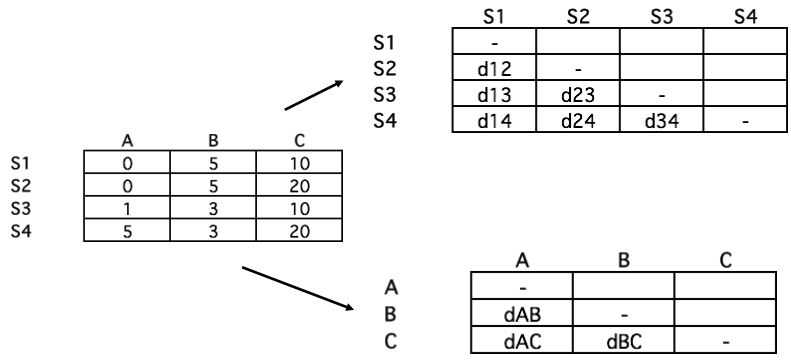
# Transformation

| Rådata | x^0.5 | x^0.25 | Pres/Abs |
|---|---|---|---|
| 10 | 3.16227766 | 1.77827941 | 1 |
| 150 | 12.2474487 | 3.49963551 | 1 |
| 0 | 0 | 0 | 0 |
| 999 | 31.6069613 | 5.62200687 | 1 |
| 40 | 6.32455532 | 2.51486686 | 1 |
| 50 | 7.07106781 | 2.65914795 | 1 |
| 3 | 1.73205081 | 1.31607401 | 1 |

Is done to make..

1. variables equally influential (some dissimilarity measures can not handle negative values)
2. improve distributional properties of data

## Dissimilarity- or distancematrices



| | A | B | C |
|---|---|---|---|
| S1 | 0 | 5 | 10 |
| S2 | 0 | 5 | 20 |
| S3 | 1 | 3 | 10 |
| S4 | 5 | 3 | 20 |

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | - | | | |
| S2 | d12 | - | | |
| S3 | d13 | d23 | - | |
| S4 | d14 | d24 | d34 | - |

| | A | B | C |
|---|---|---|---|
| A | - | | |
| B | dAB | - | |
| C | dAC | dBC | - |

## Measures of dissimilarity

- Dissimilarity among objects or variables
- Countless number!
- The chioice of measure depends on the type of data (field of research)

# Euclidian distance - normally distributed data, continuous

| | A | B | C |
|---|---|---|---|
| S1 | 0 | 5 | 10 |
| S2 | 0 | 5 | 20 |
| S3 | 1 | 3 | 10 |
| S4 | 5 | 3 | 20 |

$$D_{12} = \sqrt{\sum_{j=1}^{p} \left(y_{1j} - y_{2j}\right)^2}$$

$d_{12}$ = sqrt((0-0)$^2$+ (5-5)$^2$ + (10-20)$^2$)=sqrt( 0+0+100)=10
$d_{13}$ = sqrt((0-1)$^2$+ (5-3)$^2$ + (10-10)$^2$)= sqrt(1+4+0)=2.24
…

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | - | | | |
| S2 | 10.00 | - | | |
| S3 | 2.24 | 10.25 | - | |
| S4 | 11.36 | 5.39 | 10.77 | - |

# Bray-Curtis distance - count data with many zeros

| | A | B | C |
|---|---|---|---|
| S1 | 0 | 5 | 10 |
| S2 | 0 | 5 | 20 |
| S3 | 1 | 3 | 10 |
| S4 | 5 | 3 | 20 |

$$D = 1 - \frac{2W}{A+B} = \frac{\sum_{j=1}^{p} \left|y_{1j} - y_{2j}\right|}{\sum_{j=1}^{p} \left(y_{1j} + y_{2j}\right)}$$

$d_{12}$ = (0+0+10)/(0+10+30)=10/40=0.25
$d_{13}$ = (1+2+0)/(1+8+20)=3/29=0.10
…

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | - | | | |
| S2 | 0.25 | - | | |
| S3 | 0.10 | 0.33 | - | |
| S4 | 0.40 | 0.13 | 0.33 | - |

## ”The problem of double-zeros”

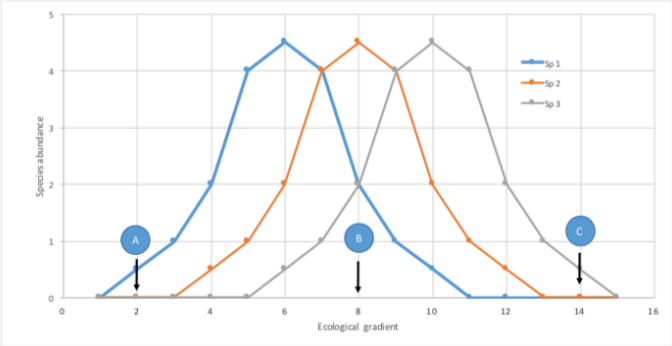



Fig. 1. Schematic environmental gradient and three species with different niches. Samples are taken at sites A, B and C.

a.

| Euclidean distance | | |
|---|---|---|
| | A | B |
| A | | |
| B | 5.15 | |
| C | 0.71 | 5.15 |

b.

| Bray-Curtis | | |
|---|---|---|
| | A | B |
| A | | |
| B | 0.89 | |
| C | 1 | 0.89 |

## Dissimilarity matrix is used for…

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | - | | | |
| S2 | 0.25 | - | | |
| S3 | 0.10 | 0.33 | - | |
| S4 | 0.40 | 0.13 | 0.33 | - |

**Cluster analysis**

**ordination**

**Hypothesis testing**

## Variance- / covariance matrix

- Measure of relationsship among variables

$$\begin{bmatrix} s_1^2 & s_{21}^2 & \dots & s_{p1}^2 \\ s_{12}^2 & s_2^2 & \dots & s_{p2}^2 \\ \dots & \dots & s_j^2 & \dots \\ s_{1p}^2 & s_1^2 & \dots & s_p^2 \end{bmatrix}$$

**Cluster analysis**

**ordination**

**Hypothesis testing**

## Variance- / covariance matrix

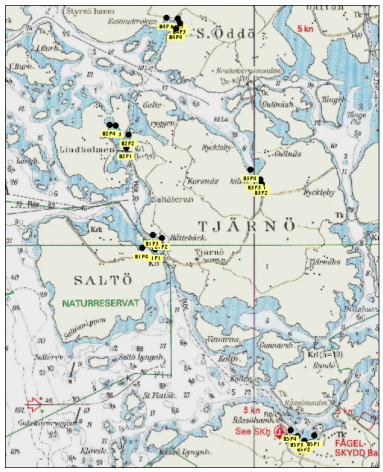
- Standardised variables results in a correlation matrix

$$\begin{bmatrix} 1 & r_{12} & \dots & r_{p1} \\ r_{12} & 1 & \dots & r_{p2} \\ \dots & \dots & 1 & \dots \\ r_{1p} & r_1 & \dots & 1 \end{bmatrix}$$

## Eigenanalysis

- Principal component analysis (PCA etc.)
- Well-known but conceptually difficult term from matrix algebra
- Results in eigenvalues and their associated eigenvectors
- No further treatment here but Erik will explain and demonstrate methods and software in Friday.

## Example-data



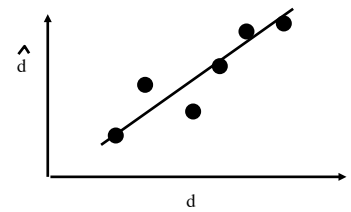

- infauna from 2 bays close to TMBL
- 4 sites per bay
- 27 species
- 9 sediment variables
- Methods based on the distance matrix

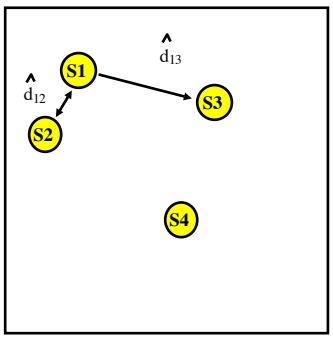
# Ordination

- Multidimensional scaling (MDS)
- 'How can we arrange a number of objects in a limited number of dimensions, so that the distances are proportional to dissimilarities in the distance matrix?'

# Multidimensional scaling (MDS)

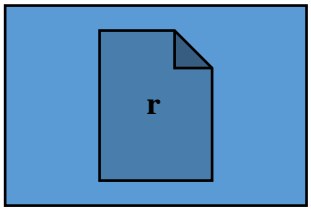
| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | - | | | |
| S2 | d12 | - | | |
| S3 | d13 | d23 | - | |
| S4 | d14 | d24 | d34 | - |



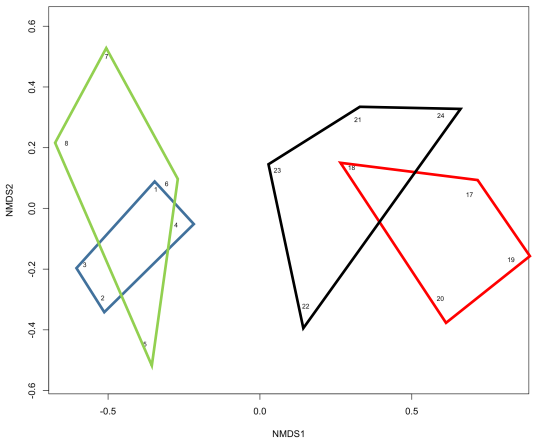
## Multidimensional scaling (MDS)

1. Ordination of sites
2. Effects of distance measure
3. Effects of transformation




## But are they different?

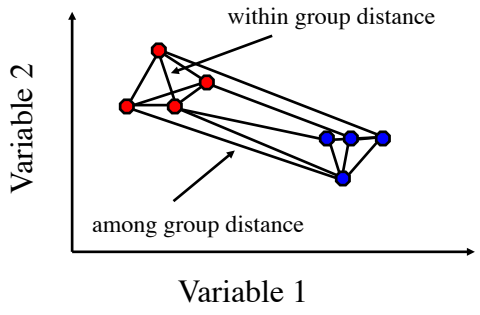
- MDS gives no answer to the question whether the bays are statistically different and if so why this is so
- Hypothesis testing!!



# Analysis of similarities (ANOSIM)

- Program package PRIMER
- Plymouth Marine Labs
- Common among benthic ecologists
- Non-parametric test of one- or two-way designs
- Based on permutation tests

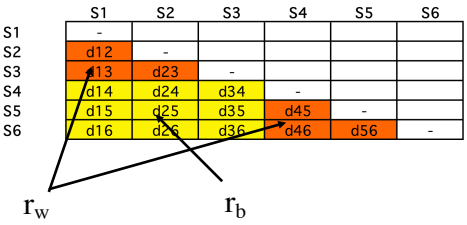
# Analysis of similarities (ANOSIM)



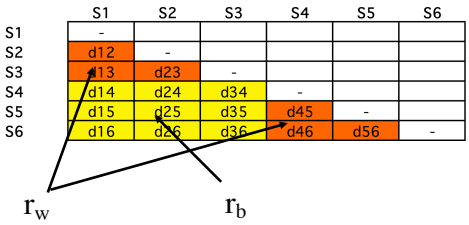
# Analysis of similarities (ANOSIM)

| | V1 | V2 |
|---|---|---|
| A | 0 | 9 |
| A | 2 | 10 |
| A | 1 | 11 |
| B | 9 | 3 |
| B | 11 | 3 |
| B | 10 | 5 |

| | S1 | S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|---|
| S1 | - | | | | | |
| S2 | d12 | - | | | | |
| S3 | d13 | d23 | - | | | |
| S4 | d14 | d24 | d34 | - | | |
| S5 | d15 | d25 | d35 | d45 | - | |
| S6 | d16 | d26 | d36 | d46 | d56 | - |




# Analysis of similarities (ANOSIM)

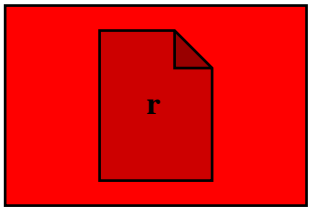



$$R = \frac{\overline{r_b} - \overline{r_w}}{n(n-1)/4}$$

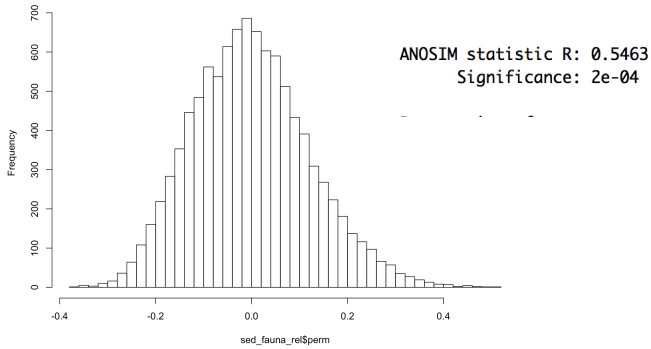
- The statistic R varies between 0-1
- Permutation tests are used to determine statistical significance

# ANOSIM

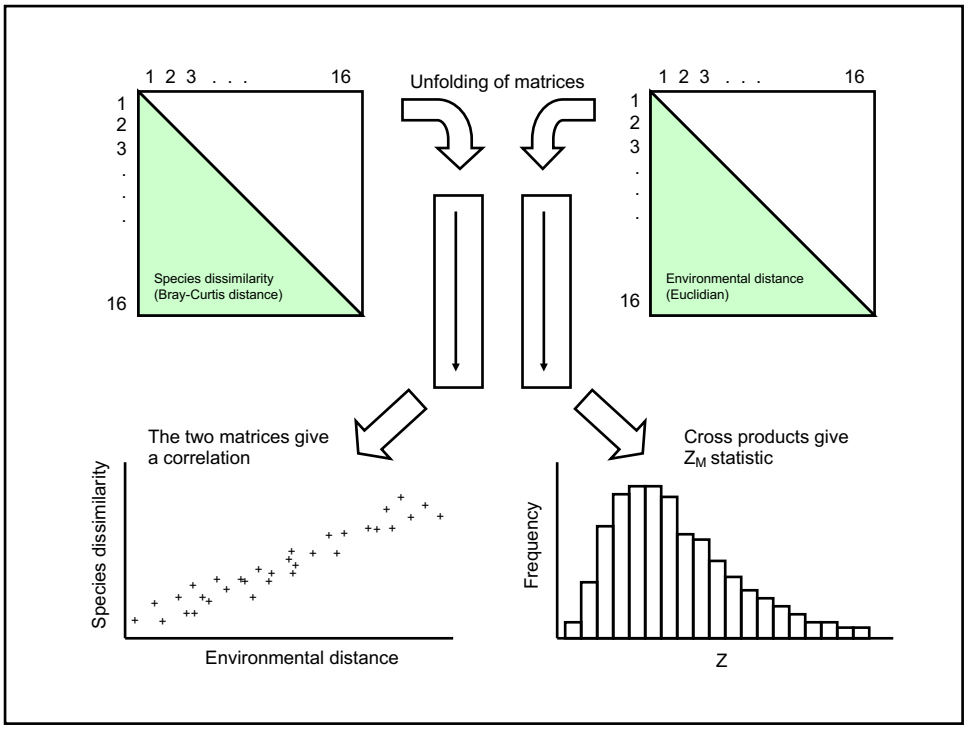
1. Test differences among bays
2. Effects of transformations




# Analysis of similarities (ANOSIM)

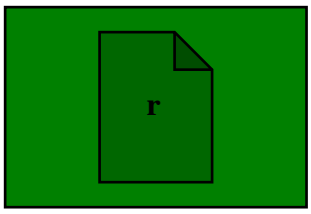



- There are differences in the composition of assemblages in different bays!
- Can these be explained by differences in sediment characteristics?
- Mantel's test!

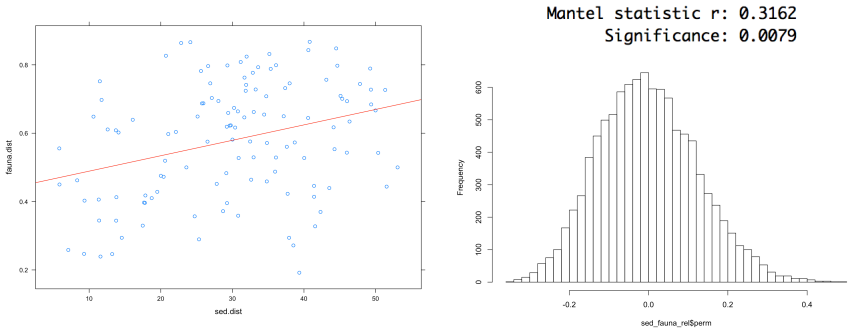



# Mantel's test

1. Correlation between matrices



# Mantels test




- Statistically significant relationship!
- Differences in sediment can to some degree explain differences in faunal composition.

# You have just heard...

- ...some examples of analyses based on distancematrices
- Conceptually simple and relatively robust
- Inferences often based on permutation tests
- Flexible with respect to measure of dissimilarity
- Non-parametric - relationsship to original variables?

## Methods based on eigenanalysis

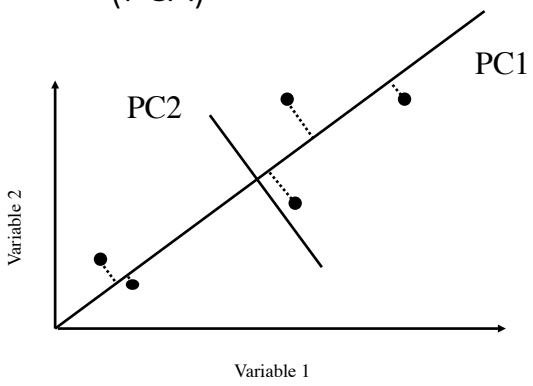
- Mathematically complicated
- Sensitive to deviations from assumptions about homogeneity of variances and multivariate normality
- Not flexible with respect to measure of dissimilarity
- Relationship to original variables!

## Principal component analysis (PCA)

- Reduces the number of dimensions by creating new, uncorrelated variables
- Suitable for continuous data

## Principal component analysis (PCA)

$$\begin{bmatrix} 2 & 3 \\ 10 & 10 \\ 7 & 10 \\ 7 & 5 \\ 3 & 2 \end{bmatrix}$$

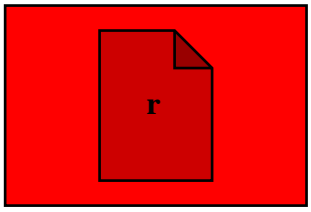



$$PC1 = c_1 * v1 + c_2 * v2$$

## Principal component analysis (PCA)

1. Calculate variance-/covariancematrix
   a. Unstandardised data (variance-/covariance matrix)
   b. Standardised data (correlationmatrix)
2. Calculate eigenvalues and -vectors
3. Eigenvalue, $\lambda_i$=variance explained by i-th component
4. Associated eigenvector=coeffecients

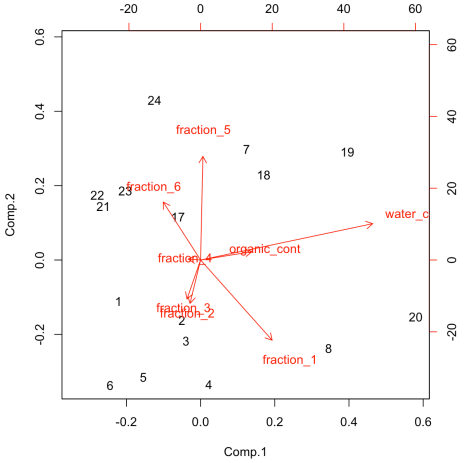
# PCA




# PCA

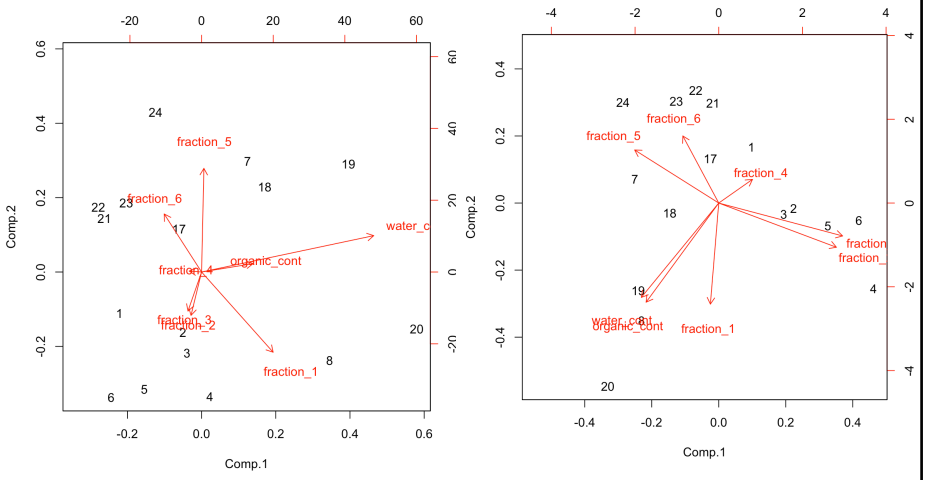
```
Call:
princomp(x = dataset[, 30:37], cor = FALSE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
17.2846160 13.8540616 12.0469009  8.6626053  5.0199753  1.3014439  0.6288494  0.0000000
```




Horisontal axis =
1st axis =
Most important axis!

# PCA



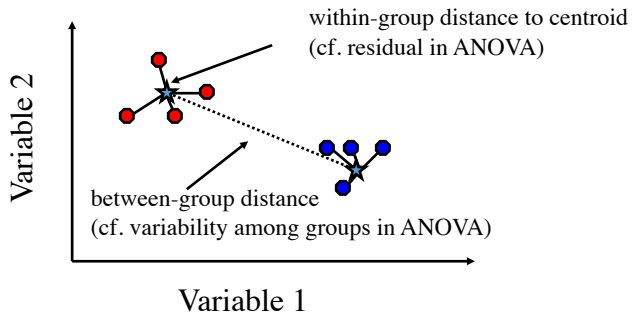

Raw data

Standardised data

# But are they different?

- PCA gives no answer to the question whether the bays are statistically different and if so why this is so
- Hypothesis testing!!
- For example MANOVA

## Multivariate Analysis of Variance (MANOVA)




## Conclusion

- There are three main purposes for multivariate analyses!!
- Methods can be based on (dis)similarity matrices or eigenanalysis
- Many methods are sensitive to violations to parametric assumptions
- Methods based on randomisation procedures are becoming increasingly popular