Training in ade4 in R - Module II: Advanced methods

Analysis of one table and one categorical variable

Stéphane Dray

2022-02-28

Introduction

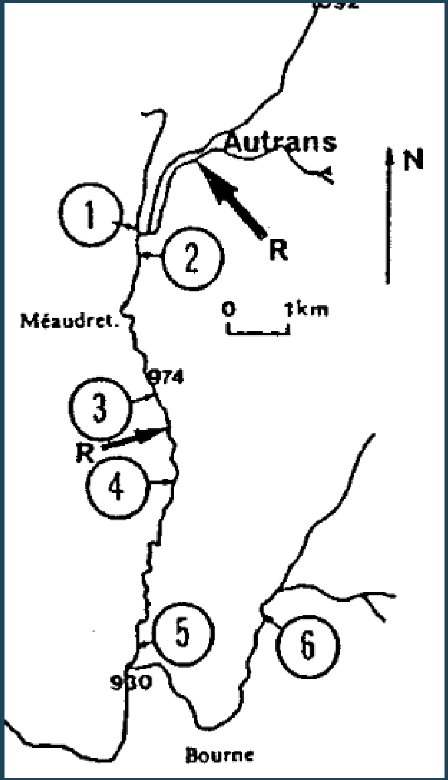
```r
library(ade4)
library(adegraphics)
adegpar(paxes.draw = TRUE, pbackground.col = "lightgrey",
    pgrid.col = "white")
data(meau)
str(meau, max.level = 1)
```

```
## List of 3
##  $ env   :'data.frame':    24 obs. of  10 variables:
##  $ design:'data.frame':    24 obs. of  2 variables:
##  $ spe   :'data.frame':    24 obs. of  13 variables:
```

- Species table: abundance of 13 Ephemeroptera species recorded for 24 sites
- Environmental table: 10 physicochemical variables for the same sites
- Expermiental design (6 sites and 4 seasons)

Introduction

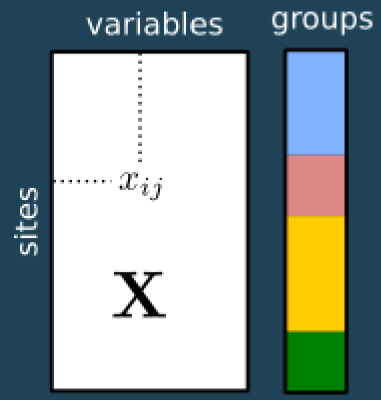
- Data table: 10 environmental variables measured for 24 samples (6 sites sampled each season) on the Méaudret river
- Categorical variable(s): 6 sites or 4 seasons
- S1-S5 on the Méaudret, S6 is a control (on the Bourne river)




```
head(meau$design)
```

```
##      season site
## sp_1 spring   S1
## sp_2 spring   S2
## sp_3 spring   S3
## sp_4 spring   S4
## sp_5 spring   S5
## sp_6 spring   S6
```

Introduction

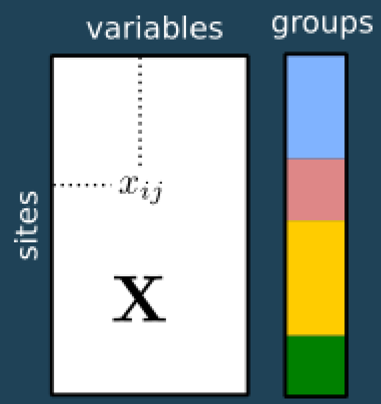


- One table with *p* variables measured on *n* individuals
- One categorical variable partitioning the *n* individuals in *g* groups (colors)

Describe the information contained in the table:

- Identify differences between individuals **belonging to different groups**
- Identify which variables best separate the groups

Introduction



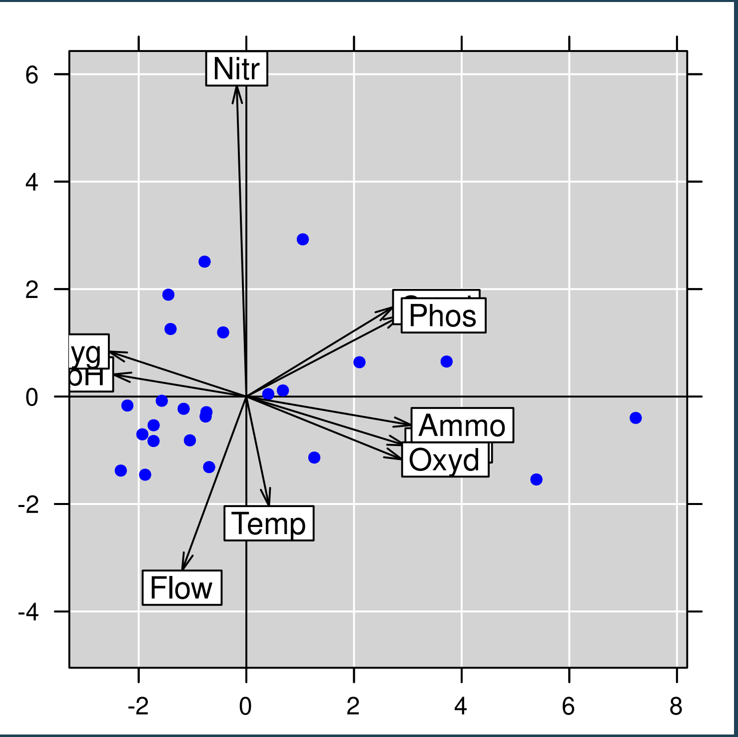
- One table with *p* variables measured on *n* individuals
- One categorical variable partitioning the *n* individuals in *g* groups (colors)

Describe the information contained in the table:

- Identify differences between individuals **after removing differences among groups**
- Identify relationships between variables

Questions

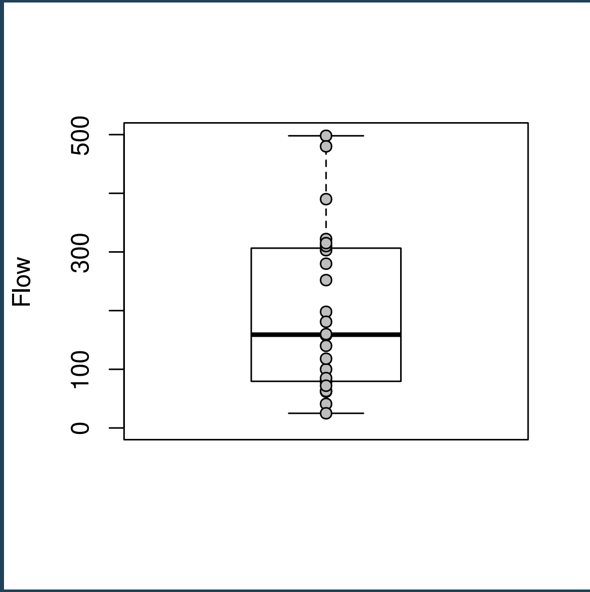
```
pca_env <- dudi.pca(meau$env, sca
biplot(pca_env, ppoints.col = "bl
    posieig = "none")
```



Which structure is due to seasonal variation?

Which part is not explained by seasonal variation?

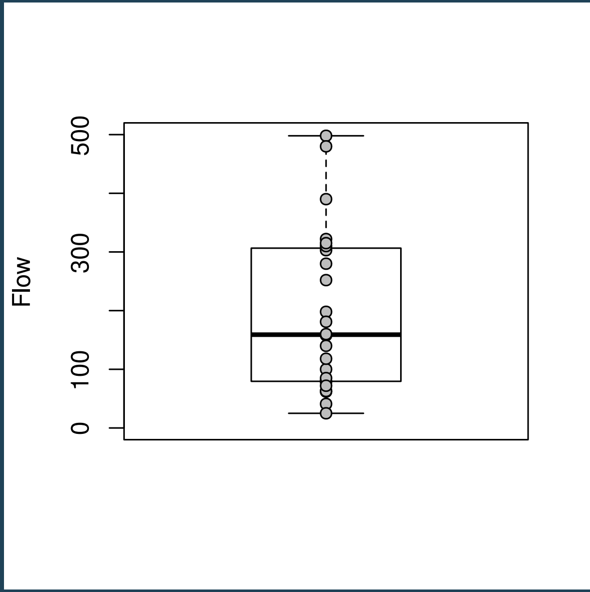
The univariate case

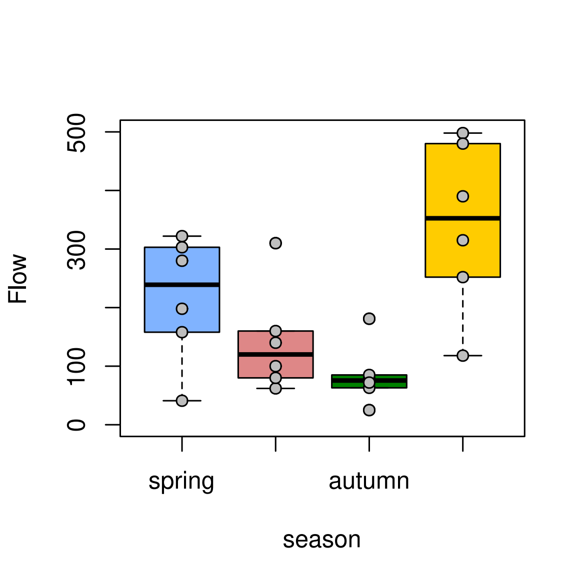


Total variation

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The univariate case





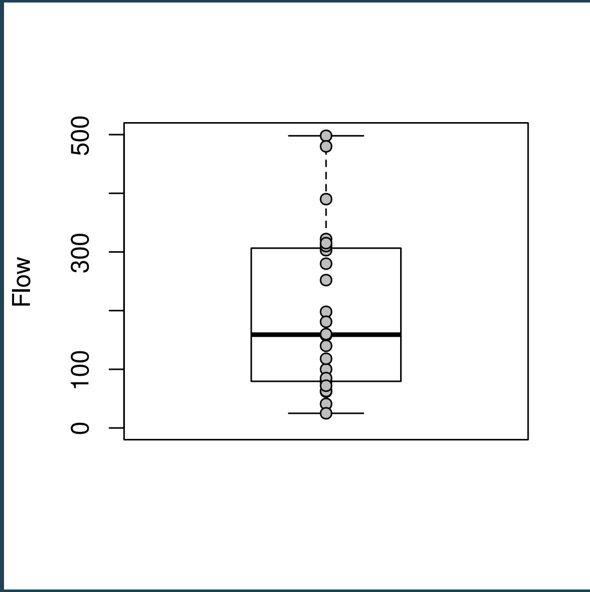
Total variation

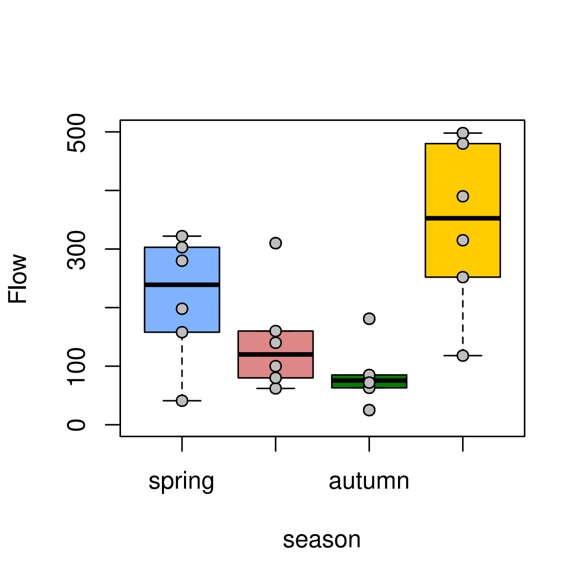
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Within-group variation

$$\sigma_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_j - \bar{x}_i)^2$$

The univariate case





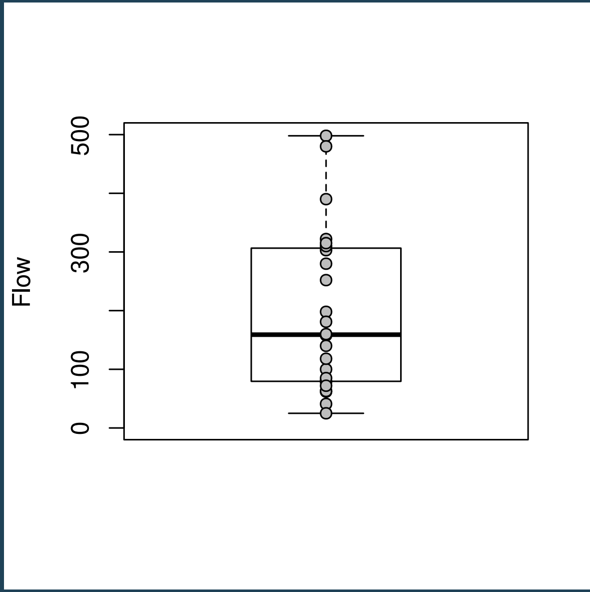
Total variation

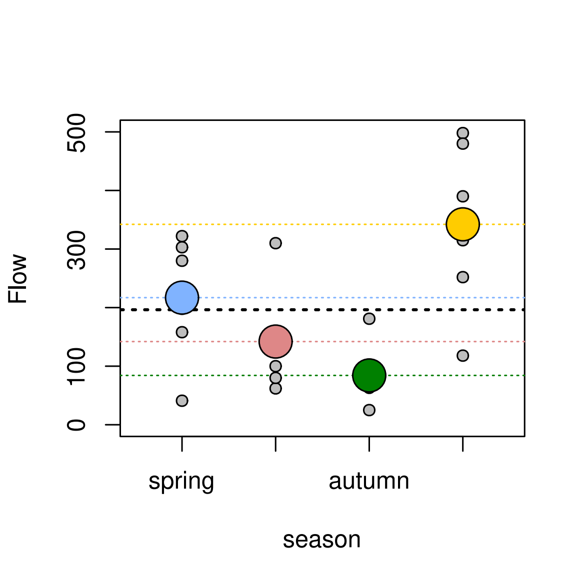
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Within-group variation

$$W = \sum_{i=1}^{k}\frac{n_i}{n}\sigma_i^2$$

The univariate case





Total variation

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Between-group variation

$$B = \sum_{i=1}^{n}\frac{n_i}{n}(\bar{x}_i - \bar{x})^2$$

The univariate case

the correlation ratio

We have

$$\sigma^2 = \sum_{i=1}^{k} \frac{n_i}{n} \sigma_i^2 + \sum_{i=1}^{k} \frac{n_i}{n} (\bar{x}_i - \bar{x})^2$$

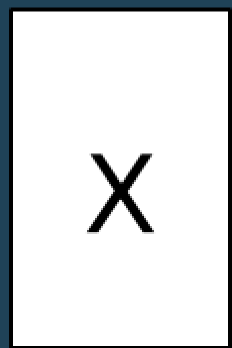
which corresponds to

$$T = W + B$$

The correlation ratio varies between 0 and 1 and is defined as

$$\eta^2 = \frac{B}{T}$$

The multivariate case



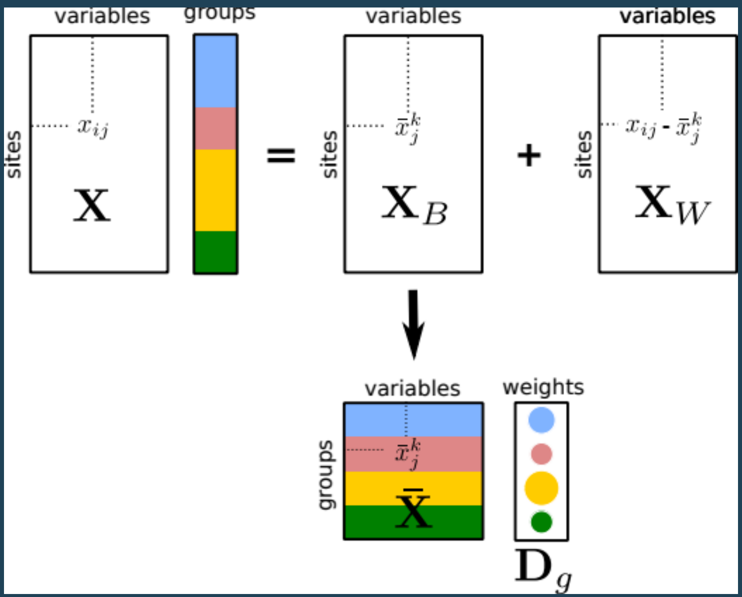
Total inertia measures the amount of variation in the data.

$$I_{(\mathbf{X},\mathbf{Q},\mathbf{D})} = Trace(\mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{Q})$$

For PCA, we have

$$I_{(\mathbf{X},\mathbf{Q},\mathbf{D})} = \sum_{j=1}^{p} var(\mathbf{x}_j)$$

ANOVA-like decomposition of a table



The analysis of $\mathbf{X}$ leads to two additive components

- Between-Class Analysis focuses on the differences between groups $(\mathbf{X}_B)$
- Within-Class Analysis focuses on the differences between individuals while removing differences between groups $(\mathbf{X}_W)$

Decomposition of total inertia

$$\begin{aligned} I_{(\mathbf{X},\mathbf{Q},\mathbf{D})} &= Trace(\mathbf{X}^\top\mathbf{D}\mathbf{X}\mathbf{Q}) \\ &= Trace((\mathbf{X}_B+\mathbf{X}_W)^\top\mathbf{D}\,(\mathbf{X}_B+\mathbf{X}_W)\,\mathbf{Q}) \\ &= Trace(\mathbf{X}_B{}^\top\mathbf{D}\mathbf{X}_B\mathbf{Q}) + Trace(\mathbf{X}_W{}^\top\mathbf{D}\mathbf{X}_W\mathbf{Q}) \end{aligned}$$

We obtain the following additive decomposition

$$I_{(\mathbf{X},\mathbf{Q},\mathbf{D})} = I_{(\mathbf{X}_B,\mathbf{Q},\mathbf{D})} + I_{(\mathbf{X}_W,\mathbf{Q},\mathbf{D})}$$

that translates into

$$\text{Total Inertia} = \text{Between-Class Inertia} + \text{Within-Class Inertia}$$

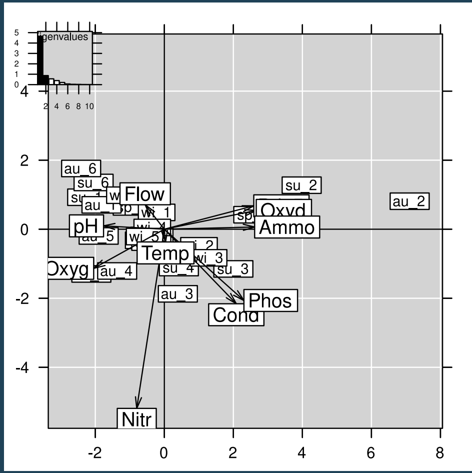
Remember that the inertia is equal to the sum of eigenvalues of the associated analysis

Removing an effect

Within-Class Analysis

WCA is simply the analysis of the table centered per group $(\mathbf{X}_W)$. It is a partial analysis that focuses on the structure removing the effect of the categorical variables.

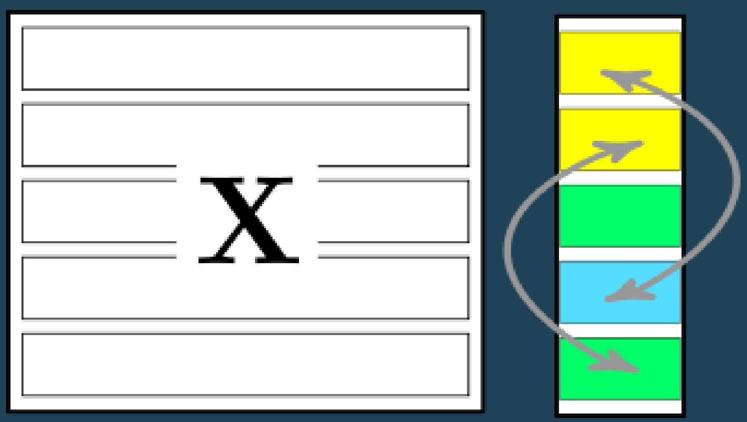
```
wca.season <- wca(pca_env, meau$design$season, scannf = FALSE)
biplot(wca.season)
```



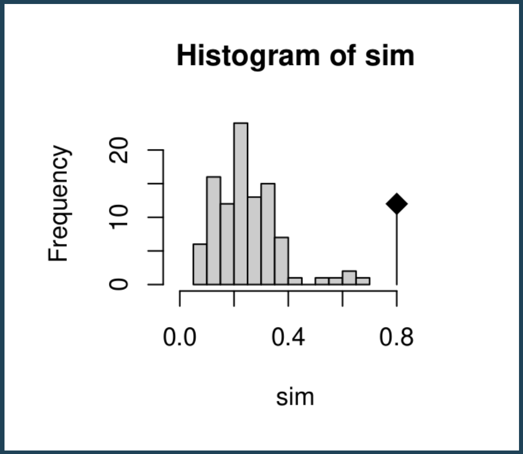
Focusing on an effect

Testing the significance

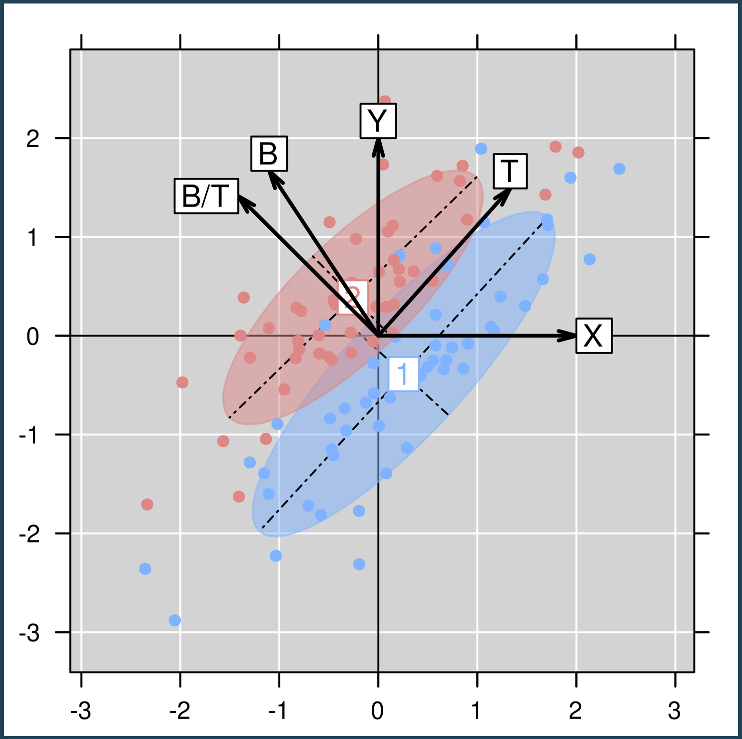
R^2 = Between-class inertia / Total inertia

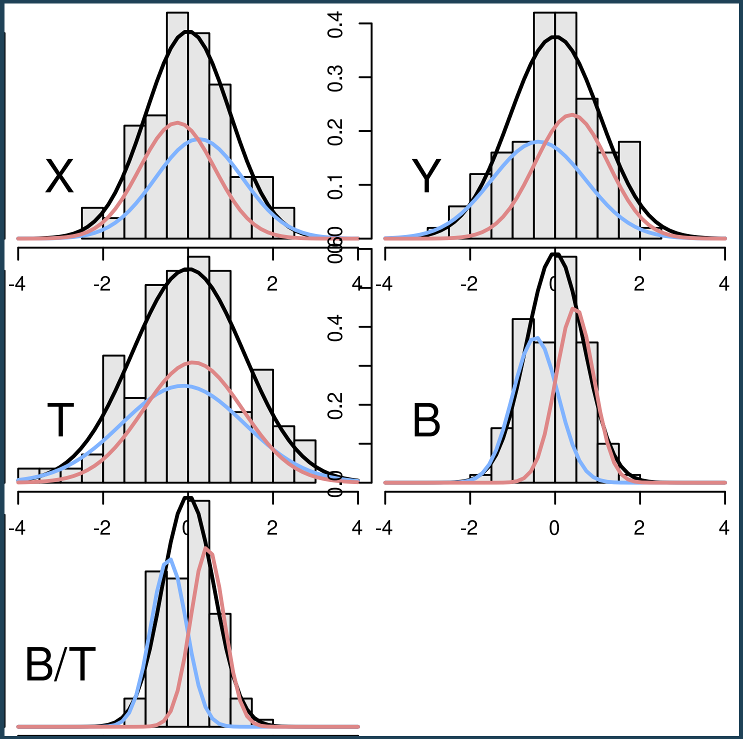


Permutation test



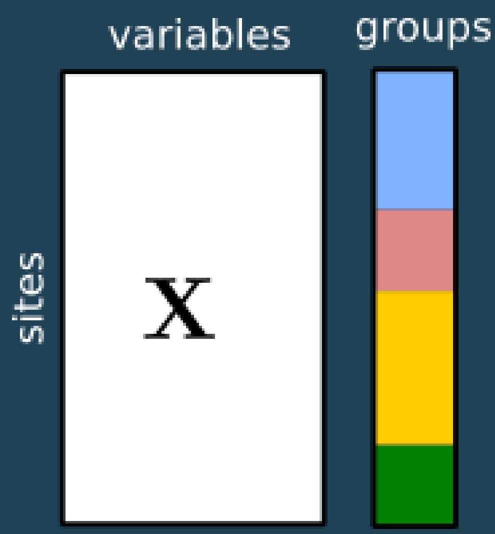
Two strategies

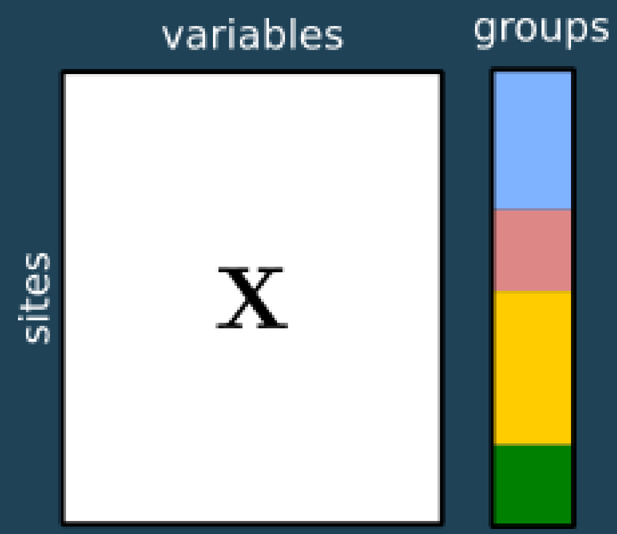




- Principal component analysis maximizes T
- Between-class analysis maximizes B
- Discriminant analysis maximizes B/T

Between-Class and Discriminant Analysis

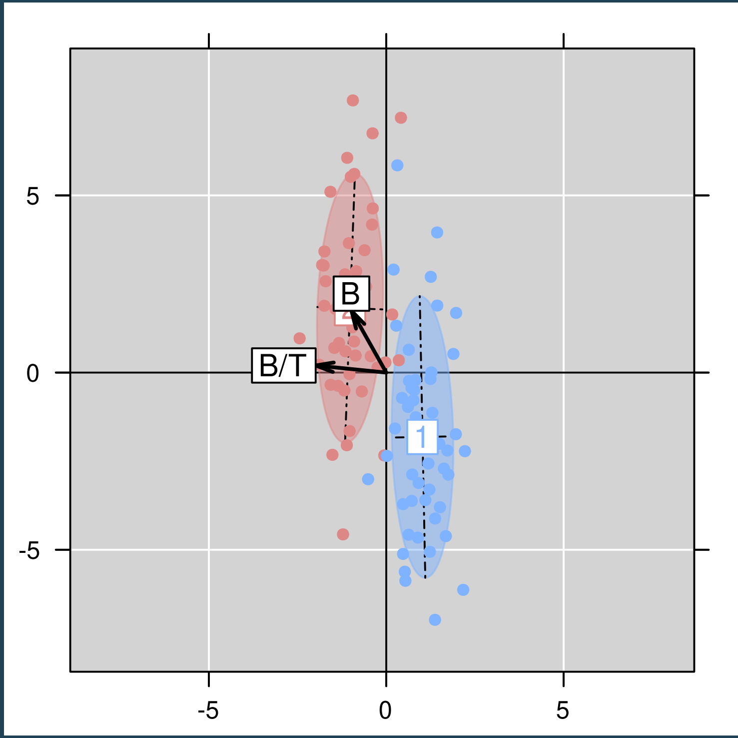


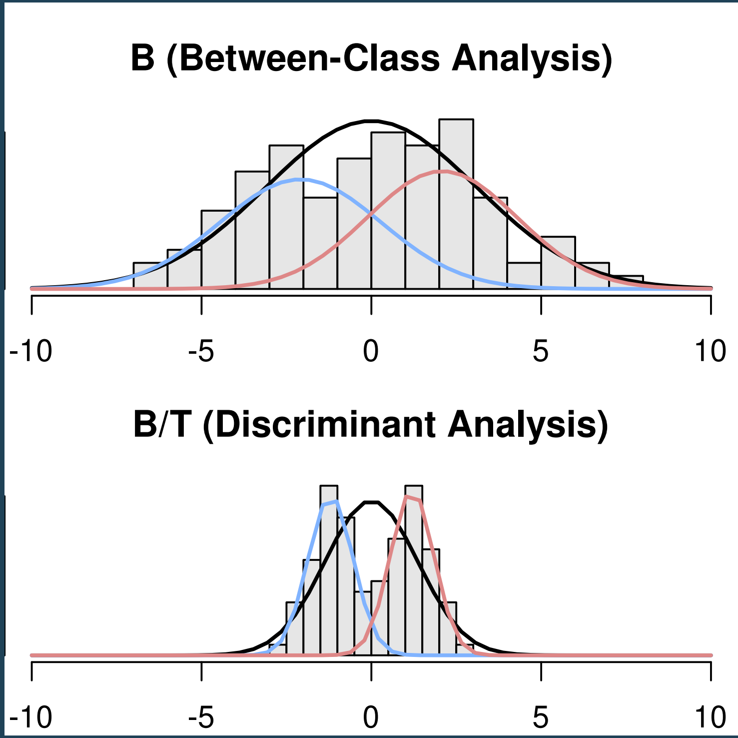



```
discrimin(pca_env, meau$design$se
```

```
bca(pca_env, meau$design$season,
```

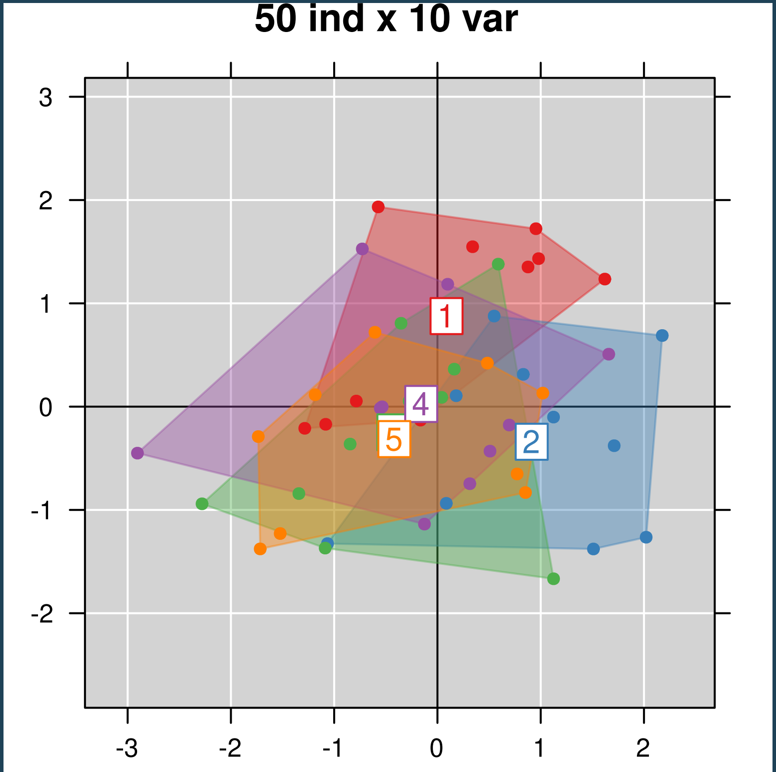
Between-Class and Discriminant Analysis





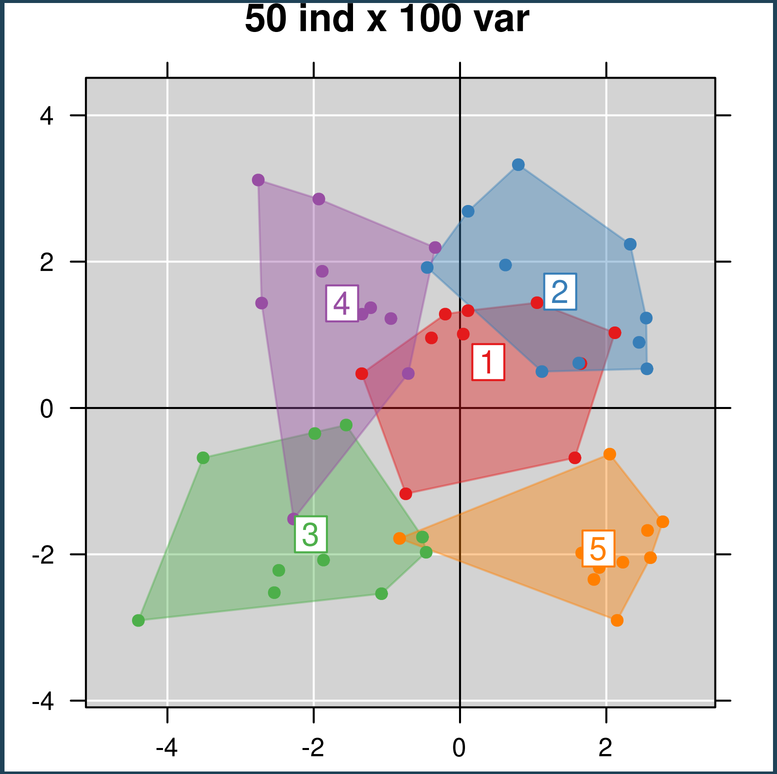
Spurious groups

BCA of random data with a moderate number of variables



Spurious groups

BCA of random data with a high number of variables



Spurious groups

- Perform permutation test even if segregation of groups is clear on the factorial map
- Cross-validation to display results

```
s.class(loocv(bca.spurious)$XValCoord, fac, col = TRUE,
    star = 0, ell = 0, chull = 1, main = "50 ind x 100 var CV")
```