


Fiche TD avec le logiciel  : course4

---

# Within PCA and Between PCA

A.B. Dufour

---

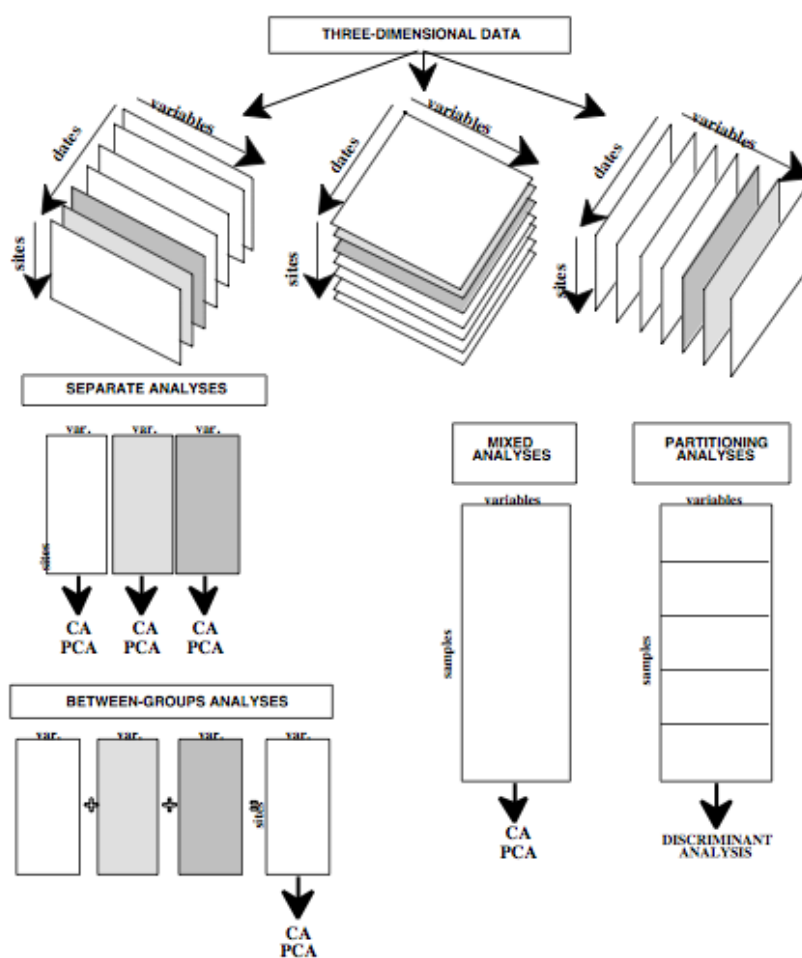
## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Classical Approach</b>	<b>3</b>
2.1	Dataset . . . . .	3
2.2	Normed Principal component analysis . . . . .	4
2.3	Usual one-way analysis of variance . . . . .	8
<b>3</b>	<b>Removing an effect: the within PCA</b>	<b>11</b>
<b>4</b>	<b>Taking an effect into account: the between PCA</b>	<b>15</b>
<b>5</b>	<b>Conclusion</b>	<b>17</b>
	<b>References</b>	<b>18</b>

# 1 Introduction

Data analyses need to take into account the ecological goals, experimental designs (space, time, ...) to answer questions such as:

1. Which variables are time-dependent? which ones display a spatial structure? which ones are structured by both time and space?
2. Which variables display a structure along the structure imposed by the sample design?

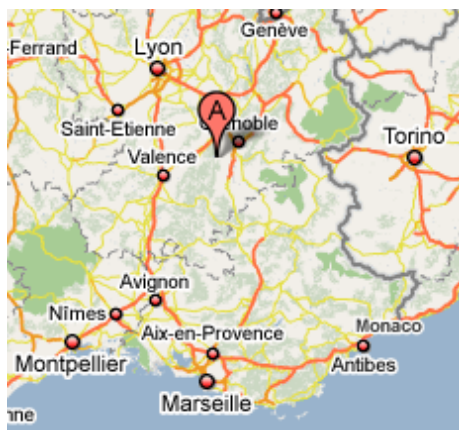


To solve such questions, four options are proposed (Dolédec et Chessel, 1991 [1]): separated analyses (1), between-groups analyses (2), mixed analyses (3) and partitioning analyses (4). Between PCA will be found in the family of between-groups analyses, within PCA in the family of partitioning analyses.

## 2 Classical Approach

### 2.1 Dataset

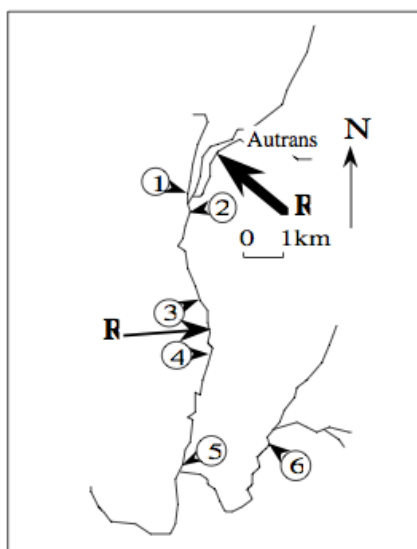
The dataset contains physico-chemical variables measured on 5 sites along the Meaudret River (a river near Grenoble) during four seasons (Pegaz-Maucet, 1980 [2]).



1. Temp Water Temperature (in Celcius degree)
2. Debit Flow (in l/s)
3. pH pH
4. Condu Conductivity (in  $\mu$  S/cm)
5. Dbo5 Biological oxygen Demand after 5 days (in mg/l)
6. Oxyd Oxygen (in mg/l of oxygen)
7. Ammo Ammonium Hydroxide (in mg/l  $NH_4^+$ )
8. Nitra Nitrate (in mg/l  $NO_3^-$ )
9. Phos Phosphate (in mg/l  $PO_4^{--}$ )

```
library(ade4)
library(xtable)
data(meaudret)
names(meaudret)
[1] "mil" "plan" "fau"
names(meaudret$mil)
[1] "Temp" "Debit" "pH" "Condu" "Dbo5" "Oxyd" "Ammo" "Nitra" "Phos"
```

`meaudret$mil` is a data frame encompassing all the environmental variables (9 variables in this case). `meaudret$plan` is a data frame detailing the experimental design. `meaudret$fau` is a data frame providing information on species richness per site, per season. Note that the experimental design is described in the data frame `meaudret$plan` by two factors: the sites (`sta`) and the seasons (`dat`).



```
summary(meaudret$plan$sta)
S1 S2 S3 S4 S5
 4  4  4  4  4

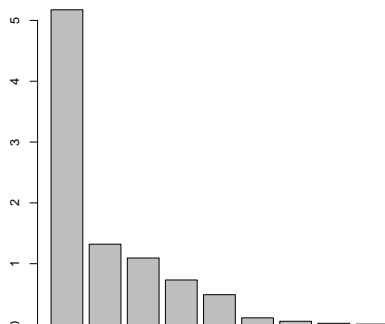
summary(meaudret$plan$dat)
autumn spring summer winter
   5     5         5         5
```

There is 4 measures (one per season) taken per site and 5 measures (one per site) for each season.

## 2.2 Normed Principal component analysis

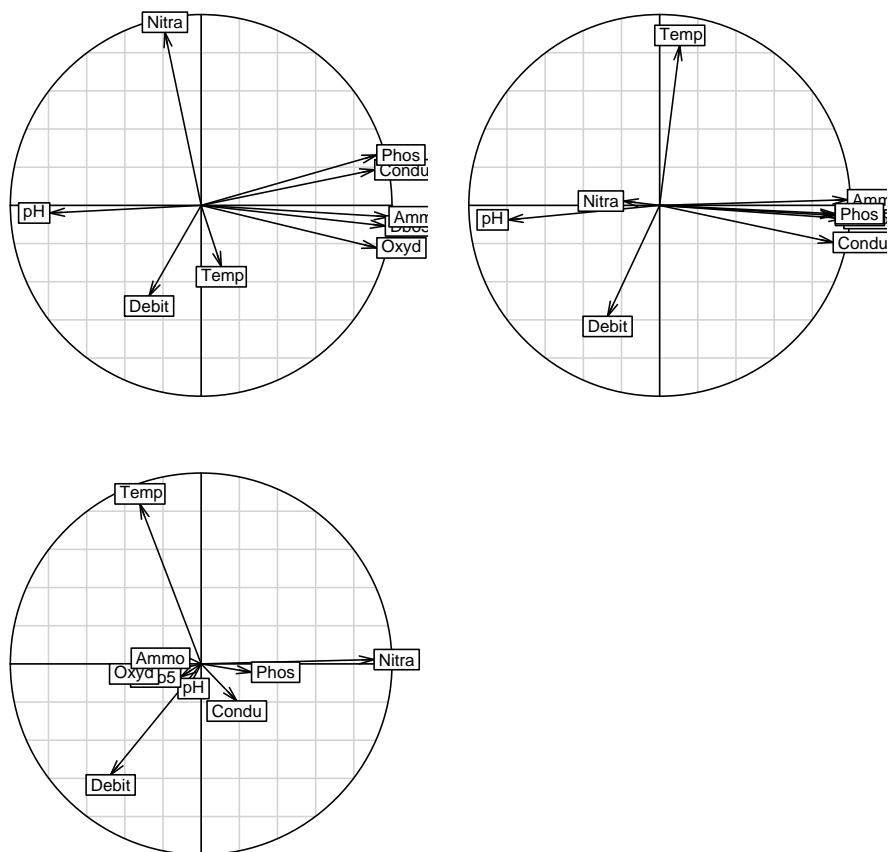
One can compute the normed P.C.A. on the 9 physico-chemical variables (NB: The default option for `dudi.pca` is a normed PCA).

```
pca1 <- dudi.pca(meaudret$mil, scann = F, nf = 3)
pca1$eig
[1] 5.174736624 1.320418552 1.093376100 0.732113258 0.490213700 0.109834881
[7] 0.052960338 0.020030611 0.006315936
sum(pca1$eig)
[1] 9
cumsum(pca1$eig)/sum(pca1$eig)
[1] 0.5749707 0.7216839 0.8431701 0.9245161 0.9789842 0.9911881 0.9970726 0.9992982
[9] 1.0000000
inertie <- cumsum(pca1$eig)/sum(pca1$eig)
barplot(pca1$eig)
```



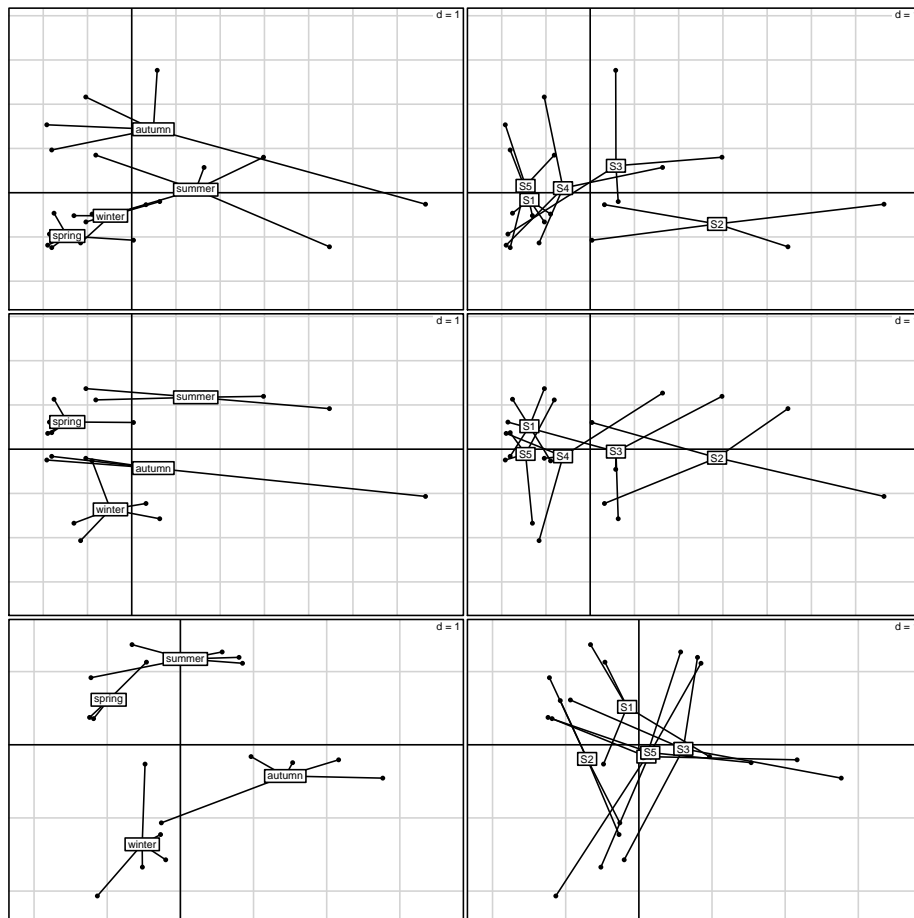
Three eigenvalues are kept (all those greater or equal to one) and one can represent the correlation circles (axes 1-2, 1-3, 2-3). These correlation circles show the redundancy between conductivity (Condu), Biological oxygen demand (Dbo5), oxygen (Oxyd), ammoniac (Ammo) and phosphate (Phos) . These variables are descriptors of organic pollution.

```
par(mfrow = c(2, 2))
s.corcircle(pca1$co, xax = 1, yax = 2)
s.corcircle(pca1$co, xax = 1, yax = 3)
s.corcircle(pca1$co, xax = 2, yax = 3)
```



Factorial maps summarize the analysis. The `s.class` function helps us to represent each group (i.e., seasons on left, sites on right) by the centre of gravity (means) and the links between the sample and its group.

```
par(mfrow = c(3, 2))
s.class(pca1$li, meaudret$plan$dat, xax = 1, yax = 2, cellipse = 0)
s.class(pca1$li, meaudret$plan$sta, xax = 1, yax = 2, cellipse = 0)
s.class(pca1$li, meaudret$plan$dat, xax = 1, yax = 3, cellipse = 0)
s.class(pca1$li, meaudret$plan$sta, xax = 1, yax = 3, cellipse = 0)
s.class(pca1$li, meaudret$plan$dat, xax = 2, yax = 3, cellipse = 0)
s.class(pca1$li, meaudret$plan$sta, xax = 2, yax = 3, cellipse = 0)
```



The three first axes of the normed PCA of physico-chemical variables help us to describe the correlations between the variables linked to the spatio-temporal structure. The first axis (57.5%) describes a high level of mineralisation i.e. the pollution of site 2 during fall.

```
rownames(meaudret$mil)[which.max(pca1$li[, 1])]
[1] "au_2"
```

NB: By doing so, we ask the name of the row where the maximum score on the first axis is.

Such a pollution leads to acidity (small pH), poor oxygen concentration, high values for biological oxygen demand. The high values for ammonium hydroxide and phosphate characterize a big organic pollution. Site 1 is not polluted and sites 3, 4 and 5 look less polluted than site 2. The pollution level varies with seasons. Seasonality correlates with temperature (axis 3).

The normed PCA mixes both the seasonal and the spatial typologies. One can decide to separate the two processes.

### 2.3 Usual one-way analysis of variance

To study the spatial (or temporal) effects, one can compute a one-way analysis of variance on each physico-chemical variable. For instance, one can study the relationship between the sites ( $g = 5$ ) and the temperature i.e. a comparison of  $g$  means.

The hypotheses of the one-way analysis of variance are:

$H_0$ : the  $g$  population means are equal.

$H_1$ : one mean is different from the others

under the assumptions that the variable follows a normal distribution in each group and that all the variances are equal.

```
options(show.signif.stars = F)
ressta <- anova(lm(meaudret$mil[, 1] ~ meaudret$plan$sta))
xtable(ressta)
```

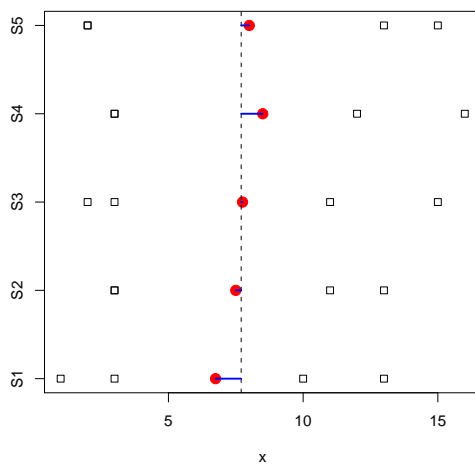
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
meaudret\$plan\$sta	4	6.70	1.67	0.04	0.9960
Residuals	15	573.50	38.23		

The one-way analysis of variance decomposes the **total** sum of the square distances to the general mean in two parts: 1- the **between** sum of square distances to the general mean [first row of the table], which is calculated as the sum of the square distances between the mean in each group and the general mean, and 2- the residuals or **within** sum of square distances [second row of the table], which is calculated as the sum of the square distances between each observation and the mean of the group where the considered observation belongs to.

```
sum((meaudret$mil[, 1] - mean(meaudret$mil[, 1]))^2)
[1] 580.2
6.7 + 573.5
[1] 580.2

graphnf <- function(x, gpe) {
  stripchart(x ~ gpe)
  points(tapply(x, gpe, mean), 1:length(levels(gpe)), col = "red",
         pch = 19, cex = 1.5)
  abline(v = mean(x), lty = 2)
  moyennes <- tapply(x, gpe, mean)
  traitnf <- function(n) segments(moyennes[n], n, mean(x), n,
                                col = "blue", lwd = 2)
  sapply(1:length(levels(gpe)), traitnf)
}
graphnf(meaudret$mil[, 1], meaudret$plan$sta)
```





The squares represent the individuals, the dotted line the general mean, the red points the means per group, the blue line the distance between a group mean and the global mean.

Let's call  $c$  the total sum of square distances and,  $b$  and  $w$  the between and within sums of square distances:

$$c = b + w$$

The statistical test of the one-way analysis of variance is based on the ratio

$$\frac{b/(g-1)}{w/(n-g)}$$

where  $n$  is the total number of individuals (i.e.  $n = 20$ ) and  $g$  is the number of groups (i.e.  $g = 5$ ).

For temperature, one can't say that means differ ( $p$ -value = 0.996).

One can compute all the one-way analyses of variance and obtain the  $p$ -values for each physico-chemical variable:

```
probasta <- rep(0, 9)
for (i in 1:9) {
  ressta <- anova(lm(meaudret$mil[, i] ~ meaudret$plan$sta))
  probasta[i] <- ressta[[1, 5]]
}
xtable(rbind(colnames(meaudret$mil), round(probasta, dig = 4)))
```

	1	2	3	4	5	6	7	8	9
1	Temp	Debit	pH	Condu	Dbo5	Oxyd	Ammo	Nitra	Phos
2	0.996	0.2366	0.3464	0.1464	0.0161	0.0022	0.0221	0.1012	0.0528

For seasonal effect and the temperature (first variable), one can obtain the following table:

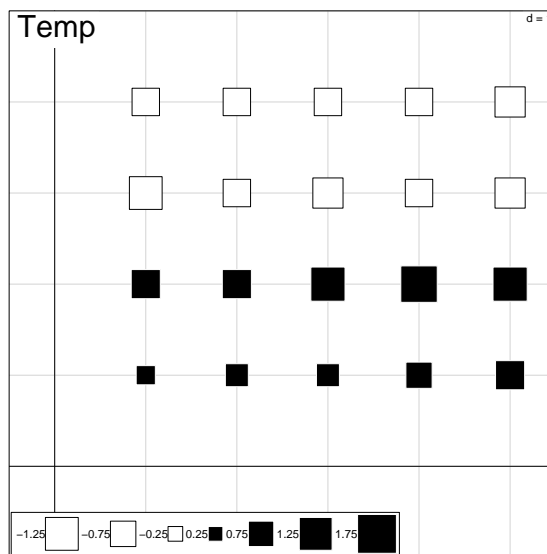
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
meaudret\$plan\$dat	3.0000	564.2000	188.0667	188.0667	0.0000
Residuals	16.0000	16.0000	1.0000		

and the results ( $p$ -values) for the all physico-chemical variables:

	1	2	3	4	5	6	7	8	9
1	Temp	Debit	pH	Condu	Dbo5	Oxyd	Ammo	Nitra	Phos
2	0	0.0032	0.0361	0.0179	0.5991	0.7795	0.3621	0.0795	0.1708

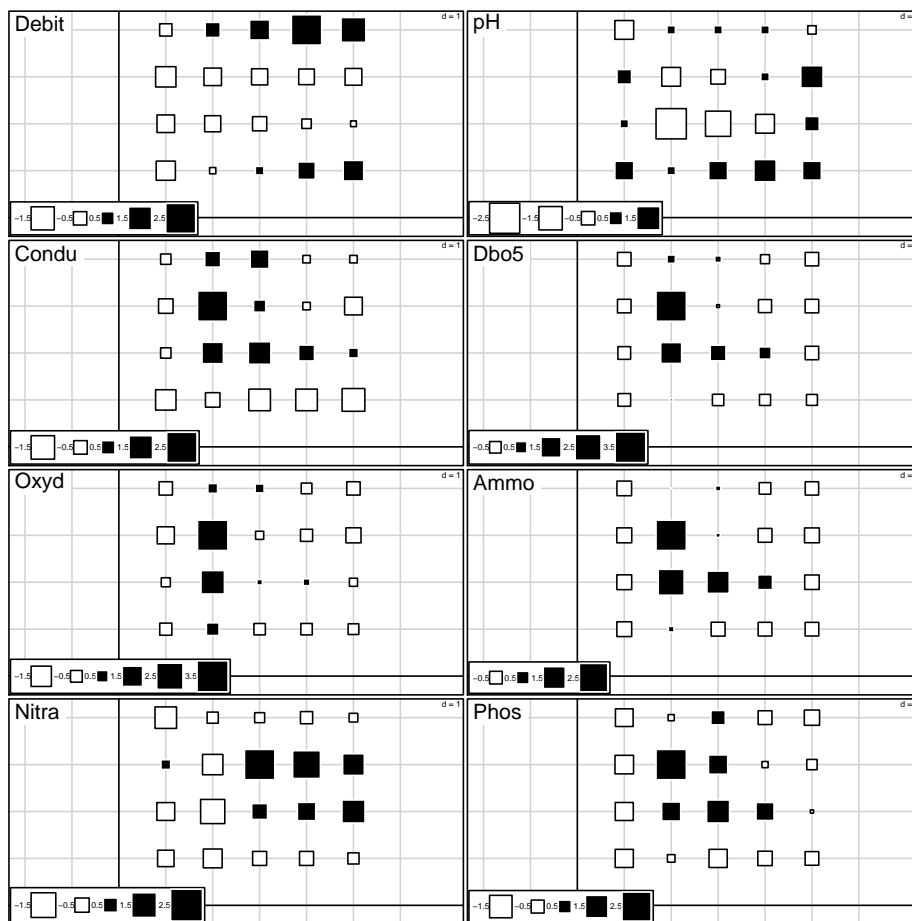
All these results can be visualized using the `s.value` function. The horizontal axis describe the 5 sites and, the vertical axis describes the seasons (from the bottom-spring to the top-winter).

```
intdat <- gl(4, 5, ordered = T)
intplan <- cbind(as.numeric(meaudret$plan$sta), intdat)
s.value(intplan, pca1$stab[, 1], sub = colnames(meaudret$mil)[1],
        possub = "topleft", csub = 2)
```



For the all physico-chemical variables:

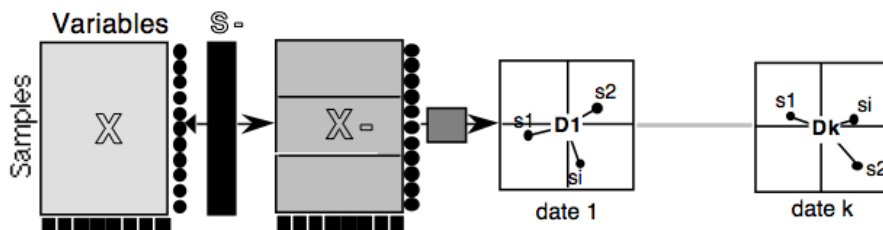
```
par(mfrow = c(4, 2))
for (i in 2:9) s.value(intplan, pca1$stab[, i], sub = colnames(meaudret$mil)[i],
                      possub = "topleft", csub = 2)
```



Eigenanalyses allow avoiding the study of each variable, one by one. The total inertia of the normed PCA can be decomposed in two inertia: one due to the within inertia and one due to the between inertia.

### 3 Removing an effect: the within PCA

The main aim of the within principal component analysis is to seek for possible seasonal (or spatial) typologies.



Let's call the statistical triplet of the classical principal component analysis

( $\mathbf{X}$ ,  $\mathbf{Q}$ ,  $\mathbf{D}$ ):

- $\mathbf{X}$  is the matrix containing the  $p$  variables measured on the  $n$  individuals,
- $\mathbf{Q}$  is the diagonal matrix of the variable weights,
- $\mathbf{D}$  is the diagonal matrix of the individual weights.

To carry out a within principal component analysis, one can add the group information i.e. the information of the categorical variable  $S$  with  $g$  modalities. In our example,  $S$  is the seasonal information divided in  $g = 4$  sites. That leads to change the  $\mathbf{X}$  matrix into  $\mathbf{X}-$  matrix.

$\mathbf{X}-$  is a matrix with  $p$  columns and  $n$  rows.

Let's call  $x_{ij}^k$  the value of the  $i$  individual for the  $j$  variable. The  $i$  individual belongs to the  $k$  modality of the  $S$  variable.

A value of the table  $\mathbf{X}-$  is then  $x_{ij}^k - \overline{x_j^k}$  where  $\overline{x_j^k}$  represents the mean of the  $j$  variable on the  $k$  group.

The means of each sub-population equal zero. All the group centers are therefore at the origin of the factorial maps and individuals are represented with the maximal variance around this origin (see maps `date 1 ... date k` on the previous figure).

In the example, the first row of the  $\mathbf{X}$  table analysed with the classical PCA contains 9 physico-chemical values, centered and normed:

```
pca1$tab[1, ]
      Temp      Debit      pH      Condu      Dbo5      Oxyd      Ammo
sp_1 0.4270257 -1.031019 0.9694584 -1.133227 -0.6055263 -0.5350257 -0.6815082
      Nitra      Phos
sp_1 -0.6636951 -0.977118
```

The first row is an individual belonging to the spring season. The means  $\overline{x_j^k}$  for all the 9 variables on this season are:

```
sepmil <- split(pca1$tab, meaudret$plan$dat)
names(sepmil)
[1] "autumn" "spring" "summer" "winter"
apply(sepmil$spring, 2, mean)
      Temp      Debit      pH      Condu      Dbo5      Oxyd      Ammo
0.6869543 0.2197512 0.8919017 -1.1332274 -0.3942822 -0.2645732 -0.4700992
      Nitra      Phos
-0.5622765 -0.6604030
meanspring <- apply(sepmil$spring, 2, mean)
```

The new data set  $\mathbf{X}-$  is therefore:

```
round(pca1$tab[1, ] - meanspring, digits = 4)
      Temp      Debit      pH      Condu      Dbo5      Oxyd      Ammo      Nitra      Phos
sp_1 -0.2599 -1.2508 0.0776      0 -0.2112 -0.2705 -0.2114 -0.1014 -0.3167
```

There is two ways of viewing the  $S$  variable: 1- by its modalities (one column) or 2- by defining a matrix with  $g$  columns and by using 0/1 to indicate presence/absence. Let's consider for instance a variable with 3 modalities observed on 7 individuals.

$$\begin{pmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 2 \\ 3 \\ 3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

From a Euclidean point of view, one can say that the  $\mathbf{X}$  matrix is projected onto the sub-space of the dummy variables (0/1) called  $A$ , so that  $\mathbf{X}$  is noted  $P_A(\mathbf{X})$ . The within principal component analysis is the analysis of the triplet  $(P_A(\mathbf{X}), \mathbf{Q}, \mathbf{D})$ .

To compute a within PCA in the `ade4` package, one uses the `within` function. Information on the classical principal component analysis (`pca1`) and the categorical variable `meaudret$plan$dat` need to be provided.

```
wit1 <- within(pca1, meaudret$plan$dat, scan = FALSE)
names(wit1)
[1] "tab" "cw" "lw" "eig" "rank" "nf" "c1" "li" "co" "l1"
[11] "call" "ratio" "ls" "as" "tabw" "fac"
```

The 11 first objects created by the `within` function are well-known (see course2 on the Principal Component Analysis). The other ones are:

- `ratio` the within group variance,
- `ls` the row projection of the normed PCA onto the new axes,
- `as` the column projection of the normed PCA onto the new axes,
- `tabw` the relative frequencies of the categorical variable,
- `fac` the categorical variable.

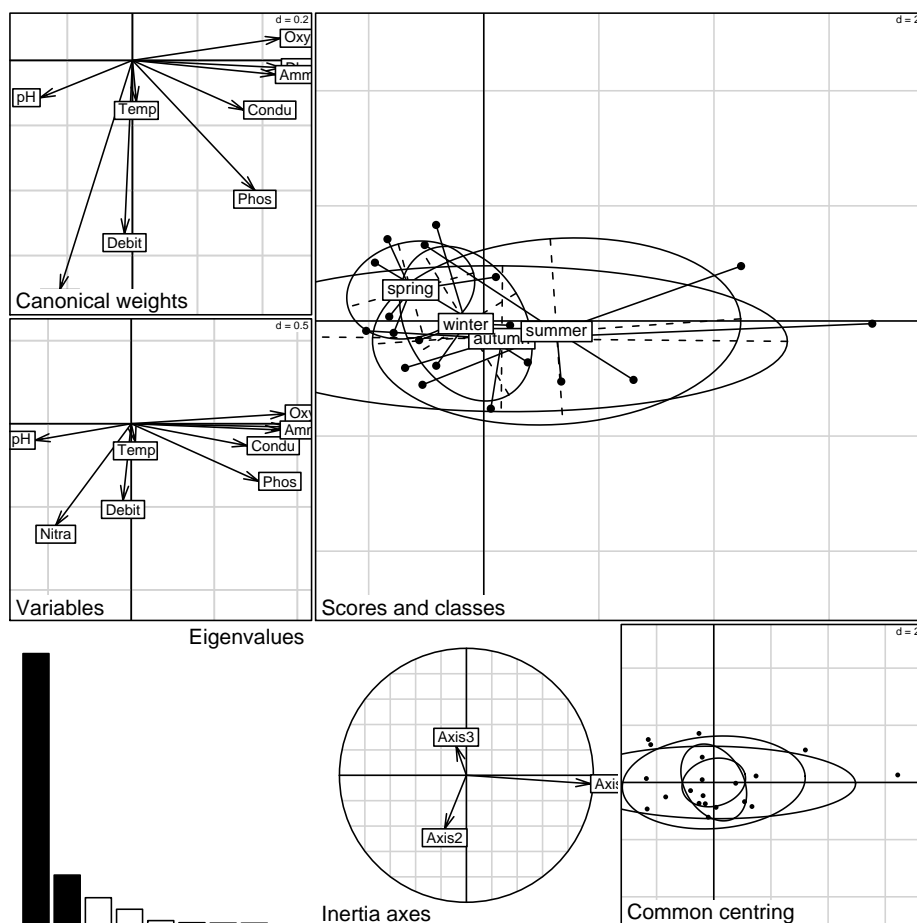
```
wit1$ratio
[1] 0.6277314
wit1$tabw
autumn spring summer winter
 0.25  0.25  0.25  0.25
wit1$fac
[1] spring spring spring spring spring summer summer summer summer summer autumn
[12] autumn autumn autumn autumn winter winter winter winter winter
Levels: autumn spring summer winter
```

`ls` and `as` are useful for more complex analyses such as the two-table or K-table analyses.

The results of the within principal component analysis can be displayed in terms of inertia i.e. explained variance using the `ratio` information. The inertia of the normed PCA of the 9 physico-chemical variables is equal to 9 (the total number of variables). The within group inertia is equal to 0.6277. In other words, 62.77% of the total inertia is due to the within PCA.

The results can be displayed using the global plot.

```
plot(wit1)
```



Six plots were built.

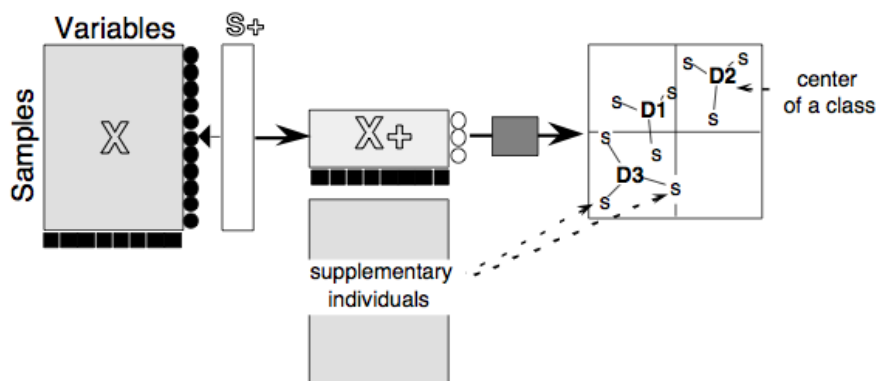
1. **Canonical weights** The first scatterplot (top left) represents the coefficients  $c1$  of the 9 variables on the two first axes of the within PCA.
2. **Variables** The second scatterplot (just below the first one) (co) represents the covariances between the 9 variables and the two first axes of the within PCA. The within PCA is meaningful IF AND ONLY IF the positions of the variables on the first plot is identical to the positions of the variables on the second plot.
3. **Eigenvalues** is the screeplot of the eigenvalues describing the contribution of each axis to the inertia.
4. **Scores and Classes** This plot shows the projections of the individuals onto the plane defined by the axes of the within PCA.
5. **Inertia axes** show the projection of the three axes kept by the normed PCA onto the two axes from the within PCA.
6. **Common centring** This plot shows the position of the individuals  $1i$  on the two first axes of the within PCA. One can see the common position centre expressed in the introduction.

To sum up the results of this analysis, one can say that during the spring (low pollution), the sites 1, 3, 4 and 5 are different from the site 2 which is still polluted. In summer, site 1 becomes different from sites 3, 4 and 5. In Autumn, the pollution grows and the site 2 becomes more different. During winter, sites 2 and 3 are always influenced by the proximity of the nearest village, Autrans.

**Exercise.** Redo the same analysis, but this time remove the site effect (instead of the season effect) from the set of physico-chemical variables.

## 4 Taking an effect into account: the between PCA

The main objective of the between PCA is to reveal the differences between groups.



Let's call the statistical triplet of the classical principal component analysis ( $\mathbf{X}$ ,  $\mathbf{Q}$ ,  $\mathbf{D}$ ):

- $\mathbf{X}$  is the matrix containing the  $p$  variables measured on the  $n$  individuals,
- $\mathbf{Q}$  is the diagonal matrix of the variable weights,
- $\mathbf{D}$  is the diagonal matrix of the individual weights.

To carry out a between principal component analysis, one can add the group information i.e. the information of the categorical variable  $S$  with  $g$  modalities. In our example,  $S$  remains the seasonal information divided in  $g = 4$  sites. That leads to change the  $\mathbf{X}$  matrix into  $\mathbf{X}^+$  matrix.

$\mathbf{X}^+$  is a matrix with  $p$  columns and  $g$  rows. That matrix contains in column the means of the  $p$  variables and in row, the groups. One can look for the dispersion between the gravity centres of the groups i.e. the means.

The weights of the columns-variables  $\mathbf{Q}$  don't change.

The weights of the rows-individuals  $\mathbf{D}$  become the relative frequencies of  $S$  i.e. the numbers of individuals per group divided by the total number.

The between principal component analysis is the analysis of the triplet  $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ .

To compute a between PCA in the `ade4` package, one uses the `between` function. Information regarding the classical principal component analysis (`pca1`) and the categorical variable `meaudret$plan$dat` need to be provided.

```
bet1 <- between(pca1, meaudret$plan$dat, scan = FALSE, nf = 2)
names(bet1)
[1] "tab" "cw" "lw" "eig" "rank" "nf" "l1" "co" "li" "c1"
[11] "call" "ratio" "ls" "as"
```

The means of all variables by group are in the dataframe `bet1$tab`:

```
bet1$tab
      Temp      Debit      pH      Condu      Dbo5      Oxyd      Ammo
autumn -1.0211483 -0.8435610  0.1163350  0.2658188  0.2778580  0.1352263  0.1864113
spring  0.6869543  0.2197512  0.8919017 -1.1332274 -0.3942822 -0.2645732 -0.4700992
summer  1.2439443 -0.5001506 -0.8919017  0.6575517  0.3666765  0.3468848  0.5943734
winter -0.9097503  1.1239604 -0.1163350  0.2098569 -0.2502522 -0.2175379 -0.3106854
      Nitra      Phos
autumn  0.9053099  0.4765227
spring -0.5622765 -0.6604030
summer  0.1058929  0.5217676
winter -0.4489263 -0.3378873
```

The weights of groups are the relative frequencies of the categorical variable.

```
bet1$lw
[1] 0.25 0.25 0.25 0.25
summary(meaudret$plan$dat)/length(meaudret$plan$dat)
autumn spring summer winter
 0.25  0.25  0.25  0.25
```

The between group variance is the inertia of the between PCA.

```
bet1$ratio
[1] 0.3722686
```

The number of eigenvalues is equal to the number of groups minus one  $g - 1$ .

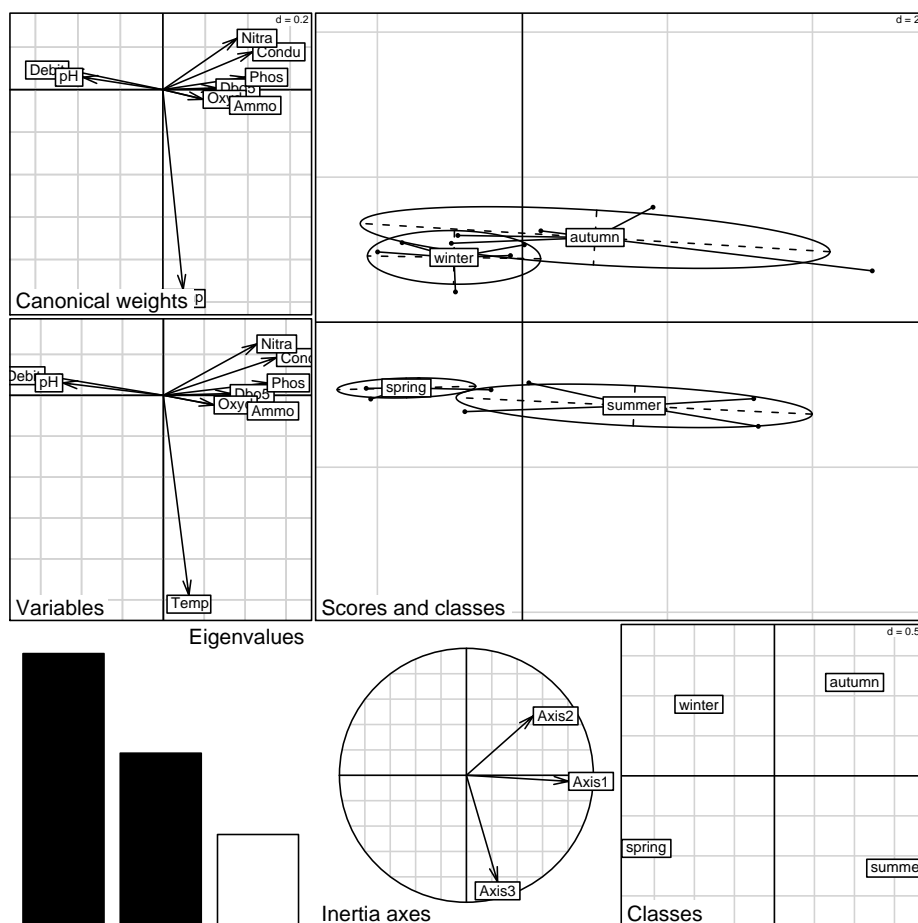
```
bet1$eig
[1] 1.7067496 1.0784279 0.5652401
```

In this analysis, the between inertia is equal to 0.3723 i.e. 37.23 % of the total inertia is due to the factor.

The results can be displayed using the global plot.

```
plot(bet1)
```





Six plots are built and their interpretations are the same than those described for the within PCA.

The first axis shows the pollution of the river: more important in summer and autumn. During winter and spring, the high flow of the river dilutes the organic pollution. The second axis describes the seasonal influence of the water temperature.

**Exercise.** Reveal the difference between sites from the set of physico-chemical variables.

## 5 Conclusion

The within and between PCA can be viewed as an exploratory generalisation of the one-way analysis of variance. The previous variance decomposition  $c = b + w$  is the inertia decomposition. For both PCA, the main objective is to maximize the projected variance.

Let's call:

-  $I_T$  the total inertia of  $\mathbf{X}$ ,

- $I_T^-$  the inertia of  $\mathbf{X}^-$  (within model: removing the effect of groups),
  - $I_T^+$  the inertia of  $\mathbf{X}^+$  (between model: revealing the effect of groups),
- one can obtain the relation:

$$I_T = I_T^+ + I_T^-$$

```
wit1$ratio
[1] 0.6277314
bet1$ratio
[1] 0.3722686
wit1$ratio + bet1$ratio
[1] 1
```

The within and between analyses, presented for the Principal Component Analysis, also exist for Correspondence Analysis in the `ade4` package. Introducing a categorical variable (i.e. a set of indicatrices) leads to the first two-table analysis. One begins to project a set of variables onto another one.

## References

- [1] S. Dolédec and D. Chessel. Recent developments in linear ordination methods for environmental sciences. *Advances in Ecology, India*, 1:133–155, 1991.
- [2] D. Pegaz-Maucet. *Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés benthiques d'un cours d'eau. Comparaison avec le benthos*. PhD thesis, 1980.