


Fiche TD avec le logiciel  : course6

Coinertia Analysis

A.B. Dufour

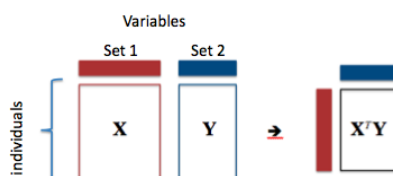
Contents

1	Introduction	2
2	Principle	2
2.1	Remembering the relationship between two variables	2
2.2	Defining the relationship between two data tables	3
3	Coinertia between two principal component analyses	5
4	Coinertia between a PCA and a Correspondence Analysis (CoA)	7
5	Your turn!	9
6	Conclusion	10
	References	11

1 Introduction

The study of the relationship between fauna (or flora) and its environment usually leads to two sets of data: (i) a faunistic array that contains the abundance or the occurrence of a number of taxa in a set of sites; and (ii) an environmental array that includes quantitative or categorical measurements from the same sites. [1]

There are several strategies to match two data tables. We have already seen three strategies, namely the within and between principal components analyses and the linear discriminant analysis. The diagram below shows a new strategy. If two tables are linked by the same individuals, one can find a structure, a **co** structure to study the relationship between the two set of variables (red and blue ones).



2 Principle

2.1 Remembering the relationship between two variables

Let's call X and Y two continuous variables measured on the same individuals. Let's call \bar{x} and \bar{y} the means of X and Y respectively.

Let's call $v(x)$ and $v(y)$ the descriptive variances of X and Y respectively.

A measure of the relationship between X and Y is provided by the descriptive covariance:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The covariance can be negative or positive: that depends on the sense of the relationship.

All the variance-covariance information can be gathered in a matrix (which can be called the covariance matrix).

$$\begin{pmatrix} v(x) & \text{cov}(x, y) \\ \text{cov}(y, x) & v(y) \end{pmatrix}$$

This matrix is symmetric: $\text{cov}(x, y) = \text{cov}(y, x)$ and $\text{cov}(x, x) = v(x)$.

One can divide the covariance between X and Y by the square roots of the variances of X and Y (i.e., by the standard deviations of X and Y). By doing so, we obtain the coefficient of correlation between X and Y .

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{v(x)}\sqrt{v(y)}}$$

The closer the coefficient to either -1 or 1 , the stronger the correlation between the two variables.

2.2 Defining the relationship between two data tables

Let's call $(\mathbf{X}, \mathbf{Q}_X, \mathbf{D})$ a statistical triplet, where \mathbf{X} is a dataset containing p variables measured on n individuals. \mathbf{Q}_X and \mathbf{D} represent the weights of the p variables and n individuals, respectively. If all the variables of \mathbf{X} are centred, the inertia I_X of $\mathbf{XQ}_X\mathbf{X}^T\mathbf{D}$ is the sum of the variances.

$\mathbf{XQ}_X\mathbf{X}^T\mathbf{D}$ is the covariance matrix. The first diagonal contains all the variances and is called the *trace*. We can write the following relationship:

$$I_X = \text{trace}(\mathbf{XQ}_X\mathbf{X}^T\mathbf{D})$$

Let's call $(\mathbf{Y}, \mathbf{Q}_Y, \mathbf{D})$ another statistical triplet, where \mathbf{Y} is a dataset containing q variables measured on the same n individuals. \mathbf{Q}_Y and \mathbf{D} represent the weights of the q variables and n individuals, respectively. If all the variables of \mathbf{Y} are centred, the inertia I_Y of $\mathbf{YQ}_Y\mathbf{Y}^T\mathbf{D}$ is also the sum of the variances.

In that case, $\mathbf{YQ}_Y\mathbf{Y}^T\mathbf{D}$ is the covariance matrix. Again, the first diagonal contains all the variances (the *trace*). We can write the following relationship:

$$I_Y = \text{trace}(\mathbf{YQ}_Y\mathbf{Y}^T\mathbf{D})$$

The following figure from Dray *et al*'s paper [2] explains the process of the coinertia analysis.

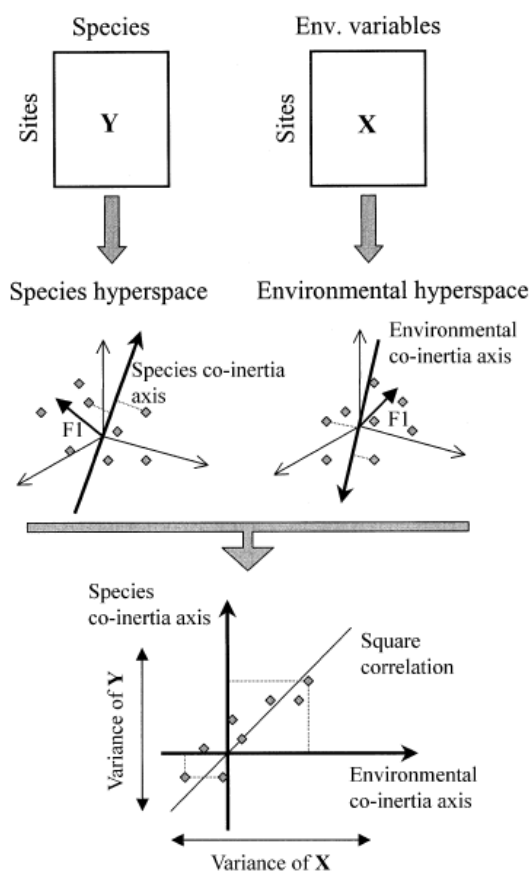


FIG. 1. Principles of co-inertia analysis (COIA). The two ecological data tables X and Y produce two representations of the sites in two hyperspaces. Separate analyses find axes maximizing inertia in each hyperspace (F1 [first factorial axis]). COIA aims to find a couple of co-inertia axes on which the sites are projected. COIA maximizes the square covariance between the projections of the sites on the co-inertia axes.

The coinertia between two hyperspaces is the sum of squares of the covariances between all the variables pairs:

$$S = \text{trace}(\mathbf{XQ}_X \mathbf{X}^T \mathbf{D} \mathbf{YQ}_Y \mathbf{Y}^T \mathbf{D})$$

$$RV = \frac{\text{coinertia}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{coinertia}(\mathbf{X}, \mathbf{X})} \sqrt{\text{coinertia}(\mathbf{Y}, \mathbf{Y})}}$$

The RV -coefficient is the coefficient of correlation between the two tables X and Y . This coefficient varies between 0 and 1: the closer the coefficient to 1, the stronger the correlation between the tables.

3 Coinertia between two principal component analyses

The dataset describes the physico-chemical characteristics (mil) and the abundance of several fish species (poi) along the Doubs river (Burgundy, France). xy contains the latitude/longitude of the sites sampled.

```
data(doubs)
names(doubs)
[1] "mil" "poi" "xy"
```

The first dataset contains the environmental variables measured (mil). One can compute a normed principal component analysis on this first dataset (the PCA is normed in this case since the variables measured have different units).

```
pcamil <- dudi.pca(doubs$mil, scale = TRUE, scan = FALSE, nf = 3)
```

The second dataset contains the abundance of fish. One can compute a centred principal component analysis on this second dataset (the PCA is only centred here since this dataset only deals with abundance).

```
pcafau <- dudi.pca(doubs$poi, scale = FALSE, scan = FALSE, nf = 2)
```

The relationship between the environmental dataset and the species abundance dataset is provided by the coinertia analysis.

```
coin1 <- coinertia(pcamil, pcafau, scan = FALSE, nf = 2)
names(coin1)
[1] "tab" "cw" "lw" "eig" "rank" "nf" "c1" "li" "co" "l1" "call"
[12] "lX" "mX" "lY" "mY" "aX" "aY" "RV"
```

```
coin1
Coinertia analysis
call: coinertia(dudiX = pcamil, dudiY = pcafau, scannf = FALSE, nf = 2)
class: coinertia dudi
$rank (rank) : 11
$nf (axis saved) : 2
$RV (RV coeff) : 0.4505569

eigen values: 119 13.87 0.7566 0.5278 0.2709 ...

vector length mode content
1 $eig 11 numeric eigen values
2 $lw 27 numeric row weights (crossed array)
3 $cw 11 numeric col weights (crossed array)

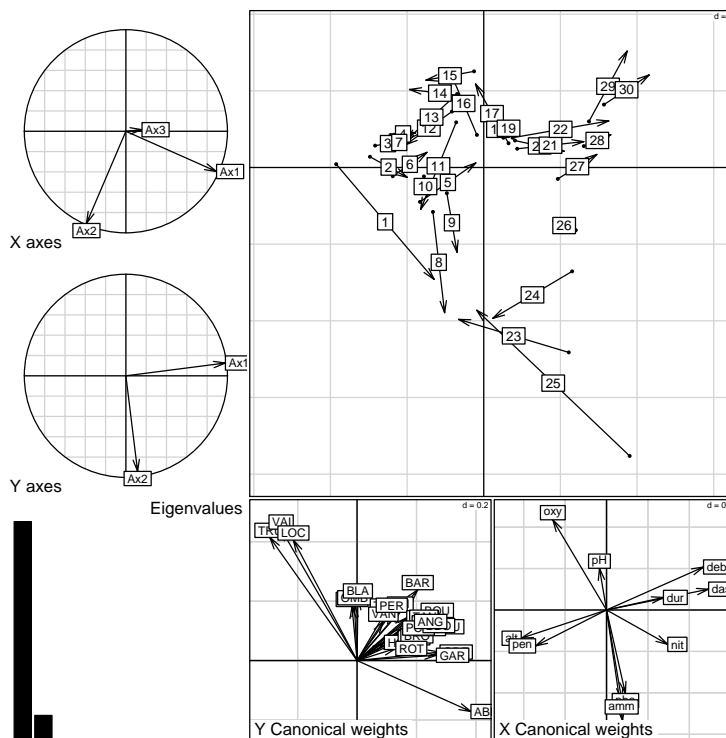
data.frame nrow ncol content
1 $tab 27 11 crossed array (CA)
2 $li 27 2 Y col = CA row: coordinates
3 $l1 27 2 Y col = CA row: normed scores
4 $co 11 2 X col = CA column: coordinates
5 $c1 11 2 X col = CA column: normed scores
6 $lX 30 2 row coordinates (X)
7 $mX 30 2 normed row scores (X)
8 $lY 30 2 row coordinates (Y)
9 $mY 30 2 normed row scores (Y)
10 $aX 3 2 axis onto co-inertia axis (X)
11 $aY 2 2 axis onto co-inertia axis (Y)

summary(coin1)
Eigenvalues decomposition:
 eig covar sdX sdY corr
1 119.01942 10.909602 2.326324 6.422570 0.7301798
2 13.87137 3.724429 1.685078 2.863743 0.7718017
Inertia & coinertia X:
 inertia max ratio
1 5.411785 6.321624 0.8560752
```

```
12 8.251272 8.553220 0.9646978
```

```
Inertia & coinertia Y:
  inertia  max  ratio
1  41.24940 42.74627 0.9649824
12 49.45042 50.90461 0.9714331
```

```
RV:
0.4505569
plot(coini)
```

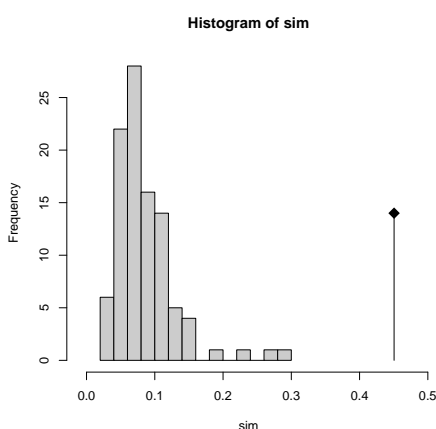


1. **X axes.** This correlation circle shows the projections of the PCA axes (from the environmental data) onto the axes of the coinertia analysis.
2. **Y axes.** This correlation circle shows the projections of the PCA axes (from the abundance data) onto the axes of the coinertia analysis. These two circles represent a view of the rotation needed to associate the two datasets.
3. **Eigenvalues.** This screeplot gives the eigenvalues of the coinertia analysis.
4. **Canonical weights.** These two scatter plots represent the coefficients of the combinations of the variables for each table to define the coinertia axes.
5. The last scatter plot with arrows is specific to the coinertia analysis, and represents the individuals (in this case, the sites sampled). The beginning of the arrow is the position of the site described by the environmental data set; the end of the arrow is the position of the site described by the abundance. For example, sites 23 to 26 are apart from the other

sites. Their being apart is linked to a pollution effect (see the X canonical weights graph). In these sites (23-26), there are also less species (see the Y canonical weights graph). Interestingly, the abundance of fish in 23-26 is similar to the abundance of fish in sites 1, 8 and 9.

We can perform a permutation test to study the strength of the relationship between the two tables, i.e. the significance of the RV coefficient. As it can be seen, this coefficient is different from what could be expected by chance.

```
rv1 <- RV.rtest(pcamil$tab, pcafauf$tab, 99)
plot(rv1)
```



4 Coinertia between a PCA and a Correspondence Analysis (CoA)

An abundance table can be viewed as a quantitative information (detailing the number of fish per site and per species) or can be viewed as a contingency table (detailing the relative abundance (

As before, the dataset (poi) contains species abundance. We now want to work with the relative frequency of each species in each site: because of this, we perform a correspondence analysis on this first dataset (instead of a PCA, as previously done):

```
coafau <- dudi.coa(doubs$poi, scannf = F, nf = 2)
```

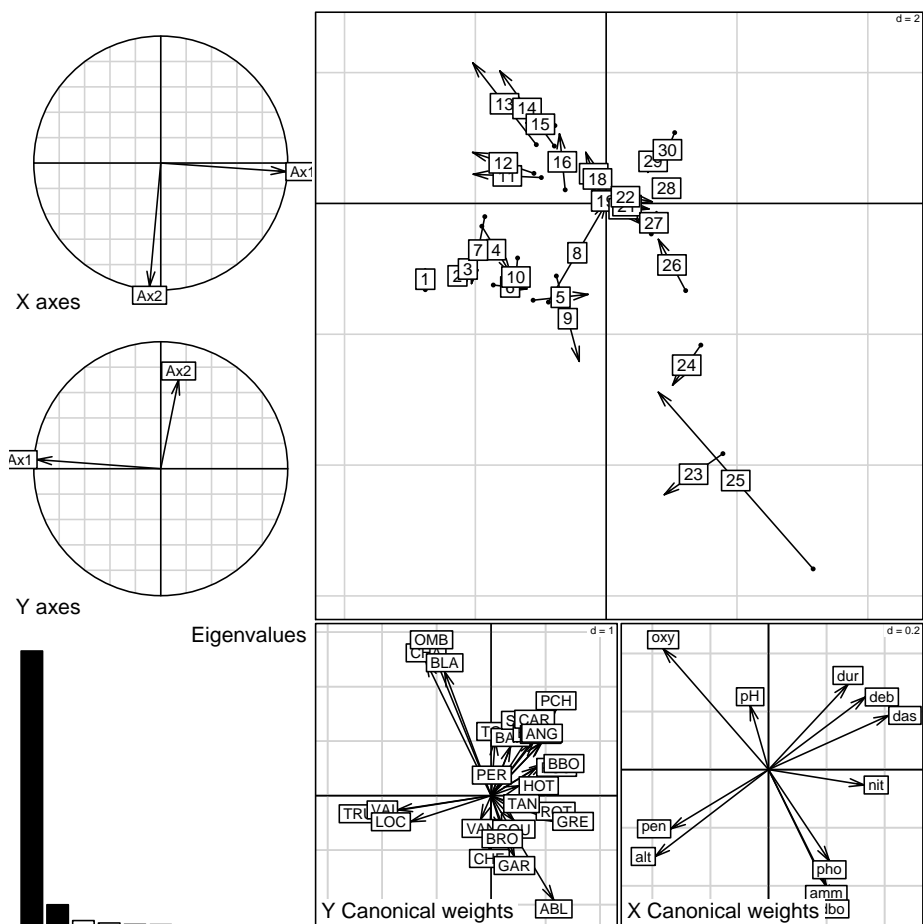
As before, the dataset (mil) contains the environmental variables. As previously done, we can compute a normed principal component analysis. However, because we decided to work with frequencies instead of row numbers of fish, we removed the information linked to some sites being richer (in terms of abundance) than others. To take into account this information, we decide to weight each site by its abundance of fish (contained in coafau\$w).

```
pcamil <- dudi.pca(doubs$mil, row.w = coafau$w, scannf = F, nf = 2)
```

The relationship between the environmental data and the species frequency is provided by the following coinertia analysis.

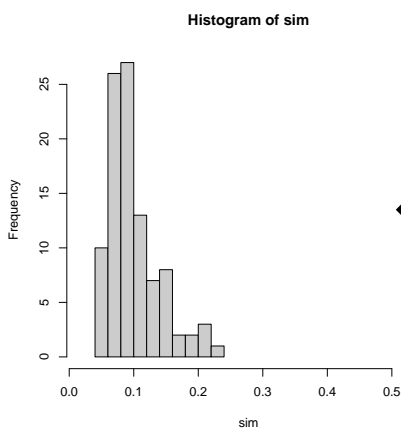
```
coin2 <- coinertia(pcamil, coafau, scannf = F, nf = 2)
```

```
plot(coin2)
```



Again, we can perform the permutation test on the RV coefficient.

```
rv2 <- RV.rtest(pcamil$tab, coafau$tab, 99)
plot(rv2)
```

5 Your turn!

The `trichometeo` dataset contains two dataframes and a factor. Insects were collected in 1959 and 1960 by J. Fontaine using light traps. The traps, located on boats, attract adult insects emerging from the river Rhône in Lyon.

1. `trichometeo$meteo` is a dataframe with 49 rows (corresponding to 49 nights) and 11 meteorological variables registered for each trapping night.

The 11 variables are:

1	T.max	Max temperature registered during day j (in Celcius degree)
2	T.soir	Crepuscular temperature = $[T18h(j) + T21h(j)]/2$ (in Celcius degree)
3	T.min	Min temperature registered during day $j + 1$ (in Celcius degree)
4	Vent	Wind Speed $[V18h(j) + T21h(j)]/2$ (in m/s)
5	Average pressure	$[P12h(j) + P18h(j) + P0h(j + 1) + P6h(j + 1)]/4$ (in mm Hg)
6	Var.Pression	Pressure variation $P6h(j + 1) - P6h(j)$ (in mm Hg)
7	Humidity	Humidity ratio $HR21h(j)$ (in %)
8	Nebu.Nuit	% of cover mean of the % at 21h (j) and 0h ($j + 1$)
9	Precip.Nuit	Night rainfall $H6h(j + 1) - H18h(j)$ (in mm)
10	Nebu.Moy	Average percentage of cover (in %)
11	Precip.Tot	Total rainfall $H6h(j + 1) - H6h(j)$ (in mm)

2. The numbers of insects captured for each species can be found in `trichometeo$fau`, and the names of the species trapped are:

1	Che	<i>Cheumatopsyche lepida</i>
2	Hyc	<i>Hydropsyche contubernalis</i>
3	Hym	<i>Hydropsyche modesta</i>
4	Hys	<i>Hydropsyche siltalai</i>
5	Psy	<i>Psychomyia pusilla</i>
6	Aga	<i>Agapetus laniger</i>
7	Glo	<i>Glossosoma boltoni</i>
8	Ath	<i>Athripsodes albifrons</i>
9	Ceam	<i>Ceraclea alboguttata</i>
10	Ced	<i>Ceraclea dissimilis</i>
11	Set	<i>Setodes punctatus</i>
12	All	<i>Allotrichia pallicornis</i>
13	Han	<i>Hydroptila angulata</i>
14	Hfo	<i>Hydroptila forcipata</i>
15	Hspm	<i>Hydroptila sparsa</i>
16	Hve	<i>Hydroptila vectis</i>
17	Sta	<i>Stactobiella risi</i>

3. `trichometeo$cla` contains a vector describing the experimental design (row information).

1	1-12	12 consecutive nights in June 59
2	13-17	5 consecutive nights in June 59
3	18-22	5 consecutive nights in June 59
4	23-26	4 consecutive nights in June 59
5	27-30	4 consecutive nights in July 59
6	31	1 night in June 60
7	32-34	3 consecutive nights in June 60
8	35-38	4 consecutive nights in June 60
9	39-43	5 consecutive nights in June 60
10	44-47	4 consecutive nights in June 60
11	48	1 night in June 60
12	49	1 night in June 60

What is the relationship between the meteorological variables and the insects captured ?

NB: To change the names of the variables in `trichometeo`, use the following command:

```
names(trichometeo$meteo) <- c("maxtemp", "eveningtemp", "mintemp",
  "windspeed", "pressure", "pressure.var", "humidity", "nightrecover",
  "night.rainfall", "average.recov", "total.rainfall")
```

6 Conclusion

A coinertia analysis is the match between two tables and their associated triplets. Each triplet may be a principal component analysis, a correspondence analysis, a multiple correspondence analysis. The choice depends on the biological question. See Dray *et al* [2] for a review on all coinertia analyses and their links with other methods found in the literature.

Matching two tables is a complex procedure and there are several ways to proceed. When possible, the best is to try different approaches, as this can sometimes lead to improved knowledge of the system.

Experience shows that in many cases different methods will give similar results, but that in particular situations the results of a study can greatly depend on the choice of the multivariate method.

References

- [1] S. Dolédec and D. Chessel. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, 31:277–294, 1994.
- [2] S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological tables. *Ecology*, 84(11):3078–3089, 2003.