Multiblock redundancy analysis

Application to drug use on farms

S. Bougeard[(1)], F. Laanaya-Tazani[(1,2)], S. Le Bouquin[(1)] & C. Chauvin[(1)]

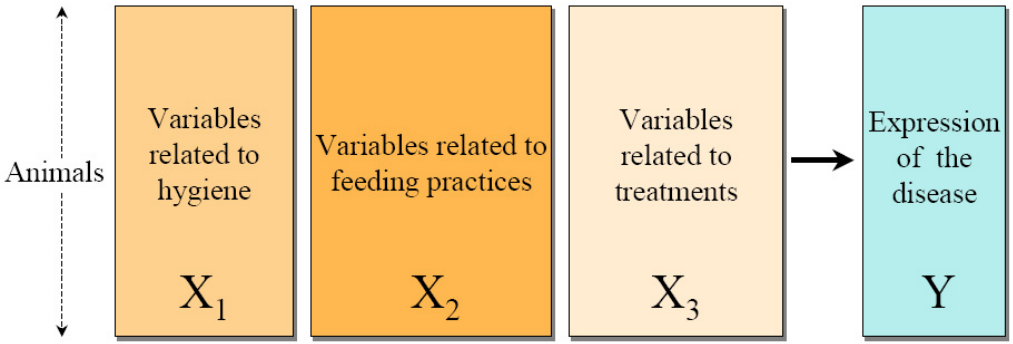
[(1)] *French Agency for food, environmental and occupational health & safety (Anses), Department of epidemiology, Ploufragan, France*

[(2)] *University of Rennes 2, Master of applied statistics, France*

anses
French agency for food, environmental and occupational health safety

Table of contents

Characteristics of epidemiological data



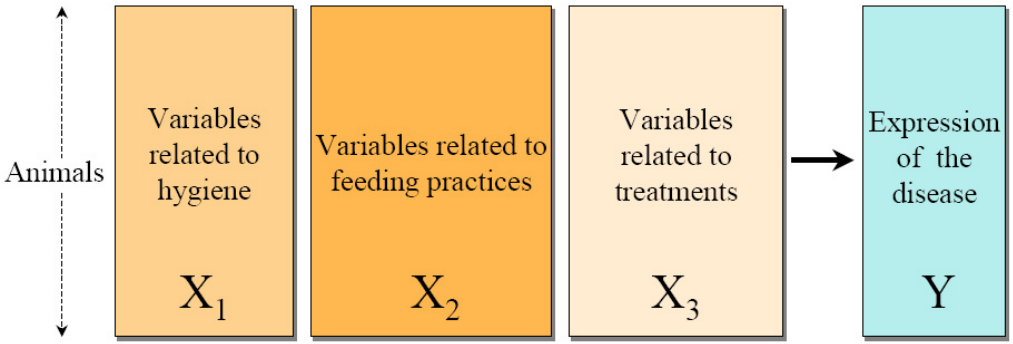
Epidemiological data features

- Large number of explanatory variables organized in meaningful blocks,
- Structural multicollinearity among explanatory variables,
- Several variables to explain.

Aims

- **Descriptive** : investigate the relationships between variables, variable blocks and individuals,
- **Predictive** : assess the risk factors for the disease, at the variable and at the block levels.

Characteristics of epidemiological data



Epidemiological data features

- Large number of explanatory variables organized in meaningful blocks,
- Structural multicollinearity among explanatory variables,
- Several variables to explain.

Aims

- **Descriptive** : investigate the relationships between variables, variable blocks and individuals,
- **Predictive** : assess the risk factors for the disease, at the variable and at the block levels.

Aims of the talk

From a methodological point of view

- Presentation of a multiblock modeling method,
- Development of useful associated interpretation tools.

From an applicative point of view

- Interpreted example from the veterinary epidemiological field,
- Associated available code program in R.

Aims of the talk

From a methodological point of view

- Presentation of a multiblock modeling method,
- Development of useful associated interpretation tools.

From an applicative point of view

- Interpreted example from the veterinary epidemiological field,
- Associated available code program in `R`.

Table of contents

Optimization problem

Aims

- Relate the partial components of each explanatory block t_k with the one of the dependent block u,
- Seek a global component t which reflects the explanatory ones t_k,
- Choose norm constraints to orient the method toward the Y explanation.

Criterion to maximize

Max. $\sum_k cov^2(u^{(1)}, t_k^{(1)})$
with $u^{(1)} = Yv^{(1)}$, $t_k^{(1)} = X_k w_k^{(1)}$
$t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$, $\sum_k a_k^{(1)^2} = 1$
and $||v^{(1)}|| = ||t_k^{(1)}|| = 1$

Aims

- Explain Y with all the explanatory variables,
- Improve the prediction ability (orthogonalized regression).

Higher order solutions

Residuals of the orthogonal projections of X_k onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)}), \ldots$

Optimization problem

Aims

- Relate the partial components of each explanatory block t_k with the one of the dependent block u,
- Seek a global component t which reflects the explanatory ones t_k,
- Choose norm constraints to orient the method toward the Y explanation.

Criterion to maximize

Max. $\sum_k cov^2(u^{(1)}, t_k^{(1)})$
with $u^{(1)} = Yv^{(1)}$, $t_k^{(1)} = X_k w_k^{(1)}$
$t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$, $\sum_k a_k^{(1)^2} = 1$
and $||v^{(1)}|| = ||t_k^{(1)}|| = 1$

Aims

- Explain Y with all the explanatory variables,
- Improve the prediction ability (orthogonalized regression).

Higher order solutions

Residuals of the orthogonal projections of X_k onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)}), \ldots$

Direct solutions

Solution

- $v^{(1)}$ is the eigenvector of $\sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$
- $t_k^{(1)} = P_{X_k} u^{(1)} / ||P_{X_k} u^{(1)}||$
- $t^{(1)} = \sum_k \frac{cov(u^{(1)}, t_k^{(1)})}{\sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})}} t_k^{(1)}$

with the projector $P_{X_k} = X_k (X_k' X_k)^{-1} X_k'$

Interpretation

- The explanatory multiblock structure is taken into account,
- The partial explanatory components t_k are derived from the projection of the dependent component u onto each X_k space,
- The more the partial components u and t_k are linked, the more they build the global component t.

Main interpretation tools

From a descriptive point of view : factorial graphical displays.

Aims

- Link all the explanatory variables with all the dependent ones,
- Sort the explanatory variables by order of priority,
- Sort the explanatory blocks by order of priority.

Predictive interpretation tools

- $\beta_{p,q}^{(1\rightarrow h_{opt})} = \sum_{h=1}^{h_{opt}} w^{(h)} * c^{(h)'}$
- $VarImp_{p}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} \dfrac{a_k^{(h)2} w_{[p]}^{(h)2}}{\sum_{p=1}^{P} a_k^{(h)2} w_{[p]}^{(h)2}}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$
- $BlockImp_{k}^{(1\rightarrow h_{opt})} = \dfrac{\sum_{h=1}^{h_{opt}} \lambda^{(h)} a_k^{(h)2}}{\sum_{h=1}^{h_{opt}} \lambda^{(h)}}$

Table of contents

Data and aims

Multiblock data and aims

- Explain the drug use on farms (Y) : 2 variables related to the ways of drug administration, *i.e.* food/water or injection.
- 177 explanatory var. $\rightarrow$ select. of 27 potential risk factors organized in 4 blocks
  - X_1 : Management and hygiene practices (8 var.)
  - X_2 : Sanitary problems (7 var.)
  - X_3 : Farm structure (5 var.)
  - X_4 : Therapeutic practices (7 var.)
- 112 randomly selected French farms.

Descriptive interpretation [1]

R code to process multiblock method (ade4)

```
resmbra_DrugU <- mbpcaiv(dudiY, ktabX, scale=T, option="uniform")
summary(resmbra_DrugU)
```

Inertia and explained variances of the various datasets by the global components

	Eig	%Iner	%Iner+	%VarYT	%VarYT+	%VarXT	%VarXT+
1	44.1	53.5	53.5	51.5	51.5	4.92	4.92
2	29.6	35.9	89.4	35.5	87	5.14	10.1
3	4.01	4.87	94.3	6.78	93.8	4.98	15
4	2.46	2.99	97.3	3.07	96.9	4.24	19.3
5	1.35	1.63	98.9	2.28	99.2	4.16	23.4
6	0.5	0.607	99.5	0.453	99.6	4.57	28
7	0.233	0.283	99.8	0.236	99.8	3.9	31.9
8	0.095	0.115	99.9	0.0899	99.9	3.5	35.4
9	0.0331	0.0402	100	0.0421	100	4.22	39.6
10	0.0117	0.0143	100	0.0125	100	3.12	42.8

	X1.%VarXkT	X1.%VarXkT+	X2.%VarXkT	X2.%VarXkT+	X3.%VarXkT	X3.%VarXkT+	X4.%VarXkT	X4.%VarXkT+
1	4.92	4.92	6.47	6.47	3.11	3.11	5.19	5.19
2	6.36	11.3	4.96	11.4	4.86	7.97	4.37	9.56
3	7.99	19.3	4.35	15.8	3.95	11.9	3.64	13.2
4	4.64	23.9	7.61	23.4	3.41	15.3	1.31	14.5
5	3.49	27.4	4.11	27.5	5.4	20.7	3.63	18.1
6	5.91	33.3	4.8	32.3	3.56	24.3	4.01	22.2
7	1.99	35.3	3.86	36.2	5.14	29.4	4.62	26.8
8	1.77	37.1	2.58	38.7	7.18	36.6	2.48	29.3
9	7.9	45	2.19	40.9	4.47	41.1	2.33	31.6
10	3.15	48.1	4.4	45.3	1.06	42.1	3.86	35.5

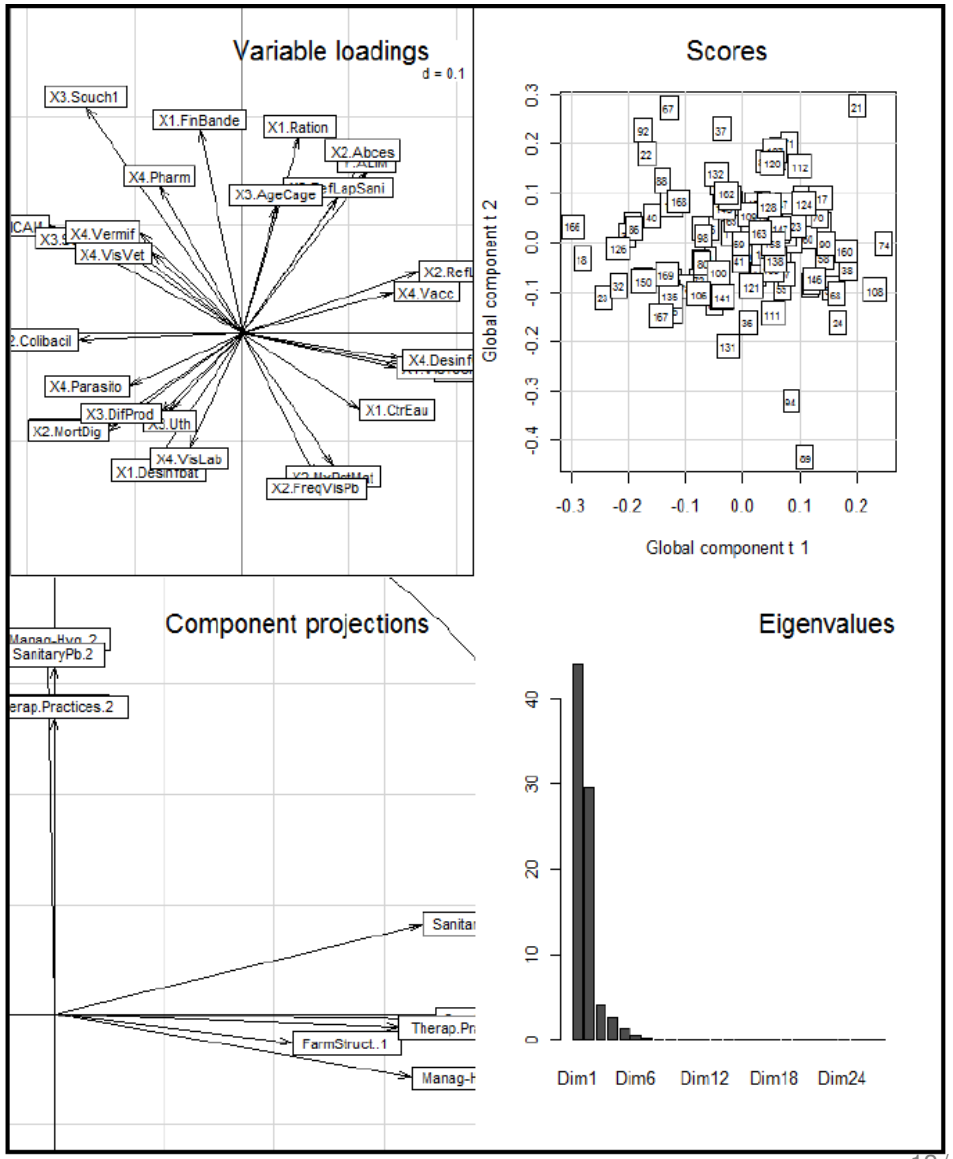
Descriptive interpretation [2]

R code to get graphical displays (ade4)

```
plot(resmbra_DrugU)
```

Interpretation

- The two dependent variables related to the ways of drug administration (food/water and injection) are not linked with each others.
- Each of them is linked with some specific risk factors.
- The first component related to drug use $u^{(1)}$ is mainly linked to the components related to the management and hygiene practices $t_1^{(1)}$, the farm structure $t_3^{(1)}$ and the therapeutic practices $t_4^{(1)}$.



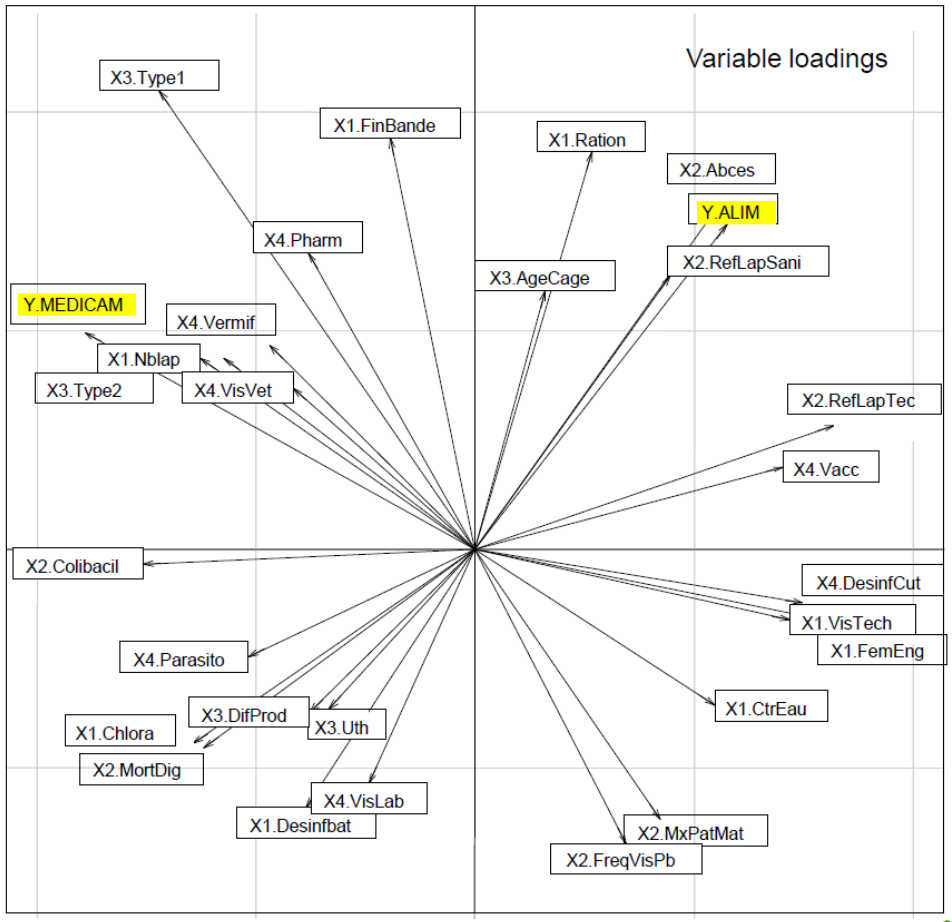
Descriptive interpretation [2]

R code to get graphical displays (ade4)

```
plot(resmbra_DrugU)
```

Interpretation

- The two dependent variables related to the ways of drug administration (food/water and injection) are not linked with each others.
- Each of them is linked with some specific risk factors.
- The first component related to drug use $u^{(1)}$ is mainly linked to the components related to the management and hygiene practices $t_1^{(1)}$, the farm structure $t_3^{(1)}$ and the therapeutic practices $t_4^{(1)}$.



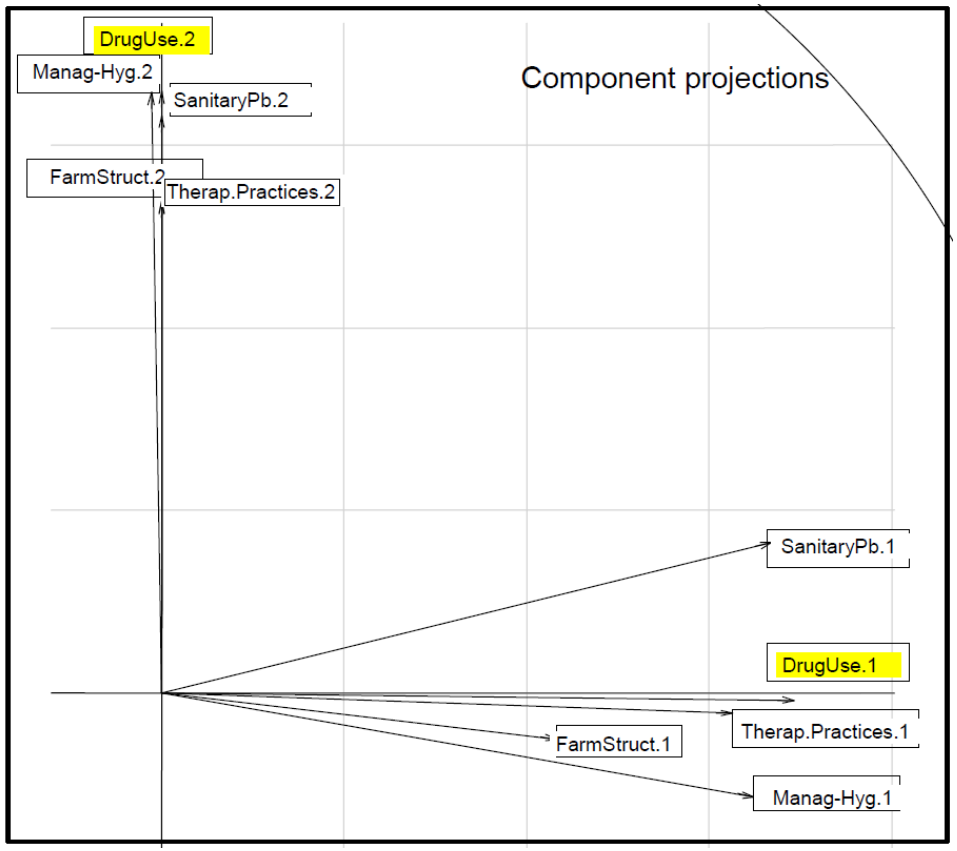
Descriptive interpretation [2]

R code to get graphical displays (ade4)

```
plot(resmbra_DrugU)
```

Interpretation

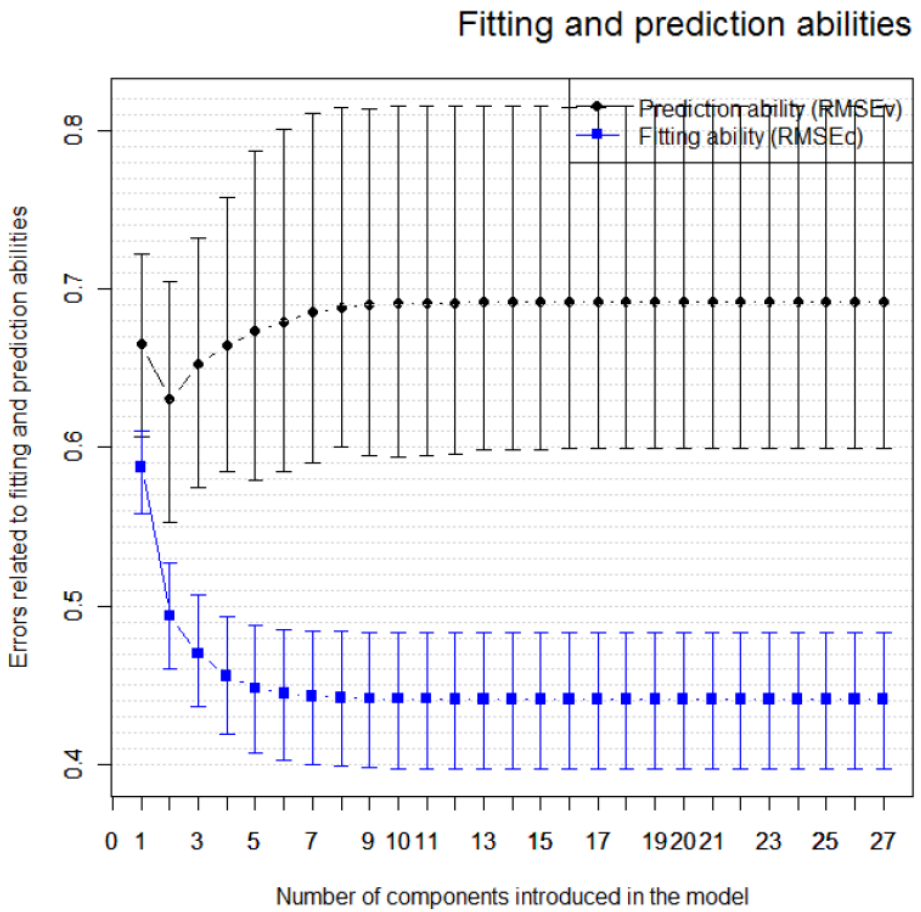
- The two dependent variables related to the ways of drug administration (food/water and injection) are not linked with each others.
- Each of them is linked with some specific risk factors.
- The first component related to drug use $u^{(1)}$ is mainly linked to the components related to the management and hygiene practices $t_1^{(1)}$, the farm structure $t_3^{(1)}$ and the therapeutic practices $t_4^{(1)}$.



Selection of the optimal number of components to introduce in the model

R code for cross-validation and graphical displays (ade4)

```
testdim_mbra_DrugU <- testdim.multiblock(resmbra_DrugU, 200)
plot(testdim_mbra_DrugU)
```



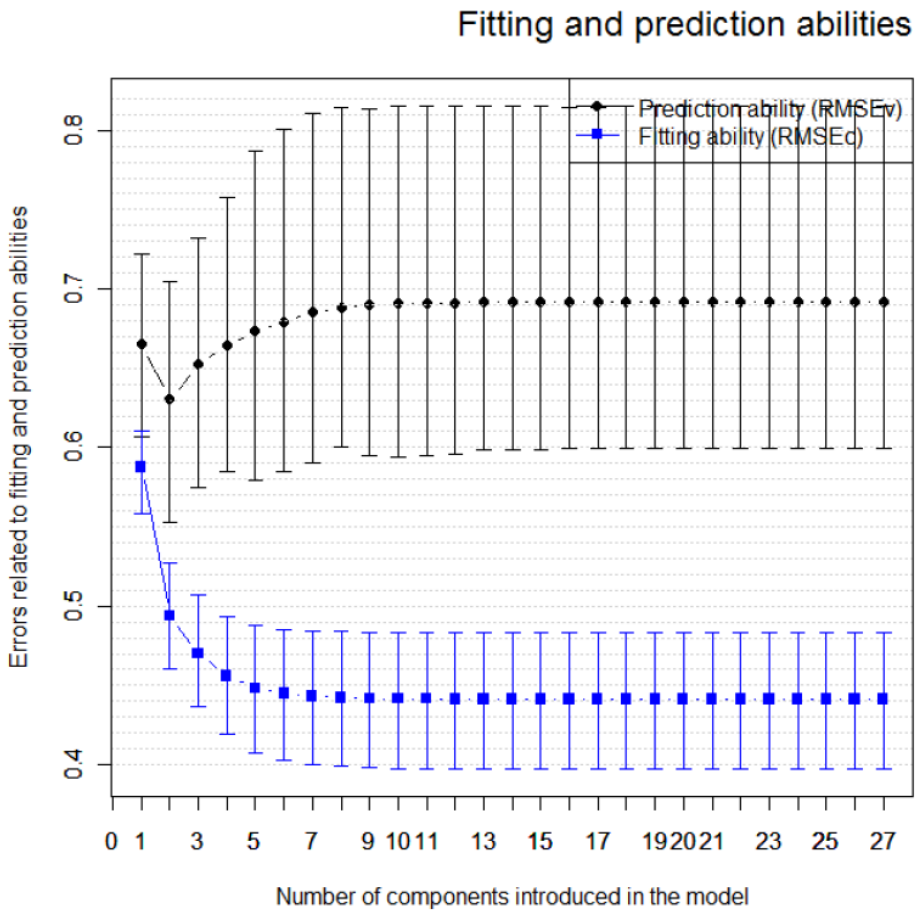
Interpretation

Selection of two components to be introduced in the model.

Selection of the optimal number of components to introduce in the model

R code for cross-validation and graphical displays (ade4)

```
testdim_mbra_DrugU <- testdim.multiblock(resmbra_DrugU, 200)
plot(testdim_mbra_DrugU)
```



Interpretation

Selection of two components to be introduced in the model.

Interpretation of the optimal model at the variable and block levels

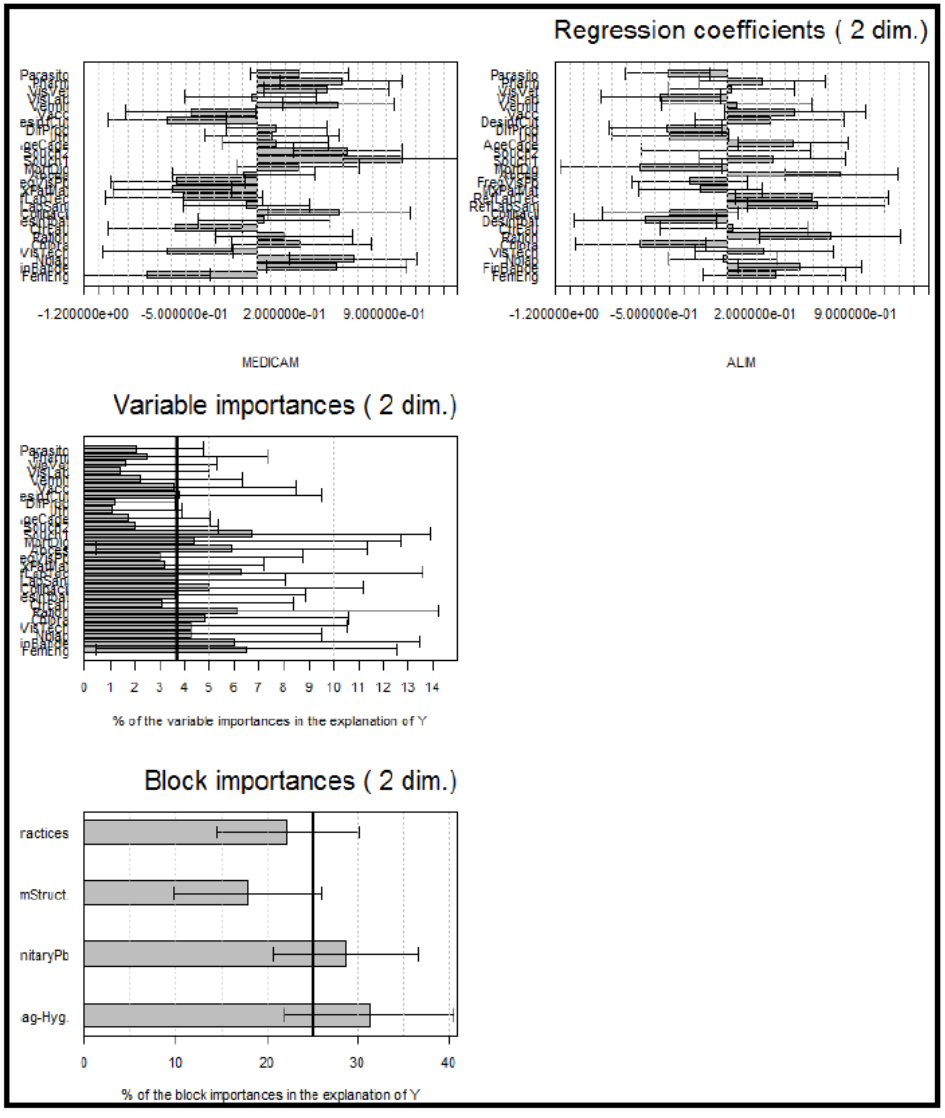
R code to get bootstrap simulations and graphical displays (ade4)

```
boot_mbra_DrugU <-
bootstrap.multiblock(resmbra_DrugU, 200, 2)
plot(boot_resmbra_DrugU)
```

Interpretation at the block level

The drug use is mainly associated with :

- . . .the variables related to the female and/or other animal presence at the end of rearing and to abscess.
- . . .the blocks related to management and hygiene practices and to sanitary problems.



Interpretation of the optimal model at the variable and block levels

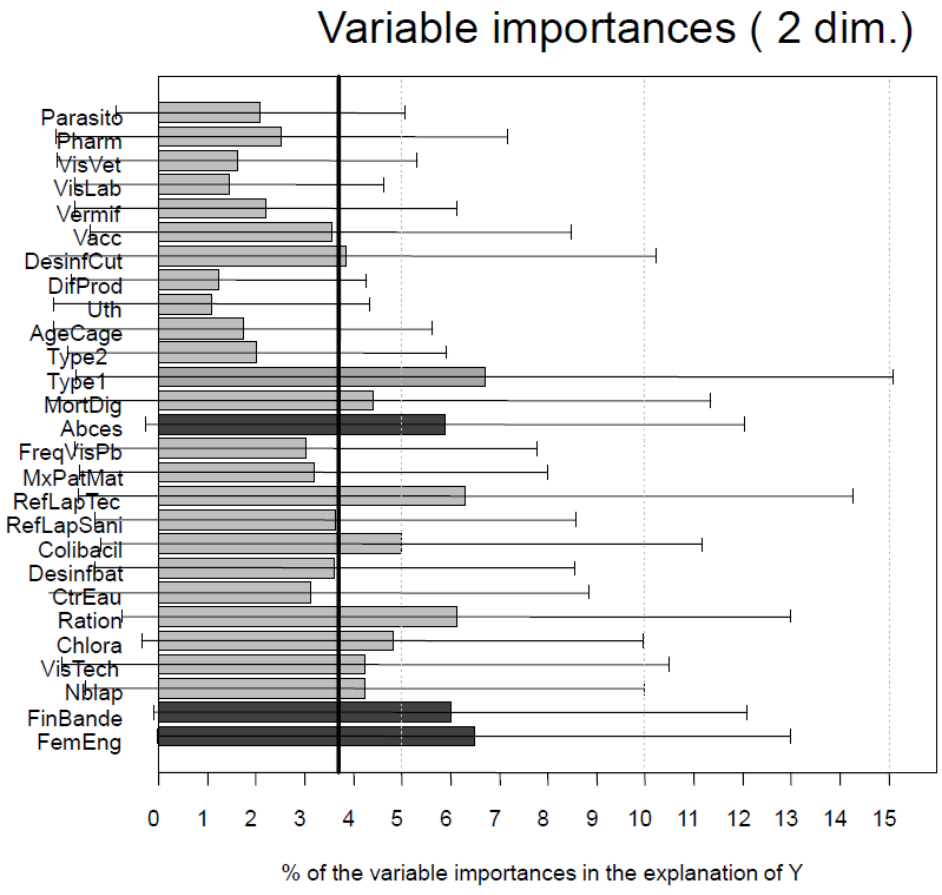
R code to get bootstrap simulations and graphical displays (ade4)

```
boot_mbra_DrugU <-
bootstrap.multiblock(resmbra_DrugU, 200, 2)
plot(boot_resmbra_DrugU)
```

Interpretation at the block level

The drug use is mainly associated with :

- . . .the variables related to the female and/or other animal presence at the end of rearing and to abscess.
- . . .the blocks related to management and hygiene practices and to sanitary problems.



Interpretation of the optimal model at the variable and block levels

R code to get bootstrap simulations and graphical displays (ade4)

```
boot_mbra_DrugU <-
bootstrap.multiblock(resmbra_DrugU, 200, 2)
plot(boot_resmbra_DrugU)
```

Interpretation at the block level

The drug use is mainly associated with :

- . . .the variables related to the female and/or other animal presence at the end of rearing and to abscess.
- . . .the blocks related to management and hygiene practices and to sanitary problems.

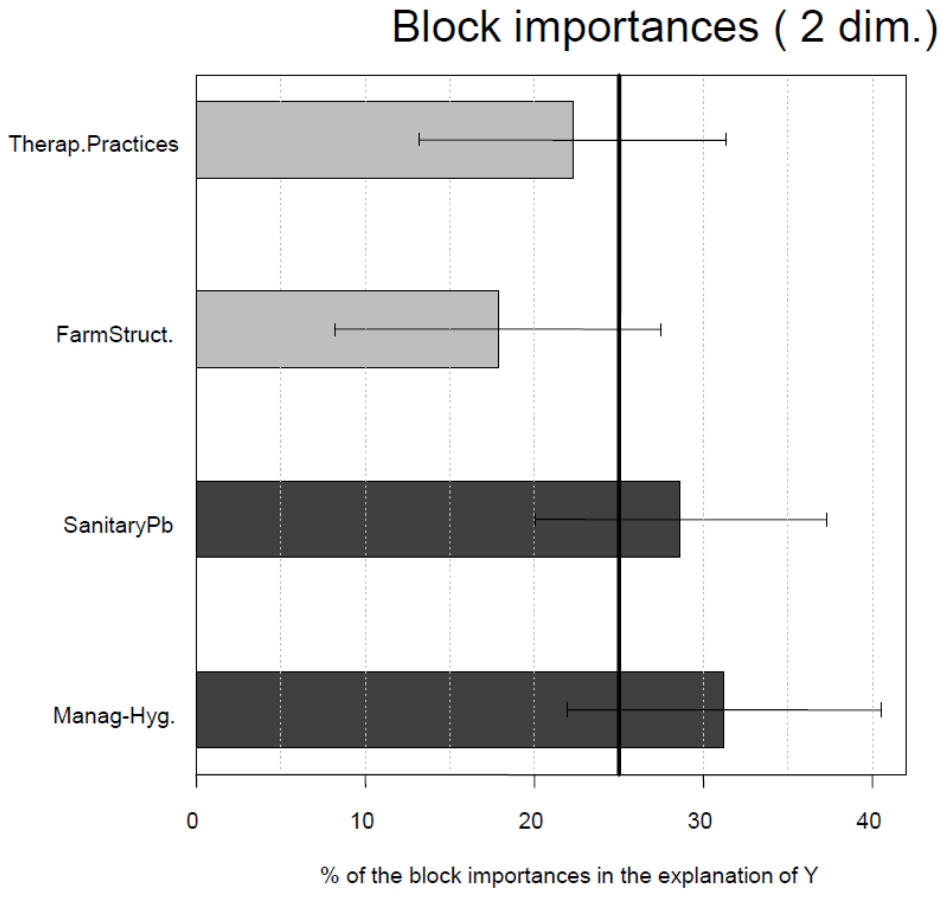


Table of contents

Conclusion & perspectives

Conclusions

- Multiblock Redundancy Analysis handles the specificity of epidemiological data,
- Enhancement of the interpretation : extensive interpretation of risk factors for a complex disease or a sanitary problem,
- Available method(s) and interpretation tools : `multiblock` package integrated in the `ade4` software (`http://pbil.univ-lyon1.fr/ade4/`).

Perspectives

- Direct extension to the explanation of several outcome blocks ($Y_1, \ldots, Y_{K'}$) or to qualitative (dummy) variables,
- New developments to handle a group structure of individuals (PhD 2010-2013 / A.Eslami).

Conclusion & perspectives

Conclusions

- Multiblock Redundancy Analysis handles the specificity of epidemiological data,
- Enhancement of the interpretation : extensive interpretation of risk factors for a complex disease or a sanitary problem,
- Available method(s) and interpretation tools : `multiblock` package integrated in the `ade4` software (`http://pbil.univ-lyon1.fr/ade4/`).

Perspectives

- Direct extension to the explanation of several outcome blocks $(Y_1, \ldots, Y_{K'})$ or to qualitative (dummy) variables,
- New developments to handle a group structure of individuals (PhD 2010-2013 / A.Eslami).

Multiblock redundancy analysis

Application to drug use on farms

S. Bougeard[(1)], F. Laanaya-Tazani[(1,2)], S. Le Bouquin[(1)] & C. Chauvin[(1)]

[(1)] *French Agency for food, environmental and occupational health & safety (Anses), Department of epidemiology, Ploufragan, France*

[(2)] *University of Rennes 2, Master of applied statistics, France*

anses
French agency for food, environmental and occupational health safety