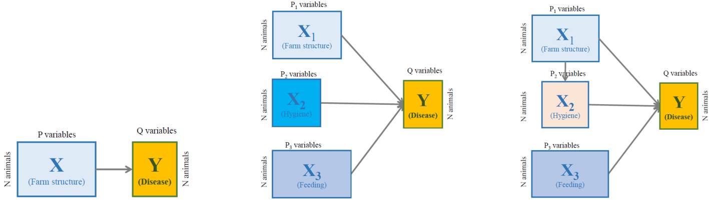
Supervised multiblock analyses

Cases of two-blocks, (K+1)-blocks, (K+K')-blocks

Stéphanie Bougeard

French Agency for Food, Environmental, Occupational Health & Safety (Anses), Ploufragan, France



Journée Analyses Factorielles
March 30 2023, INRAe Jouy-en-Josas

Outline

1 Introduction

2 Supervised two-block analyses

3 Supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses

5 Conclusion & perspectives

Outline

1 Introduction

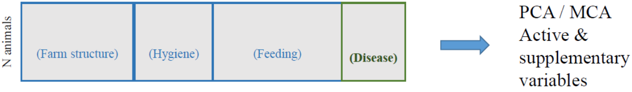
2 Supervised two-block analyses

3 Supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses

5 Conclusion & perspectives

From factorial analyses to multiblock factorial analyses



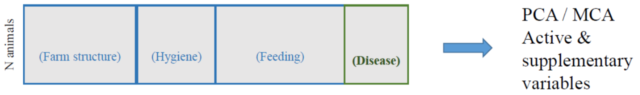
Data features

1. Blocks of variables
   - Of known structure,
   - Links between blocks are known.
2. Block features
   - Large dimension (nb var. > nb obs.),
   - Quantitative and quasi-collinear variables,
   - No distributional assumptions.

   → Ill-conditioned (multidimensional) blocks.
3. Observations (same for all the variables)

anses

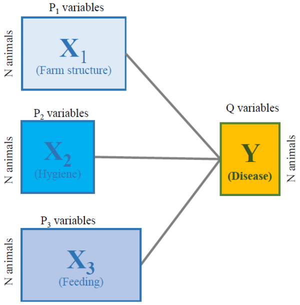
From factorial analyses to multiblock factorial analyses



Data features

1. Blocks of variables
   - Of known structure,
   - Links between blocks are known.
2. Block features
   - Large dimension (nb var. > nb obs.),
   - Quantitative and quasi-collinear variables,
   - No distributional assumptions.

   → Ill-conditioned (multidimensional) blocks.
3. Observations (same for all the variables)



anses

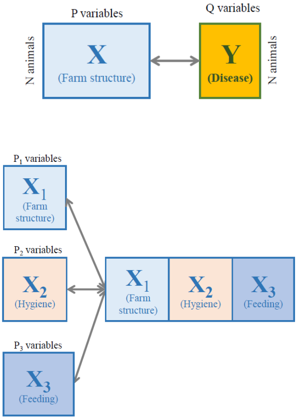
From unsupervised to supervised analyses

Unsupervised or supervised (two-block case)

- Unsup.: Study the relationships between **X** and **Y** or between $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$
- Supervised: Explain **Y** with **X**

Supervised cases

- Two-block case: $\mathbf{X} \rightarrow \mathbf{Y}$
- K+1 case: $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$
- K+K' case: e.g., $\mathbf{X}_1 \rightarrow \mathbf{X}_2$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$



anses

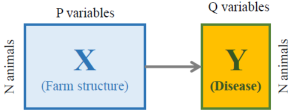
From unsupervised to supervised analyses

Unsupervised or supervised (two-block case)

- Unsup.: Study the relationships between **X** and **Y** or between $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$
- Supervised: Explain **Y** with **X**

Supervised cases

- Two-block case: $\mathbf{X} \rightarrow \mathbf{Y}$
- K+1 case: $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$
- K+K' case: e.g., $\mathbf{X}_1 \rightarrow \mathbf{X}_2$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$



anses

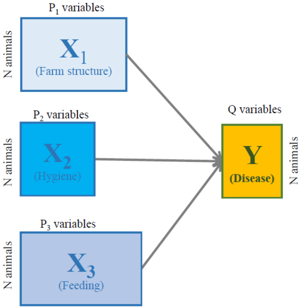
From unsupervised to supervised analyses

Unsupervised or supervised (two-block case)

- Unsup.: Study the relationships between $\mathbf{X}$ and $\mathbf{Y}$ or between $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$
- Supervised: Explain $\mathbf{Y}$ with $\mathbf{X}$

Supervised cases

- Two-block case: $\mathbf{X} \to \mathbf{Y}$
- K+1 case: $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \to \mathbf{Y}$
- K+K' case: e.g., $\mathbf{X}_1 \to \mathbf{X}_2$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \to \mathbf{Y}$



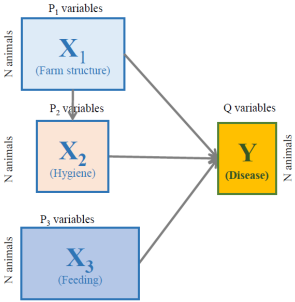
From unsupervised to supervised analyses

Unsupervised or supervised (two-block case)

- Unsup.: Study the relationships between **X** and **Y** or between $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$
- Supervised: Explain **Y** with **X**

Supervised cases

- Two-block case: $\mathbf{X} \rightarrow \mathbf{Y}$
- K+1 case: $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$
- K+K' case: e.g., $\mathbf{X}_1 \rightarrow \mathbf{X}_2$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \rightarrow \mathbf{Y}$



Outline

1 Introduction

2 Supervised two-block analyses
- Methods
- Application
- Doing my own supervised two-block analyses

3 Supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses

5 Conclusion & perspectives

anses

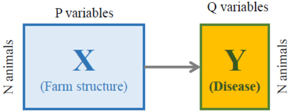
Relate two-block data sets with a criterion

Aim

Explore/Explain **Y** with **X**

How blocks are linked?

- Raw data sets . . .
- Are summarized with components . . .
- Which are linked by a criterion



Two-block case criterion (first-order solution)

Maximize $\text{cov}^2(\mathbf{t}, \mathbf{u})$
with $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$
with specific constraints (associated with methods)

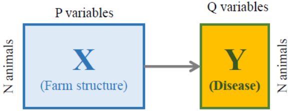
Relate two-block data sets with a criterion

Aim

Explore/Explain **Y** with **X**

How blocks are linked?

- Raw data sets . . .
- Are summarized with components . . .
- Which are linked by a criterion



Two-block case criterion (first-order solution)

Maximize $\text{cov}^2(\mathbf{t}, \mathbf{u})$
with $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$
with specific constraints (associated with methods)

anses

Relate two-block data sets with a criterion

Aim

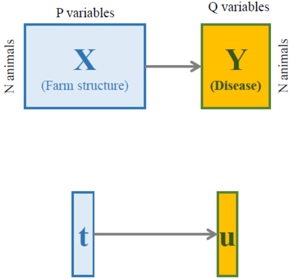
Explore/Explain **Y** with **X**

How blocks are linked?

- Raw data sets . . .
- Are summarized with components . . .
- Which are linked by a criterion

Two-block case criterion (first-order solution)

Maximize $\text{cov}^2(\mathbf{t}, \mathbf{u})$
with $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$
with specific constraints (associated with methods)



How to be a supervised two-block method? Constraints and deflation

Criterion

Maximize $\text{cov}^2(\mathbf{t}, \mathbf{u})$ with $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yv}$

Constraints and deflation

Method	**Constraints**	**w eigenvector of**	**Deflation**
Canonical an. [Hotelling, 36]	$\|\|\mathbf{t}\|\| = \|\|\mathbf{u}\|\| = 1$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$	No deflation (DVS)
Redundancy an. [Rao, 64]	$\|\|\mathbf{t}\|\| = \|\|\mathbf{v}\|\| = 1$	$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{YY}'\mathbf{X})$	No deflation (DVS) Or deflation on **t**
PLS regression* [Wold, 66]	$\|\|\mathbf{w}\|\| = \|\|\mathbf{v}\|\| = 1$	$\mathbf{X}'\mathbf{YY}'\mathbf{X}$	Deflation on **t**

* Co-inertia an. [Chessel, 93], concordance an. [Lafosse, 97]: close criteria, different deflation.

Supervised two-block methods

- RA: supervised constraint-based method
- PLS: supervised deflation-based method

anses

How to be a supervised two-block method? Constraints and deflation

Criterion

Maximize $\text{cov}^2(\mathbf{t}, \mathbf{u})$ with $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$

Constraints and deflation

Method	Constraints	w eigenvector of	Deflation
Canonical an. [Hotelling, 36]	$\|\|\mathbf{t}\|\| = \|\|\mathbf{u}\|\| = 1$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$	No deflation (DVS)
Redundancy an. [Rao, 64]	$\|\|\mathbf{t}\|\| = \|\|\mathbf{v}\|\| = 1$	$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$	No deflation (DVS) Or deflation on **t**
PLS regression* [Wold, 66]	$\|\|\mathbf{w}\|\| = \|\|\mathbf{v}\|\| = 1$	$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$	Deflation on **t**

* Co-inertia an. [Chessel, 93], concordance an. [Lafosse, 97]: close criteria, different deflation.

Supervised two-block methods

- RA: supervised constraint-based method
- PLS: supervised deflation-based method

anses

Supervised two-block analyses: advices for application

Choice according to your aim (and knowledge)

- My first two-block analysis: PCA(**X**) with **Y** as supplementary variables
- Unsupervised: Canonical analysis, co-inertia analysis, concordance analysis
- Supervised: RA or PLS
  - Explain: RA better explains the inertia of **Y**,
  - Predict: PLS leads to more stable results thus best predictions.

Choice according to the data features

- Limited within-correlation in **X** $\rightarrow$ RA
- High within-correlation in **X** $\rightarrow$ PLS

Not able to choose

- Trade-off with regularization on the norm-constraint: $\gamma||\mathbf{w}||^2 + (1-\gamma)||\mathbf{t}||^2 = 1$
- Solution: DVS of $[\gamma\mathbf{I} + (1-\gamma)(\mathbf{X}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$
- Optimise $0 \leq \gamma \leq 1$ while, e.g., minimizing the prediction error.

anses

Supervised two-block analyses: advices for application

Choice according to your aim (and knowledge)

- My first two-block analysis: PCA(**X**) with **Y** as supplementary variables
- Unsupervised: Canonical analysis, co-inertia analysis, concordance analysis
- Supervised: RA or PLS
  - Explain: RA better explains the inertia of **Y**,
  - Predict: PLS leads to more stable results thus best predictions.

Choice according to the data features

- Limited within-correlation in **X** $\rightarrow$ RA
- High within-correlation in **X** $\rightarrow$ PLS

Not able to choose

- Trade-off with regularization on the norm-constraint: $\gamma||\mathbf{w}||^2 + (1-\gamma)||\mathbf{t}||^2 = 1$
- Solution: DVS of $[\gamma\mathbf{I} + (1-\gamma)(\mathbf{X}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$
- Optimise $0 \leq \gamma \leq 1$ while, e.g., minimizing the prediction error.

anses

Supervised two-block analyses: advices for application

Choice according to your aim (and knowledge)

- My first two-block analysis: PCA(**X**) with **Y** as supplementary variables
- Unsupervised: Canonical analysis, co-inertia analysis, concordance analysis
- Supervised: RA or PLS
  - Explain: RA better explains the inertia of **Y**,
  - Predict: PLS leads to more stable results thus best predictions.

Choice according to the data features

- Limited within-correlation in **X** $\rightarrow$ RA
- High within-correlation in **X** $\rightarrow$ PLS

Not able to choose

- Trade-off with regularization on the norm-constraint: $\gamma||\mathbf{w}||^2 + (1-\gamma)||\mathbf{t}||^2 = 1$
- Solution: DVS of $[\gamma\mathbf{I} + (1-\gamma)(\mathbf{X}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$
- Optimise $0 \leq \gamma \leq 1$ while, e.g., minimizing the prediction error.

anses

Supervised two-block analyses: advices for application

Choice according to your aim (and knowledge)

- My first two-block analysis: PCA(**X**) with **Y** as supplementary variables
- Unsupervised: Canonical analysis, co-inertia analysis, concordance analysis
- Supervised: RA or PLS
  - Explain: RA better explains the inertia of **Y**,
  - Predict: PLS leads to more stable results thus best predictions.

Choice according to the data features

- Limited within-correlation in **X** $\rightarrow$ RA
- High within-correlation in **X** $\rightarrow$ PLS

Not able to choose

- Trade-off with regularization on the norm-constraint: $\gamma||\mathbf{w}||^2 + (1-\gamma)||\mathbf{t}||^2 = 1$
- Solution: DVS of $[\gamma\mathbf{I} + (1-\gamma)(\mathbf{X}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$
- Optimise $0 \leq \gamma \leq 1$ while, e.g., minimizing the prediction error.

anses

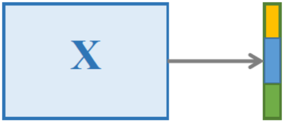
(Some) extensions for supervised two-block analyses

Extensions according to data features

- **Y** is a single nominal variable: discriminant PLS (PLS-DA) [Barker, 2003]
- **X** contains a very large number of variables: sparse PLS (sPLS) [Lê Cao, 2008]

Extensions according to observation-structure

- Known group-structure of observations: multigroup PLS [Eslami, 2014]
- Unknown group-structure of observations: clusterwise PLS [Vinzi, 05; Preda, 05]



anses

(Some) extensions for supervised two-block analyses

Extensions according to data features

- **Y** is a single nominal variable: discriminant PLS (PLS-DA) [Barker, 2003]
- **X** contains a very large number of variables: sparse PLS (sPLS) [Lê Cao, 2008]

Extensions according to observation-structure

- Known group-structure of observations: multigroup PLS [Eslami, 2014]
- Unknown group-structure of observations: clusterwise PLS [Vinzi, 05; Preda, 05]

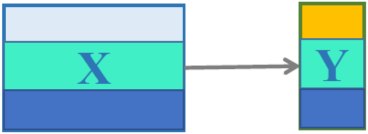
(Some) extensions for supervised two-block analyses

Extensions according to data features

- **Y** is a single nominal variable: discriminant PLS (PLS-DA) [Barker, 2003]
- **X** contains a very large number of variables: sparse PLS (sPLS) [Lê Cao, 2008]

Extensions according to observation-structure

- Known group-structure of observations: multigroup PLS [Eslami, 2014]
- Unknown group-structure of observations: clusterwise PLS [Vinzi, 05; Preda, 05]



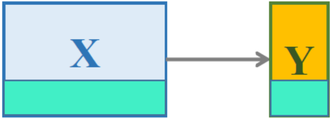
(Some) extensions for supervised two-block analyses

Extensions according to data features

- **Y** is a single nominal variable: discriminant PLS (PLS-DA) [Barker, 2003]
- **X** contains a very large number of variables: sparse PLS (sPLS) [Lê Cao, 2008]

Extensions according to observation-structure

- Known group-structure of observations: multigroup PLS [Eslami, 2014]
- Unknown group-structure of observations: clusterwise PLS [Vinzi, 05; Preda, 05]



Extensions for supervised two-block analyses: multigroup PLS

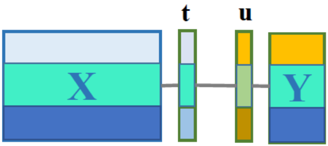
Main aim

Explore the links between **X** and **Y** while taking into account their multigroup structure, i.e.:

- Observations with a structure in known groups (which they should be freed).

Sub-aims

1. Summarize each block of variables by components adjusted to the data features (i.e., ill-conditioned multidimensional blocks),
2. Study links between variables in a space common to all groups,
3. Understand the group-particularities in relation to the common structure.



Extensions for supervised two-block analyses: multigroup PLS

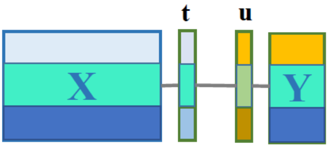
Main aim

Explore the links between **X** and **Y** while taking into account their multigroup structure, i.e.:

- Observations with a structure in known groups (which they should be freed).

Sub-aims

1. Summarize each block of variables by components adjusted to the data features (i.e., ill-conditioned multidimensional blocks),
2. Study links between variables in a space common to all groups,
3. Understand the group-particularities in relation to the common structure.



Extensions for supervised two-block analyses: multigroup PLS [Eslami, 2013, 2014]

Criterion to maximize (first-order solution)

$$\sum_{m=1}^{M} N_m \operatorname{cov}(\mathbf{t}_m, \mathbf{u}_m)$$

$$\text{s.t. } \mathbf{t}_m = \mathbf{X}_m \mathbf{a},\ \mathbf{u}_m = \mathbf{Y}_m \mathbf{b},\ \|\mathbf{a}\| = \|\mathbf{b}\| = 1$$

Main features

- Solved by a monotonous convergent algorithm,
- Robust to within-block multicollinearity,
- Group-components specific to each group,
- Common axes and components (vertical concatenation of the group components) to all the groups.

anses

Extensions for supervised two-block analyses: multigroup PLS [Eslami, 2013, 2014]

Criterion to maximize (first-order solution)

$$\sum_{m=1}^{M} N_m \operatorname{cov}(\mathbf{t}_m, \mathbf{u}_m)$$

$$\text{s.t. } \mathbf{t}_m = \mathbf{X}_m\mathbf{a},\ \mathbf{u}_m = \mathbf{Y}_m\mathbf{b},\ \|\mathbf{a}\| = \|\mathbf{b}\| = 1$$

Main features

- Solved by a monotonous convergent algorithm,
- Robust to within-block multicollinearity,
- Group-components specific to each group,
- Common axes and components (vertical concatenation of the group components) to all the groups.

anses

Extensions for supervised two-block analyses: multigroup PLS [Eslami, 2013, 2014]

Criterion to maximize (first-order solution)

$$\sum_{m=1}^{M} N_m \operatorname{cov}(\mathbf{t}_m, \mathbf{u}_m)$$

$$\text{s.t. } \mathbf{t}_m = \mathbf{X}_m \mathbf{a},\ \mathbf{u}_m = \mathbf{Y}_m \mathbf{b},\ \|\mathbf{a}\| = \|\mathbf{b}\| = 1$$

Main features

- Solved by a monotonous convergent algorithm,
- Robust to within-block multicollinearity,
- Group-components specific to each group,
- Common axes and components (vertical concatenation of the group components) to all the groups.

Extensions for supervised two-block analyses: multigroup PLS [Eslami, 2013, 2014]

Criterion to maximize (first-order solution)

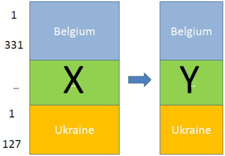
$$\sum_{m=1}^{M} N_m \operatorname{cov}(\mathbf{t}_m, \mathbf{u}_m)$$

$$\text{s.t. } \mathbf{t}_m = \mathbf{X}_m \mathbf{a},\ \mathbf{u}_m = \mathbf{Y}_m \mathbf{b},\ \|\mathbf{a}\| = \|\mathbf{b}\| = 1$$

Main features

- Solved by a monotonous convergent algorithm,
- Robust to within-block multicollinearity,
- Group-components specific to each group,
- Common axes and components (vertical concatenation of the group components) to all the groups.

Mg-PLS: 'European School Survey Project on Alcohol and other Drugs' data



Individuals

N=5204 teenagers from M=13 countries

Belgium (331), Cyprus (177), Czech Republic (1013), France (723), Germany (365), Italy (617), Kosovo (55), Latvia (292), Lichtenstein (52), Poland (1113), Romania (93), Slovak Republic (246) and Ukraine (127).

X dataset: use and context

P=9 questions

Cannabis consumption in the last year or month (c25b, c25c), age they first take cannabis (c26), number of smoked cigarettes in the last month (c09), number of times they were drunk in their life or in the last year (c19a, c19b), facility to get cannabis (c24), number of friends who take cannabis (c34d) and perceived risk of taking cannabis (c36h)

Y dataset: drug consumption (CAST)

Q=6 questions

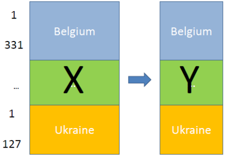
Non-recreational use (rcast1, rcast2), memory disorder (rcast3), reproaches from family or friends (rcast4), unsuccessful quit attempts (rcast5) and problems associated with cannabis consumption (rcast6)

Aims

- Investigate the relationships between the cannabis consumption variables (**Y**),
- Explain the cannabis consumption (**Y**) by the use and context variables (**X**).

anses

Mg-PLS: 'European School Survey Project on Alcohol and other Drugs' data



Individuals

N=5204 teenagers from M=13 countries

Belgium (331), Cyprus (177), Czech Republic (1013), France (723), Germany (365), Italy (617), Kosovo (55), Latvia (292), Lichtenstein (52), Poland (1113), Romania (93), Slovak Republic (246) and Ukraine (127).

X dataset: use and context

P=9 questions

Cannabis consumption in the last year or month (c25b, c25c), age they first take cannabis (c26), number of smoked cigarettes in the last month (c09), number of times they were drunk in their life or in the last year (c19a, c19b), facility to get cannabis (c24), number of friends who take cannabis (c34d) and perceived risk of taking cannabis (c36h)

Y dataset: drug consumption (CAST)

Q=6 questions

Non-recreational use (rcast1, rcast2), memory disorder (rcast3), reproaches from family or friends (rcast4), unsuccessful quit attempts (rcast5) and problems associated with cannabis consumption (rcast6)

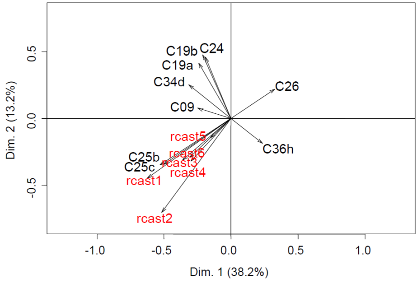
Aims

- Investigate the relationships between the cannabis consumption variables (**Y**),
- Explain the cannabis consumption (**Y**) by the use and context variables (**X**).

Mg-PLS: Common relationships between consumption, use and context

Pre-processing

- Variables are centred and scaled globally $\rightarrow$ Variables have the same weights,
- Variables are centred and scaled by group $\rightarrow$ Groups have the same weights,
- Group effect=11% of inertia (discarded) $\rightarrow$ Focus on the within-group analysis.



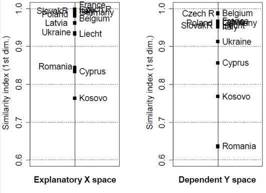
Interpretation

- All the CAST variables (**Y**) are linked and explained with the cannabis consumption in the last year or month (c25b, c25c) and the age they first take cannabis (c26)
- The non-recreational use (rcast1, rcast2) are the variables which are more linked to c25b, c25c and c26.

Y: Non-recreational use (rcast1, rcast2), memory disorder (rcast3), reproaches from family or friends (rcast4), unsuccessful quit attempts (rcast5) and problems associated with cannabis consumption (rcast6) - **X**: Cannabis consumption in the last year or month (c25b, c25c), age they first take cannabis (c26), number of smoked cigarettes in the last month (c09), number of times they were drunk in their life or in the last year (c19a, c19b), facility to get cannabis (c24), number of friends who take cannabis (c34d) and perceived risk of taking cannabis (c36h)

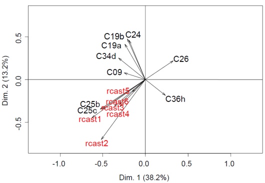
Mg-PLS: Group specificities in comparison with the common structure

Similarities between groups



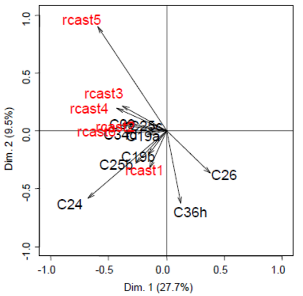
Most of the countries are similar to the common structure, except Kosovo, Cyprus and Romania.

Common loadings



Y: Non-recreational use (rcast1, rcast2), memory disorder (rcast3), reproaches from family or friends (rcast4), unsuccessful quit attempts (rcast5) and problems associated with cannabis consumption (rcast6) - **X**: Cannabis consumption in the last year or month (c25b, c25c), age they first take cannabis (c26), number of smoked cigarettes in the last month (c09), number of times they were drunk in their life or in the last year (c19a, c19b), facility to get cannabis (c24), number of friends who take cannabis (c34d) and perceived risk of taking cannabis (c36h)

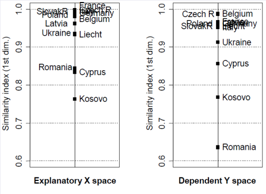
Group loadings: Kosovo



The relationships between the variables from Kosovo are really different than those from the common structure.

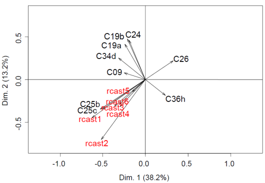
Mg-PLS: Group specificities in comparison with the common structure

Similarities between groups



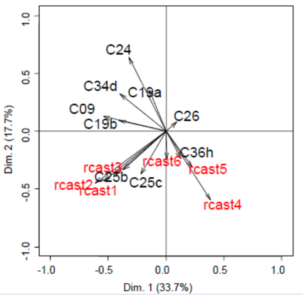
Most of the countries are similar to the common structure, except Kosovo, Cyprus and Romania.

Common loadings



Y: Non-recreational use (rcast1, rcast2), memory disorder (rcast3), reproaches from family or friends (rcast4), unsuccessful quit attempts (rcast5) and problems associated with cannabis consumption (rcast6) - **X**: Cannabis consumption in the last year or month (c25b, c25c), age they first take cannabis (c26), number of smoked cigarettes in the last month (c09), number of times they were drunk in their life or in the last year (c19a, c19b), facility to get cannabis (c24), number of friends who take cannabis (c34d) and perceived risk of taking cannabis (c36h)

Group loadings: Romania



The variables rcast1, rcast2 and rcast3 are linked and explained with different explanatory variables than rcast4 and rcast5.

Supervised two-block analyses with R

Standard supervised two-block analyses

- RA: `pcaiv` function in the `ade4` package, `rda` function in the `vegan` package
- PLS regression: e.g., `plsr` function in the `pls` package, `pls` function in the `MixOmics` package
- Regularized-RA: `cw.multiblock` function with 'mbregular' option, a single-block **X** and a single-cluster ('G=1') in the `mbclusterwise` package
- Regularized-CCA: `rcca` function in the `MixOmics` package

Extensions of supervised two-block analyses

- Discriminant PLS: e.g., `plsda` function in the `mdatools` or `MixOmics` package
- Sparse PLS: e.g., `spls/splsda` functions in the `MixOmics` package, `spls` package
- Multigroup PLS: `mgPLS` function in the `multigroup` package, `mint` functions (mint.pls, mint.spls, mint.plsda, mint.splsda) in the `MixOmics` package
- Multigroup RA: `within` and `pcaiv` functions in the `ade4` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised two-block analyses with R

Standard supervised two-block analyses

- RA: `pcaiv` function in the `ade4` package, `rda` function in the `vegan` package
- PLS regression: e.g., `plsr` function in the `pls` package, `pls` function in the `MixOmics` package
- Regularized-RA: `cw.multiblock` function with 'mbregular' option, a single-block **X** and a single-cluster ('G=1') in the `mbclusterwise` package
- Regularized-CCA: `rcca` function in the `MixOmics` package

Extensions of supervised two-block analyses

- Discriminant PLS: e.g., `plsda` function in the `mdatools` or `MixOmics` package
- Sparse PLS: e.g., `spls/splsda` functions in the `MixOmics` package, `spls` package
- Multigroup PLS: `mgPLS` function in the `multigroup` package, `mint` functions (mint.pls, mint.spls, mint.plsda, mint.splsda) in the `MixOmics` package
- Multigroup RA: `within` and `pcaiv` functions in the `ade4` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised two-block analyses with R

Standard supervised two-block analyses

- RA: `pcaiv` function in the `ade4` package, `rda` function in the `vegan` package
- PLS regression: e.g., `plsr` function in the `pls` package, `pls` function in the `MixOmics` package
- Regularized-RA: `cw.multiblock` function with 'mbregular' option, a single-block **X** and a single-cluster ('G=1') in the `mbclusterwise` package
- Regularized-CCA: `rcca` function in the `MixOmics` package

Extensions of supervised two-block analyses

- Discriminant PLS: e.g., `plsda` function in the `mdatools` or `MixOmics` package
- Sparse PLS: e.g., `spls/splsda` functions in the `MixOmics` package, `spls` package
- Multigroup PLS: `mgPLS` function in the `multigroup` package, `mint` functions (mint.pls, mint.spls, mint.plsda, mint.splsda) in the `MixOmics` package
- Multigroup RA: `within` and `pcaiv` functions in the `ade4` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised two-block analyses with R

Standard supervised two-block analyses

- RA: `pcaiv` function in the `ade4` package, `rda` function in the `vegan` package
- PLS regression: e.g., `plsr` function in the `pls` package, `pls` function in the `MixOmics` package
- Regularized-RA: `cw.multiblock` function with 'mbregular' option, a single-block **X** and a single-cluster ('G=1') in the `mbclusterwise` package
- Regularized-CCA: `rcca` function in the `MixOmics` package

Extensions of supervised two-block analyses

- Discriminant PLS: e.g., `plsda` function in the `mdatools` or `MixOmics` package
- Sparse PLS: e.g., `spls/splsda` functions in the `MixOmics` package, `spls` package
- Multigroup PLS: `mgPLS` function in the `multigroup` package, `mint` functions (mint.pls, mint.spls, mint.plsda, mint.splsda) in the `MixOmics` package
- Multigroup RA: `within` and `pcaiv` functions in the `ade4` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised two-block analyses with R

Standard supervised two-block analyses

- RA: `pcaiv` function in the `ade4` package, `rda` function in the `vegan` package
- PLS regression: e.g., `plsr` function in the `pls` package, `pls` function in the `MixOmics` package
- Regularized-RA: `cw.multiblock` function with 'mbregular' option, a single-block **X** and a single-cluster ('G=1') in the `mbclusterwise` package
- Regularized-CCA: `rcca` function in the `MixOmics` package

Extensions of supervised two-block analyses

- Discriminant PLS: e.g., `plsda` function in the `mdatools` or `MixOmics` package
- Sparse PLS: e.g., `spls/splsda` functions in the `MixOmics` package, `spls` package
- Multigroup PLS: `mgPLS` function in the `multigroup` package, `mint` functions (mint.pls, mint.spls, mint.plsda, mint.splsda) in the `MixOmics` package
- Multigroup RA: `within` and `pcaiv` functions in the `ade4` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Outline

1 Introduction

2 Supervised two-block analyses

3 Supervised (K+1)-block analyses
- Methods
- Applications
- Doing my own supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses

5 Conclusion & perspectives

anses

Relate (K+1) blocks with a criterion

Aim

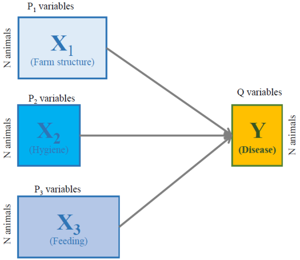
Explore/Explain $\mathbf{Y}$ with $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$

How blocks are linked?

- Raw data sets . . .
- Are summarized with block-components . . .
- Which are linked by a criterion*

(K+1)-block case criterion (first-order solution)

Maximize $\sum_{k=1}^{K} \text{cov}^2(\mathbf{t}_k, \mathbf{u})$
with $\mathbf{t}_k = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yv}$
with specific constraints (associated with methods)



anses

Relate (K+1) blocks with a criterion

Aim

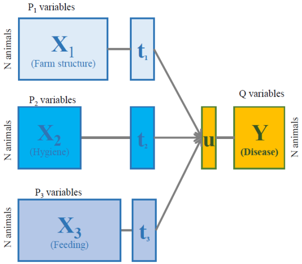
Explore/Explain $\mathbf{Y}$ with $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$

How blocks are linked?

- Raw data sets ...
- Are summarized with block-components ...
- Which are linked by a criterion*

(K+1)-block case criterion (first-order solution)

Maximize $\sum_{k=1}^{K} \mathrm{cov}^2(\mathbf{t}_k, \mathbf{u})$
with $\mathbf{t}_k = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$
with specific constraints (associated with methods)



anses

Relate (K+1) blocks with a criterion

Aim

Explore/Explain $\mathbf{Y}$ with $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$

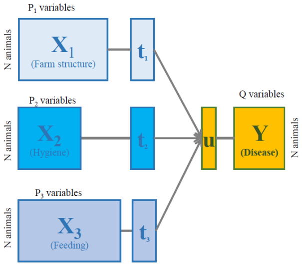
How blocks are linked?

- Raw data sets ...
- Are summarized with block-components ...
- Which are linked by a criterion*

* Methods which are not based on a criterion are not given here.

(K+1)-block case criterion (first-order solution)

Maximize $\sum_{k=1}^{K} \text{cov}^2(\mathbf{t}_k, \mathbf{u})$
with $\mathbf{t}_k = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$
with specific constraints (associated with methods)



anses

How to be a supervised (K+1)-block method? Constraints, deflation, pre-proc.

Criterion

Maximize $\sum_{k=1}^{K} \text{cov}^2(\mathbf{t}_k, \mathbf{u})$ with $\mathbf{t}_k = \mathbf{X}_k \mathbf{w}_k$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$

Constraints and deflation

Method	Constraints	$\mathbf{v}$ eigenvector of	Deflation
ACG-TR [Kissita, 2003]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{u}\|\| = 1$	$\sum_k (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	No deflation (DVS)
mb-Redund. an. [Bougeard, 2011]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\sum_k \mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	Deflation on $\mathbf{t}$
mb-PLS * [Wold, 1984]	$\|\|\mathbf{w}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}$	Deflation on $\mathbf{t}$

* ACIMO [Vivien, 2002], ConcorG [Lafosse, 1997]: close criteria, different deflation.

Supervised (K+1)-block methods

- mb-RA: supervised constraint-based method; multiblock solution
- mb-PLS: supervised deflation-based method; multiblock solution/pre-processing

anses

How to be a supervised (K+1)-block method? Constraints, deflation, pre-proc.

Criterion

Maximize $\sum_{k=1}^{K} \text{cov}^2(\mathbf{t}_k, \mathbf{u})$ with $\mathbf{t}_k = \mathbf{X}_k \mathbf{w}_k$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$

Constraints and deflation

Method	Constraints	v eigenvector of	Deflation
ACG-TR [Kissita, 2003]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{u}\|\| = 1$	$\sum_k (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	No deflation (DVS)
mb-Redund. an. [Bougeard, 2011]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\sum_k \mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	Deflation on $\mathbf{t}$
mb-PLS * [Wold, 1984]	$\|\|\mathbf{w}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}$	Deflation on $\mathbf{t}$

* ACIMO [Vivien, 2002], ConcorG [Lafosse, 1997]: close criteria, different deflation.

Supervised (K+1)-block methods

- mb-RA: supervised constraint-based method; multiblock solution
- mb-PLS: supervised deflation-based method; multiblock solution/pre-processing

anses

How to be a supervised (K+1)-block method? Constraints, deflation, pre-proc.

Criterion

Maximize $\sum_{k=1}^{K} \text{cov}^2(\mathbf{t}_k, \mathbf{u})$ with $\mathbf{t}_k = \mathbf{X}_k\mathbf{w}_k$ and $\mathbf{u} = \mathbf{Y}\mathbf{v}$

Constraints and deflation

Method	Constraints	v eigenvector of	Deflation
ACG-TR [Kissita, 2003]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{u}\|\| = 1$	$\sum_k (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	No deflation (DVS)
mb-Redund. an. [Bougeard, 2011]	$\|\|\mathbf{t}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\sum_k \mathbf{Y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$	Deflation on **t**
mb-PLS * [Wold, 1984]	$\|\|\mathbf{w}_k\|\| = \|\|\mathbf{v}\|\| = 1$	$\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}$	Deflation on **t**

* ACIMO [Vivien, 2002], ConcorG [Lafosse, 1997]: close criteria, different deflation.

Supervised (K+1)-block methods

- mb-RA: supervised constraint-based method; multiblock solution
- mb-PLS: supervised deflation-based method; multiblock solution/pre-processing

anses

Pre-processing: an important step

Variable-centering

- Centering = All the variable-means are equal to 0
- Variables are supposed to be centered (without loss of generality)

Variable-reduction

- Reduction = All the variable-standard deviations are equal to 1
  $\rightarrow$ All the variable have the same importance in the analysis
- No reduction $\rightarrow$ The variables with the largest variances are the most important

Block-scaling (variables are supposed to be standardized)

- Scaling / $\lambda_k^{(1)}$ = = Variable-sd are equal to $1/\lambda_k^{(1)}$ $\rightarrow$ Block-inertia are equal to $P_k/\lambda_k^{(1)}$
  $\rightarrow$ Blocks with a small number of variables and low within-block correlation are the most important
- Scaling / Inertia($\mathbf{X}_k$) = Variable-sd are equal to $1/P_k$ $\rightarrow$ All the block-inertia are equal to 1
  $\rightarrow$ All the blocks have the same importance in the analysis
- No Scaling $\rightarrow$ The blocks with the largest number of variables are the most important.

anses

Pre-processing: an important step

Variable-centering

- Centering = All the variable-means are equal to 0
- Variables are supposed to be centered (without loss of generality)

Variable-reduction

- Reduction = All the variable-standard deviations are equal to 1
  $\rightarrow$ All the variable have the same importance in the analysis
- No reduction $\rightarrow$ The variables with the largest variances are the most important

Block-scaling (variables are supposed to be standardized)

- Scaling / $\lambda_k^{(1)}$ = = Variable-sd are equal to $1/\lambda_k^{(1)}$ $\rightarrow$ Block-inertia are equal to $P_k/\lambda_k^{(1)}$
  $\rightarrow$ Blocks with a small number of variables and low within-block correlation are the most important
- Scaling / Inertia($\mathbf{X}_k$) = Variable-sd are equal to $1/P_k$ $\rightarrow$ All the block-inertia are equal to 1
  $\rightarrow$ All the blocks have the same importance in the analysis
- No Scaling $\rightarrow$ The blocks with the largest number of variables are the most important.

anses

Pre-processing: an important step

Variable-centering

- Centering = All the variable-means are equal to 0
- Variables are supposed to be centered (without loss of generality)

Variable-reduction

- Reduction = All the variable-standard deviations are equal to 1
  $\rightarrow$ All the variable have the same importance in the analysis
- No reduction $\rightarrow$ The variables with the largest variances are the most important

Block-scaling (variables are supposed to be standardized)

- Scaling / $\lambda_k^{(1)}$ = = Variable-sd are equal to $1/\lambda_k^{(1)}$ $\rightarrow$ Block-inertia are equal to $P_k/\lambda_k^{(1)}$
  $\rightarrow$ Blocks with a small number of variables and low within-block correlation are the most important
- Scaling / Inertia($\mathbf{X}_k$) = Variable-sd are equal to $1/P_k$ $\rightarrow$ All the block-inertia are equal to 1
  $\rightarrow$ All the blocks have the same importance in the analysis
- No Scaling $\rightarrow$ The blocks with the largest number of variables are the most important.

anses

Supervised (K+1)-block analyses: prediction model

(K+1)-prediction model

- Aim: Explain $\mathbf{Y}$ with $\mathbf{X} = [\mathbf{X}_1 | \ldots | \mathbf{X}_K]$ (regression coefficients)
- Method:
  - Build a global-component $\mathbf{t} = \mathbf{X}\mathbf{w}$
  - Deflation on $\mathbf{t}$ (orthogonal)
  - NB: $\mathbf{t}$ is also a summary of the block-components: $\mathbf{t} = \sum_k \mathbf{a}_k \mathbf{t}_k$
- Solution: $\mathbf{Y} = \sum_h \mathbf{t}^{(h)} (\mathbf{c}^{(h)})' = \mathbf{X} \left[\sum_h (\mathbf{w}^{(h)})^* (\mathbf{c}^{(h)})' \right]$

Limits of the (K+1)-prediction model

- Deflation of $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ on $\mathbf{t}$ 'mix' the block-information
- Consideration of the same number of dimensions for all blocks
- The criterion maximize symmetrical links ($\sum_k \text{cov}^2(\mathbf{t}_k, \mathbf{u})$) whereas the prediction model is based on asymmetrical ones ($\mathbf{u} = f(\mathbf{t}_1, \ldots, \mathbf{t}_K)$)

Supervised (K+1)-block analyses: prediction model

(K+1)-prediction model

- Aim: Explain $\mathbf{Y}$ with $\mathbf{X} = [\mathbf{X}_1 | \dots | \mathbf{X}_K]$ (regression coefficients)
- Method:
  - Build a global-component $\mathbf{t} = \mathbf{Xw}$
  - Deflation on $\mathbf{t}$ (orthogonal)
  - NB: $\mathbf{t}$ is also a summary of the block-components: $\mathbf{t} = \sum_k \mathbf{a}_k \mathbf{t}_k$
- Solution: $\mathbf{Y} = \sum_h \mathbf{t}^{(h)} (\mathbf{c}^{(h)})' = \mathbf{X} \left[\sum_h (\mathbf{w}^{(h)})^* (\mathbf{c}^{(h)})' \right]$

Limits of the (K+1)-prediction model

- Deflation of $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ on $\mathbf{t}$ 'mix' the block-information
- Consideration of the same number of dimensions for all blocks
- The criterion maximize symmetrical links $(\sum_k \mathrm{cov}^2(\mathbf{t}_k, \mathbf{u}))$ whereas the prediction model is based on asymmetrical ones $(\mathbf{u} = f(\mathbf{t}_1, \dots, \mathbf{t}_K))$
  $\mathrm{cor}^2(\mathbf{X}_1, \mathbf{Y})$ and $\mathrm{cor}^2(\mathbf{X}_2, \mathbf{Y})$ *versus* $\mathbf{Y} = a_1 \mathbf{X}_1 + a_2 \mathbf{X}_2$

anses

Supervised (K+1)-block analyses: interpretation tools

Optimal dimension

- Select the optimal number of dimension H to be taken into account
- E.g., minimization of the cross-validated prediction error

Block-importance [Vivien, 2005; Bougeard, 2011]

- Obtained from the $\mathbf{a}_k$ coefficients which reflect the links between the block-components $\mathbf{t}_k$ and $\mathbf{u}$
- Can be computed for each and several dimensions

Variable-importance [Wold, 1994; Gosselin , 2010; Bougeard, 2011]

- Obtained from the $\mathbf{w}^*$ coefficients which reflect the importance of the variables to build the global components $\mathbf{t}$
- Can be computed for each and several dimensions

Supervised (K+1)-block analyses: interpretation tools

Optimal dimension

- Select the optimal number of dimension H to be taken into account
- E.g., minimization of the cross-validated prediction error

Block-importance [Vivien, 2005; Bougeard, 2011]

- Obtained from the $\mathbf{a}_k$ coefficients which reflect the links between the block-components $\mathbf{t}_k$ and $\mathbf{u}$
- Can be computed for each and several dimensions

Variable-importance [Wold, 1994; Gosselin , 2010; Bougeard, 2011]

- Obtained from the $\mathbf{w}^*$ coefficients which reflect the importance of the variables to build the global components $\mathbf{t}$
- Can be computed for each and several dimensions

Supervised (K+1)-block analyses: interpretation tools

Optimal dimension

- Select the optimal number of dimension H to be taken into account
- E.g., minimization of the cross-validated prediction error

Block-importance [Vivien, 2005; Bougeard, 2011]

- Obtained from the $\mathbf{a}_k$ coefficients which reflect the links between the block-components $\mathbf{t}_k$ and $\mathbf{u}$
- Can be computed for each and several dimensions

Variable-importance [Wold, 1994; Gosselin , 2010; Bougeard, 2011]

- Obtained from the $\mathbf{w}^*$ coefficients which reflect the importance of the variables to build the global components $\mathbf{t}$
- Can be computed for each and several dimensions

Extensions for supervised (K+1)-block analyses: clusterwise (r-)multiblock RA

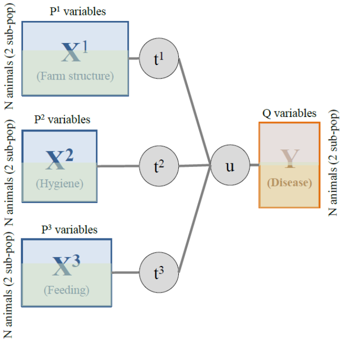
Main aim

Explore the links between the blocks while taking into account their complex structure, i.e.:

- Known block-structure and block-links,
- Unknown sub-populations of observations.

Sub-aims

1. Summarize each block of variables by components adjusted to the data features (i.e., ill-conditioned multidimensional blocks),
2. Get the partition of the observations into clusters,
3. Get (multiblock) regression models for each cluster.



Extensions for supervised (K+1)-block analyses: clusterwise (r-)multiblock RA

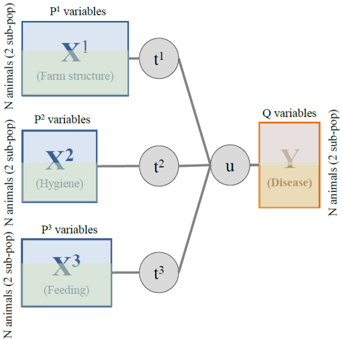
Main aim

Explore the links between the blocks while taking into account their complex structure, i.e.:

- Known block-structure and block-links,
- Unknown sub-populations of observations.

Sub-aims

1. Summarize each block of variables by components adjusted to the data features (i.e., ill-conditioned multidimensional blocks),
2. Get the partition of the observations into clusters,
3. Get (multiblock) regression models for each cluster.



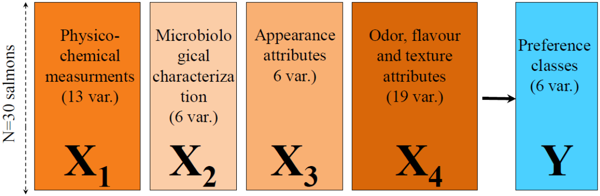
Extensions for supervised (K+1)-block analyses: clusterwise (r-)multiblock RA

[Bougeard, 2017, 2018]

Algorithm

1. Start from an initialization of the N observations into G clusters
2. For each observation n
   - Compute R-MBRA where n belongs alternatively to each of the G clusters
   - For each of the G solutions, compute the criterion $C = \sum_g ||\mathbf{Y}_g - \sum_h \mathbf{t}_g^{(h)}(\mathbf{c}_g^{(h)})'||^2$
   - Update the assignment of n to the cluster which minimize C
   - Update the regression coefficients
3. Repeat the procedure for several initializations and select the best one.

Standard mbRA: Eurosalmon data [Cardinal et al., 2004]



Salmon data features

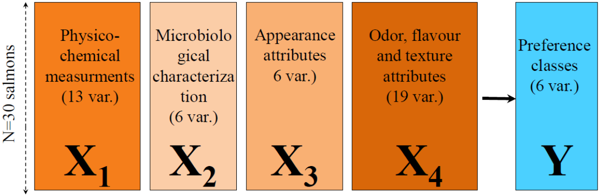
- **Y**: 6 preference classes from 1063 consumers [Semenou et al., 2007],
- **X**: 44 potential preference drivers organized into 4 blocks,

Aims

- **Descriptive**: explain the consumer preferences with the explanatory variables and blocks in relation with the tasted salmons,
- **Predictive**: assess the key drivers of preference at the variable and block levels.

anses

Standard mbRA: Eurosalmon data [Cardinal et al., 2004]



Salmon data features

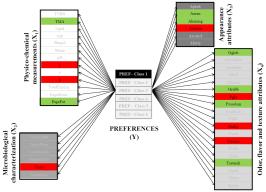
- **Y**: 6 preference classes from 1063 consumers [Semenou et al., 2007],
- **X**: 44 potential preference drivers organized into 4 blocks,

Aims

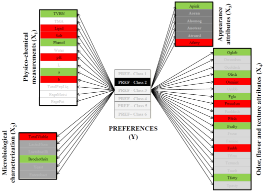
- **Descriptive**: explain the consumer preferences with the explanatory variables and blocks in relation with the tasted salmons,
- **Predictive**: assess the key drivers of preference at the variable and block levels.

Key drivers of preference at the variable level (2)

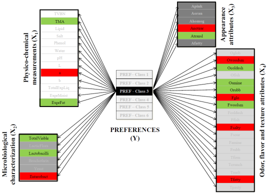
Regression coefficients and bootstraped tolerance interval. Optimal model with 4 components.



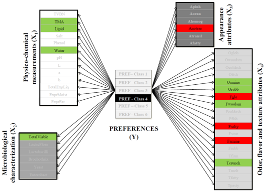
Pref 1 (N=121)



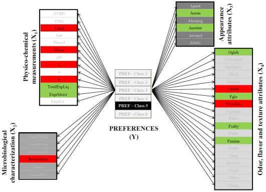
Pref 2 (N=74)



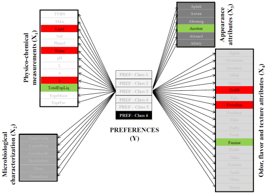
Pref 3 (N=349)



Pref 4 (N=78)



Pref 5 (N=404)

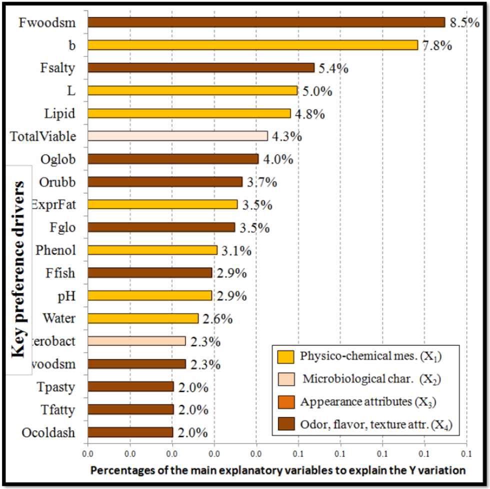


Pref 6 (N=37)

Results are difficult to sum up → Difficulties to get overall interpretation of key drivers.

Standard mbRA: Key drivers of preference at the variable-level

Variable Importance expressed as percentage and bootstraped tolerance interval. Optimal model with 4 components.



Interpretation for overall preference

The model explains 82% of the variation in **Y** which is significantly explained by:

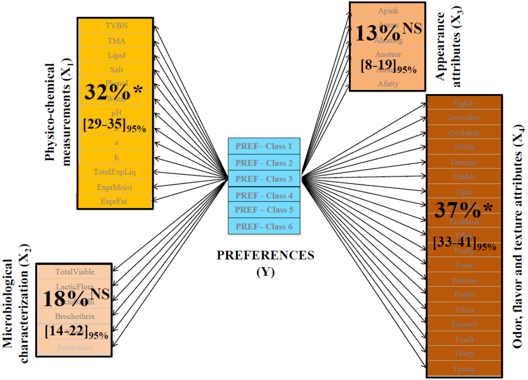
- The wood smoked flavor ("++" for classes 1, 3 and **4**, "–" for classes 2, **5** and 6),
- The hue parameter b^* (yellow) ("–" for classes 1, 2, **5** and 6),

  $\rightarrow$ Both these variables explain 14.3% of the overall preference.

anses

Standard mbRA: Key drivers of preference at the block-level

Block Importance expressed as percentage and bootstraped tolerance interval. Optimal model with 4 components.



Interpretation for overall preference

The model explains 82% of the variation in **Y**, which is significantly explained by:

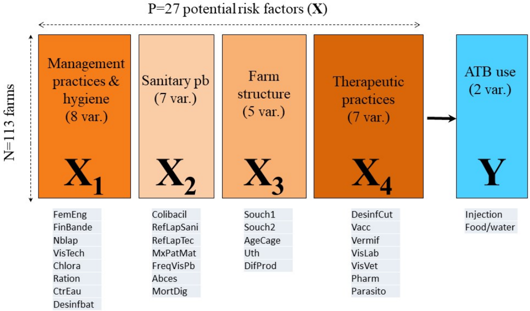
- The odor, flavor and texture attributes (37%),
- The physico-chemical measurements (32%),

anses

Clusterwise mbRA: 'antibiotic consumption in rabbit farms' data

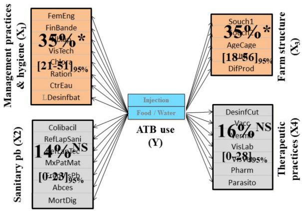
Data & aim

- Data: Retrospective survey conducted in 2010 in 113 French rabbit farms
- Aim: Identify risk markers for antibiotic use in rabbit farming

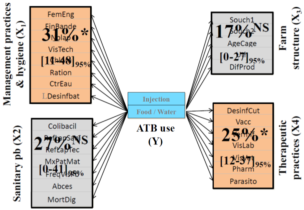


Clusterwise mbRA: Risk markers for each cluster [blocks]

Cluster 1
N_1=52 farms; R^2=0.56



Cluster 2
N_2=61 farms; R^2=0.65



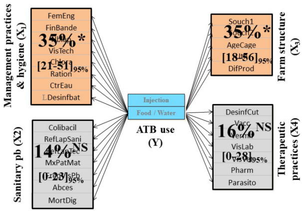
Interpretation

- Cluster 1: importance of management and hygiene practices ($\mathbf{X}_1$) and of the farm structure ($\mathbf{X}_3$)
- Cluster 2: importance of management and hygiene practices ($\mathbf{X}_1$) and of therapeutic practices ($\mathbf{X}_4$)
- NB: For all observations: R^2=0.25 ; importance of $\mathbf{X}_2$ (32%) and $\mathbf{X}_4$ (25%).

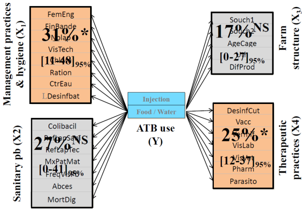
anses

Clusterwise mbRA: Risk markers for each cluster [blocks]

Cluster 1
N_1=52 farms; R^2=0.56



Cluster 2
N_2=61 farms; R^2=0.65



Interpretation

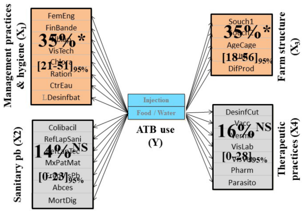
- Cluster 1: importance of management and hygiene practices ($\mathbf{X}_1$) and of the farm structure ($\mathbf{X}_3$)
- Cluster 2: importance of management and hygiene practices ($\mathbf{X}_1$) and of therapeutic practices ($\mathbf{X}_4$)
- NB: For all observations: R^2=0.25 ; importance of $\mathbf{X}_2$ (32%) and $\mathbf{X}_4$ (25%).

anses

Clusterwise mbRA: Risk markers for each cluster [blocks]

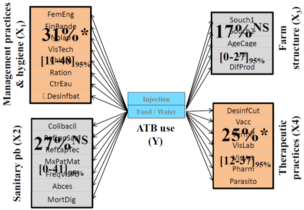
Cluster 1

N_1=52 farms; R^2=0.56



Cluster 2

N_2=61 farms; R^2=0.65



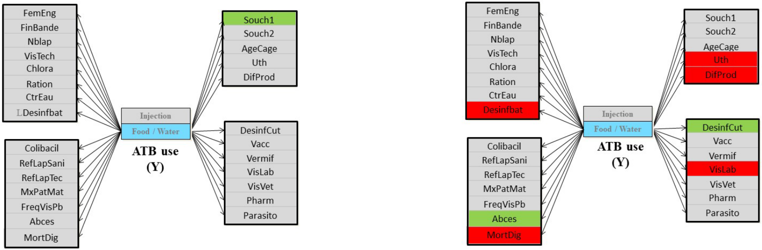
Interpretation

- Cluster 1: importance of management and hygiene practices ($\mathbf{X}_1$) and of the farm structure ($\mathbf{X}_3$)
- Cluster 2: importance of management and hygiene practices ($\mathbf{X}_1$) and of therapeutic practices ($\mathbf{X}_4$)
- NB: For all observations: R^2=0.25 ; importance of $\mathbf{X}_2$ (32%) and $\mathbf{X}_4$ (25%).

Clusterwise mbRA: Risk markers for each cluster [variables]

Cluster 1 (|Reg.coef.|>0.5)
N_1=52 farms; R^2=0.56

Cluster 2 (|Reg.coef.|>0.5)
N_2=61 farms; R^2=0.65



Grey : Not significant / Green : significant (positive link) & coef. >0.5 / Red : significant (negative link) & coef. <-0.5

Interpretation

- Cluster 1: importance of the rabbit strain,
- Cluster 2: importance of disinfection of the building, abscesses, digestive pb, . . .
- NB: For all observations: R^2=0.25; importance of the digestive pb.

Supervised (K+1)-block analyses with R

Standard supervised (K+1)-block analyses

- mbRA: `mbpcaiv` function in the `ade4` package (thus `mbrda` in in the `multiblock` package),
- mbPLS: `mbpls` function in the `ade4` package, `block.pls` function in the `MixOmics` package, `mbpls` function in the `multiblock` package
- Regularized-mbRA: `cw.multiblock` function with 'mbregular' option and a single-cluster ('G=1') in the `mbclusterwise` package

Extensions of supervised (K+1)-block analyses

- Discriminant mbPLS: `block.plsda` function `MixOmics` package, `mbplsda` function `packMBPLSDA` package
- Sparse mbPLS: `block.spls/block.splsda` functions in the `MixOmics` package, `smbpls` function in the `multiblock` package
- Multigroup mbPLS: `mint` functions (mint.block.pls, mint.block.spls, mint.block.plsda, mint.block.splsda) in the `MixOmics` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised (K+1)-block analyses with R

Standard supervised (K+1)-block analyses

- mbRA: `mbpcaiv` function in the `ade4` package (thus `mbrda` in in the `multiblock` package),
- mbPLS: `mbpls` function in the `ade4` package, `block.pls` function in the `MixOmics` package, `mbpls` function in the `multiblock` package
- Regularized-mbRA: `cw.multiblock` function with 'mbregular' option and a single-cluster ('G=1') in the `mbclusterwise` package

Extensions of supervised (K+1)-block analyses

- Discriminant mbPLS: `block.plsda` function `MixOmics` package, `mbplsda` function `packMBPLSDA` package
- Sparse mbPLS: `block.spls`/`block.splsda` functions in the `MixOmics` package, `smbpls` function in the `multiblock` package
- Multigroup mbPLS: `mint` functions (mint.block.pls, mint.block.spls, mint.block.plsda, mint.block.splsda) in the `MixOmics` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Supervised (K+1)-block analyses with R

Standard supervised (K+1)-block analyses

- mbRA: `mbpcaiv` function in the `ade4` package (thus `mbrda` in in the `multiblock` package),
- mbPLS: `mbpls` function in the `ade4` package, `block.pls` function in the `MixOmics` package, `mbpls` function in the `multiblock` package
- Regularized-mbRA: `cw.multiblock` function with 'mbregular' option and a single-cluster ('G=1') in the `mbclusterwise` package

Extensions of supervised (K+1)-block analyses

- Discriminant mbPLS: `block.plsda` function `MixOmics` package, `mbplsda` function `packMBPLSDA` package
- Sparse mbPLS: `block.spls/block.splsda` functions in the `MixOmics` package, `smbpls` function in the `multiblock` package
- Multigroup mbPLS: `mint` functions (mint.block.pls, mint.block.spls, mint.block.plsda, mint.block.splsda) in the `MixOmics` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

anses

Supervised (K+1)-block analyses with R

Standard supervised (K+1)-block analyses

- mbRA: `mbpcaiv` function in the `ade4` package (thus `mbrda` in in the `multiblock` package),
- mbPLS: `mbpls` function in the `ade4` package, `block.pls` function in the `MixOmics` package, `mbpls` function in the `multiblock` package
- Regularized-mbRA: `cw.multiblock` function with 'mbregular' option and a single-cluster ('G=1') in the `mbclusterwise` package

Extensions of supervised (K+1)-block analyses

- Discriminant mbPLS: `block.plsda` function `MixOmics` package, `mbplsda` function `packMBPLSDA` package
- Sparse mbPLS: `block.spls/block.splsda` functions in the `MixOmics` package, `smbpls` function in the `multiblock` package
- Multigroup mbPLS: `mint` functions (mint.block.pls, mint.block.spls, mint.block.plsda, mint.block.splsda) in the `MixOmics` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

anses

Supervised (K+1)-block analyses with R

Standard supervised (K+1)-block analyses

- mbRA: `mbpcaiv` function in the `ade4` package (thus `mbrda` in in the `multiblock` package),
- mbPLS: `mbpls` function in the `ade4` package, `block.pls` function in the `MixOmics` package, `mbpls` function in the `multiblock` package
- Regularized-mbRA: `cw.multiblock` function with 'mbregular' option and a single-cluster ('G=1') in the `mbclusterwise` package

Extensions of supervised (K+1)-block analyses

- Discriminant mbPLS: `block.plsda` function `MixOmics` package, `mbplsda` function `packMBPLSDA` package
- Sparse mbPLS: `block.spls/block.splsda` functions in the `MixOmics` package, `smbpls` function in the `multiblock` package
- Multigroup mbPLS: `mint` functions (mint.block.pls, mint.block.spls, mint.block.plsda, mint.block.splsda) in the `MixOmics` package
- Clusterwise: `cw.multiblock` function with 'mbpls'/'mbpcaiv'/'mbregular' options and a single-block **X** in the `mbclusterwise` package

Outline

1 Introduction

2 Supervised two-block analyses

3 Supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses
- Methods
- Application
- Doing my own supervised (K+K')-block analyses

5 Conclusion & perspectives

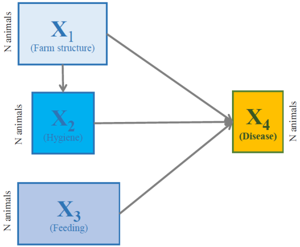
Relate (K+K')-blocks with a criterion

Aim

- Explore the relationships between blocks
- Blocks connected by the user (*a priori* information)

How blocks are linked?

- Raw data sets . . .
- Are summarized with block-components . . .
- Which are linked by a criterion*



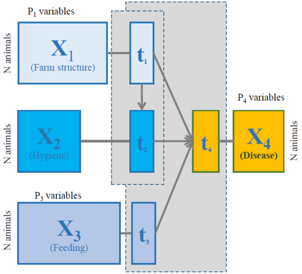
Relate (K+K')-blocks with a criterion

Aim

- Explore the relationships between blocks
- Blocks connected by the user (*a priori* information)

How blocks are linked?

- Raw data sets . . .
- Are summarized with block-components . . .
- Which are linked by a criterion*



Relate (K+K')-blocks with a criterion

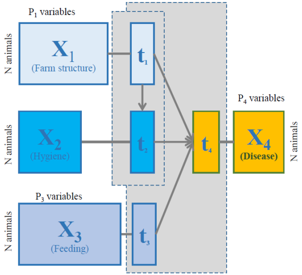
Aim

- Explore the relationships between blocks
- Blocks connected by the user (*a priori* information)

How blocks are linked?

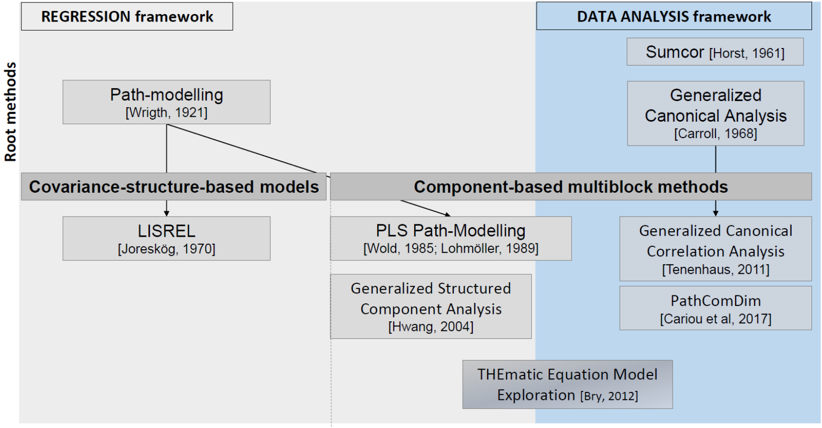
- Raw data sets . . .
- Are summarized with block-components . . .
- Which are linked by a criterion*

* Methods which are not based on a criterion are not given here.



Supervised (K+K')-block analyses: methods

(K+K')-block analyses come from two different frameworks.



In the following, only component-based multiblock methods (with criterion) will be studied.

(Three) supervised (K+K')-block analyses: criteria (in a nutshell)

Regularized Generalized Canonical Correlation Analysis (rGCCA) [Tenenhaus, 2011]

$$\max \quad \sum_{k,l=1,k\neq l}^{K} d_{kl}\,\mathrm{cov}^2(\mathbf{X_k w_k}, \mathbf{X_l w_l}) \quad \text{s.t.} \quad \tau_\mathbf{k}\|\mathbf{w_k}\|^2 + (\mathbf{1}-\tau_\mathbf{k})\,\mathrm{var}(\mathbf{X_k w_k}) = \mathbf{1}$$

- Symmetrical links
- Several components per block (block-dim. are supposed to be identical)

Regularized Generalized Structured Component Analysis (rGSCA) [Hwang, 2004]

$$\min \quad \|\mathbf{XW_M} - \mathbf{XWB}\|^2 + \|\mathbf{XI_R} - \mathbf{XWC}\|^2 + \lambda_1\|\mathbf{B}\|^2 + \lambda_2\|\mathbf{W}\|^2 + \lambda_3\|\mathbf{C}\|^2 \quad \text{s.t.} \quad \mathrm{diag}(\mathbf{W^T X^T X W}) = \mathbf{I}, \quad \lambda_1 \geq \mathbf{0}, \quad \lambda_2 \geq \mathbf{0}, \quad \lambda_3 \geq \mathbf{0}$$

- Asymmetrical links (regression)
- Blocks are supposed to be unidimensional

(Simplified) THEmatic Equation Model Exploration (THEME) [Bry, 2015]

$$\max \quad \prod_{m=1}^{M}\left(1 - \frac{\|\mathbf{Xw_m} - \mathbf{XWb_m}\|^2}{\|\mathbf{Xw_m}\|^2}\right)\prod_{k=1}^{K}\left(\sum_{p_k=1}^{P_k}\mathrm{cor}^2(\mathbf{X_k w_k}, \mathbf{x_{p_k}})\right) \quad \text{s.t.} \quad \|\mathbf{X_k w_k}\|^2 = \mathbf{1}$$

- Asymmetrical links (regression)
- Higher-rank solutions (selection of the relevant number of components per block)

anses

(Three) supervised (K+K')-block analyses: criteria (in a nutshell)

Regularized Generalized Canonical Correlation Analysis (rGCCA) [Tenenhaus, 2011]

$$\max \quad \sum_{k,l=1,k\neq l}^{K} d_{kl}\,\mathrm{cov}^2(\mathbf{X_k w_k}, \mathbf{X_l w_l}) \quad \text{s.t.} \quad \tau_\mathbf{k}\|\mathbf{w_k}\|^2 + (\mathbf{1}-\tau_\mathbf{k})\,\mathrm{var}(\mathbf{X_k w_k}) = \mathbf{1}$$

- Symmetrical links
- Several components per block (block-dim. are supposed to be identical)

Regularized Generalized Structured Component Analysis (rGSCA) [Hwang, 2004]

$$\min \quad \|\mathbf{XW_M} - \mathbf{XWB}\|^2 + \|\mathbf{XI_R} - \mathbf{XWC}\|^2 + \lambda_1\|\mathbf{B}\|^2 + \lambda_2\|\mathbf{W}\|^2 + \lambda_3\|\mathbf{C}\|^2 \quad \text{s.t.} \quad \mathrm{diag}(\mathbf{W^T X^T X W}) = \mathbf{I}, \quad \lambda_1 \geq \mathbf{0}, \quad \lambda_2 \geq \mathbf{0}, \quad \lambda_3 \geq \mathbf{0}$$

- Asymmetrical links (regression)
- Blocks are supposed to be unidimensional

(Simplified) THEmatic Equation Model Exploration (THEME) [Bry, 2015]

$$\max \quad \prod_{m=1}^{M}\left(1 - \frac{\|\mathbf{Xw_m} - \mathbf{XWb_m}\|^2}{\|\mathbf{Xw_m}\|^2}\right)\prod_{k=1}^{K}\left(\sum_{p_k=1}^{P_k}\mathrm{cor}^2(\mathbf{X_k w_k}, \mathbf{x_{p_k}})\right) \quad \text{s.t.} \quad \|\mathbf{X_k w_k}\|^2 = \mathbf{1}$$

- Asymmetrical links (regression)
- Higher-rank solutions (selection of the relevant number of components per block)

anses

(Three) supervised (K+K')-block analyses: criteria (in a nutshell)

Regularized Generalized Canonical Correlation Analysis (rGCCA) [Tenenhaus, 2011]

$$\max \quad \sum_{k,l=1,k\neq l}^{K} d_{kl}\, \mathrm{cov}^2(\mathbf{X_k w_k}, \mathbf{X_l w_l}) \quad \text{s.t.} \quad \tau_\mathbf{k} \|\mathbf{w_k}\|^2 + (\mathbf{1} - \tau_\mathbf{k})\, \mathrm{var}(\mathbf{X_k w_k}) = \mathbf{1}$$

- Symmetrical links
- Several components per block (block-dim. are supposed to be identical)

Regularized Generalized Structured Component Analysis (rGSCA) [Hwang, 2004]

$$\min \quad \|\mathbf{XW_M} - \mathbf{XWB}\|^2 + \|\mathbf{XI_R} - \mathbf{XWC}\|^2 + \lambda_\mathbf{1}\|\mathbf{B}\|^2 + \lambda_\mathbf{2}\|\mathbf{W}\|^2 + \lambda_\mathbf{3}\|\mathbf{C}\|^2 \quad \text{s.t.} \quad \mathrm{diag}(\mathbf{W^T X^T X W}) = \mathbf{I}, \quad \lambda_\mathbf{1} \geq \mathbf{0}, \quad \lambda_\mathbf{2} \geq \mathbf{0}, \quad \lambda_\mathbf{3} \geq \mathbf{0}$$

- Asymmetrical links (regression)
- Blocks are supposed to be unidimensional

(Simplified) THEmatic Equation Model Exploration (THEME) [Bry, 2015]

$$\max \quad \prod_{m=1}^{M} \left(1 - \frac{\|\mathbf{Xw_m} - \mathbf{XWb_m}\|^2}{\|\mathbf{Xw_m}\|^2}\right) \prod_{k=1}^{K} \left(\sum_{p_k=1}^{P_k} \mathrm{cor}^2(\mathbf{X_k w_k}, \mathbf{x_{p_k}})\right) \quad \text{s.t.} \quad \|\mathbf{X_k w_k}\|^2 = \mathbf{1}$$

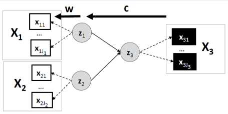
- Asymmetrical links (regression)
- Higher-rank solutions (selection of the relevant number of components per block)

Supervised (K+K')-block analyses: prediction model

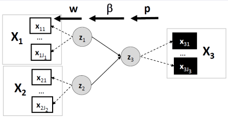
Work in progress with M. Hanafi - Application to PLS Path Modelling

Two proposed estimation of the regression coefficients $\mathbf{B}$ such as $\mathbf{X} = \mathbf{X}\mathbf{B} + \mathbf{R}$

$$\hat{\mathbf{B}}_{lk} = \begin{cases} \mathbf{w}_l \mathbf{c}_{lk}^T & \text{for the PLSR-like estimation} \\ \beta_{lk} \mathbf{w}_l \mathbf{p}_k^T & \text{for the PLSPM-like estimation (=PLSpredict) [Shmueli, 2016]} \end{cases}$$



(a) PLSR-like estimation.



(b) PLSPM-like estimation.

Property: These estimations are reformulations of the structural model.

Deflation

- Explanatory blocks are deflated with respect to their measurement model ($\mathbf{w}_k \mathbf{p}_k^T$)
- Intermediate and blocks to be explained are deflated with respect to the prediction model ($\mathbf{B}$)

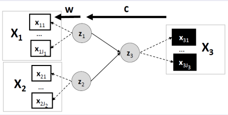
anses

Supervised (K+K')-block analyses: prediction model

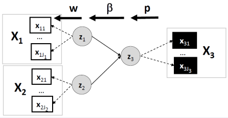
Work in progress with M. Hanafi - Application to PLS Path Modelling

Two proposed estimation of the regression coefficients B such as $\mathbf{X} = \mathbf{XB} + \mathbf{R}$

$$\hat{\mathbf{B}}_{lk} = \begin{cases} \mathbf{w}_l \mathbf{c}_{lk}^T & \text{for the PLSR-like estimation} \\ \beta_{lk} \mathbf{w}_l \mathbf{p}_k^T & \text{for the PLSPM-like estimation (=PLSpredict) [Shmueli, 2016]} \end{cases}$$



(c) PLSR-like estimation.



(d) PLSPM-like estimation.

Property: These estimations are reformulations of the structural model.

Deflation

- Explanatory blocks are deflated with respect to their measurement model ($\mathbf{w_k p_k^T}$)
- Intermediate and blocks to be explained are deflated with respect to the prediction model (**B**)

Supervised (K+K')-block analyses: Advices for application

Explain or predict?

- First explain (rGCCA, Path-Comdim)
- If the explanation is good enough, model and predict (rGSCA, THEME)

Uni or multidimensional blocks?

- In practice, multidimensional blocks

Within-block multicollinearity?

- Data analysis framework: regularization of the block-norm constraints
- Regression framework: elastic-net (=lasso + ridge) regularization
- Both: Data summary with component(s)

Supervised (K+K')-block analyses: Advices for application

Explain or predict?

- First explain (rGCCA, Path-Comdim)
- If the explanation is good enough, model and predict (rGSCA, THEME)

Uni or multidimensional blocks?

- In practice, multidimensional blocks

Within-block multicollinearity?

- Data analysis framework: regularization of the block-norm constraints
- Regression framework: elastic-net (=lasso + ridge) regularization
- Both: Data summary with component(s)

Supervised (K+K')-block analyses: Advices for application

Explain or predict?

- First explain (rGCCA, Path-Comdim)
- If the explanation is good enough, model and predict (rGSCA, THEME)

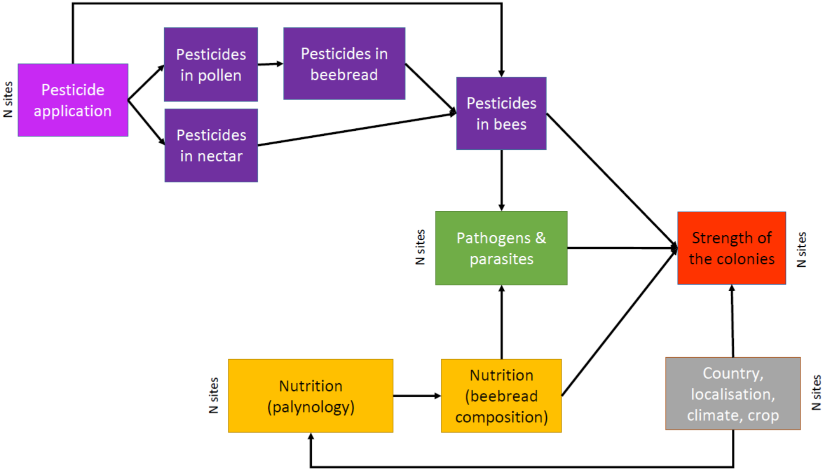
Uni or multidimensional blocks?

- In practice, multidimensional blocks

Within-block multicollinearity?

- Data analysis framework: regularization of the block-norm constraints
- Regression framework: elastic-net (=lasso + ridge) regularization
- Both: Data summary with component(s)

Explain the bee-mortality (In progress)



Supervised (K+K')-block analyses with R

Standard supervised (K+K')-block analyses

- PLS-PM: `SEMinR` package
- GSCA: `gsca` package or https://www.gscapro.com/
- GCCA: `rgcca` function in the `RGCCA` package or https://github.com/rgcca-factory/RGCCA
- THEME: `SCGLR` package
- PathComDim: `MBAnalysis` package (In progress)

Extension of supervised (K+K')-block analyses

- Sparse: `sgcca` function in the `RGCCA` package
- 'Clusterwise': `rebus.pls` function in the `plspm` package
- 'Multigroup', 'mixed variables' with https://www.gscapro.com/

Supervised (K+K')-block analyses with R

Standard supervised (K+K')-block analyses

- PLS-PM: `SEMinR` package
- GSCA: `gsca` package or https://www.gscapro.com/
- GCCA: `rgcca` function in the `RGCCA` package or https://github.com/rgcca-factory/RGCCA
- THEME: `SCGLR` package
- PathComDim: `MBAnalysis` package (In progress)

Extension of supervised (K+K')-block analyses

- Sparse: `sgcca` function in the `RGCCA` package
- 'Clusterwise': `rebus.pls` function in the `plspm` package
- 'Multigroup', 'mixed variables' with https://www.gscapro.com/

Outline

1 Introduction

2 Supervised two-block analyses

3 Supervised (K+1)-block analyses

4 Supervised (K+K')-block analyses

5 Conclusion & perspectives

Conclusion related to multiblock methods

From data ...

- Data that answer complex questions come from different sources → Multiblock
- Numerous blocks with complex links → (K+K')-block methods
- Blocks are multidimensional → Component-based methods with several dimensions
- Users usually seek to explain block(s) → Supervised with models

... To methods

- Multiblock methods are increasingly applied
- Development of multiblock methods from 2-block to (K+1)-blocks and (K+K')-blocks
- But many points remain to be clarified / developed (new methods)

anses

Conclusion related to multiblock methods

From data . . .

- Data that answer complex questions come from different sources → Multiblock
- Numerous blocks with complex links → (K+K')-block methods
- Blocks are multidimensional → Component-based methods with several dimensions
- Users usually seek to explain block(s) → Supervised with models

. . . To methods

- Multiblock methods are increasingly applied
- Development of multiblock methods from 2-block to (K+1)-blocks and (K+K')-blocks
- But many points remain to be clarified / developed (new methods)

anses

Perspectives related to multiblock methods

(Some) extensions related to the data features

- Structure of observations in known (covariables / multigroup) or unknown groups (clusterwise)
- Temporal structure of blocks
- Large number of variables (e.g., regularization, sparse)
- Mixed data (numeric, nominal ordinal)

Other extensions

- Prediction model (write model, relevant deflation, component selection, elastic-net regularization)
- Link with IA / machine learning
  - Integrate IA in multiblock prediction models (e.g., neural networks)
  - Integrate multiblock methods in machine learning packages/softwares

Perspectives related to multiblock methods

(Some) extensions related to the data features

- Structure of observations in known (covariables / multigroup) or unknown groups (clusterwise)
- Temporal structure of blocks
- Large number of variables (e.g., regularization, sparse)
- Mixed data (numeric, nominal ordinal)

Other extensions

- Prediction model (write model, relevant deflation, component selection, elastic-net regularization)
- Link with IA / machine learning
  - Integrate IA in multiblock prediction models (e.g., neural networks)
  - Integrate multiblock methods in machine learning packages/softwares

anses

Perspectives related to the application of multiblock methods

Know and be able to (link between developers and users)

- Train and disseminate methods
- Give advices for application to users
- Develop packages or softwares with interpretation tools

Apply and publish

- Multi-source data come from all fields
- Apply to different fields (psychometry $\rightarrow$ chimiometry $\rightarrow$ biology (e.g., sensometry, epidemiology, omic) $\rightarrow$ all fields)
- Publish in journals of application

Perspectives related to the application of multiblock methods

Know and be able to (link between developers and users)

- Train and disseminate methods
- Give advices for application to users
- Develop packages or softwares with interpretation tools

Apply and publish

- Multi-source data come from all fields
- Apply to different fields (psychometry $\rightarrow$ chimiometry $\rightarrow$ biology (e.g., sensometry, epidemiology, omic) $\rightarrow$ all fields)
- Publish in journals of application

Working together

A multiblock joined program

- Métaprogramme INRAe DIGIT-BIO 2022 « Biologie Numérique pour explorer et prédire le vivant » / Consortium inter-disciplinaire 'MIMS' (Regards Méthodologiques Croisés pour l'Intégration de données Multi-sources)
  `Contact : mohamed.hanafi@oniris-nantes.fr`

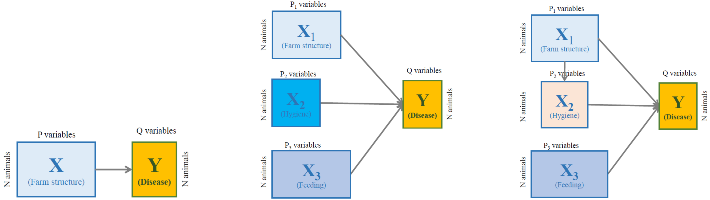
A part of the "French multiblock team"! Join us!

Supervised multiblock analyses

Cases of two-blocks, (K+1)-blocks, (K+K')-blocks

Stéphanie Bougeard

French Agency for Food, Environmental, Occupational Health & Safety (Anses), Ploufragan, France



Journée Analyses Factorielles
March 30 2023, INRAe Jouy-en-Josas