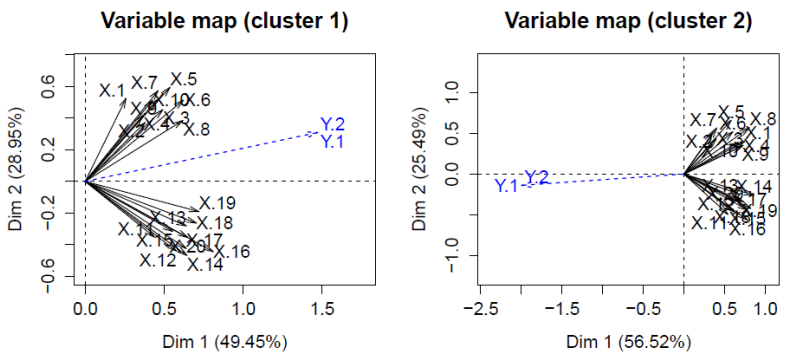
Clusterwise multiblock PLS regression

N'Dèye Niang[1], Stéphanie Bougeard[2] & Gilbert Saporta[1]

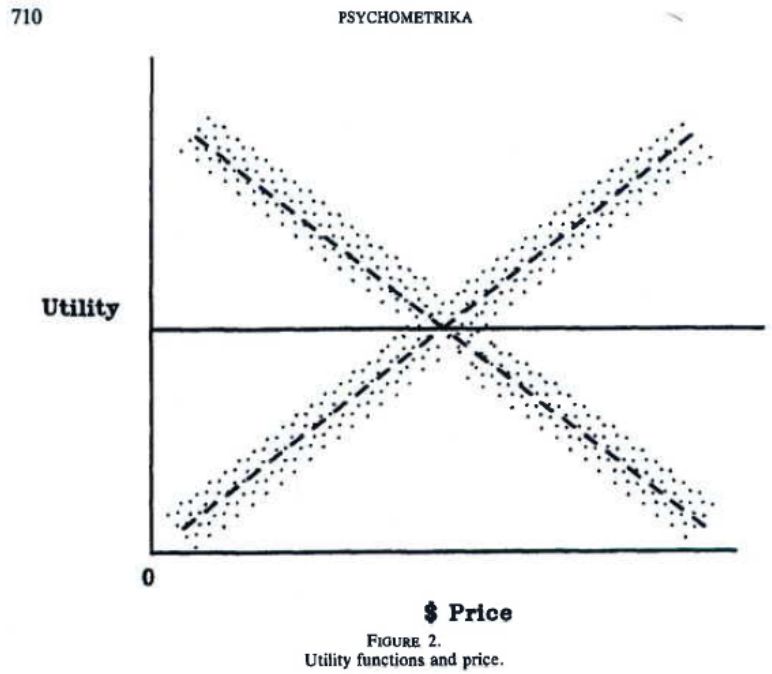
[1] *CEDRIC CNAM, Paris, France*
[2] *French Agency for Food, Environmental, Occupational Health & Safety (Anses), Ploufragan, France*



The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, 12-14 December 2015, London, UK

Table of contents

Context of clusterwise regression



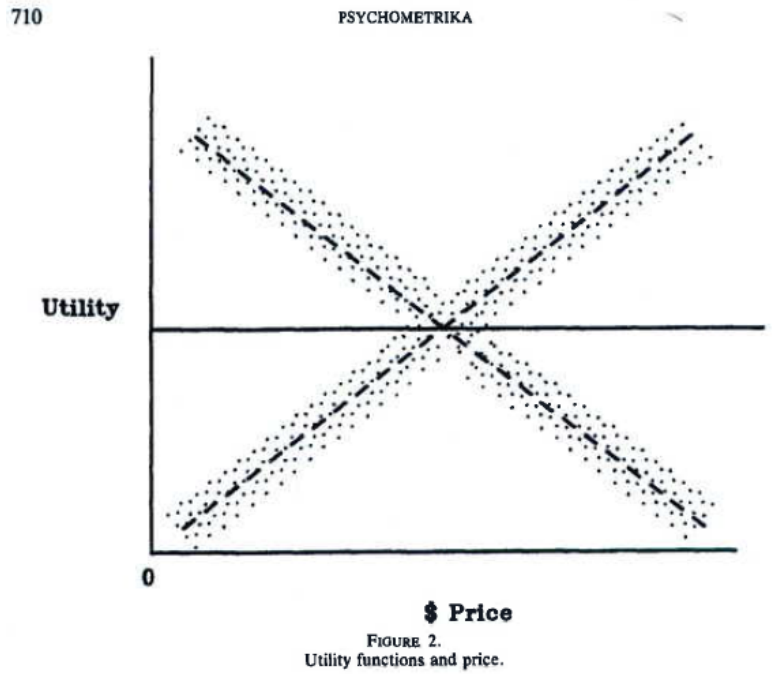
From DeSarbo et al. (1989), Psychometrika

- "For one group of consumers, higher utility corresponds to lower prices . . .
- . . . while the opposite is true for the other group."

Clusterwise principle

- **Multiple regression**: No such grahical display for easy clustering detection.
- **Clusterwise regression**: cluster the data and simultaneously compute these cluster regression coefficients.

Context of clusterwise regression



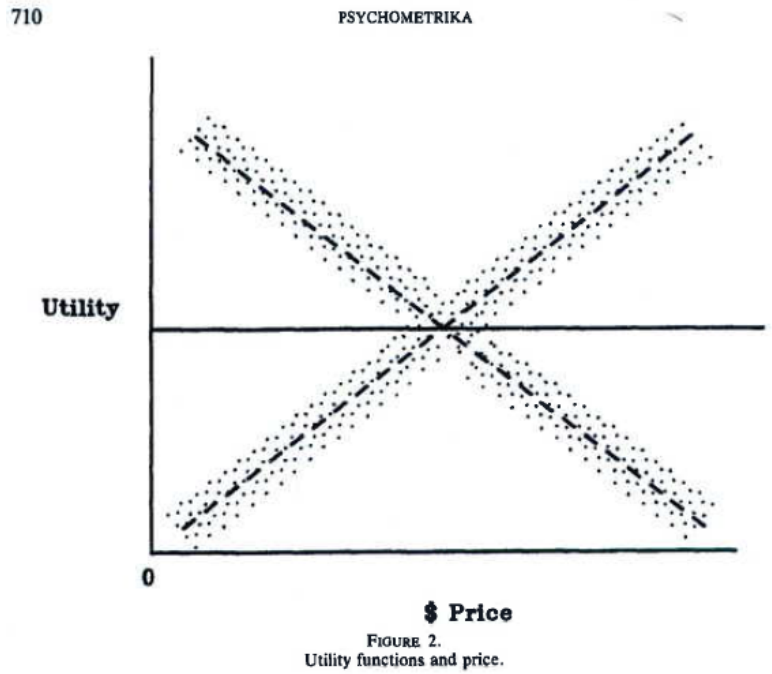
From DeSarbo et al. (1989), Psychometrika

- "For one group of consumers, higher utility corresponds to lower prices . . .
- . . . while the opposite is true for the other group."

Clusterwise principle

- **Multiple regression**: No such grahical display for easy clustering detection.
- **Clusterwise regression**: cluster the data and simultaneously compute these cluster regression coefficients.

Context of clusterwise regression



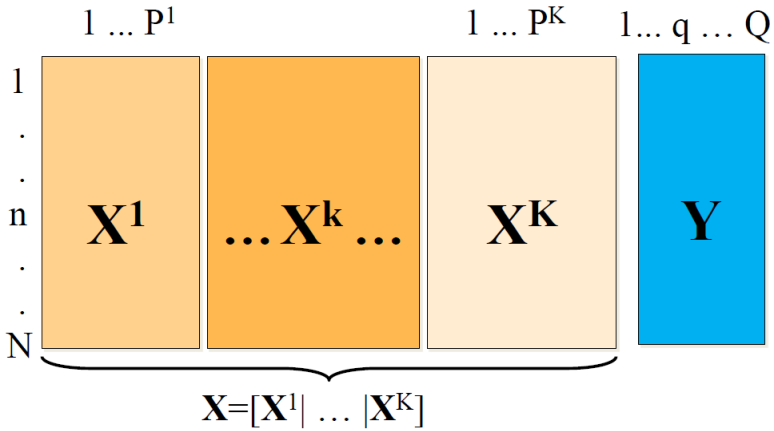
From DeSarbo et al. (1989), Psychometrika

- "For one group of consumers, higher utility corresponds to lower prices . . .
- . . . while the opposite is true for the other group."

Clusterwise principle

- **Multiple regression**: No such grahical display for easy clustering detection.
- **Clusterwise regression**: cluster the data and simultaneously compute these cluster regression coefficients.

Extension of clusterwise regression to clusterwise multiblock regression



Example of individual clustering in G=2 clusters

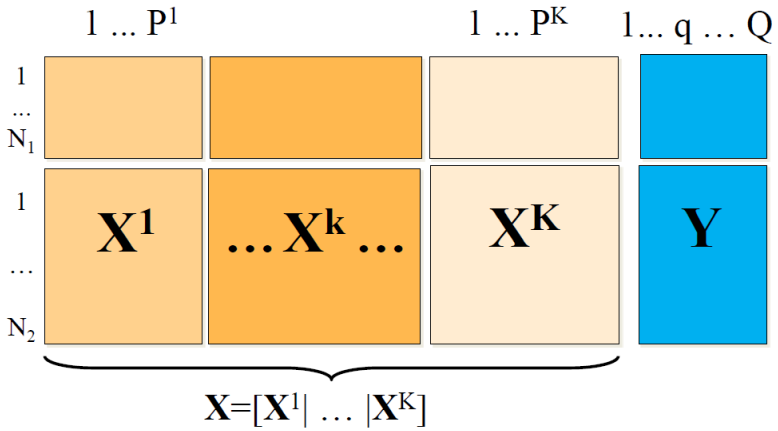
Clusterwise multiblock data features

- Large number of explanatory variables organized in meaningful blocks $[\mathbf{X}^1, \ldots, \mathbf{X}^K]$ (**known** block structure),
- Several variables **Y** to explain and predict,
- **Unknown** cluster structure of the *N* individuals (in *G* clusters).

Aims

- **Clustering**: Get an optimal clustering of individuals (cluster number defined in advance) . . . ,
- **Multiblock regression**: . . . and compute the cluster multiblock regression coefficients within each cluster in order to improve the **Y** explanation and prediction.

Extension of clusterwise regression to clusterwise multiblock regression



Example of individual clustering in G=2 clusters

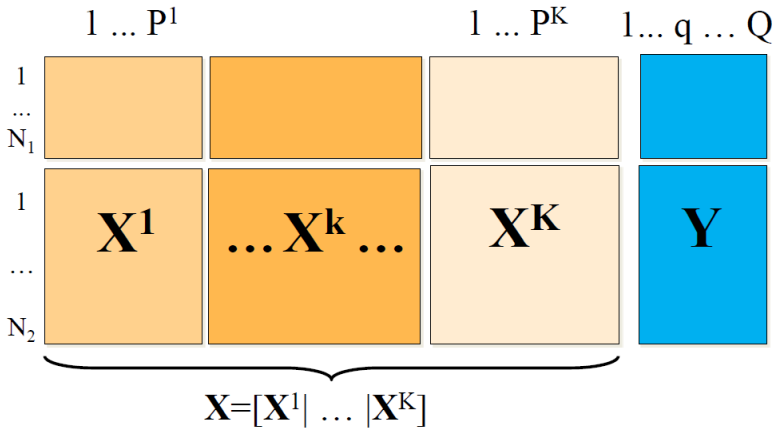
Clusterwise multiblock data features

- Large number of explanatory variables organized in meaningful blocks $[\mathbf{X}^1, \ldots, \mathbf{X}^K]$ (**known** block structure),
- Several variables **Y** to explain and predict,
- **Unknown** cluster structure of the *N* individuals (in *G* clusters).

Aims

- **Clustering**: Get an optimal clustering of individuals (cluster number defined in advance) . . . ,
- **Multiblock regression**: . . . and compute the cluster multiblock regression coefficients within each cluster in order to improve the **Y** explanation and prediction.

Extension of clusterwise regression to clusterwise multiblock regression



Example of individual clustering in G=2 clusters

Clusterwise multiblock data features

- Large number of explanatory variables organized in meaningful blocks $[\mathbf{X}^1, \ldots, \mathbf{X}^K]$ (**known** block structure),
- Several variables **Y** to explain and predict,
- **Unknown** cluster structure of the *N* individuals (in *G* clusters).

Aims

- **Clustering**: Get an optimal clustering of individuals (cluster number defined in advance) . . . ,
- **Multiblock regression**: . . . and compute the cluster multiblock regression coefficients within each cluster in order to improve the **Y** explanation and prediction.

Table of contents

Available clusterwise regression methods: finite mixture models

Mixture model field (clusterwise methods)

- Earliest works: DeSarbo & Cron, 1988; Hennig, 2000
- Minimize $\sum_{g=1}^{G} \|\mathbf{Y}_g - \mathbf{X}_g\mathbf{B}_g\|^2$ by means of an EM algorithm,
  Assumption: response variable is distributed as a finite mixture of conditional normal densities.

Advantages and limits

- Fast algorithm, available programs (commercial softwares, `Flexmix` R package)
- Necessary conditions
  - Multivariate normal distribution of dependent variable $\mathbf{Y}$,
  - N_g (nb of individuals within the gth cluster) > P (nb of explanatory variables)
- Oriented towards modelling and not towards prediction.

Available clusterwise regression methods: finite mixture models

Mixture model field (clusterwise methods)

- Earliest works: DeSarbo & Cron, 1988; Hennig, 2000
- Minimize $\sum_{g=1}^{G} \|\mathbf{Y}_g - \mathbf{X}_g\mathbf{B}_g\|^2$ by means of an EM algorithm,
  Assumption: response variable is distributed as a finite mixture of conditional normal densities.

Advantages and limits

- Fast algorithm, available programs (commercial softwares, `Flexmix` R package)
- Necessary conditions
  - Multivariate normal distribution of dependent variable $\mathbf{Y}$,
  - N_g (nb of individuals within the gth cluster) > P (nb of explanatory variables)
- Oriented towards modelling and not towards prediction.

Available clusterwise regression methods: geometrical approach

Distance-based model field (typological methods)

- Earliest works: Diday, 1974 (typological factorial analysis); Charles, 1977; Späth, 1979
- Minimize $\sum_{g=1}^{G} \|\mathbf{Y}_g - \mathbf{X}_g\mathbf{B}_g\|^2$ by means of a K-means like algorithm,
- Interesting extensions to high-dimensional data: typological PCR (Charles, 1977), typological PLS (Vinzi, 2005; Preda, 2005).

Advantages and limits

- No available programs,
- Non-monotonicity decrease of the criterion (batch *versus* stochastic algorithm),
- Oriented towards both modelling and prediction.

Available clusterwise regression methods: geometrical approach

Distance-based model field (typological methods)

- Earliest works: Diday, 1974 (typological factorial analysis); Charles, 1977; Späth, 1979
- Minimize $\sum_{g=1}^{G} \|\mathbf{Y}_g - \mathbf{X}_g\mathbf{B}_g\|^2$ by means of a K-means like algorithm,
- Interesting extensions to high-dimensional data: typological PCR (Charles, 1977), typological PLS (Vinzi, 2005; Preda, 2005).

Advantages and limits

- No available programs,
- Non-monotonicity decrease of the criterion (batch *versus* stochastic algorithm),
- Oriented towards both modelling and prediction.

Available clusterwise regression methods: geometrical approach

Distance-based model field (typological methods)

- Earliest works: Diday, 1974 (typological factorial analysis); Charles, 1977; Späth, 1979
- Minimize $\sum_{g=1}^{G} \|\mathbf{Y}_g - \mathbf{X}_g\mathbf{B}_g\|^2$ by means of a K-means like algorithm,
- Interesting extensions to high-dimensional data: typological PCR (Charles, 1977), typological PLS (Vinzi, 2005; Preda, 2005).

Advantages and limits

- No available programs,
- Non-monotonicity decrease of the criterion (batch *versus* stochastic algorithm),
- Oriented towards both modelling and prediction.

Proposition: Extend this approach to multiblock regression with high-dimensional data (+ guarantee of monotonicity, prediction and program availability).

Table of contents

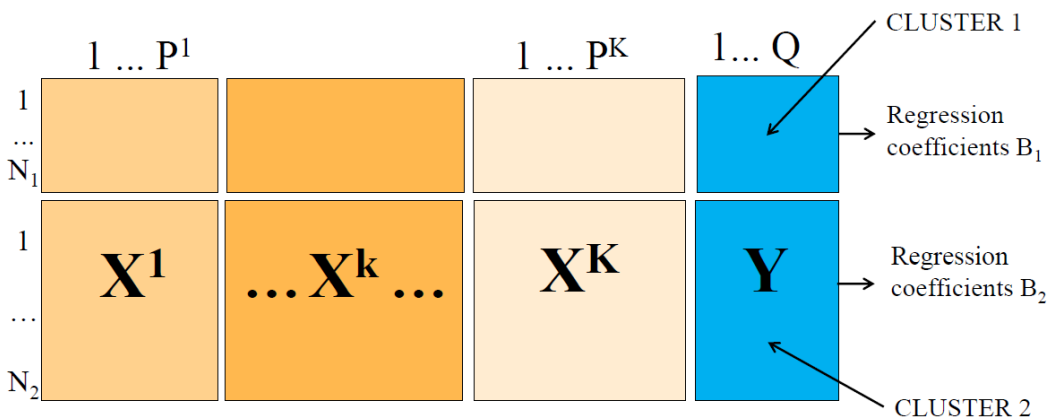
Clusterwise multiblock PLS: aims

Context: clusterwise multiblock regression.

Aims

1. **Clustering**: Get an optimal clustering of individuals (cluster number defined in advance) . . . ,
2. **Multiblock regression**: . . . and compute the cluster multiblock regression coefficients within each cluster in order to improve the **Y** explanation and prediction.

Two parameters are chosen: the number of clusters G and of components H.

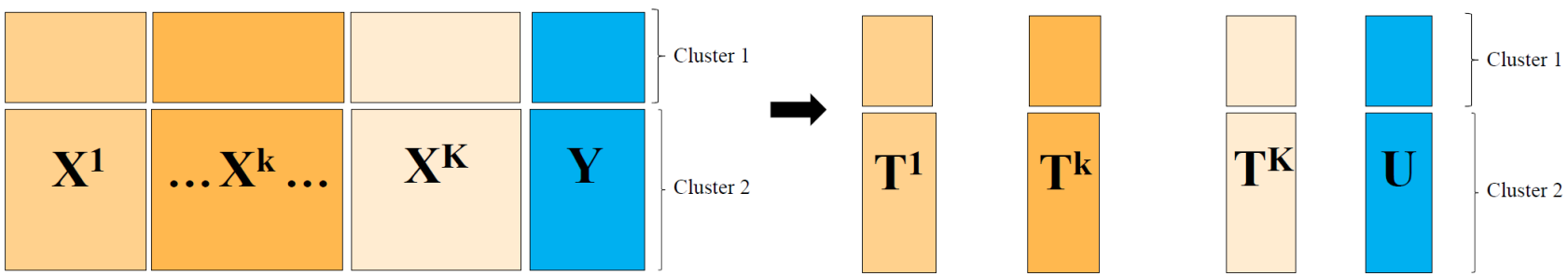


Clusterwise multiblock PLS: criterion

Get the optimal clusters on components which summarize **X**

Min. the sum of the residuals $\sum_{g=1}^{G} \|\mathbf{Y}_g - \underbrace{\sum_k \mathbf{T}_g^{k(H)} (a_g^{k(H)} \mathbf{C}_g^{(H)'})}_{\text{Predicted } \mathbf{Y}_g} \|^2$ where:

- $\mathbf{T}_g^{k(H)}$ are the matrices of block components with H (given) dimensions obtained from the G mbPLS, such as $\mathbf{T}_g^{k(H)} = [t_g^{k(1)} | \ldots | t_g^{k(H)}]$
- $a_g^{k(H)}$ are the normalized covariances between components from $\mathbf{X}_g^k$ and $\mathbf{Y}_g$
- $(\mathbf{C}_1^{(H)'}, \ldots, \mathbf{C}_G^{(H)'})$ are the regression coefficients of $\mathbf{Y}_g$ on $(\sum_k a_g^{k(H)} \mathbf{T}_g^{k(H)})$



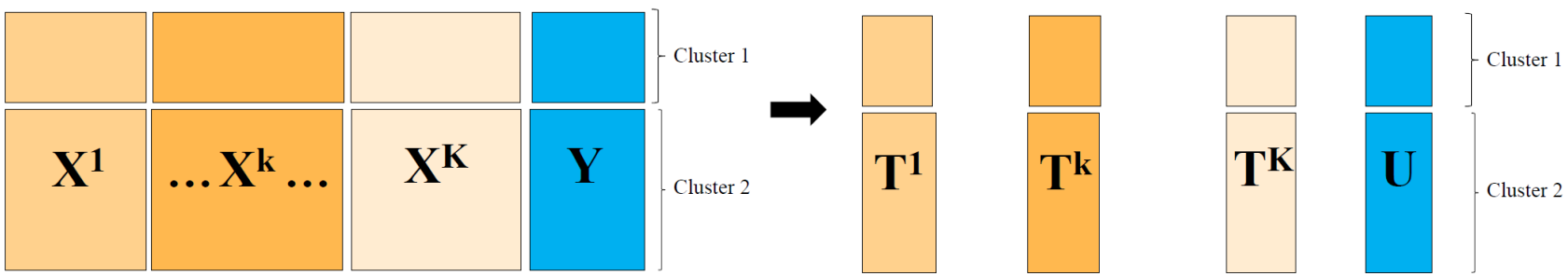
From the original data … … to the partial components

Clusterwise multiblock PLS: criterion

Get the optimal clusters on components which summarize **X**

Min. the sum of the residuals $\sum_{g=1}^{G} \|\mathbf{Y}_g - \underbrace{\sum_k \mathbf{T}_g^{k(H)} (a_g^{k(H)} \mathbf{C}_g^{(H)'})}_{\text{Predicted } \mathbf{Y}_g} \|^2$ where:

- $\mathbf{T}_g^{k(H)}$ are the matrices of block components with H (given) dimensions obtained from the G mbPLS, such as $\mathbf{T}_g^{k(H)} = [t_g^{k(1)} | \ldots | t_g^{k(H)}]$
- $a_g^{k(H)}$ are the normalized covariances between components from $\mathbf{X}_g^k$ and $\mathbf{Y}_g$
- $(\mathbf{C}_1^{(H)'}, \ldots, \mathbf{C}_G^{(H)'})$ are the regression coefficients of $\mathbf{Y}_g$ on $(\sum_k a_g^{k(H)} \mathbf{T}_g^{k(H)})$



From the original data … … to the partial components

Clusterwise multiblock PLS: stochastic algorithm

Algorithm

1. Start from a random initialization of N individuals into G clusters
2. For each individual n
   - Compute mbPLS where n belongs alternatively to each of the G clusters
   - For each of the G solutions, compute the residual sums $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$
   - Assign n to the cluster which minimize the former residual sums.
3. Repeat the former procedure for several random initializations of the clusters and select the best solution.
4. Compute the G mbPLS within each optimal cluster.

Comments

- Decreasing monotonicity of the criterion $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$ according to n
- Avoid local optimum
  Get the G clusters.
- Get the G regression coefficient matrices.

Clusterwise multiblock PLS: stochastic algorithm

Algorithm

1. Start from a random initialization of N individuals into G clusters
2. For each individual n
   - Compute mbPLS where n belongs alternatively to each of the G clusters
   - For each of the G solutions, compute the residual sums $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$
   - Assign n to the cluster which minimize the former residual sums.
3. Repeat the former procedure for several random initializations of the clusters and select the best solution.
4. Compute the G mbPLS within each optimal cluster.

Comments

- Decreasing monotonicity of the criterion $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$ according to n
- Avoid local optimum
  Get the G clusters.
- Get the G regression coefficient matrices.

Clusterwise multiblock PLS: stochastic algorithm

Algorithm

1. Start from a random initialization of N individuals into G clusters
2. For each individual n
   - Compute mbPLS where n belongs alternatively to each of the G clusters
   - For each of the G solutions, compute the residual sums $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$
   - Assign n to the cluster which minimize the former residual sums.
3. Repeat the former procedure for several random initializations of the clusters and select the best solution.
4. Compute the G mbPLS within each optimal cluster.

Comments

- Decreasing monotonicity of the criterion $\sum_{g=1}^{G} \|\mathbf{Y}_g - \hat{\mathbf{Y}}_g\|^2$ according to n
- Avoid local optimum
  Get the G clusters.
- Get the G regression coefficient matrices.

Clusterwise multiblock PLS: parameter selection (10-fold cross-validation)

Aims

- Parameter selection: minimize the $\text{RMSE}_{G,H}$ processed for several model dimensions H and several cluster numbers $G \rightarrow$ get H_{opt} and G_{opt},

Algorithm: 10-fold cross-validation

- Split the data into ten folds,
- Apply clusterwise multiblock PLS on the train data (9/10 folds),
- For each individual n, choose the cluster yielding the lowest error (1/10 fold),
- Compare the observed and the predicted dependent values $\rightarrow$ RMSE.

Clusterwise multiblock PLS: parameter selection (10-fold cross-validation)

Aims

- Parameter selection: minimize the $\text{RMSE}_{G,H}$ processed for several model dimensions H and several cluster numbers $G \rightarrow$ get H_{opt} and G_{opt},

Algorithm: 10-fold cross-validation

- Split the data into ten folds,
- Apply clusterwise multiblock PLS on the train data (9/10 folds),
- For each individual n, choose the cluster yielding the lowest error (1/10 fold),
- Compare the observed and the predicted dependent values $\rightarrow$ RMSE.

Clusterwise multiblock PLS: prediction of new individuals

Problem: Assign a new individual to its optimal cluster and predict its dependent variable values.

Bayesian affectation to clusters

- Project the new individual into each of the G component spaces from the cw.mbpls model ($\mathbf{X}_{new} \rightarrow \mathbf{T}_{g,new}^{k(H)}$),
- Compute the Bayesian Mahalanobis distances between the new individual and the G cluster gravity centres; transform them into probabilities,
- Maximize the G probabilities ($proba_g$) $\rightarrow$ cluster membership of the new individual

Model averaging prediction

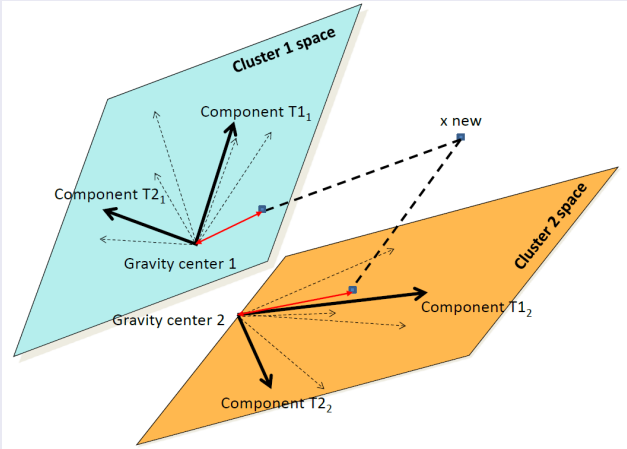
- $\hat{\mathbf{Y}}_{new}^{(H)} = \sum_{g=1}^{G} proba_g^{(H)} \times \underbrace{[\sum_k \mathbf{T}_{g,new}^{k(H)} (a_g^{k(H)} \mathbf{C}_g^{(H)'})]}_{\text{Predicted } \mathbf{Y}_g^{(H)}}$
- This prediction takes into account the predictions to the G clusters.

Clusterwise multiblock PLS: prediction of new individuals

Problem: Assign a new individual to its optimal cluster and predict its dependent variable values.

Bayesian affectation to clusters

- Project the new individual into each of the G component spaces from the cw.mbpls model ($\mathbf{X}_{new} \rightarrow \mathbf{T}_{g,new}^{k(H)}$),
- Compute the Bayesian Mahalanobis distances between the new individual and the G cluster gravity centres; transform them into probabilities,



- Maximize the G probabilities ($proba_g$) $\rightarrow$ cluster membership of the new individual

Model averaging prediction

Clusterwise multiblock PLS: prediction of new individuals

Problem: Assign a new individual to its optimal cluster and predict its dependent variable values.

Bayesian affectation to clusters

- Project the new individual into each of the G component spaces from the cw.mbpls model ($\mathbf{X}_{new} \rightarrow \mathbf{T}_{g,new}^{k(H)}$),
- Compute the Bayesian Mahalanobis distances between the new individual and the G cluster gravity centres; transform them into probabilities,
- Maximize the G probabilities ($proba_g$) $\rightarrow$ cluster membership of the new individual

Model averaging prediction

- $$\hat{\mathbf{Y}}_{new}^{(H)} = \sum_{g=1}^{G} proba_g^{(H)} \times \underbrace{[\sum_k \mathbf{T}_{g,new}^{k(H)} (a_g^{k(H)} \mathbf{C}_g^{(H)'})]}_{\text{Predicted } \mathbf{Y}_g^{(H)}}$$
- This prediction takes into account the predictions to the G clusters.

Clusterwise multiblock PLS: prediction of new individuals

Problem: Assign a new individual to its optimal cluster and predict its dependent variable values.

Bayesian affectation to clusters

- Project the new individual into each of the G component spaces from the cw.mbpls model ($\mathbf{X}_{new} \rightarrow \mathbf{T}_{g,new}^{k(H)}$),
- Compute the Bayesian Mahalanobis distances between the new individual and the G cluster gravity centres; transform them into probabilities,
- Maximize the G probabilities ($proba_g$) $\rightarrow$ cluster membership of the new individual

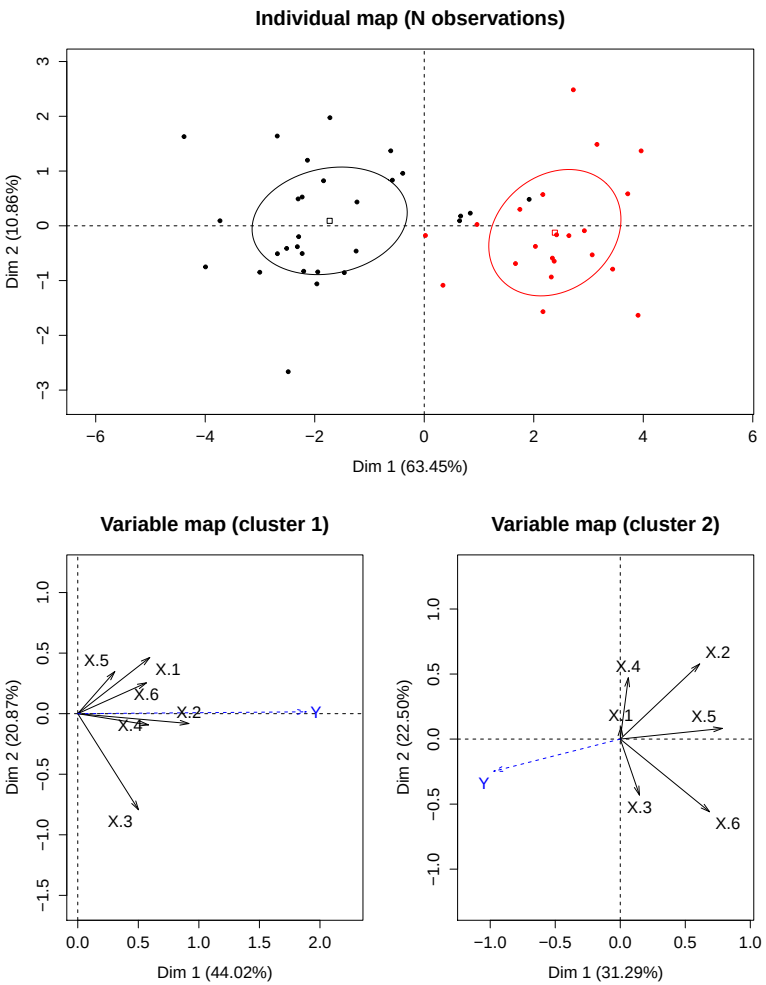
Model averaging prediction

- $\hat{\mathbf{Y}}_{new}^{(H)} = \sum_{g=1}^{G} proba_g^{(H)} \times \underbrace{[\sum_k \mathbf{T}_{g,new}^{k(H)} (a_g^{k(H)} \mathbf{C}_g^{(H)'})]}_{\text{Predicted } \mathbf{Y}_g^{(H)}}$
- This prediction takes into account the predictions to the G clusters.

Table of contents

Simulated data

Comparison of cw.PLS and Flexmix

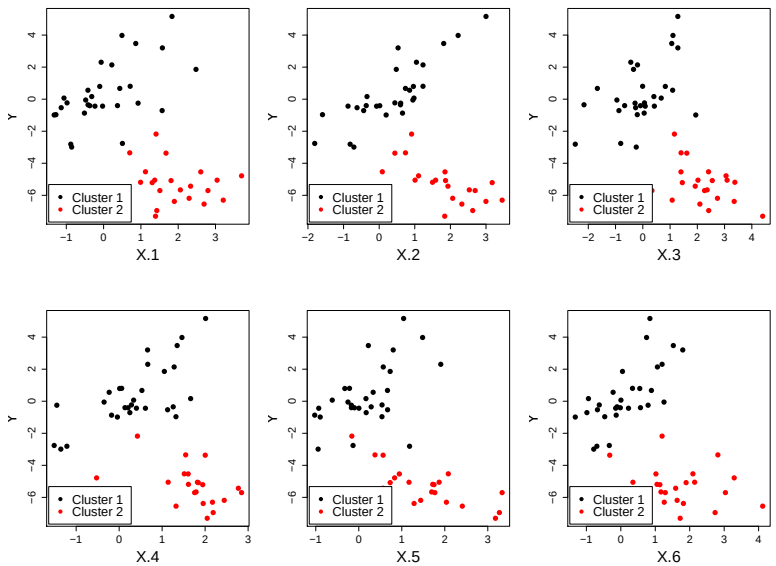


Simulated data features

- G=2 clusters with the same size, volume, orientation and shape and correctly separated,
- N=50 individuals,
- P=6 explanatory variables **X** moderately correlated (0.30) with no block structure,
- Q=1 dependent variable **Y** positively linked with **X** for cluster 1 ($\beta_1 = 0.5$) and negatively for cluster 2 ($\beta_2 = -0.5$).

Simulated data

Comparison of cw.PLS and Flexmix



Simulated data features

- G=2 clusters with the same size, volume, orientation and shape and correctly separated,
- N=50 individuals,
- P=6 explanatory variables **X** moderately correlated (0.30) with no block structure,
- Q=1 dependent variable **Y** positively linked with **X** for cluster 1 ($\beta_1 = 0.5$) and negatively for cluster 2 ($\beta_2 = -0.5$).

Simulated data

Comparison of cw.PLS and Flexmix

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	1.52 [1.17;1.80]	1.41 [1.033;1.71]

- $G = 2$ clusters

Index	Flexmix	cw.PLS (h=1)
Adj. Rand	**1** [1;1]	0.76 [0.47;1]
RMSE	**6.2e-16** [3.9e-16;7.8e-16]	0.18 [0.10;0.24]

Interpretation

- Performance improvement while taking account the G=2 clusters,
- Excellent performance of Flexmix and correct one for cw.PLS.

Simulated data

Comparison of cw.PLS and Flexmix

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	1.52 [1.17;1.80]	1.41 [1.033;1.71]

- $G = 2$ clusters

Index	Flexmix	cw.PLS (h=1)
Adj. Rand	**1** [1;1]	0.76 [0.47;1]
RMSE	**6.2e-16** [3.9e-16;7.8e-16]	0.18 [0.10;0.24]

Interpretation

- Performance improvement while taking account the G=2 clusters,
- Excellent performance of Flexmix and correct one for cw.PLS.

Simulated data

Comparison of cw.PLS and Flexmix

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	1.52 [1.17;1.80]	1.41 [1.033;1.71]

- $G = 2$ clusters

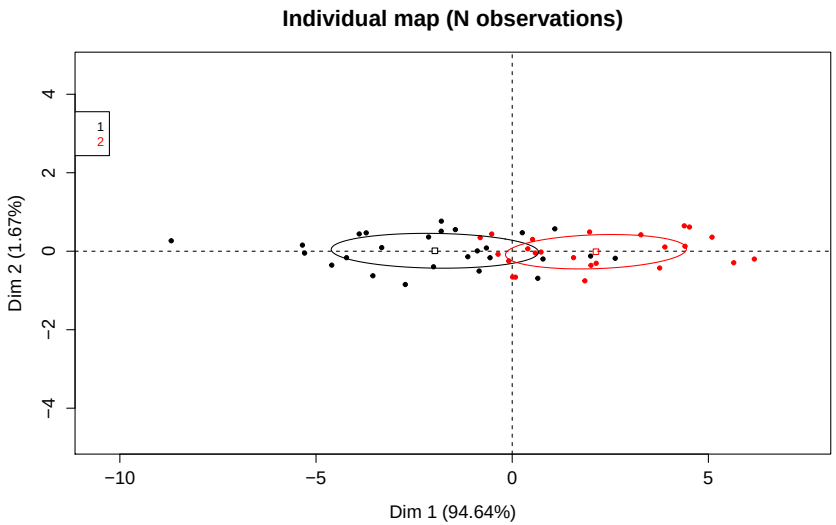
Index	Flexmix	cw.PLS (h=1)
Adj. Rand	**1** [1;1]	0.76 [0.47;1]
RMSE	**6.2e-16** [3.9e-16;7.8e-16]	0.18 [0.10;0.24]

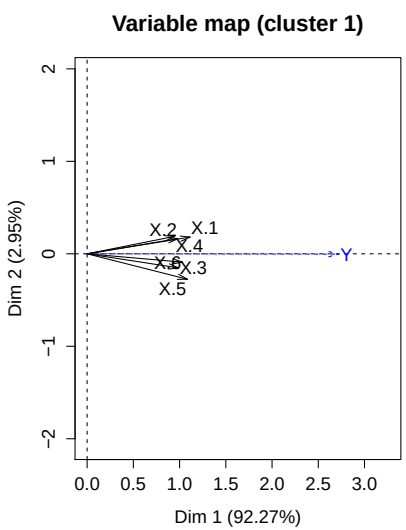
Interpretation

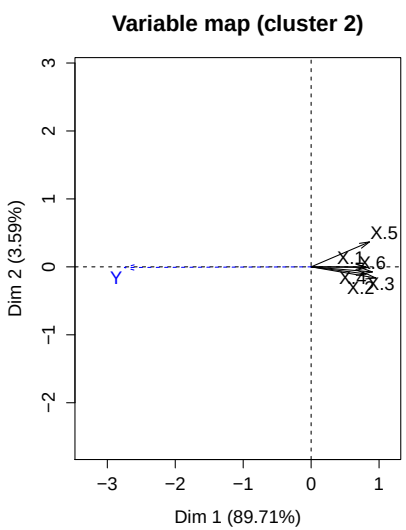
- Performance improvement while taking account the G=2 clusters,
- Excellent performance of Flexmix and correct one for cw.PLS.

Simulated data n°2

Comparison of cw.PLS and Flexmix with a reduced number of individuals and higher correlation





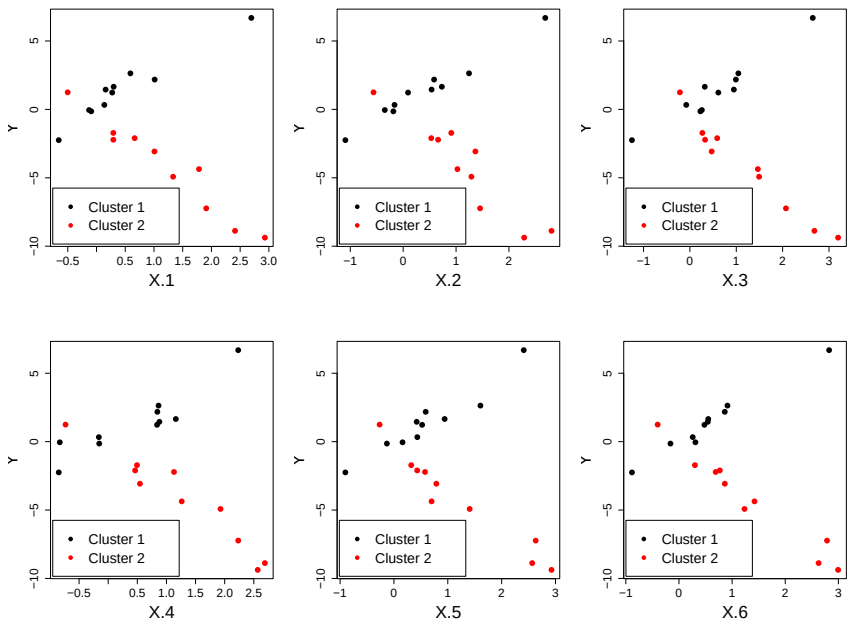


Simulated data features

- G=2 clusters with the same size, volume, orientation and shape and little separated,
- N=20 individuals,
- P=6 explanatory variables **X** correlated (0.90) with no block structure,
- Q=1 dependent variable **Y** positively linked with **X** for cluster 1 ($\beta_1 = 0.5$) and negatively for cluster 2 ($\beta_2 = -0.5$).

Simulated data n°2

Comparison of cw.PLS and Flexmix with a reduced number of individuals and higher correlation



Simulated data features

- G=2 clusters with the same size, volume, orientation and shape and little separated,
- N=20 individuals,
- P=6 explanatory variables **X** correlated (0.90) with no block structure,
- Q=1 dependent variable **Y** positively linked with **X** for cluster 1 ($\beta_1 = 0.5$) and negatively for cluster 2 ($\beta_2 = -0.5$).

Simulated data n°2

Comparison of cw.PLS and Flexmix with a reduced number of individuals

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	3.68 [1.92;4.84]	1.68 [0.89;2.20]

- $G = 2$ clusters

Index	Flexmix	cw.PLS (h=1)
Adj. Rand	0.56 [0.15;0.96]	0.83 [0.40;1]
RMSE	5.21 [0;79]	0.085 [0.042;0.11]

Interpretation

- Performance improvement while taking account the G=2 clusters,
- Good performance of cw.PLS especially for prediction error,
- High instability of Flexmix results.

Simulated data n°2

Comparison of cw.PLS and Flexmix with a reduced number of individuals

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	3.68 [1.92;4.84]	1.68 [0.89;2.20]

- $G = 2$ clusters

Index	Flexmix	cw.PLS (h=1)
Adj. Rand	0.56 [0.15;0.96]	0.83 [0.40;1]
RMSE	5.21 [0;79]	0.085 [0.042;0.11]

Interpretation

- Performance improvement while taking account the G=2 clusters,
- Good performance of cw.PLS especially for prediction error,
- High instability of Flexmix results.

Simulated data n°2

Comparison of cw.PLS and Flexmix with a reduced number of individuals

Results (ten-fold cross-validation, prediction purpose)

- No cluster

Index	Linear reg.	PLS reg. (h=1)
RMSE	3.68 [1.92;4.84]	1.68 [0.89;2.20]

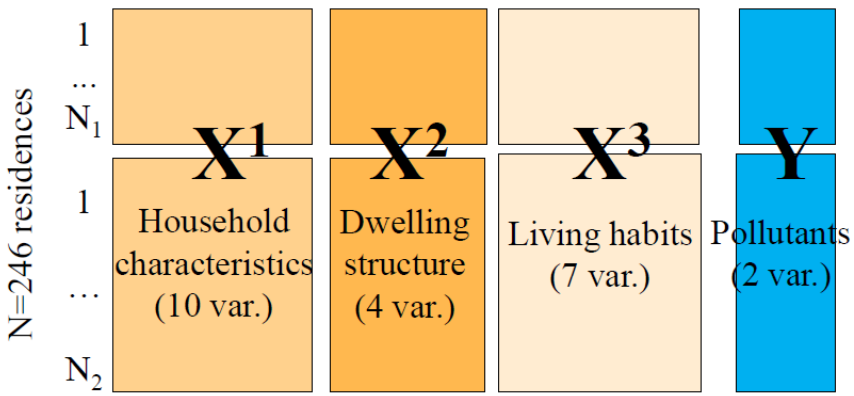
- $G = 2$ clusters

Index	Flexmix	cw.PLS (h=1)
Adj. Rand	0.56 [0.15;0.96]	0.83 [0.40;1]
RMSE	5.21 [0;79]	0.085 [0.042;0.11]

Interpretation

- Performance improvement while taking account the G=2 clusters,
- Good performance of cw.PLS especially for prediction error,
- High instability of Flexmix results.

Indoor air quality data: multiblock data & aims



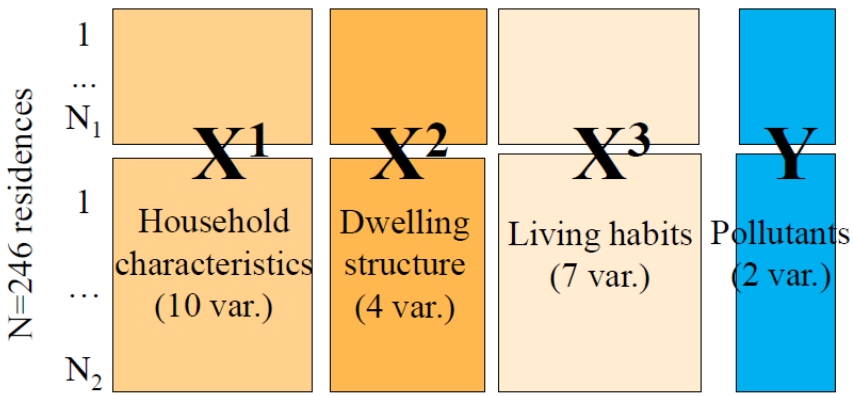
Data features [Kirchner et al., 2007]

- **Y**: 2 pollutants: formaldehyde, acrolein
- **X**: 21 potential risk factors of pollution organized into 3 blocks: household characteristics (10 var.), dwelling structure (4 var.), living habits (7 var.)
- **Individuals**: 246 main residences
- **Pre-processing**: Variables are centred and scaled; block are weighted to have the same weight.

Aims

- **Clustering**: Get the optimal clustering of residences to improve the pollutant prediction . . . ,
- **Regression**: . . . and compute the multiblock regression coefficients within each residence cluster.

Indoor air quality data: multiblock data & aims



Data features [Kirchner et al., 2007]

- **Y**: 2 pollutants: formaldehyde, acrolein
- **X**: 21 potential risk factors of pollution organized into 3 blocks: household characteristics (10 var.), dwelling structure (4 var.), living habits (7 var.)
- **Individuals**: 246 main residences
- **Pre-processing**: Variables are centred and scaled; block are weighted to have the same weight.

Aims

- **Clustering**: Get the optimal clustering of residences to improve the pollutant prediction . . . ,
- **Regression**: . . . and compute the multiblock regression coefficients within each residence cluster.

Indoor air quality data: comparison to the global model & parameter selection

	H=1	H=2	H=3	H=4
G=1	0,83	0,79	0,77	0,77
G=2	0,57	0,50	0,46	0,44
G=3	0,41	0,34	0,29	0,27
G=4	0,32	0,24	0,19	0,19
G=5	0,26	0,18	0,15	0,13
G=6	0,21	0,15	0,12	0,10

Table : $RMSE^2$ for calibration

	H=1	H=2	H=3	H=4
G=1	0,84	0,86	0,86	0,87
G=2	0,66	0,70	0,73	0,73
G=3	0,60	0,58	0,54	0,61
G=4	0,53	0,44	0,48	0,49
G=5	0,40	0,43	0,46	0,45
G=6	0,37	0,39	0,41	0,40

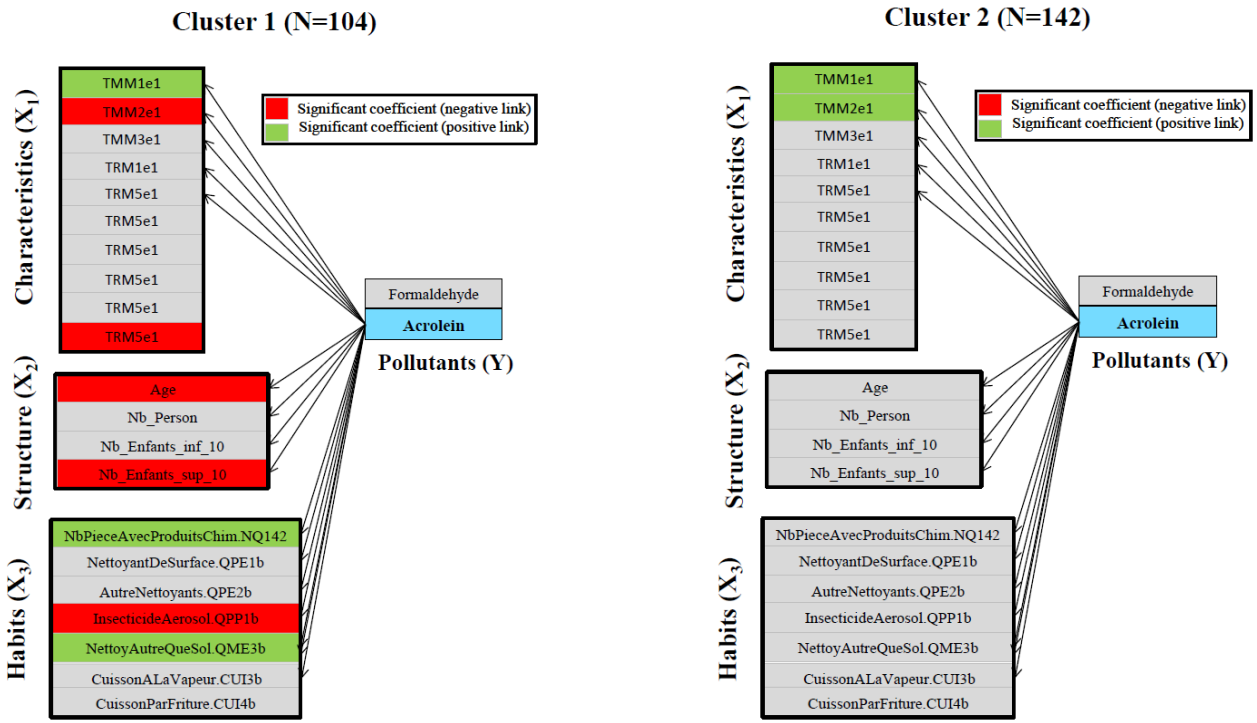
Table : $RMSE^2$ for prediction

Interpretation

- Performance improvement while taking account several clusters,
- Use of the $Q^2 = \sqrt{\frac{RMSE^2_{pred}(G)}{RMSE^2_{cal}(G-1)}} < 0.95$ (Tenenhaus, 1998) to select the optimal number of clusters G and dimensions H,
- Selection of G=2 clusters with H=2 dimensions ($Q^2 = 0.94$).

Indoor air quality data: cluster regression coefficients

Dependent variable: acrolein, significant regression coefficients by means of bootstrap simulations

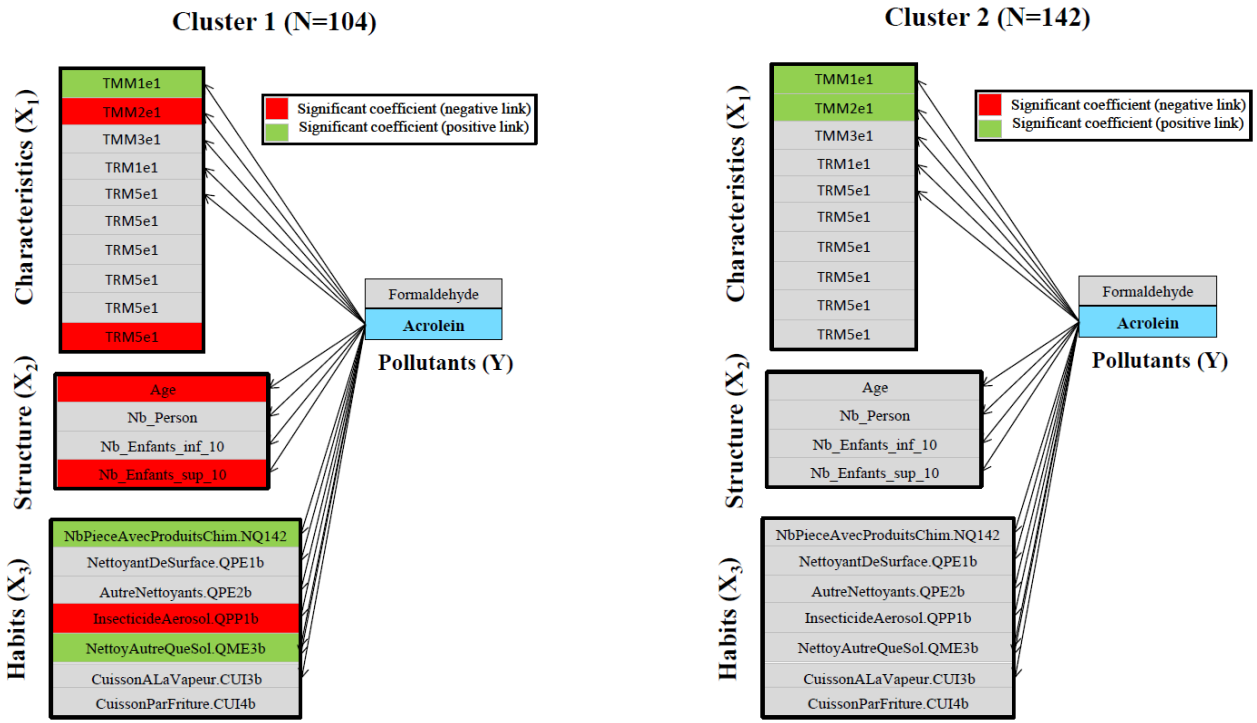


Interpretation

- **Cluster 1**: significantly linked to several variables (e.g., age of inhabitants, number and the age of children),
- **Cluster 2**: significantly and mainly linked to the rates of wood and PVC carpentry

Indoor air quality data: cluster regression coefficients

Dependent variable: acrolein, significant regression coefficients by means of bootstrap simulations



Interpretation

- **Cluster 1**: significantly linked to several variables (e.g., age of inhabitants, number and the age of children),
- **Cluster 2**: significantly and mainly linked to the rates of wood and PVC carpentry

Table of contents

Conclusion & perspectives

Conclusion

- Clusterwise multiblock PLS handles the specificity supervised multiblock data with an unknown structure on individuals
  - Aim: improve the prediction,
  - Meaningful criterion to minimize,
  - Parameters (number of dimensions and clusters) obtained through cross-validation,
  - Prediction of new individuals (clusters, predicted **Y** values).
- Useful tool to deal with real data especially in biology (*e.g.*, different risk factors of a disease according to sub-populations).

Perspectives

- Any other supervised multiblock method can be included in the algorithm,
- Next step 1: develop an index select the optimal numbers of clusters and dimensions by means of cross-validation,
- Next step 2: allow specific clusters (and dimensions) for each block,
- Programs will be transformed into a `R` package.

Conclusion & perspectives

Conclusion

- Clusterwise multiblock PLS handles the specificity supervised multiblock data with an unknown structure on individuals
  - Aim: improve the prediction,
  - Meaningful criterion to minimize,
  - Parameters (number of dimensions and clusters) obtained through cross-validation,
  - Prediction of new individuals (clusters, predicted **Y** values).
- Useful tool to deal with real data especially in biology (*e.g.*, different risk factors of a disease according to sub-populations).

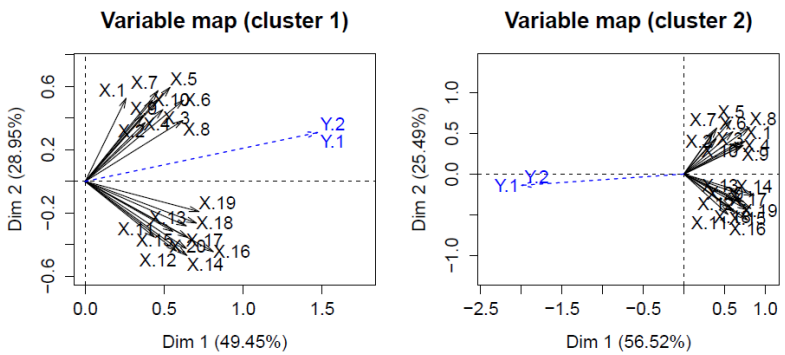
Perspectives

- Any other supervised multiblock method can be included in the algorithm,
- Next step 1: develop an index select the optimal numbers of clusters and dimensions by means of cross-validation,
- Next step 2: allow specific clusters (and dimensions) for each block,
- Programs will be transformed into a `R` package.

Clusterwise multiblock PLS regression

N'Dèye Niang[1], Stéphanie Bougeard[2] & Gilbert Saporta[1]

[1] *CEDRIC CNAM, Paris, France*
[2] *French Agency for Food, Environmental, Occupational Health & Safety (Anses), Ploufragan, France*



The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, 12-14 December 2015, London, UK