Multiblock Method for Categorical Variables

Application to the study of antibiotic resistance

S. Bougeard[1], E.M. Qannari[2] & C. Chauvin[1]

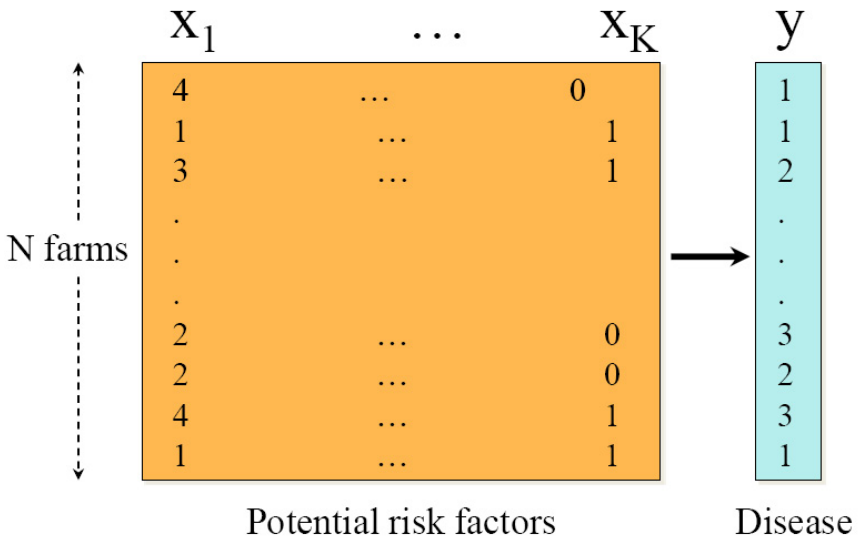
[1] *French agency for food, environmental and occupational health safety (Anses), Department of Epidemiology, Ploufragan, France*

[2] *Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering (Oniris), Department of Chemometrics and Sensometrics, Nantes, France*

19*th* International Conference on Computational Statistics, Paris, August 22 – 27, 2010

Table of contents

Statistical issues for epidemiological surveys



2. Expectations

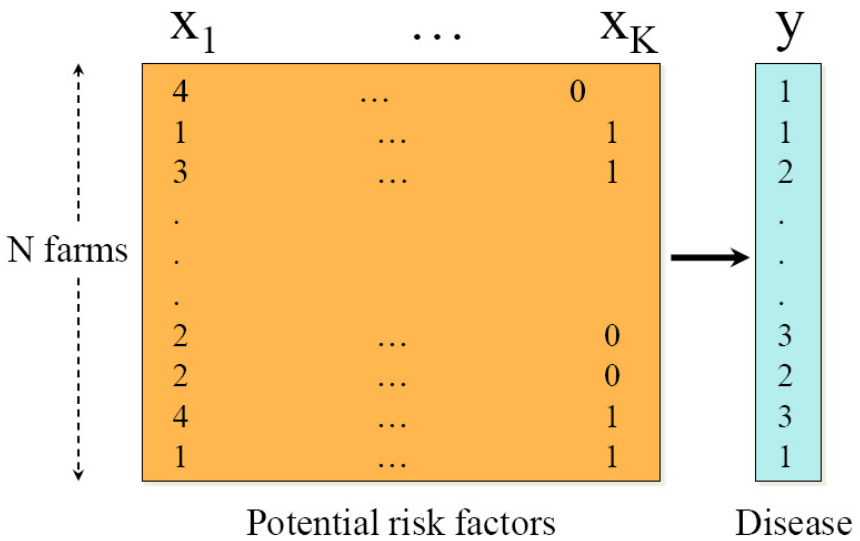
- Global optimization criterion with eigensolution,
- Assessement of the risk factors,
- Factorial representation of data.

→ **Multiblock modelling** extended to categorical data.

1. Advantages & limits of usual procedures

- Generalized linear models
  - Well-adapted for categorical variables,
  - Limited number of explanatory variables,
  - Constraints when *y* consists of more than 2 categories.
- Decision trees, Random Forest
  - Small misclassification errors,
  - Variables sorted in order of magnitude,
  - No regression coefficients.
- Boosting, bagging, SVM
  - Small misclassification errors,
  - No link with explanatory variables.

Statistical issues for epidemiological surveys



2. Expectations

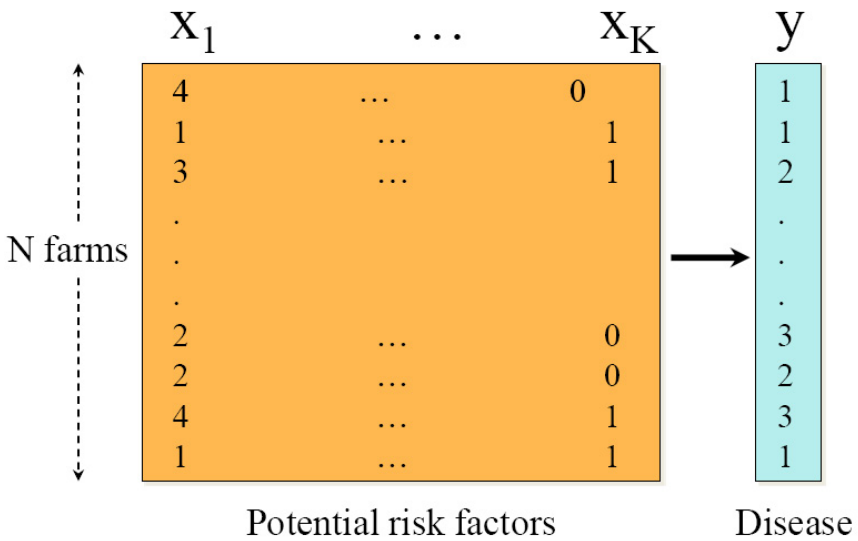
- Global optimization criterion with eigensolution,
- Assessement of the risk factors,
- Factorial representation of data.

→ **Multiblock modelling** extended to categorical data.

1. Advantages & limits of usual procedures

- Generalized linear models
  - Well-adapted for categorical variables,
  - Limited number of explanatory variables,
  - Constraints when *y* consists of more than 2 categories.
- Decision trees, Random Forest
  - Small misclassification errors,
  - Variables sorted in order of magnitude,
  - No regression coefficients.
- Boosting, bagging, SVM
  - Small misclassification errors,
  - No link with explanatory variables.

Statistical issues for epidemiological surveys



2. Expectations

- Global optimization criterion with eigensolution,
- Assessement of the risk factors,
- Factorial representation of data.

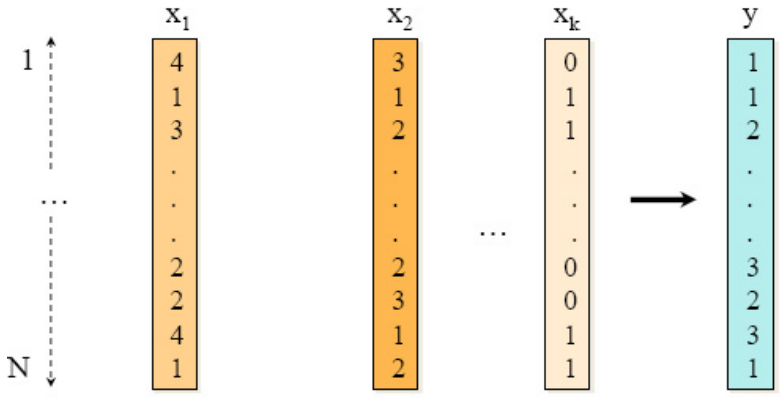
→ **Multiblock modelling** extended to categorical data.

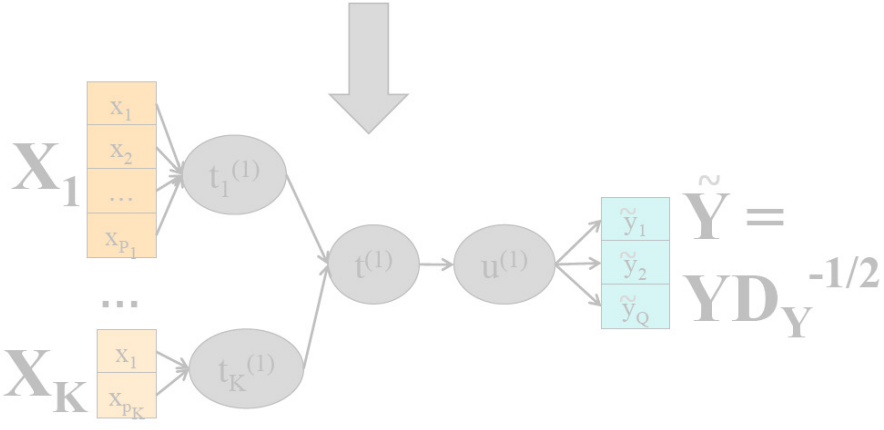
1. Advantages & limits of usual procedures

- Generalized linear models
  - Well-adapted for categorical variables,
  - Limited number of explanatory variables,
  - Constraints when *y* consists of more than 2 categories.
- Decision trees, Random Forest
  - Small misclassification errors,
  - Variables sorted in order of magnitude,
  - No regression coefficients.
- Boosting, bagging, SVM
  - Small misclassification errors,
  - No link with explanatory variables.

Table of contents

Categorical multiblock Redundancy Analysis





The latent variables represent the categorical variable coding : $t_k^{(1)} = X_k w_k^{(1)}$, $u^{(1)} = \tilde{Y} v^{(1)}$

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.

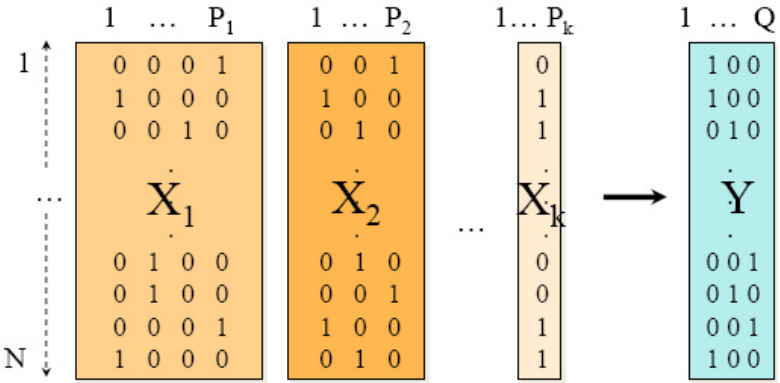
Criterion to maximize

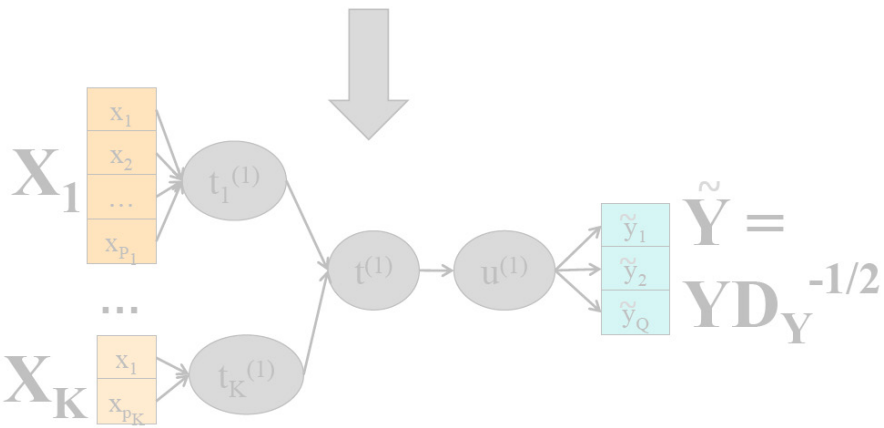
- $\sum_k cov^2(u^{(1)}, t_k^{(1)})$, with $||t_k^{(1)}|| = ||v^{(1)}|| = 1$
- $\sum_k ||P_{X_k} u^{(1)}||^2 = v^{(1)'} \tilde{Y}' \sum_k P_{X_k} \tilde{Y} v^{(1)}$ with $||v^{(1)}|| = 1$

First order solution

$v^{(1)}$ is the eigenvector of $\sum_k \tilde{Y}' P_{X_k} \tilde{Y}$ associated with the largest eigenvalue $\lambda^{(1)} = \sum_k ||P_{X_k} u^{(1)}||^2$

Categorical multiblock Redundancy Analysis





The latent variables represent the categorical variable coding : $t_k^{(1)} = X_k w_k^{(1)}$, $u^{(1)} = \tilde{Y} v^{(1)}$

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.

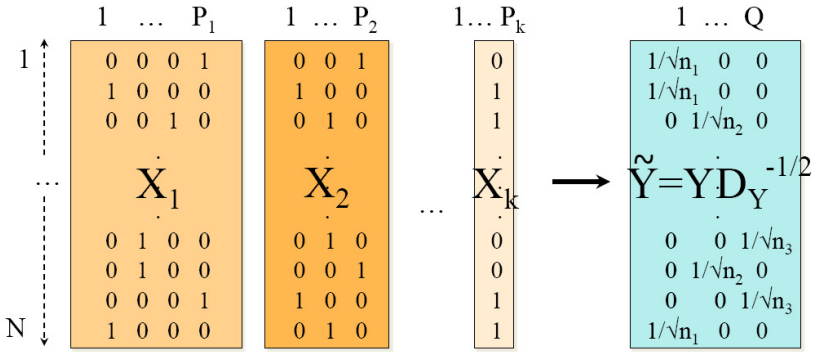
Criterion to maximize

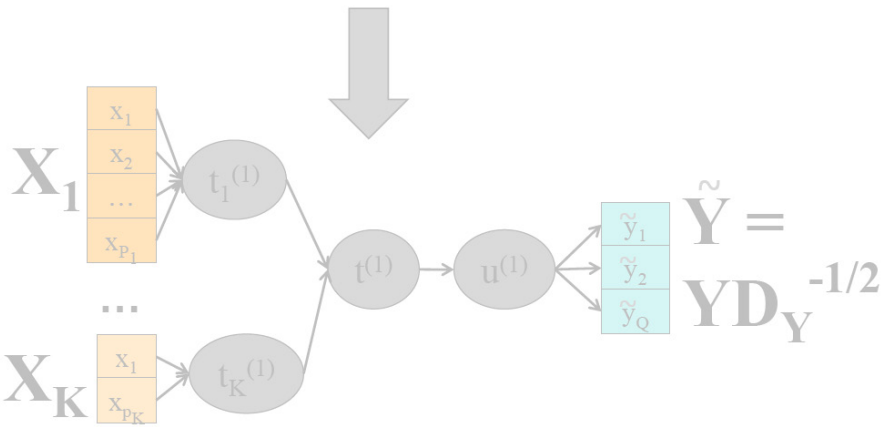
- $\sum_k cov^2(u^{(1)}, t_k^{(1)})$, with $||t_k^{(1)}|| = ||v^{(1)}|| = 1$
- $\sum_k ||P_{X_k} u^{(1)}||^2 = v^{(1)'} \tilde{Y}' \sum_k P_{X_k} \tilde{Y} v^{(1)}$ with $||v^{(1)}|| = 1$

First order solution

$v^{(1)}$ is the eigenvector of $\sum_k \tilde{Y}' P_{X_k} \tilde{Y}$ associated with the largest eigenvalue $\lambda^{(1)} = \sum_k ||P_{X_k} u^{(1)}||^2$

Categorical multiblock Redundancy Analysis





The latent variables represent the categorical variable coding : $t_k^{(1)} = X_k w_k^{(1)}$, $u^{(1)} = \tilde{Y} v^{(1)}$

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.

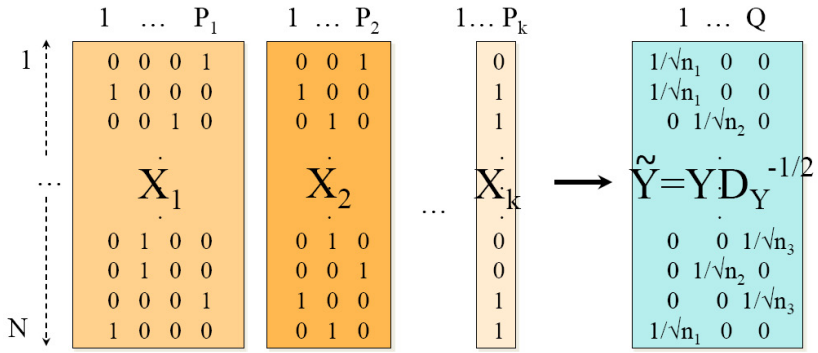
Criterion to maximize

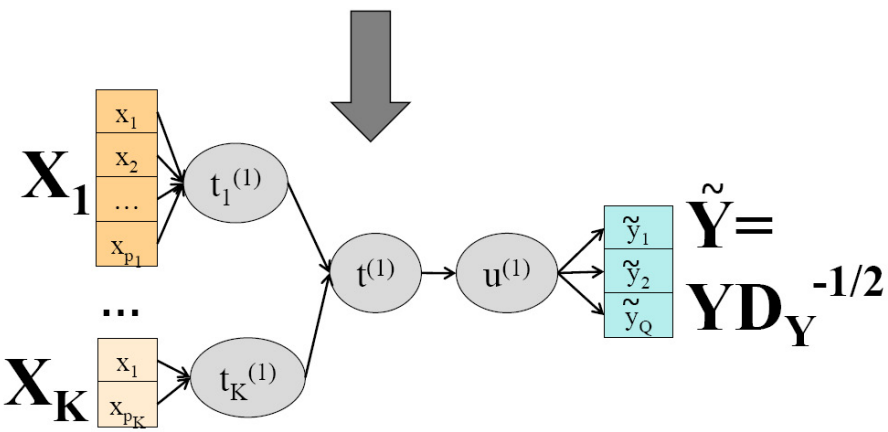
- $\sum_k cov^2(u^{(1)}, t_k^{(1)})$, with $||t_k^{(1)}|| = ||v^{(1)}|| = 1$
- $\sum_k ||P_{X_k} u^{(1)}||^2 = v^{(1)'} \tilde{Y}' \sum_k P_{X_k} \tilde{Y} v^{(1)}$ with $||v^{(1)}|| = 1$

First order solution

$v^{(1)}$ is the eigenvector of $\sum_k \tilde{Y}' P_{X_k} \tilde{Y}$ associated with the largest eigenvalue $\lambda^{(1)} = \sum_k ||P_{X_k} u^{(1)}||^2$

Categorical multiblock Redundancy Analysis





The latent variables represent the categorical variable coding : $t_k^{(1)} = X_k w_k^{(1)}$, $u^{(1)} = \tilde{Y} v^{(1)}$

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.

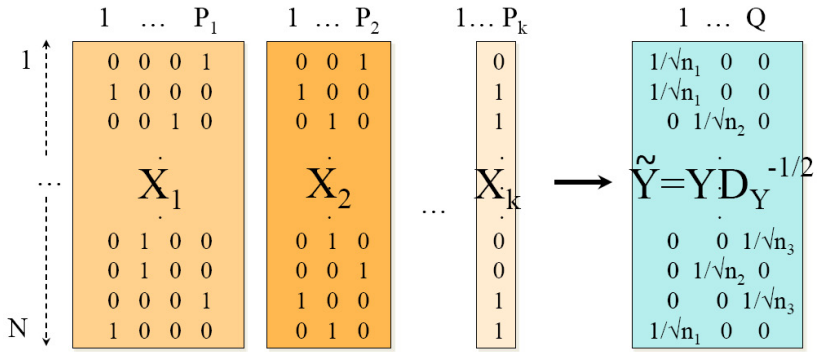
Criterion to maximize

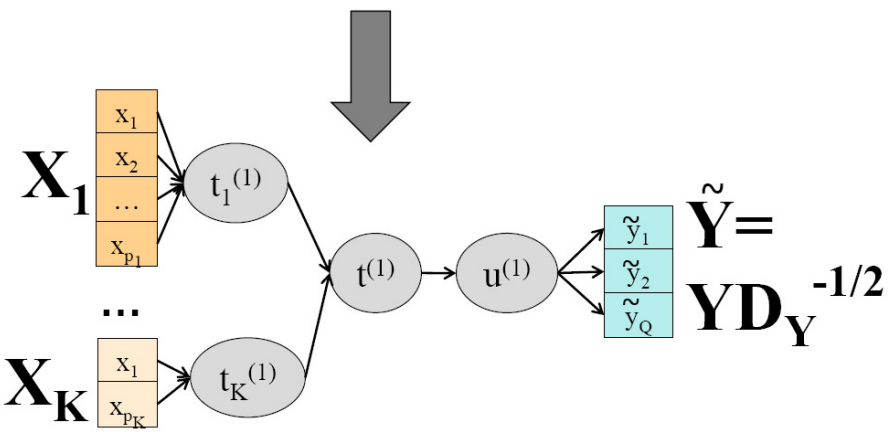
- $\sum_k cov^2(u^{(1)}, t_k^{(1)})$, with $||t_k^{(1)}|| = ||v^{(1)}|| = 1$
- $\sum_k ||P_{X_k} u^{(1)}||^2 = v^{(1)'} \tilde{Y}' \sum_k P_{X_k} \tilde{Y} v^{(1)}$ with $||v^{(1)}|| = 1$

First order solution

$v^{(1)}$ is the eigenvector of $\sum_k \tilde{Y}' P_{X_k} \tilde{Y}$ associated with the largest eigenvalue $\lambda^{(1)} = \sum_k ||P_{X_k} u^{(1)}||^2$

Categorical multiblock Redundancy Analysis





The latent variables represent the categorical variable coding : $t_k^{(1)} = X_k w_k^{(1)}$, $u^{(1)} = \tilde{Y} v^{(1)}$

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.

Criterion to maximize

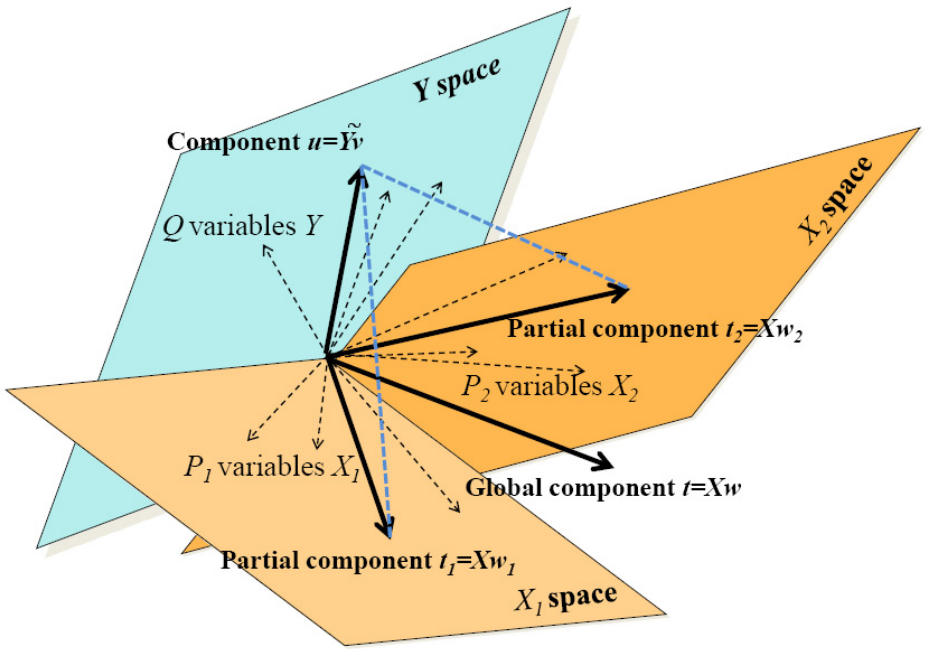
- $\sum_k cov^2(u^{(1)}, t_k^{(1)})$, with $||t_k^{(1)}|| = ||v^{(1)}|| = 1$
- $\sum_k ||P_{X_k} u^{(1)}||^2 = v^{(1)'} \tilde{Y}' \sum_k P_{X_k} \tilde{Y} v^{(1)}$ with $||v^{(1)}|| = 1$

First order solution

$v^{(1)}$ is the eigenvector of $\sum_k \tilde{Y}' P_{X_k} \tilde{Y}$ associated with the largest eigenvalue $\lambda^{(1)} = \sum_k ||P_{X_k} u^{(1)}||^2$

Categorical multiblock Redundancy Analysis (Cat-mbRA)

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.



Partial components $(t_1, \ldots, t_K)$

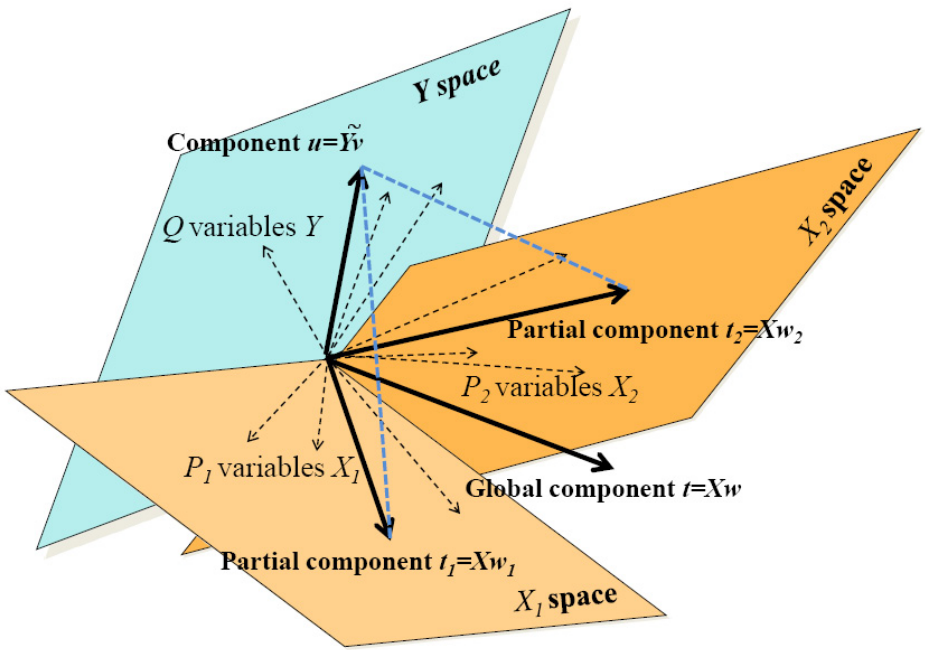
Projection of $u^{(1)}$ onto each subspace spanned by $X_k \rightarrow t_k^{(1)} = \frac{P_{X_k} u^{(1)}}{||P_{X_k} u^{(1)}||}$

Synthesis with a global component t

- $t^{(1)}$ sums up all the partial codings : $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$ with $\sum_k a_k^{(1)^2} = 1$,
- $t^{(1)} = \sum_k \frac{||P_{X_k} u^{(1)}||}{\sqrt{\sum_l ||P_{X_l} u^{(1)}||^2}} t_k^{(1)} = \frac{\sum_k P_{X_k} u^{(1)}}{\sqrt{\sum_l ||P_{X_l} u^{(1)}||^2}}$

Categorical multiblock Redundancy Analysis (Cat-mbRA)

P_{X_k} is the projector onto the subspace spanned by the dummy variables associated with x_k.



Partial components $(t_1, \ldots, t_K)$

Projection of $u^{(1)}$ onto each subspace spanned by $X_k \rightarrow t_k^{(1)} = \frac{P_{X_k} u^{(1)}}{||P_{X_k} u^{(1)}||}$

Synthesis with a global component t

- $t^{(1)}$ sums up all the partial codings : $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)}$ with $\sum_k a_k^{(1)^2} = 1$,
- $t^{(1)} = \sum_k \frac{||P_{X_k} u^{(1)}||}{\sqrt{\sum_l ||P_{X_l} u^{(1)}||^2}} t_k^{(1)} = \frac{\sum_k P_{X_k} u^{(1)}}{\sqrt{\sum_l ||P_{X_l} u^{(1)}||^2}}$

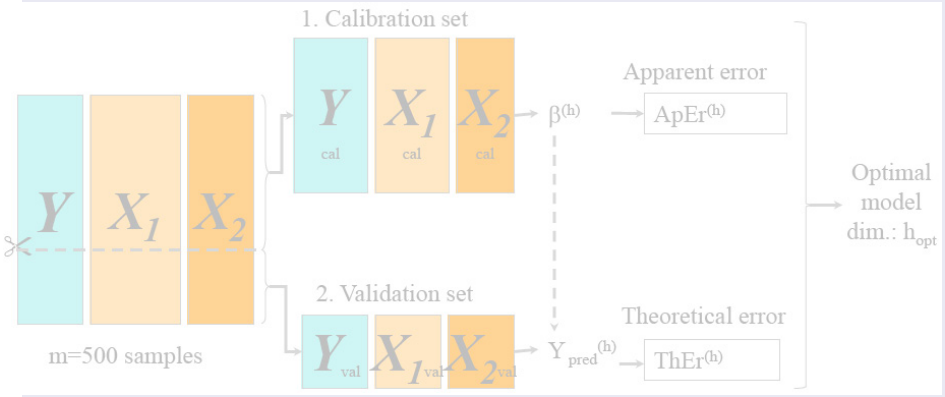
Higher order solutions and optimal Cat-mbRA model

Higher order solutions

Aim : Orthogonalised regressions which take into account all the explanatory variables, *i.e.* orthogonal components $(t^{(1)}, \ldots, t^{(H)})$.

$\longrightarrow$ Consider the residuals of the orthogonal projections of $(X_1, \ldots, X_K)$ onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)})$, ...

Selection of the optimal model



Additional information :

- Confusion matrix,
- ROC (=Receiver Operating Characteristic) curve.

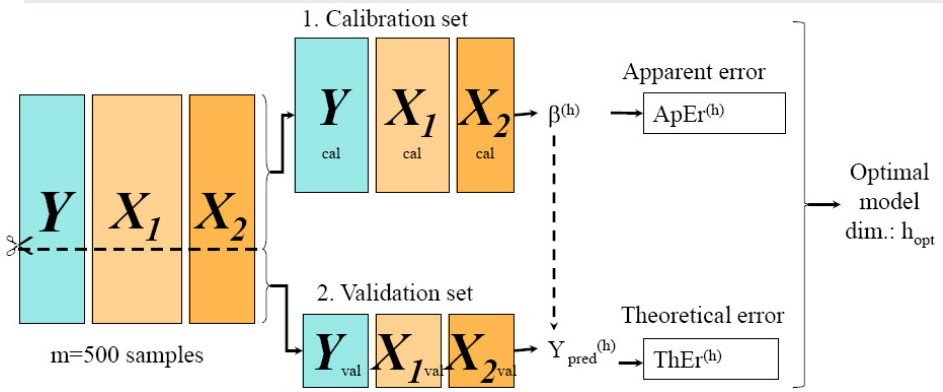
Higher order solutions and optimal Cat-mbRA model

Higher order solutions

Aim : Orthogonalised regressions which take into account all the explanatory variables, *i.e.* orthogonal components $(t^{(1)}, \ldots, t^{(H)})$.

$\longrightarrow$ Consider the residuals of the orthogonal projections of $(X_1, \ldots, X_K)$ onto the subspaces spanned by $t^{(1)}$, $(t^{(1)}, t^{(2)})$, ...

Selection of the optimal model



Additional information :

- Confusion matrix,
- ROC (=Receiver Operating Characteristic) curve.

Alternative methods for qualitative discrimination

Robust Generalized Linear Model framework

- Ridge logistic regression [Barker & Brown, 2001], principal component logistic regression [Aguilera *et al.*, 2006],
- PLS generalized regression (*e.g.* PLS logistic regression) [Marx, 1996 ; Bastien *et al.*, 2005].

Factorial analysis framework

- *Disqual* procedure [Saporta & Niang, 2006],
- Multiple non Symmetrical Correspondence Analysis [Lauro & Balbi, 1999].

Multiblock and Structural Equation Modelling framework

- Categorical extension of GCA-RT, *i.e.* MCA-RT [Kissita, 2003] and of multiblock PLS, *i.e.* MCOI-catPLS [D'Ambra et al., 2002],
- Categorical extension of SEM [Skrondal & Rabe-Hesketh, 2005] and of PLS-PM [Jakobowicz & Derquenne, 2007 ; Russolillo, 2009].

Alternative methods for qualitative discrimination

Robust Generalized Linear Model framework

- Ridge logistic regression [Barker & Brown, 2001], principal component logistic regression [Aguilera *et al.*, 2006],
- PLS generalized regression (*e.g.* PLS logistic regression) [Marx, 1996 ; Bastien *et al.*, 2005].

Factorial analysis framework

- *Disqual* procedure [Saporta & Niang, 2006],
- Multiple non Symmetrical Correspondence Analysis [Lauro & Balbi, 1999].

Multiblock and Structural Equation Modelling framework

- Categorical extension of GCA-RT, *i.e.* MCA-RT [Kissita, 2003] and of multiblock PLS, *i.e.* MCOI-catPLS [D'Ambra et al., 2002],
- Categorical extension of SEM [Skrondal & Rabe-Hesketh, 2005] and of PLS-PM [Jakobowicz & Derquenne, 2007 ; Russolillo, 2009].

Alternative methods for qualitative discrimination

Robust Generalized Linear Model framework

- Ridge logistic regression [Barker & Brown, 2001], principal component logistic regression [Aguilera *et al.*, 2006],
- PLS generalized regression (*e.g.* PLS logistic regression) [Marx, 1996 ; Bastien *et al.*, 2005].

Factorial analysis framework

- *Disqual* procedure [Saporta & Niang, 2006],
- Multiple non Symmetrical Correspondence Analysis [Lauro & Balbi, 1999].

Multiblock and Structural Equation Modelling framework

- Categorical extension of GCA-RT, *i.e.* MCA-RT [Kissita, 2003] and of multiblock PLS, *i.e.* MCOI-catPLS [D'Ambra et al., 2002],
- Categorical extension of SEM [Skrondal & Rabe-Hesketh, 2005] and of PLS-PM [Jakobowicz & Derquenne, 2007 ; Russolillo, 2009].

Table of contents

Epidemiological data

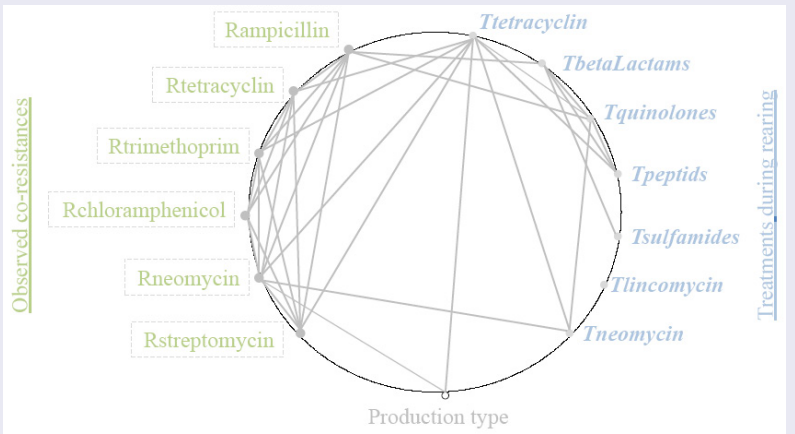
Epidemiological survey

- Part of the French antimicrobial resistance monitoring program (1999 – 2002),
- Study of the relationships between antibiotic consumption and resistance in healthy poultry.
- Screening of *E. coli* for antimicrobial resistances.

Data description

- Dependent variable : resistance to Nalidixic Acid,
- 14 explanatory variables : production type, previous antimicrobial treatments (7 var.), observed co-resistances (6 var.),
- $N = 554$ broiler chicken flocks.

Highly correlated explanatory variables



Epidemiological data

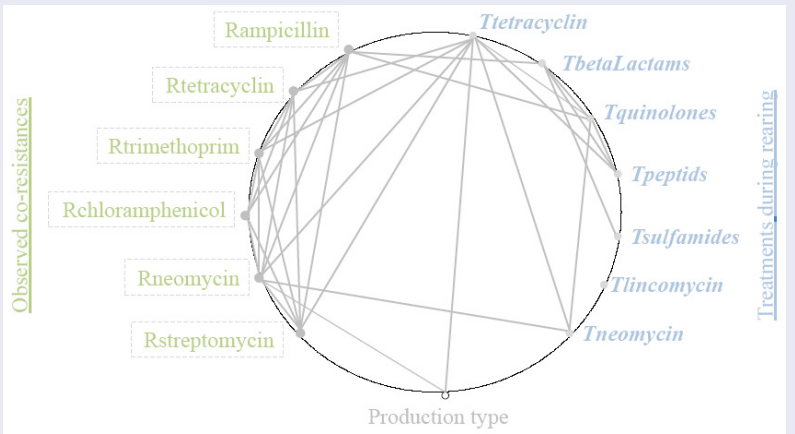
Epidemiological survey

- Part of the French antimicrobial resistance monitoring program ($1999 - 2002$),
- Study of the relationships between antibiotic consumption and resistance in healthy poultry.
- Screening of *E. coli* for antimicrobial resistances.

Data description

- Dependent variable : resistance to Nalidixic Acid,
- 14 explanatory variables : production type, previous antimicrobial treatments (7 var.), observed co-resistances (6 var.),
- $N = 554$ broiler chicken flocks.

Highly correlated explanatory variables



Epidemiological data

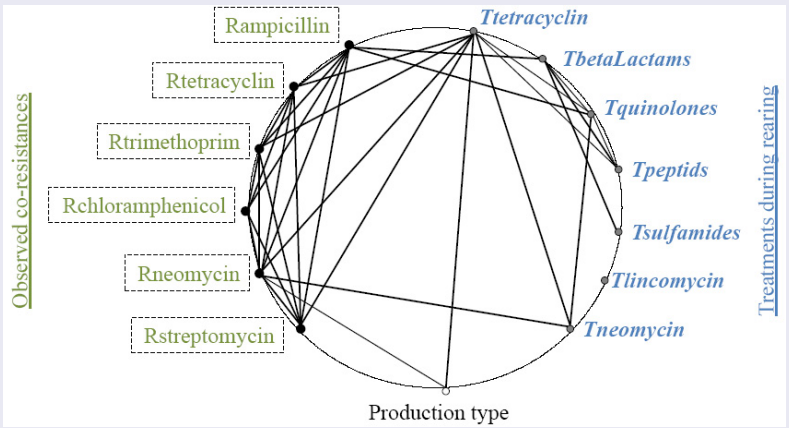
Epidemiological survey

- Part of the French antimicrobial resistance monitoring program (1999 – 2002),
- Study of the relationships between antibiotic consumption and resistance in healthy poultry.
- Screening of *E. coli* for antimicrobial resistances.

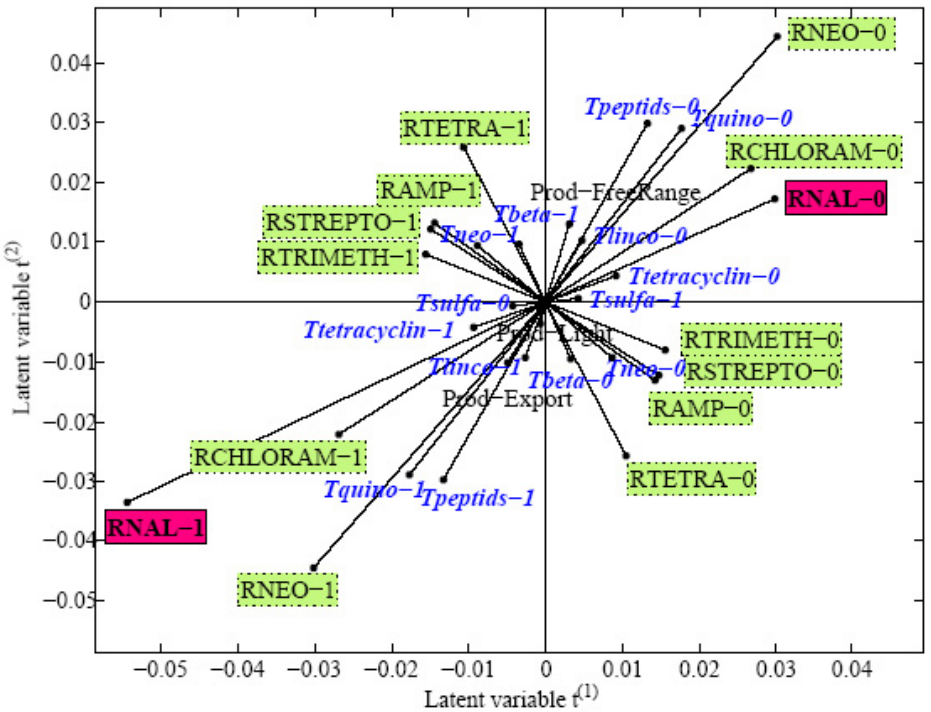
Data description

- Dependent variable : resistance to Nalidixic Acid,
- 14 explanatory variables : production type, previous antimicrobial treatments (7 var.), observed co-resistances (6 var.),
- $N = 554$ broiler chicken flocks.

Highly correlated explanatory variables



Plot of the variable loadings on the first two latent variables of cat-mbRA



- Dependent variable
- Observed co-resistances (explanatory variables),
- Previous antimicrobial treatments (explanatory variables),
- Production type (explanatory variables).

Interpretation

The resistance to Nalidixic Acid ($RNAL = 1$) is mainly associated with :

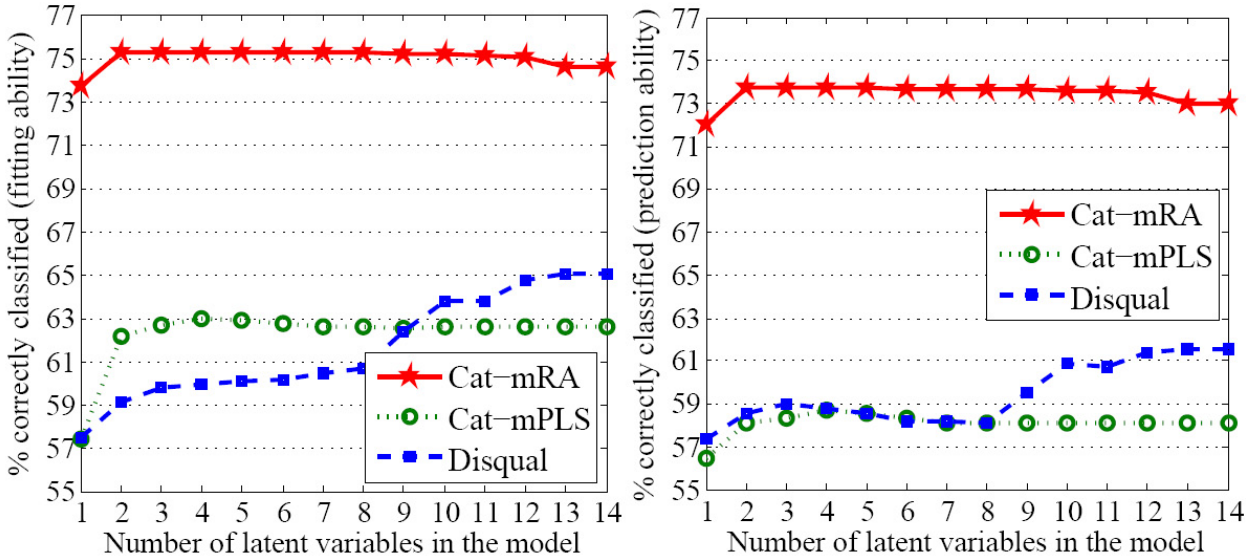
- Two other co-resistances (Chloramphenicol and Neomycin),
- Two antimicrobial treatments during rearing (Quinolones and Peptides).

Risk factors for Nalidixic Acid resistance

Results obtained from cat-mbRA with ($h_{opt} = 2$) latent variables, significant regression corfficients

Explanatory variables	**Number of cases**	**Nalidixic Acid resistance**
Treatments during rearing :		
Tetracyclin	153/554 (27.6%)	NS
Beta-lactams	75/554 (13.5%)	NS
Quinolones	93/554 (16.8%)	0.0058 [0.0015-0.0101]
Peptides	48/554 (8.7%)	NS
Sulfonamides	38/554 (6.9%)	NS
Lincomycin	33/554 (6.0%)	NS
Neomycin	26/554 (4.7%)	NS
Observed co-resistances :		
Ampicillin	278/554 (50.2%)	NS
Tetracyclin	462/554 (83.4%)	NS
Trimethoprim	284/554 (51.3%)	NS
Chloramphenicol	86/554 (15.5%)	0.0066 [0.0012-0.0119]
Neomycin	62/554 (11.2%)	0.0094 [0.0037-0.0151]
Streptomycin	297/554 (53.6%)	NS
Production :		
Export	192/554 (34.6%)	NS
Free-range	63/554 (11.4%)	NS
Light	299/554 (54.0%)	NS

Comparison with alternative methods



Additional information

- Cat-mbRA : good performance due to $Se = 96.5\%$, whereas $Sp = 17.7\%$ (fitting ab.),
- Logistic regression : surprising good performance, with $Se = 95.7\%$ and $Sp = 21.4\%$ (fitting ab.),
- Cat-mbPLS (resp. *Disqual*) : average performance with $Se = 61.2\%$ (resp. 56.4%) and $Sp = 65.2\%$ (resp. 66.2%) (fitting ab.),
- No real differences between the methods on the ROC curves.

Table of contents

Concluding remarks

Conclusion

- Proposition of a new and successful method for qualitative discrimination (categorical multiblock Redundancy Analysis, cat-mbRA),
- Extension in the field of multiblock modelling framework,
- Application to a real epidemiological survey,
- Code programs and interpretation tools developed in Matlab®.

Perspectives

- Comparison with other methods (*e.g.* PLS logistic regression, M-NSCA, MCA-RT, ...) [working paper],
- Simulation study to better compare the method performances,
- Extension to the prediction of several categorical variables.

Concluding remarks

Conclusion

- Proposition of a new and successful method for qualitative discrimination (categorical multiblock Redundancy Analysis, cat-mbRA),
- Extension in the field of multiblock modelling framework,
- Application to a real epidemiological survey,
- Code programs and interpretation tools developed in Matlab®.

Perspectives

- Comparison with other methods (*e.g.* PLS logistic regression, M-NSCA, MCA-RT, ...) [working paper],
- Simulation study to better compare the method performances,
- Extension to the prediction of several categorical variables.

Multiblock Method for Categorical Variables

Application to the study of antibiotic resistance

S. Bougeard[1], E.M. Qannari[2] & C. Chauvin[1]

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

[2] *Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering (Oniris), Department of Chemometrics and Sensometrics, Nantes, France*

19*th* International Conference on Computational Statistics, Paris, August 22 – 27, 2010

anses