Generalizing Partial Least Squares and Correspondence Analysis to Predict Categorical (and Heterogeneous) Data

Hervé Abdi, Derek Beaton, & Gilbert Saporta

Outline

1. Introduction
2. Pioneering works
3. PLS-CA
4. PLSR-CA
5. Application
6. Conclusions and perspectives

1. Introduction

- Two data matrices, usually of indicator variables $\mathbf{X}=[\mathbf{X}_1|..|\mathbf{X}_J]$ and $\mathbf{Y}=[\mathbf{Y}_1|...|\mathbf{Y}_K]$ describing the same set of n observations by two different sets of *J* and *K* nominal variables. In PLSR-CA, the first matrix is used to “predict” the other matrix.

- Motivation (Beaton et al., 2015):
  - ADNI (Alzheimer's Disease NeuroImaging Initiative) project data base
  - Question: How single nucleotide polymorphisms (SNPs) can be used to predict a set of variables measuring cognitive impairment and depression in Alzheimer's Disease?

- Just like standard PLSR, PLSR-CA first computes a pair of latent variables—which are linear combinations of the original variables—that have maximal covariance.
- The coefficients of these latent variables are obtained from the (generalized) singular value decomposition (equivalent to correspondence analysis of the matrix **X’Y**)

- The latent variables are obtained by projecting the original data matrices **X** and **Y** as supplementary rows and columns in the analysis of **X'Y** data table.
- The latent variable from the first matrix **X** is then used (after an adequate normalization) to predict the second matrix **Y.**

2. Pioneering works

- Focused on the case of a **single** categorical response Y and p categorical predictors $X_1,\ldots, X_p$
- CA of $\mathbf{Y'X}=(\mathbf{Y'X_1}\,|\ldots\,|\,\mathbf{Y'X_p})$, concatenation of the p contingency tables $\mathbf{Y'X_j}$
  - a "bande de Burt" (Lebart et al. , 2006) or "Burt's strip"
  - "Stacked tables" (Greenacre, 2007) . Bourdieu (1979) did a famous application of CA on stacked tables (Greenacre and Blasius, 2007)
- For general properties of this CA see Leclerc, 1975, Leclerc, 1976 and Cazes, 1980

- CA of **Y'X** as a method for predicting the categories of Y
  - Masson, 1974
  - *average mean squares* (Saporta, 1975)
  - *barycentric discrimination* (Nakache, 1977)
  - *BADIA* (Abdi & Williams, 2010).
- New observations are projected as supplementary barycenters of their categories and classified to the nearest barycenter of Y categories

- Barycentric discrimination ignores relations between predictors
- Experiments show surprisingly good performances in a similar way as naive Bayes prediction but with an additive rule instead of a multiplicative rule.
- A contradiction with Beh & Lombardo, 2014?
  - Page 397: "when stacking or concatenating, we do neglect some of the interactions that are of order 2 and 3 among the variables. Even if it is a very simple way of analyzing multiple tables, it is generally not recommended to perform correspondence analysis in this way."

3.PLS-CA

- Integrate PLSC & Correspondence Analysis
  – Beaton, Filbey, & Abdi (2013)
  – Beaton, Dunlop, ADNI, Abdi (2015, in press)

- What is PLSC (Partial Least Squares Correlation) for 2 tables of numerical data **X** and **Y**?
  - Inter-battery Factor Analysis (Tucker, 1958)
  - Canonical-PLS (Tenenhaus, 1998)
  - Co-Inertia Analysis (Dray, 2014)
  - And many more, actually (Krishnan *et al.,* 2010)

$\mathbf{X}$ and $\mathbf{Y}$ centered and normalized

$$\mathbf{R} = \mathbf{X}'\mathbf{Y}$$

first pair of "latent" variables

$$\ell_{\mathbf{x}} = \mathbf{Xu} \quad \ell_{\mathbf{y}} = \mathbf{Yv}$$

max $\mathrm{cov}(\ell_{\mathbf{x}}, \ell_{\mathbf{y}})$ or max $\mathbf{u}'\mathbf{R}\mathbf{v}$ with $\mathbf{u}'\mathbf{u} = \mathbf{v}'\mathbf{v}=1$

- Successive pairs under orthogonality constraints

$$\mathbf{L_x} = \mathbf{XU} \quad \mathbf{L_y} = \mathbf{YV}$$

$$\max diag(\mathbf{L}_{\mathbf{x}}^{'}\mathbf{L}_{\mathbf{y}})$$

$$\mathbf{R} = \mathbf{U\Delta V'} \quad \text{SVD}$$

- **U** and **V** are called "saliences"
- $\mathbf{U\Delta}$ and $\mathbf{V\Delta}$ component scores

- PLSC for categorical data: PLS-CA

Y

INCORRECT	CORRECT	INCORRECT	CORRECT	ID
1	0	0	1	AK
1	0	1	0	JD
1	0	1	0	MR
1	0	1	0	GB
0	1	1	0	NH
1	0	1	0	AL
0	1	0	1	DO
0	1	0	1	JM
1	0	1	0	ED
0	1	0	1	CB
1	0	0	1	CF
0	1	1	0	OD
0	1	0	1	DM
1	0	1	0	NS
0	1	0	1	JR

X'

GG	1	0	0	1	0	1	0	0	0	0	1	0	0	1	0
GT	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
TT	0	0	0	0	1	0	1	1	1	1	0	1	1	0	0
CC	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
AC	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0
AA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	AK	JD	MR	GB	NH	AL	DO	JM	ED	CB	CF	OD	DM	NS	JR

X'Y=

5	0	3	2	GG
2	1	2	1	GT
1	6	3	4	TT
4	3	5	2	CC
3	3	3	3	AC
1	1	0	2	AA
INCORRECT	CORRECT	INCORRECT	CORRECT	

- $\mathbf{X'Y}$: a matrix of counts
- $\mathbf{X}$ and $\mathbf{Y}$ disjunctive (binary) tables not centered, nor standardised
- 1st step: center and normalize

$$\mathbf{Z}_{\mathbf{X}} = \frac{\mathbf{X}\left(\mathbf{I} - \frac{\mathbf{11'}}{n}\right)}{J\sqrt{n}} \quad \mathbf{Z}_{\mathbf{Y}} = \frac{\mathbf{Y}\left(\mathbf{I} - \frac{\mathbf{11'}}{n}\right)}{K\sqrt{n}}$$

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{Z}_{\mathbf{X}}\mathbf{Z}_{\mathbf{Y}}$$

- Second step: compute the chi-square metric
  - Column mass vectors $\mathbf{m_X}$ and $\mathbf{m_Y}$

$$\mathbf{m_X} = \left(\mathbf{1'X1}\right)^{-1}\mathbf{1'X} \quad \mathbf{m_Y} = \left(\mathbf{1'Y1}\right)^{-1}\mathbf{1'Y}$$

$$\mathbf{D_x} = diag\left(\mathbf{m_X}\right)^{-1} \quad \mathbf{D_y} = diag\left(\mathbf{m_Y}\right)^{-1}$$

- Third step : maximisation

$\max_{\mathbf{U,V}} diag\left(\mathbf{L'_X L_Y}\right)$ where $\mathbf{L_X} = \mathbf{Z_X D_X V} \quad \mathbf{L_Y} = \mathbf{Z_Y D_Y V}$

with $\mathbf{U'D_X U} = \mathbf{V'D_Y V} = 1$

- Generalized (weighted) SVD

$$\mathbf{Z_R} = \mathbf{U\Delta V'}$$

- Equivalent to Correspondence Analysis of $\mathbf{Z_R}$

- Factor scores: $\mathbf{F_X}=\mathbf{D_X U\Delta}$ $\quad \mathbf{F_Y}=\mathbf{D_Y V\Delta}$

- The original $\mathbf{X}$ and $\mathbf{Y}$ matrices can be projected as supplementary elements on their respective factor scores.
- These supplementary factors scores—denoted respectively $\mathbf{G_X}$ and $\mathbf{G_Y}$—are computed as

$$\mathbf{G_X} = \frac{1}{J}\mathbf{X}\mathbf{F_X}\boldsymbol{\Delta}^{-1} \quad \mathbf{G_Y} = \frac{1}{K}\mathbf{Y}\mathbf{F_Y}\boldsymbol{\Delta}^{-1}$$

- Each row represents frequencies (this is called a row profile in correspondence analysis) and so each row now sums to one. This last equation shows that an observation is positioned as the barycenter of the coordinates of its variables. […] Both PLS and CA contribute to the interpretation of PLSCA. PLS shows that the latent variables have maximum covariance, CA shows that factors scores have maximal variance and that this variance "explains" a proportion of the Inertia associated to the contingency table. (Beaton et al., 2013)

4. PLSR-CA

- PLS-CA is symmetric and thus not fitted for "prediction."
- PLSR-CA starts from the same 1st order solution (pair of latent variables), but uses different standardizations and deflations

- Step 1: obtain $\ell_{\mathbf{x}} = \mathbf{Xu}$ $\ell_{\mathbf{y}} = \mathbf{Yv}$ as before

- Step 2 Predicting $\mathbf{X}$ and $\mathbf{Y}$ from $\mathbf{X}$
  - Define $\mathbf{t} = \dfrac{\ell_{\mathbf{x}}}{\|\ell_{\mathbf{x}}\|}$
  - Predict linearly $\mathbf{X}$ from $\mathbf{t}$

$$\hat{\mathbf{X}} = \mathbf{tp'} \text{ with } \mathbf{p} = \mathbf{t'X}$$

  - Predict linearly $\mathbf{Y}$ from $\mathbf{t}$

$$\hat{\mathbf{Y}} = \mathbf{t}\beta\mathbf{v'} \text{ with } \beta = \mathbf{t'YD_y v}$$

- Step 3 : deflating $\mathbf{X}$ and $\mathbf{Y}$
  - Residual matrices

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad \mathbf{F} = \mathbf{Y} - \hat{\mathbf{Y}}$$

  - Iterate the process on $\mathbf{E}$ and $\mathbf{F}$ instead of $\mathbf{X}$ and $\mathbf{Y}$ until a specific number of latent variables has been extracted or when the first matrix (i.e., **X**) is completely decomposed
  - If *L* denotes the number of components kept, then the components for the $\mathbf{X}$ set are stored in the $n \times L$ matrix $\mathbf{T}$ and the $J \times L$ matrix $\mathbf{W}$. Similarly, the components for the $\mathbf{Y}$ set are stored in the $n \times L$ matrix $\mathbf{T}$, the $L \times L$ diagonal matrix $\mathbf{B}$ of the β's, and the $K \times L$ matrix $\mathbf{C}$.

- Predicting $\mathbf{Y}$ (and $\mathbf{X}$)
  - The predicted matrix of the dependent variables is decomposed as:

$$\hat{\mathbf{Y}} = \mathbf{TBC'}$$

- This equation can be used to predict $\mathbf{Y}$ from its decomposition (and therefore from $\mathbf{X}$) but is not usable to predict $\mathbf{Y}$ for a new observation where only the x’s are known

- Solution: use a PLS regression

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}_{PLS} \quad \text{where } \mathbf{B}_{PLS} = \left(\mathbf{W}'\right)^{+}\mathbf{B}\mathbf{C}'$$

- However "prediction" should be understood as an estimation of the deviation from the distribution under independence, not as a classification procedure. $\mathbf{Z}_{\mathbf{R}}$ is doubly centered with respect to the product of marginal frequencies.

5. Application

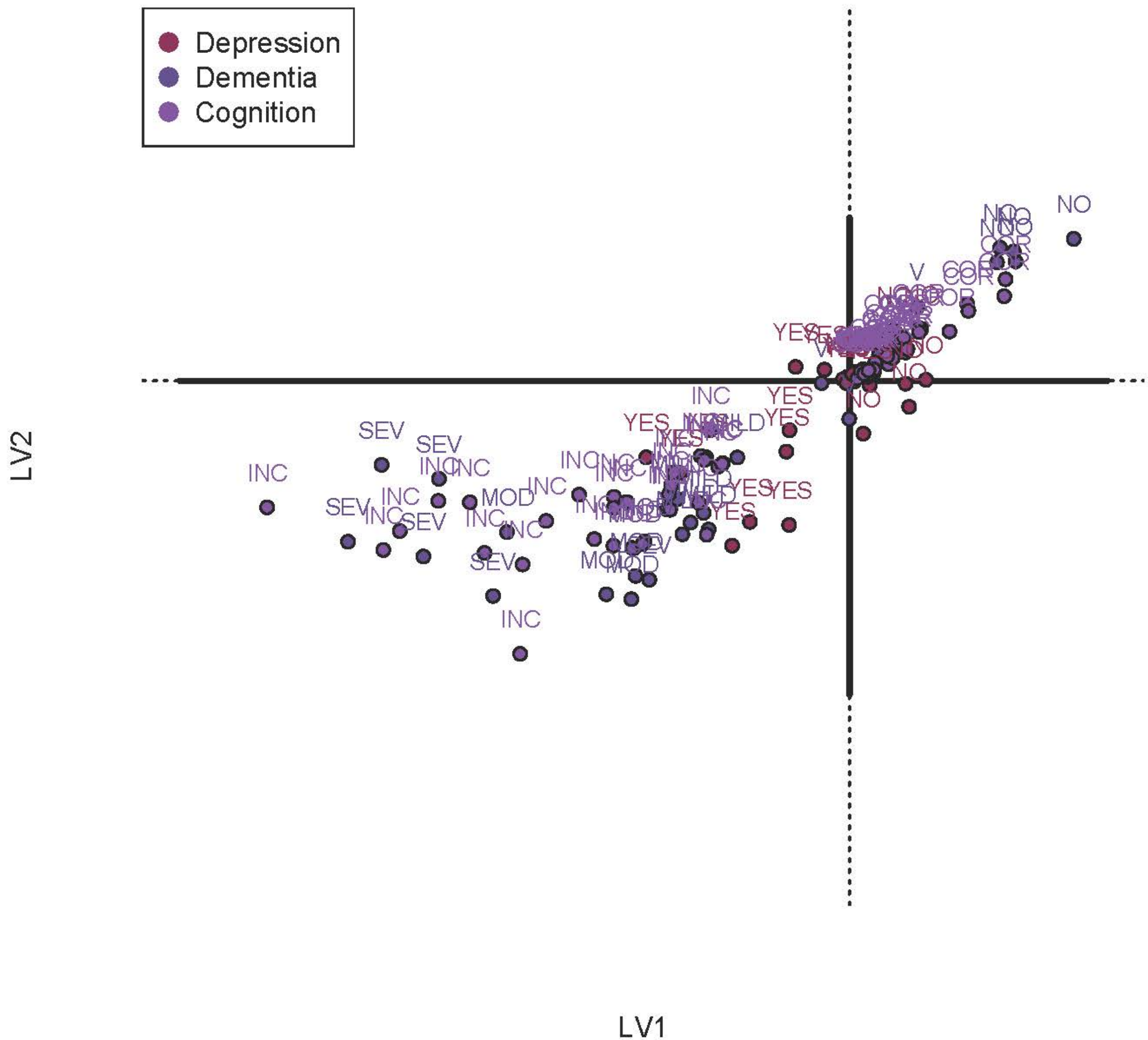
- In the ADNI project data base, try to predict *K* behavioural variables $\mathbf{Y} = [\mathbf{Y}_1 | \ldots | \mathbf{Y}_K]$ from *J* SNPs (Single Nucleotide Polymorphisms)

  $\mathbf{X} = [\mathbf{X}_1 | .. | \mathbf{X}_J]$

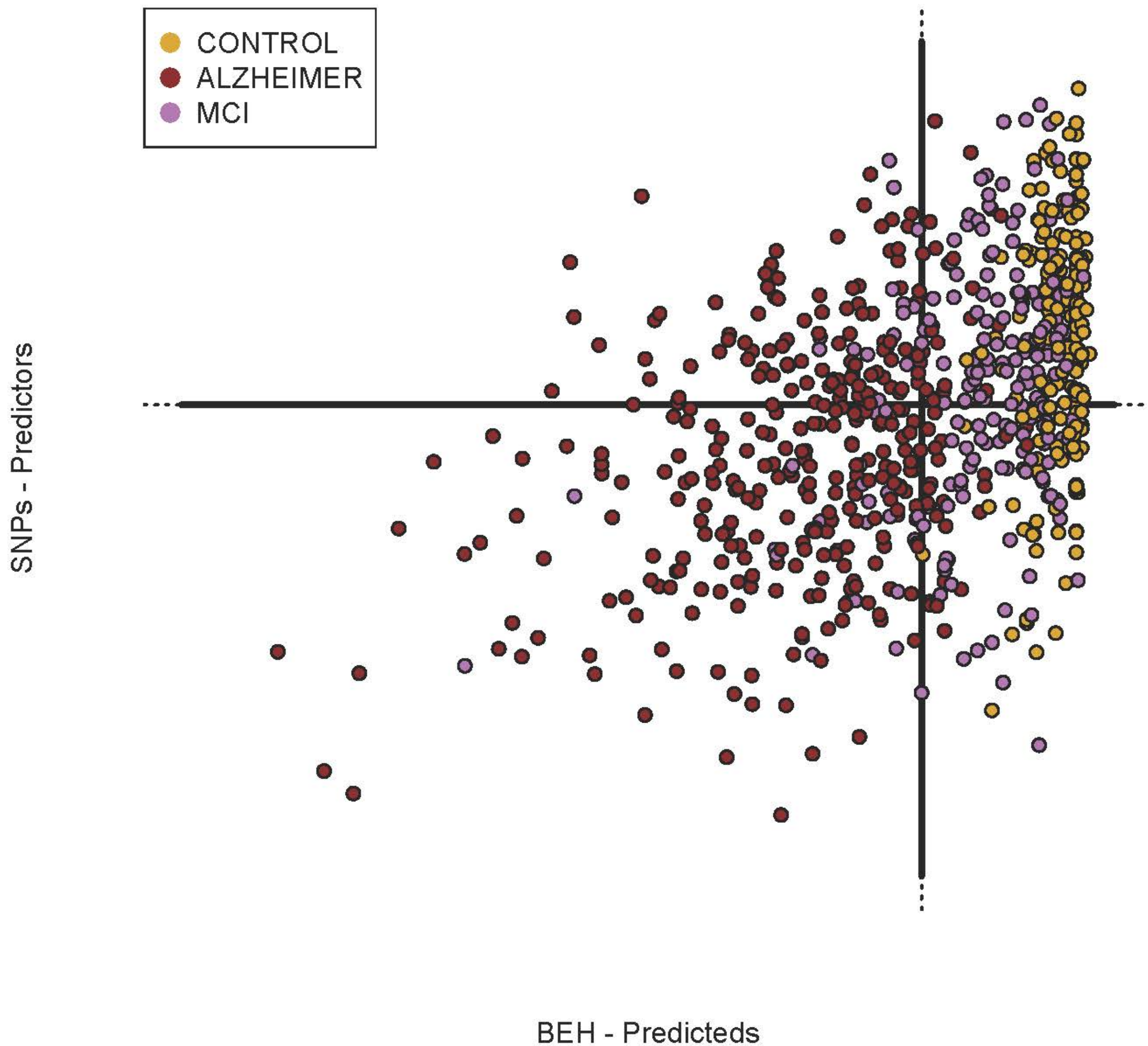
- n =756 observations
  - Alzheimer's = 337;

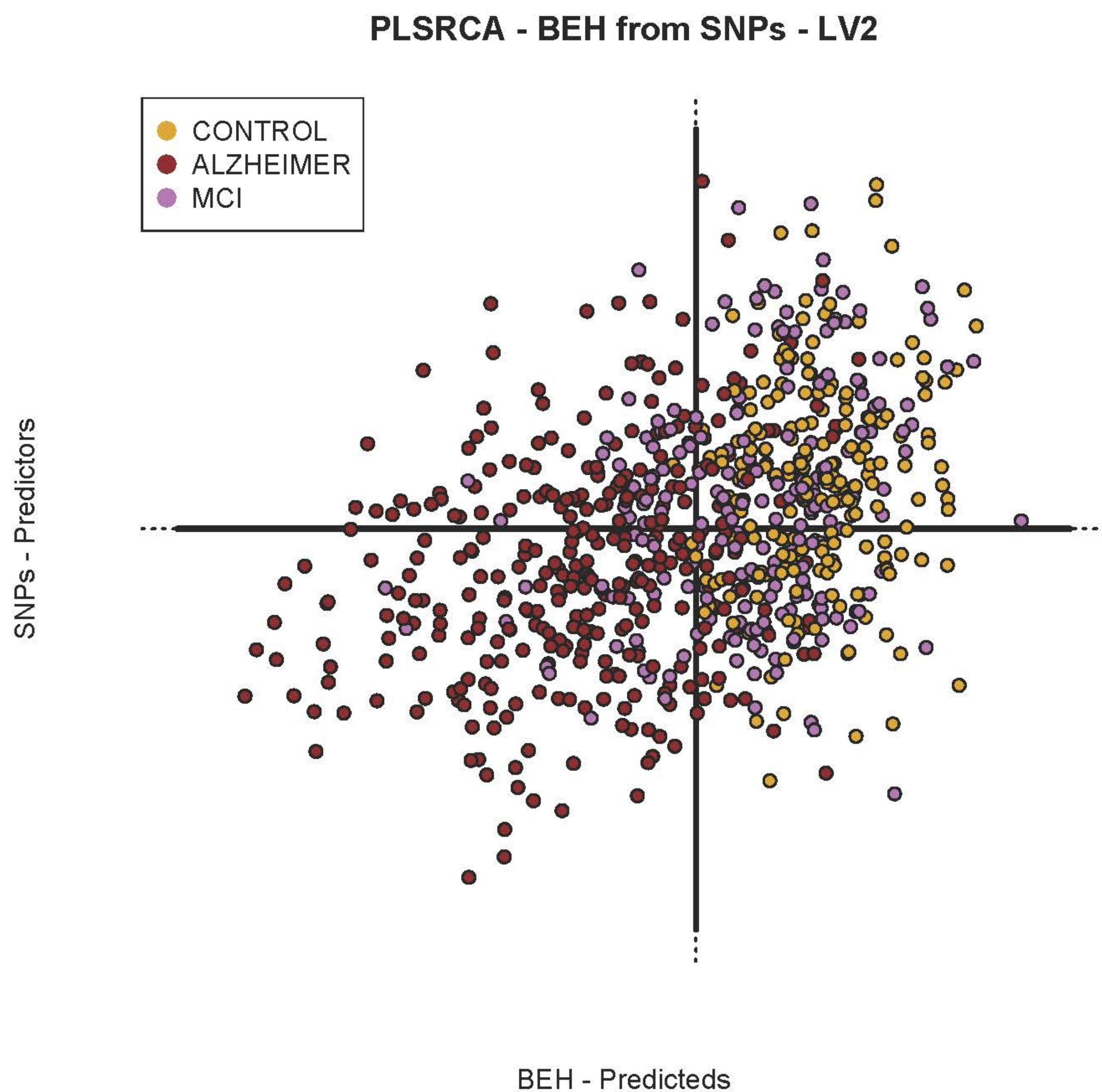
    Mild Cognitive Impairment = 210; Control = 210
- K = 145 SNPs from more than 500 000
  - $\mathbf{Y}$ has 435 (disjunctive) columns
- J = 51 questions from 3 scales (depression, dementia, cognition)
  - $\mathbf{X}$ has 119 (disjunctive) columns

PLSRCA (BEH from SNPs) - BEH

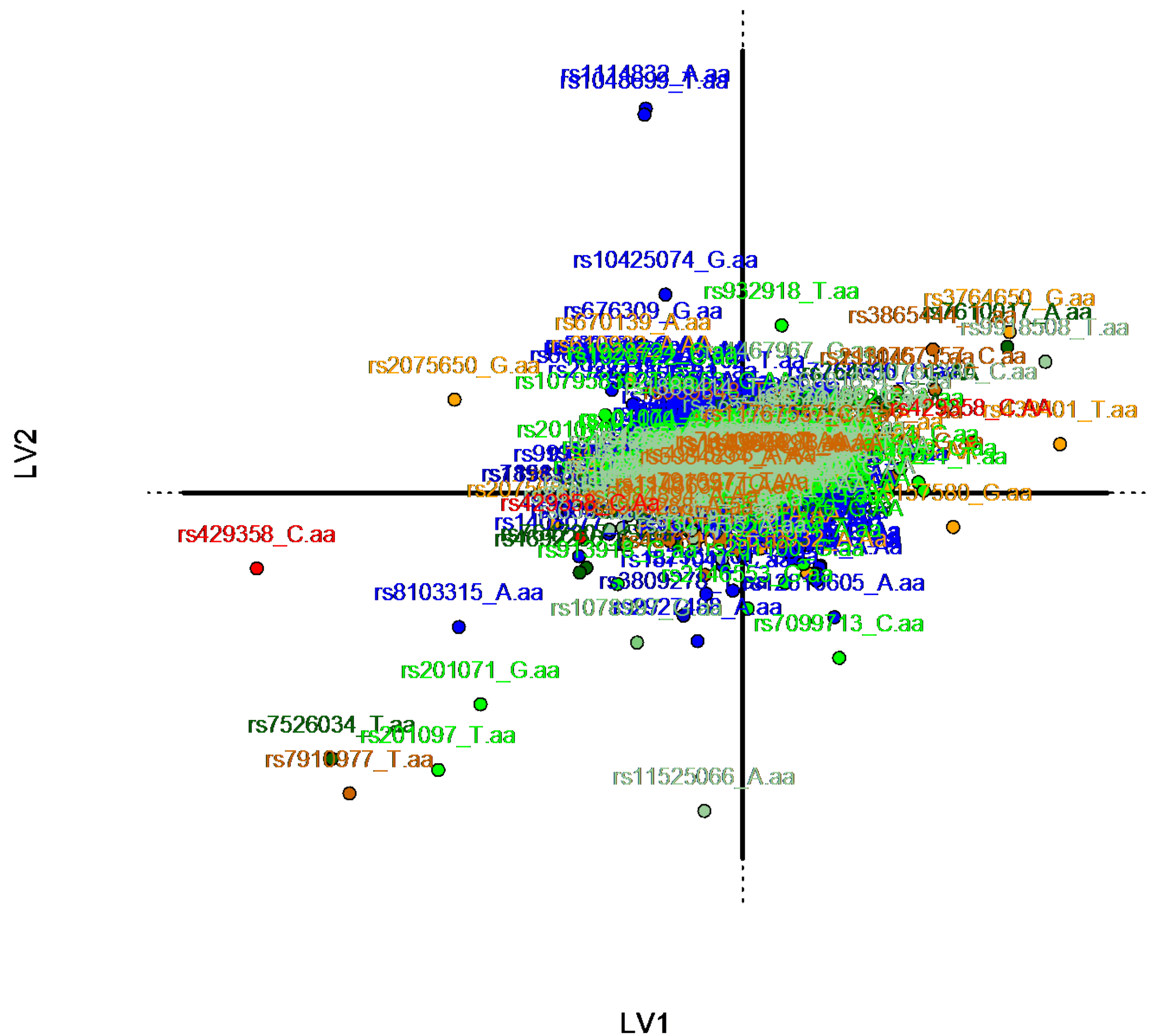


PLSRCA - BEH from SNPs - LV1




PLSRCA - BEH from SNPs - LV2
CONTROL
ALZHEIMER
MCI
SNPs - Predictors
BEH - Predicteds

PLSRCA (BEH from SNPs) - SNPs



6. Conclusions and perspectives

- **Summary**
  - CA of a subtable of a Burt’s matrix as a symmetrical PLS (or Inter Battery analysis) between 2 disjunctive tables
  - PLSR-CA as a modified version and its relationship with CA
  - Successful application to the explanation of behavioural data by SNPs for Alzheimer’s disease

- **Perspectives**
  - Multi-block PLSR-CA: can be used to answer the following question: how does each behavioral measure *uniquely* contributes to the results?...
  - Classification into categories of the Y variables
  - Heterogenous data with eg Escofier transformation

(b) Escofier-style transform

	$-\mathbf{x}_1$	$+\mathbf{x}_1$	$-\mathbf{x}_2$	$+\mathbf{x}_2$
Subj.1	$\frac{1-x_{1,1}}{2}$	$\frac{1+x_{1,1}}{2}$	$\frac{1-x_{1,2}}{2}$	$\frac{1+x_{1,2}}{2}$
Subj.2	$\frac{1-x_{2,1}}{2}$	$\frac{1+x_{2,1}}{2}$	$\frac{1-x_{2,2}}{2}$	$\frac{1+x_{2,2}}{2}$
...	...	...	...	...
Subj.I-1	$\frac{1-x_{I-1,1}}{2}$	$\frac{1+x_{I-1,1}}{2}$	$\frac{1-x_{I-1,2}}{2}$	$\frac{1+x_{I-1,2}}{2}$
Subj.I	$\frac{1-x_{I,1}}{2}$	$\frac{1+x_{I,1}}{2}$	$\frac{1-x_{I,2}}{2}$	$\frac{1+x_{I,2}}{2}$

References

- Abdi, H. & Williams, L.J. (2010). Barycentric discriminant analysis (BADIA). In N.J. Salkind, D.M., Dougherty, & B. Frey (Eds.): *Encyclopedia of research design*. Thousand Oaks (CA): Sage. pp. 64–75.
- Beaton, D., Dunlop, J., ADNI, & Abdi, H. (in press, 2015). Partial Least Squares-Correspondence Analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, **20**.
- Beaton, D., Filbey, F., & Abdi H. (2013). Integrating partial least squares correlation and correspondence analysis for nominal data. In Abdi, H., Chin, W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.), *New Perspectives in Partial Least Squares and Related Methods*. New York: Springer Verlag. pp. 81–94.
- Beh, E., Lombardo, R. (2014) : *Correspondence analysis*. London: Wiley
- Bourdieu, P. (1979) : *La distinction*, Editions de Minuit, (English translation 1984). Harvard University Press

- Cazes, P. (1980). L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. *Les cahiers de l'analyse des données*, **5**, 145–161.
- Celeux, P., & Nakache, J.P. (1994). *Analyse discriminante sur variables qualitatives*. Paris: Polytechnica.
- Dray, S. (2014). Analyzing a pair of tables: co-inertia analysis and duality diagrams. In J. Blasius & M. Greenacre, (Eds.), *Visualization of verbalization of data* (p. 289–300). Boca Raton, FL: CRC Press.
- Greenacre, M. (2007). *Correspondence analysis in practice*, 2nd edition, Chapman & Hall/CRC
- Lebart, L., Piron, M., & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle* (4th edition). Paris: Dunod.
- Leclerc, A. (1975). L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, **23**, 5–16.
- Leclerc, A. (1976). Une Étude de la Relation Entre une Variable Qualitative et un Groupe de Variables Qualitatives, *International Statistical Review / Revue Internationale de Statistique*, **44**, 241–248.

- Masson, M. (1974). *Processus linéaires et analyses de données non-linéaires*, Ph.D. Université Paris 6,
- Nakache, J.-P., Lorente, P., Benzécri, J.-P., & Chastang, J.-F. (1977). Aspects pronostiques et thérapeutiques de l'infarctus myocardique aigu compliqué d'une défaillance sévère de la pompe cardiaque. Application des méthodes de discrimination. *Les cahiers de l'analyse des données*, **2**, 415–434.
- Nakache, J.P., & Confais, J. (2003). *Statistique explicative appliquée*. Paris: Technip.
- Nakache, J.-P. (1981). Some methods in discriminant analysis on binary variables. In: *Perspective in Medical Statistics*. New York: Academic Press. pp 133–135.
- Saporta, G. (1975). *Liaisons entre plusieurs ensembles de variables et codages de données qualitatives*, Ph.D. , Université Paris 6, https://tel.archives-ouvertes.fr/tel-00649127/document
- Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Paris: Technip.
- Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*, 111–136.