Clusterwise Sparse PLS

Stéphanie Bougeard [1], Ndeye Niang-Keita [2],
Cristian Preda [3], Gilbert Saporta [2]

[1] French Agency for Food, Environmental, Occupational Health & Safety (Anses) France
[2] Conservatoire National des Arts et Métiers, France
[3] Université Lille 1 , France

Outline

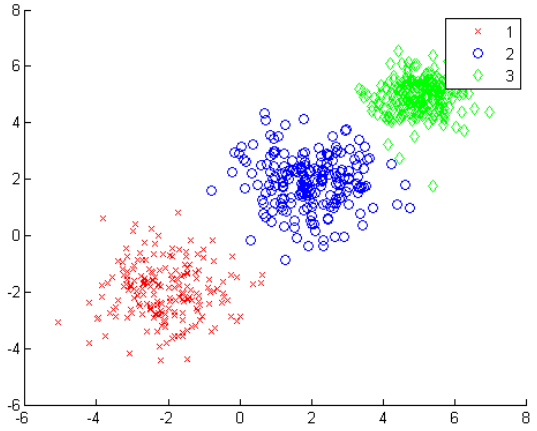
1. Introduction
2. Sparse regression and sparse PLS regression
3. Clusterwise regression
4. Clusterwise sparse PLS regression
5. Conclusion

1.Introduction

- PLS regression successful for : $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$
  - multicollinearty
  - p>n
- PLS keeps all variables but for high dimensional data p>>n (gene expression data, chemometrics, …) : lack of interpretabilty, non robust results
- Sparse regression: provides combinations of a small number of variables

An other issue: Big Data are usually heterogeneous

- When the cluster structure is unknown, clusterwise regression provides groups and local models



This talk: **clusterwise sparse PLS regression**

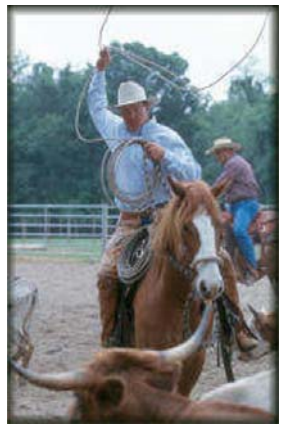
A focus on prediction

- To predict new observations.
- To get (unknown) parameters (coefficients) and hyperparameters (number of clusters, number of components) by cross-validation

2. Sparse regression and sparse PLS regression

- Keeping all predictors is a drawback for high dimensional data: combinations of too many variables cannot be interpreted
- Sparse methods simultaneously shrink coefficients and select variables, hence better predictions

2.1 Lasso



The Lasso (Tibshirani, 1996) is a shrinkage and selection method. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients (L_1 penalty).

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{with} \sum_{j=1}^{p} |\beta_j| < c$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

- Looks similar to ridge regression (L_2 penalty)

$$\min \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 \quad \text{with} \quad \left\| \boldsymbol{\beta} \right\|^2 < c$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 - \lambda \left\| \boldsymbol{\beta} \right\|^2$$

- Lasso continuously shrinks the coefficients towards zero when c decreases
- Convex optimisation; no explicit solution
- If $c > \sum_{j=1}^{p} \left| b_{jols} \right|$ Lasso identical to OLS

- Finding the optimal parameter
  - Cross validation if optimal prediction is needed
  - BIC when the sparsity is the main concern

$$\lambda_{opt} = \arg\min_{\lambda} \left(\frac{\left\| y - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda) \right\|^2}{n\sigma^2} + \frac{\log(n)}{n} \hat{df}(\lambda) \right)$$

a good unbiased estimate of df is the number of nonzero coefficients . (Zou et al., 2007)

2.2. Sparse PLS

Several solutions:

a. Le Cao et al. (2008), Liquet et al. (2016)

b. **Chun & Keles (2010)**

Framework: PLS2 for a block of responses Y

- Solution a (Package sgPLS)

$$\min\{\left\| \mathbf{X}'\mathbf{Y} - \mathbf{u}'\mathbf{v} \right\|^2 + \lambda_1 |\mathbf{u}| + \lambda_2 |\mathbf{v}|\},$$

$$\text{subject to } \left\|\mathbf{u}\right\| = \left\|\mathbf{v}\right\| = 1 \quad where\ |\mathbf{u}| = \sum_{j=1}^{p} |u_j|$$

Equivalent to:

$$\max\{\text{cov}(\mathbf{Y}\mathbf{v}, \mathbf{X}\mathbf{u}) + \lambda_1 |\mathbf{u}| + \lambda_2 |\mathbf{v}|\},$$

$$\text{subject to } \left\|\mathbf{u}\right\| = \left\|\mathbf{v}\right\| = 1$$

variant (Package `mixOmics`)

$$\min_{\mathbf{u},\mathbf{v}} \left\| \mathbf{X}'\mathbf{Y} - \mathbf{u}\mathbf{v}' \right\| + P_{\lambda_1}(\mathbf{u}) + P_{\lambda_2}(\mathbf{v})$$

$$P_\lambda(x) = \left(|x| - \lambda\right)_+ sign(x)$$

« Soft thresholding function »

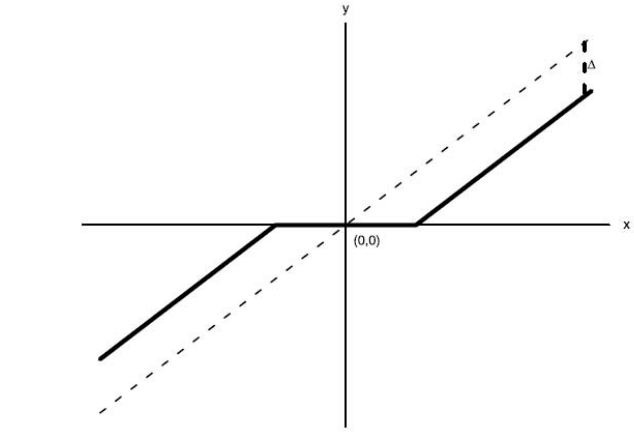


Figure 1. An illustration of soft-thresholding rule $y = (|x| - \Delta)_+ Sign(x)$ *with* $\Delta = 1$.

- **Solution b** (package `spls`)

inspired by SIMPLS: $\max_{\mathbf{w}} \left(\mathbf{w'X'YY'Xw} \right)$ avec $\left\| \mathbf{w} \right\|^2 = 1$

sparse factor **c** close to initial PLS solution **α**

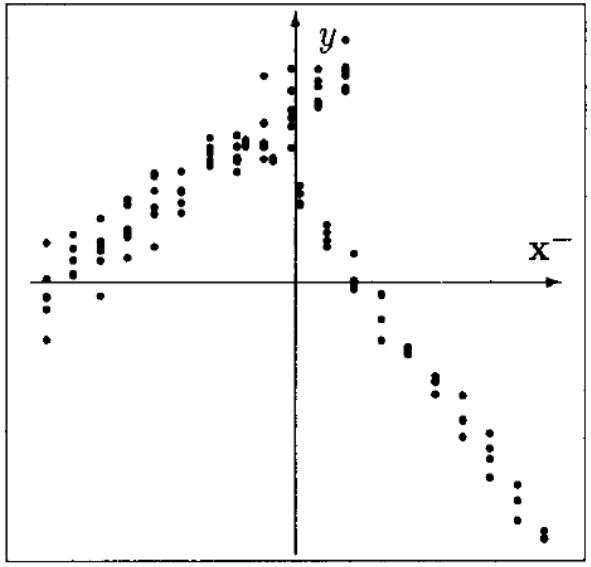
$$\min_{\alpha, c} \left(-k\boldsymbol{\alpha}'\mathbf{M}\boldsymbol{\alpha} + (1-k)(\mathbf{c}-\boldsymbol{\alpha})'\mathbf{M}(\mathbf{c}-\boldsymbol{\alpha}) + \lambda_1 \left\| \mathbf{c} \right\|_1 + \lambda_2 \left\| \mathbf{c} \right\|^2 \right)$$

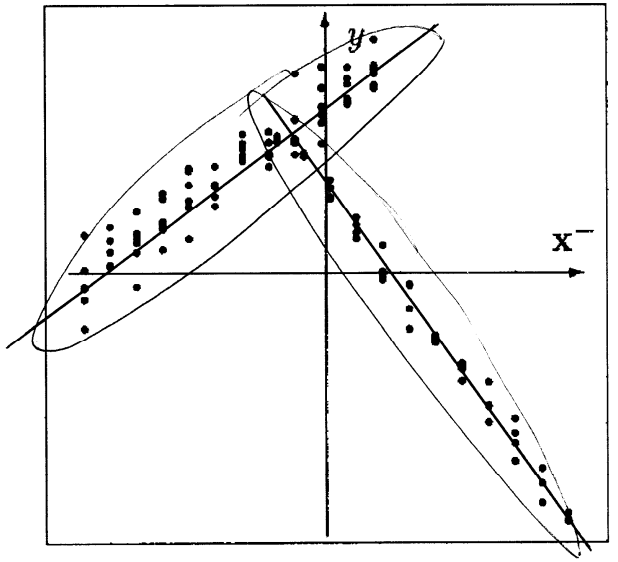
$$\text{with } \boldsymbol{\alpha}'\boldsymbol{\alpha} = \mathbf{c}'\mathbf{c} = 1 \,, \;\; \mathbf{M} = \mathbf{X'YY'X}$$

Iterative algorithm, alternating between **c** and **α**

3. Clusterwise regression

- Local models *versus* global model





Hennig, 2000

- Unknown partition *:*
  - *Unobserved heterogeneity ; latent classes*
  - *Simultaneous search for classes and models for each class*
- For each class, fit a linear model by least squares.

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbf{1}_k(i)\left(y_i-(\alpha_k+\beta_k x_i)\right)^2$$

  - The fit is always better than with the global model

Two algorithms

- " Batch " algorithm (Charles, 1977):
  - An application of dynamic clustering (Diday, 1974)
    - Step 1: define an initial partition and estimate K local models.
    - Step 2: Each observation is reallocated to the cluster giving the smallest regression residual ie the best prediction. Once all observations being reallocated (or not) , replace the initial partition
    - Iterate step 1 and 2 until convergence
- "Stochastic" algorithm
    - Step 2: Update criterium and partition after each reallocation (true k-means)
    - Späth (1979) coined the expression « clusterwise regression »

How to predict new observations knowing only the predictors?

- « Hard » rule: allocate to the nearest cluster and apply the relevant model
  - How? **PLS-DA**, nearest neighbours etc.
- « Soft » rule: weighted average of the K predictions
  - Close to Bayesian Model Averaging : weight = posterior probabilities of each model given the observation
- Random rule: choose at random one prediction according to posterior probabilities

A few comments on mixture models or latent class regression:

Proposed by DeSarbo & Cron (1988)

We assume y_i is distributed as a finite sum or mixture of conditional univariate normal densities:

$$y_i \sim \sum_{k=1}^{K} \lambda_k f_{ik}(y_i \mid X_{ij}, \sigma_k^2, b_{jk}) \tag{9}$$

$$= \sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{x}_i \mathbf{b}_k)^2}{2\sigma_k^2}\right], \tag{10}$$

- Mainstream methodology: maximum likelihood through EM algorithm
  - Extensions to Poisson regression (Wedel & al. 1993), generalized local linear models (Wedel & DeSarbo, 1995)
- However prediction is generally not possible
  - In `flexmix`, posterior probabilities need the true class

$$\mathrm{P}(j|x,y,\psi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_k \pi_k f(y|x,\theta_k)}$$

"We have no solution for this problem: Without y you cannot determine the likelihood and hence not into which cluster the observation belongs. You could calculate predictions for each cluster, but then you have K answers, not one. " (F.Leisch , personal communication)

Clusterwise PLS regression

- Some kind of regularized regression should be used for small size clusters where n<p and no OLS solution exists
  - Clusterwise Ridge regression (Charles, 1977)
  - Clusterwise PLS regression :
    - Functional data (Preda & Saporta, 2005)
    - Symbolic data (De Carvalho et al, 2010)

4. Clusterwise sparse PLS regression

- **Context**
  - High dimensional explanatory data **X** (with p>>n),
  - One or several variables **Y** to explain,
  - Unknown clusters of the n observations.

- **Twofold aim**
  - Clustering: get the optimal clustering of observations,
  - Sparse regression: compute the cluster sparse regression coefficients within each cluster to improve the **Y** prediction.

  $\rightarrow$ Improve high-dimensional data interpretation.

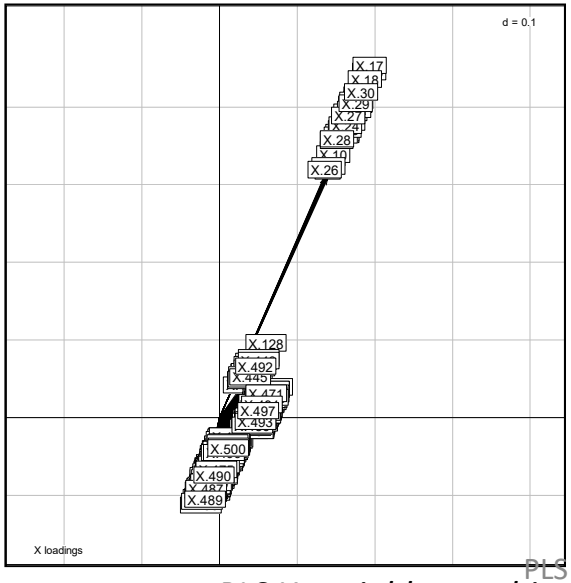
Clusterwise sparse PLS: stochastic algorithm

- **Start from an initialization of the n observations into K clusters**
- **For each observation i ∈ [1;n]**
  - Compute sparse PLS regressions where i belongs alternatively to each of the K clusters (select the optimal sparse parameter λ and the dimension H through 10-fold CV)
  - For each of the K solutions, compute the criterion C=$\sum_{k=1}^{K} \|y_k - \widehat{y_k}\|^2$
  - Update the assignment of i to the cluster which minimizes C
  - Update the sparse PLS regression coefficients
- **Iterate until convergence**
- **Repeat the procedure for several initializations; select the best one**
  - Get the assignment of the n observations into K clusters
  - Get the sparse PLS regression coefficients for each cluster
- **Process the prediction model**
  - PLS-DA applied on the concatetaned **X** variables selected from the sparse PLS regression coefficients from each cluster.

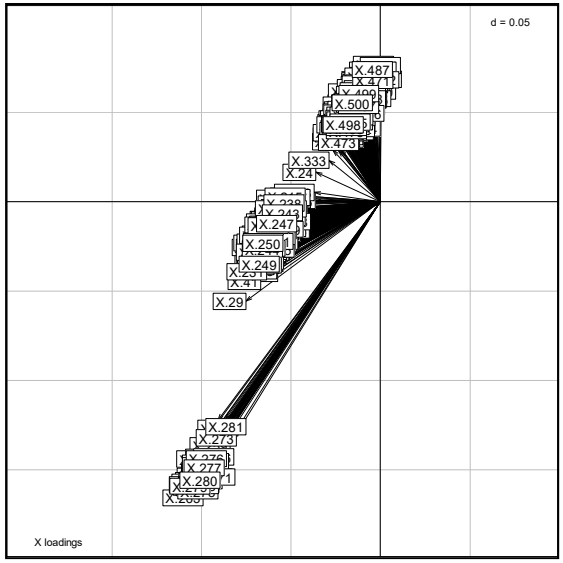
Application: simulated data

- **Data features**
  - 500 explanatory variables (**X**) and 1 dependent one (y),
  - 50 observations organized into K=2 well-separated clusters of equal size
  - Cluster 1: 30 **X** variables positively linked to y (β=2), correlated (cor=0.9) and non-correlated (cor=0) with the 470 other X variables,
  - Cluster 2: 30 other **X** variables negatively linked to y (β=-2) , correlated (cor=0.9) and non-correlated (cor=0) with the 470 other X variables.

Cluster 1 (N=23)

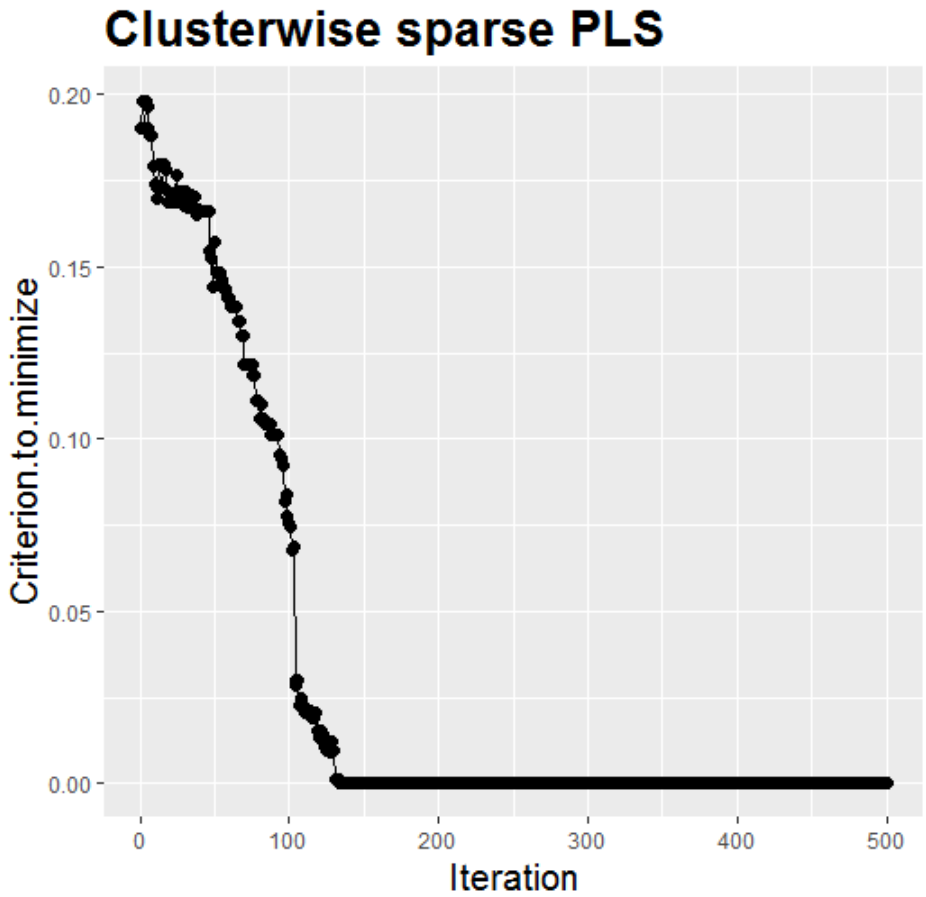


Cluster 2 (N=27)



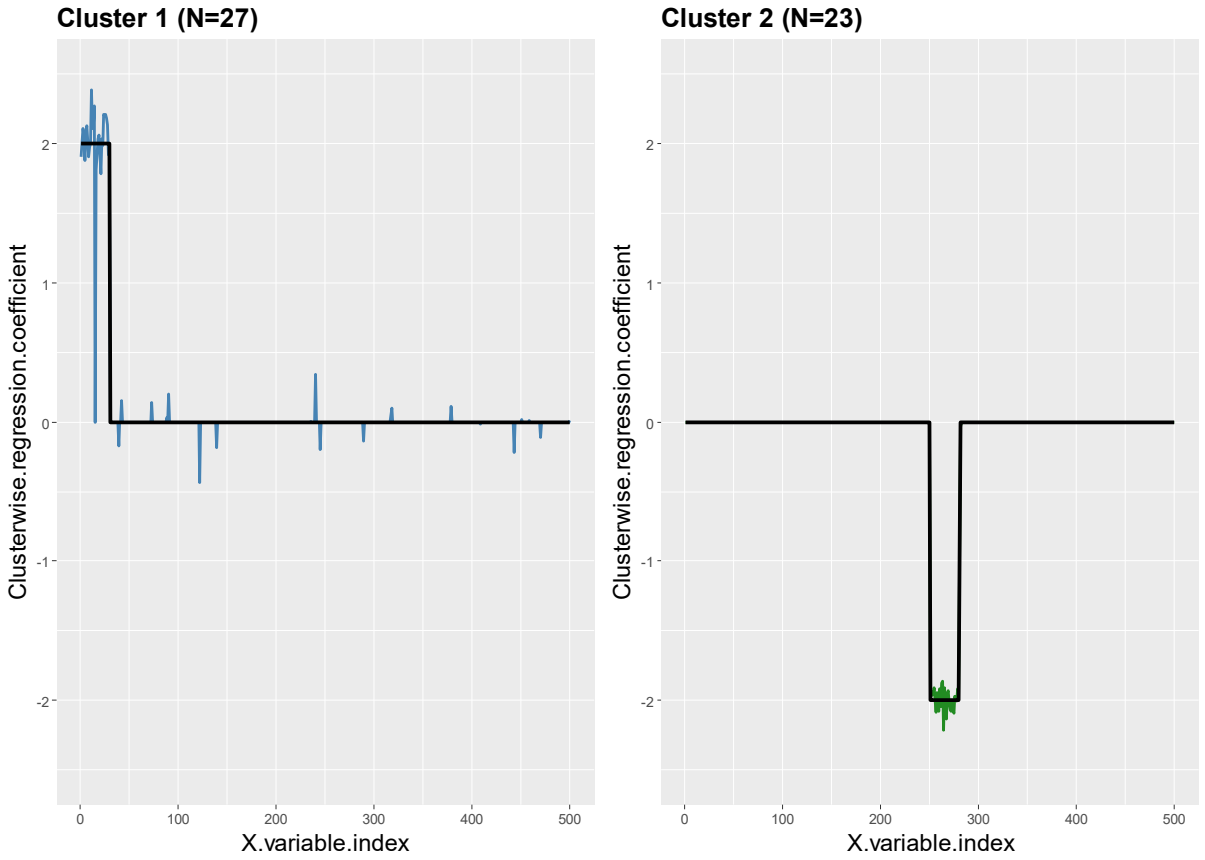
PLS X-variable graphical display for the first two components

Simulated data: convergence criterium



- Number of iterations = Number of observations (50) * Number of passes (10)
- Stable convergence after 3 passes

Simulated data (retrieving the true betas)



- λ = 0.8, H= 2 (for both clusters)
- The method cw.spls perfectly retrieve the actual clusters (AdjRand index = 1)
- The actual regression coefficients are correctly retrieved in cluster 1 and well retrieved in cluster 2.

Application: genomic data (Bushel et al., 2007)

- **Data features** (liver toxicity from `mixOmics` R package)
  - **X**: 3116 genes (mRNA extracted from liver)
  - y: clinical chemistry marker for liver injury (serum enzymes level)
  - Observations: 64 rats
    - Each rat is subject to different more or less toxic acetaminophen doses (50, 150, 1500, 2000)
    - Necropsy is performed at 6, 18, 24 and 48 hours after exposure (i.e., time when the mRNA is extracted from liver)

- **Data pre-processing**
  - y and **X** are centered and scaled

- **Aims**
  - Improve the sparse PLS link between genes (**X**) and liver injury marker (y)…
  - … while seeking K unknown clusters of the 64 rats.

Application: optimal number of clusters

- A sparse model without cluster (K=1) leads to more stable prediction but is less explanatory,
- A clusterwise sparse model with (K=3) clusters is a correct trade-off between explanation and prediction.

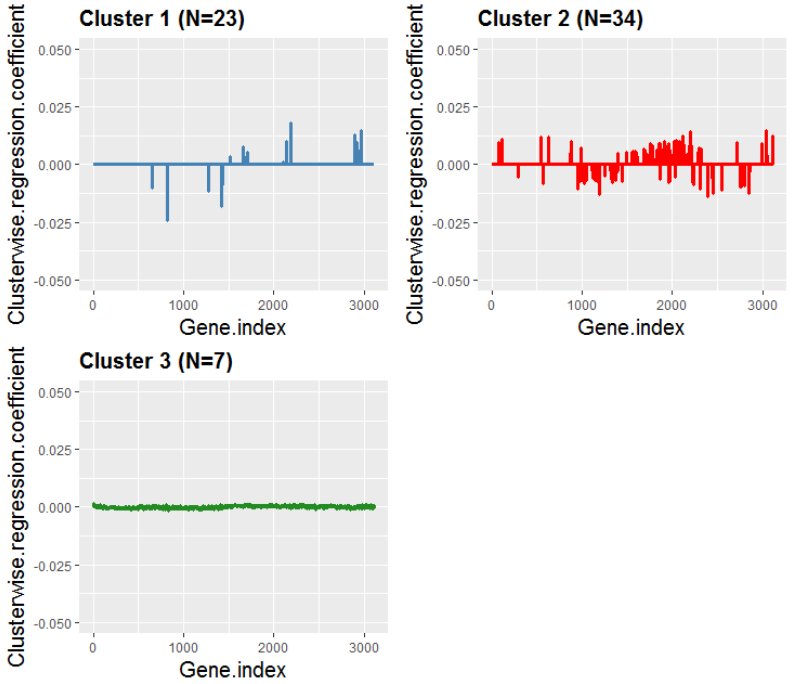
Application: cw.spls regression coefficients

No cluster K=1

λ = 0.9 (sparse parameter)
→ 25 selected genes

H = 2 dimensions

K=3 clusters of rats



Links between genes (**X**) and marker for liver injury (y) are specific to each cluster

Cluster 1: 25 selected genes (λ=0.8, H=2)
Cluster 2: 279 selected genes (λ=0.7, H=2)
Cluster 3: 3016 selected genes (λ=0.1, H=2)

Conclusions and perspectives

- Clusterwise sparse PLS regression useful for the Big Data case
  - Find clusters and local models with a good prediction ability and interpretability

- Research in progress:
  - Extension to the case where predictors have a block structure
  - Enhancing the R package `mbclusterwise`

Thank for your attention

References

- Bougeard,S. (2016): R package mbclusterwise, CRAN
- Charles, C. (1977): *Régression Typologique et Reconnaissance des Formes.* Thèse de doctorat, Université Paris-Dauphine.
- Chun, H. and Keles, S. (2010): Sparse partial least squares for simultaneous dimension reduction and variable selection*, Journal of the Royal Statistical Society - Series B*, Vol. 72, pp. 3–25.
- De Carvalho, F., Saporta, G., Queiroz, D. (2010): A Clusterwise Center and Range Regression Model for Interval-Valued , *COMPSTAT'2010, 19th International Conference on Computational Statistics,* pp.461-468,
- Diday, E. (1974): Introduction à l'analyse factorielle typologique, *Revue de Statistique Appliquée*, 22, 4, pp.29-38
- Hennig, C. (1999): Models and methods for clusterwise linear regression. In*:Classification in the Information Age*, Springer, pp.179-187.

- Lê Cao K.-A., Rossouw, D., Robert-Granié C., Besse, P. (2008) :A Sparse PLS for Variable Selection when Integrating Omics data *Statistical Applications in Genetics and Molecular Biology:* Vol. **7** : Iss. 1, Article 35.
- Leisch, F. (2004) : FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(8).
- Liquet, B., Lafaye de Michaux, P., Hejblum,B. , Thiebaut, R. (2016) : Group and sparse group partial least square approaches applied in genomics context, *Bioinformatics,* 32(1), 35–42
- Niang, N., Bougeard, S., Saporta,G., Abdi, H. (2015) : Clusterwise multiblock PLS, *CARME 2015*, pp.58, Neaples
- Preda, C. and Saporta, G. (2005): Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 49, pp.99–108.
- Späth, H. (1979): Clusterwise linear regression, *Computing*, 22, pp.367-373
- Tibshirani, R. (1996) : Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288