Classification de variables et modèle de prédiction

Evelyne Vigneau, Véronique Cariou

StatSC

Oniris
École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation
Nantes Atlantique

Plan

- Contexte / Objectif
- La méthode CLV pour la classification de variables
- Adaptation de CLV dans un cadre supervisé
- Application à des données de RMN en contrôle qualité

Contexte / Objectif

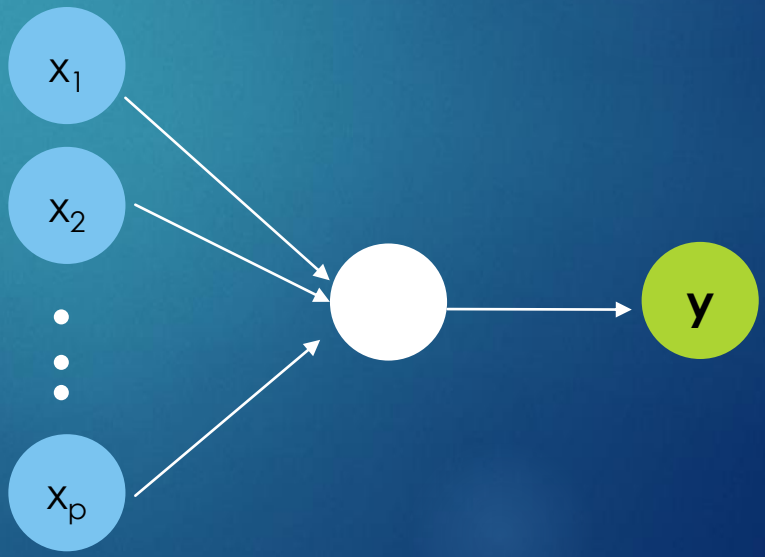
Modèle pour la prédiction

d'une variable de réponse (quantitative)

en fonction

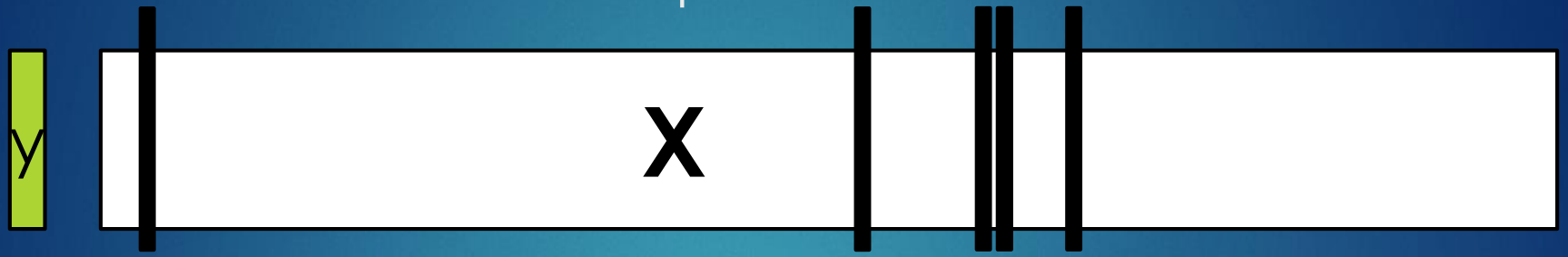
d'un grand nombre de prédicteurs,

souvent très colinéaires.



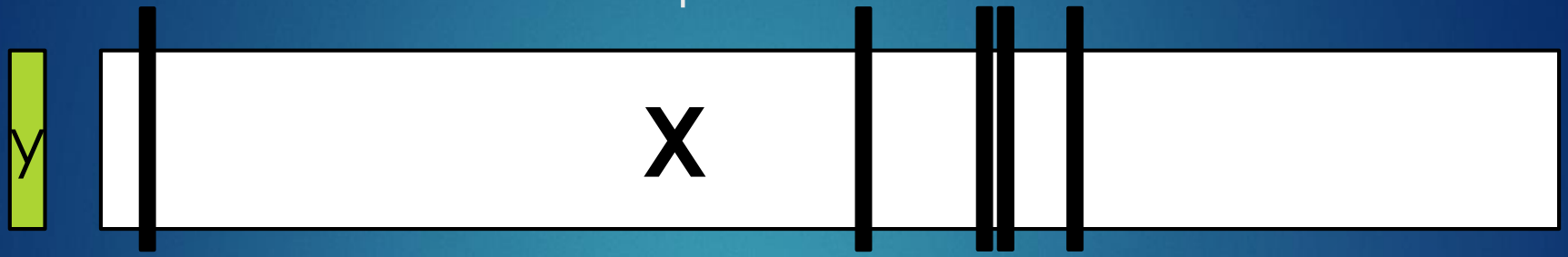
Contexte / Objectif

Sélection de variables prédictrices

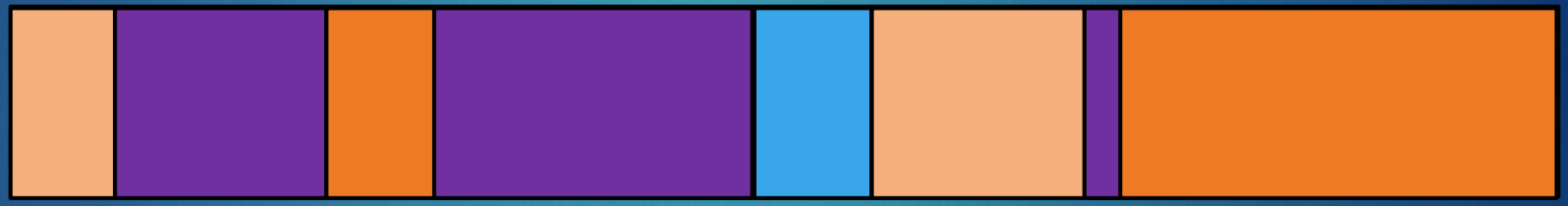


Contexte/ Objectif

Sélection de variables prédictrices

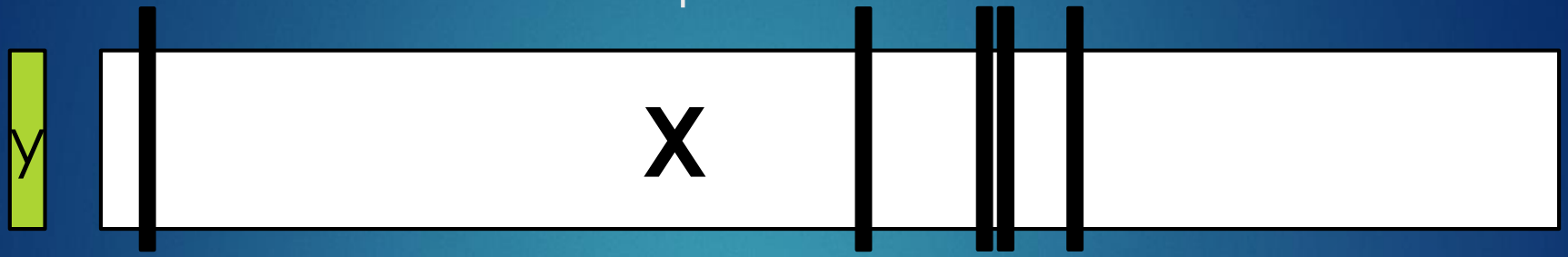


Classification de variables

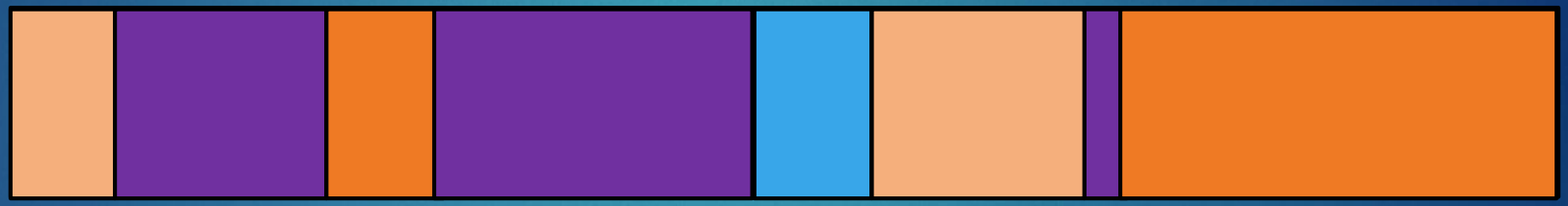


Contexte/ Objectif

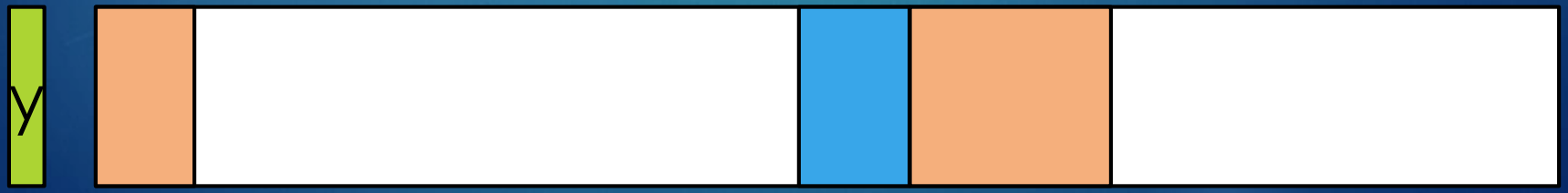
Sélection de variables prédictrices



Classification de variables

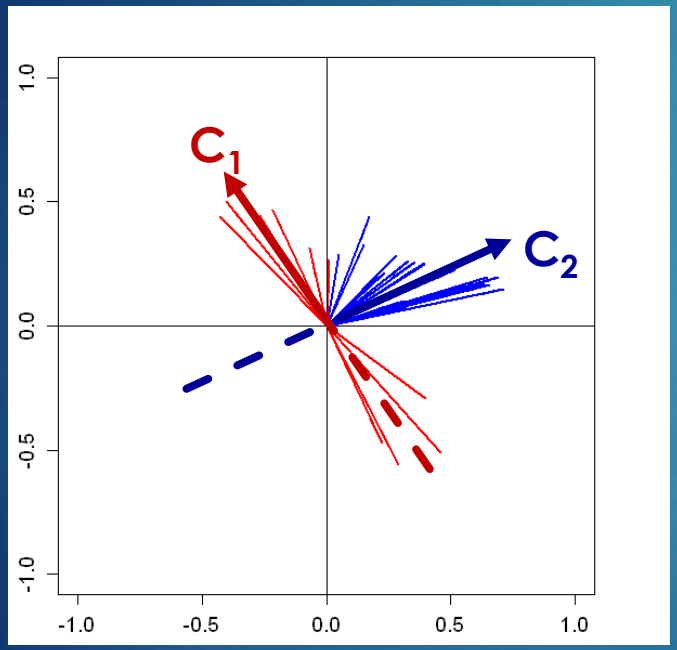


Proposition : Combinaison



La méthode CLV pour la classification de variables

Cas de groupes directionnels

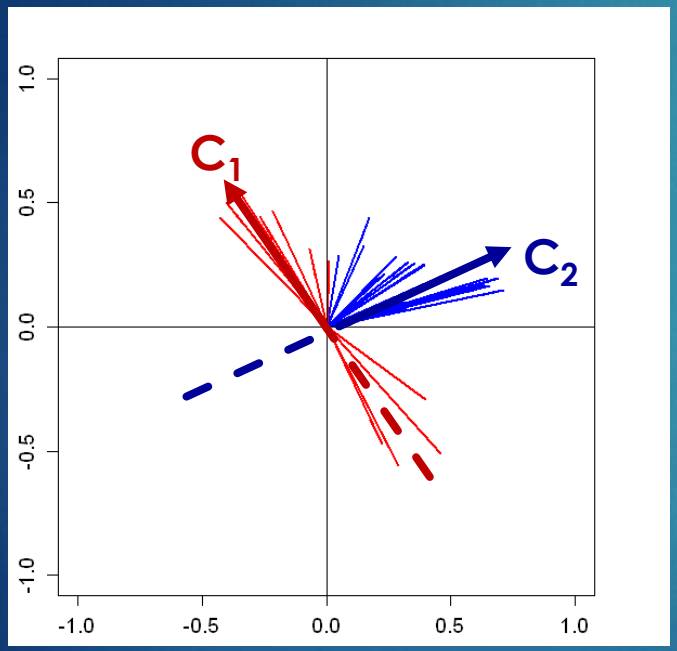


Identifier K groupes de variables
et, en même temps,
K composantes latentes
tel que
chaque variable dans un groupe G_k soit fortement corrélée, negativement ou positivement (groupes directionnels) à la variable latente correspondante, $\mathbf{c_k}$.

Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent component. *Communications in Statistics – Simulation and Computation*, **32**,1131–1150.

La méthode CLV pour la classification de variables

Cas de groupes directionnels

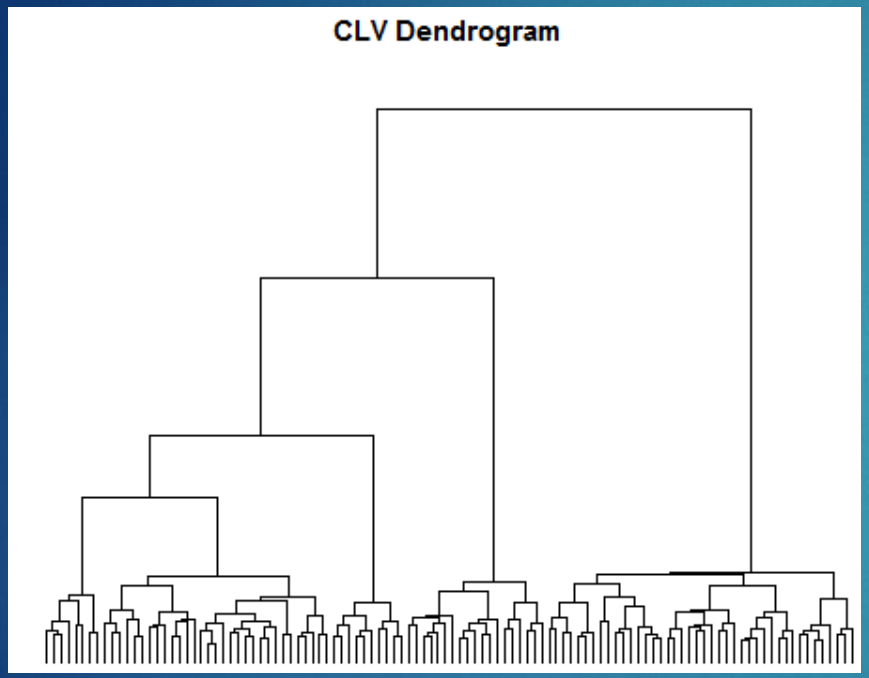


maximiser

$$T^{(K)} = \sum_{k=1}^{K} \sum_{j=1}^{q} \delta_{kj}\, cov^2(\mathbf{x_j}, \mathbf{c_k})$$

avec $var(\mathbf{c_k}) = 1$

La méthode CLV pour la classification de variables



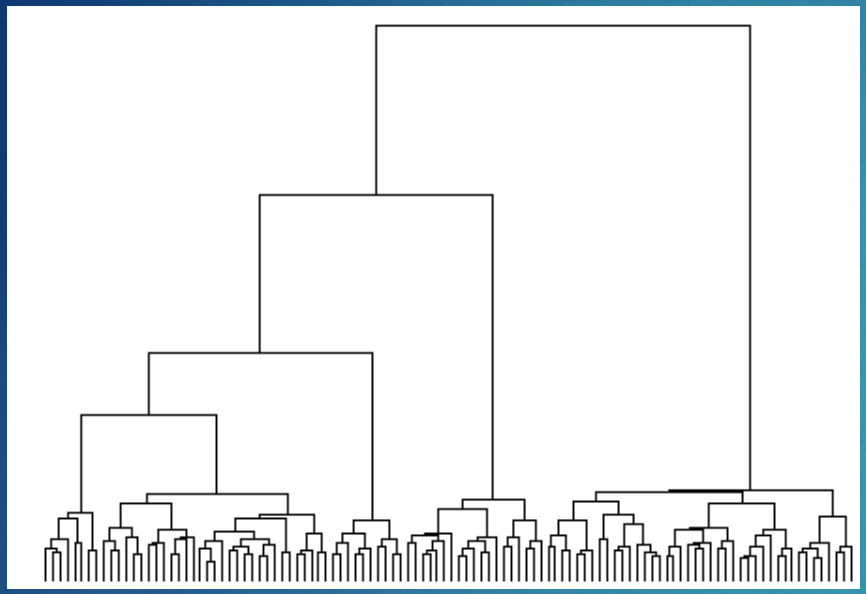
Mise en oeuvre d'un algorithme hiérarchique avec,

à chaque niveau ($\leftrightarrow$ K groupes),

une phase de consolidation (optimisation de $T^{(K)}$)

Fonction CLV du package R ClustVarLV

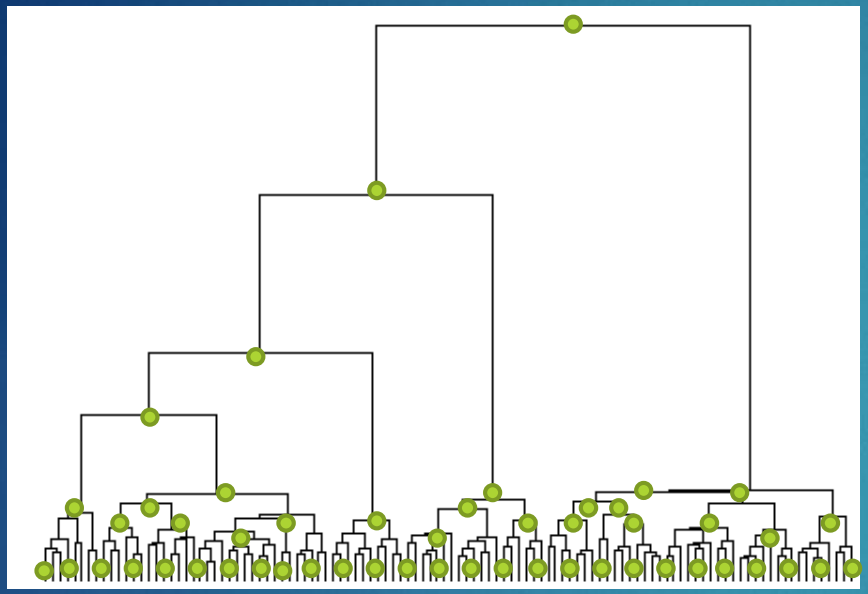
Classification et prédiction



y **X**

- Combien de groupes de variables **X** former ?
- Quel niveau de coupure du dendrogramme choisir ?

Classification et prédiction



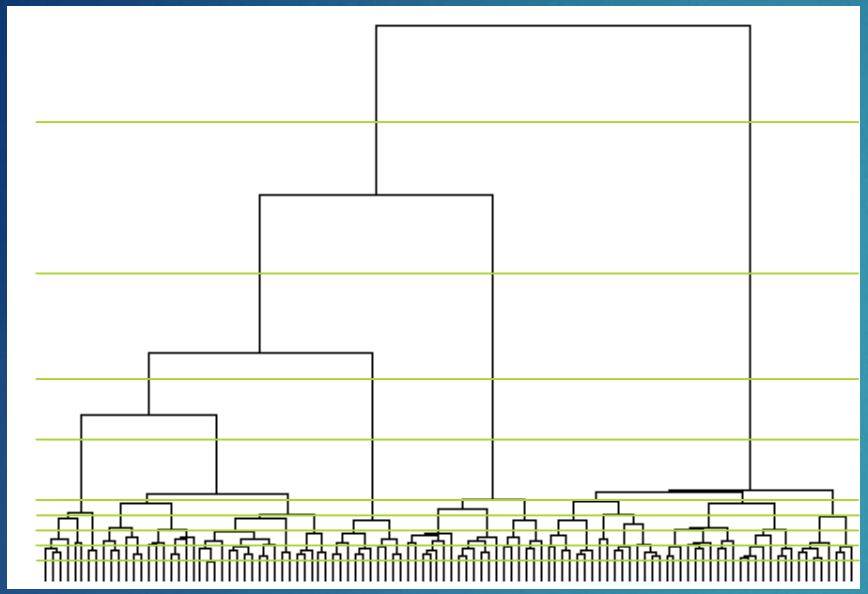
y **X**

Hastie et al., 2001

- Variables candidates : *super-*variables calculées aux $(2p - 1)$ noeuds
- Stratégie ascendante progressive

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biol*, 2(1); research0003.1–0003.12.

Classification et prédiction



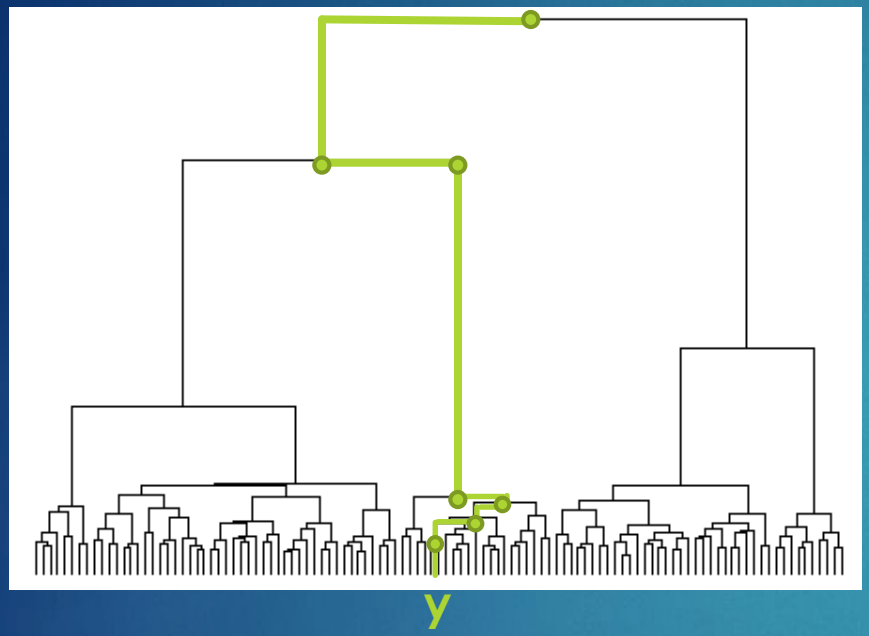
y **X**

Park et al. 2007

- Régression Lasso ajustée à chaque niveau de la hiérarchie
- Choix du niveau optimum par cross-validation

Park, M. Y., Hastie, T., and Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, 8(2), 212-227.

Proposition : 1ère étape



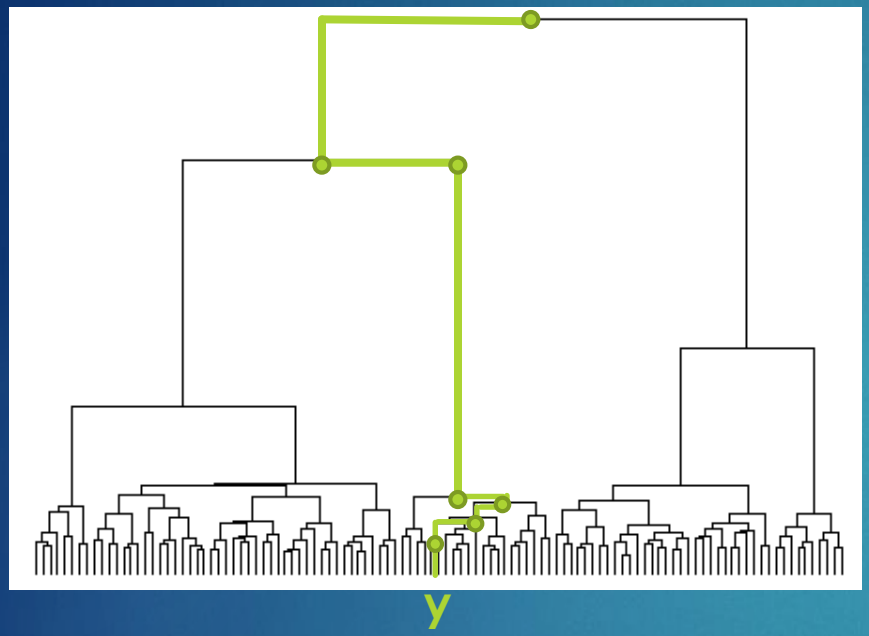
Chen & Vigneau, 2016

- Classification de [X | y]

$$[\mathbf{X} \mid \mathbf{y}]$$
$$\text{trace}(\mathbf{y}^T\mathbf{y}) = \text{trace}(\mathbf{X}^T\mathbf{X})/p$$

Chen, M. et E. Vigneau (2016). Supervised clustering of variables. *Advanced in Data Analysis and Classification* 10(1), 85–101.

1ère étape



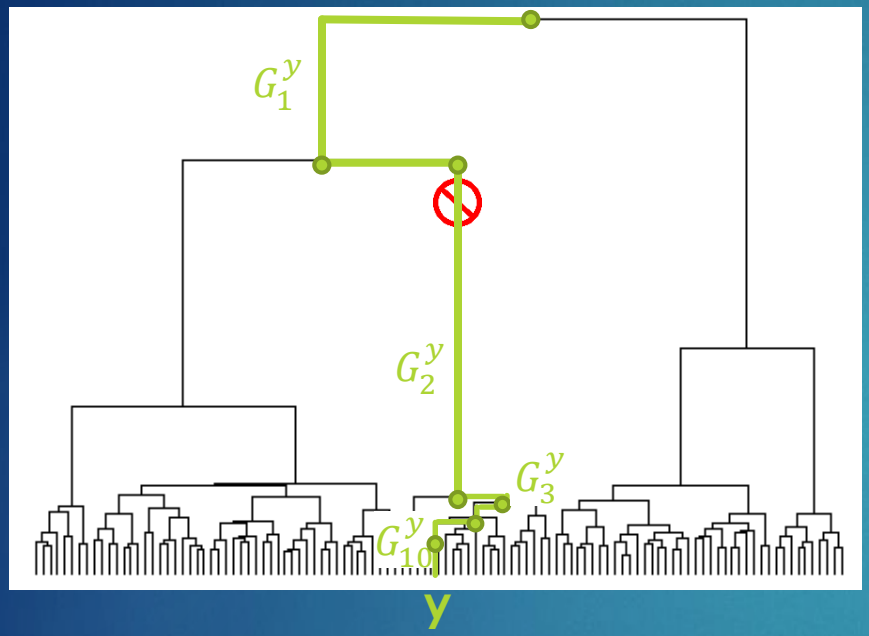
[X | y]

trace($\mathbf{y}^T\mathbf{y}$)=trace($\mathbf{X}^T\mathbf{X}$)/p

- A chaque niveau q de l'ensemble des niveaux de la hiérarchie, on repère le groupe de variables incluant **y** : G_q^y

1ère étape

discussion

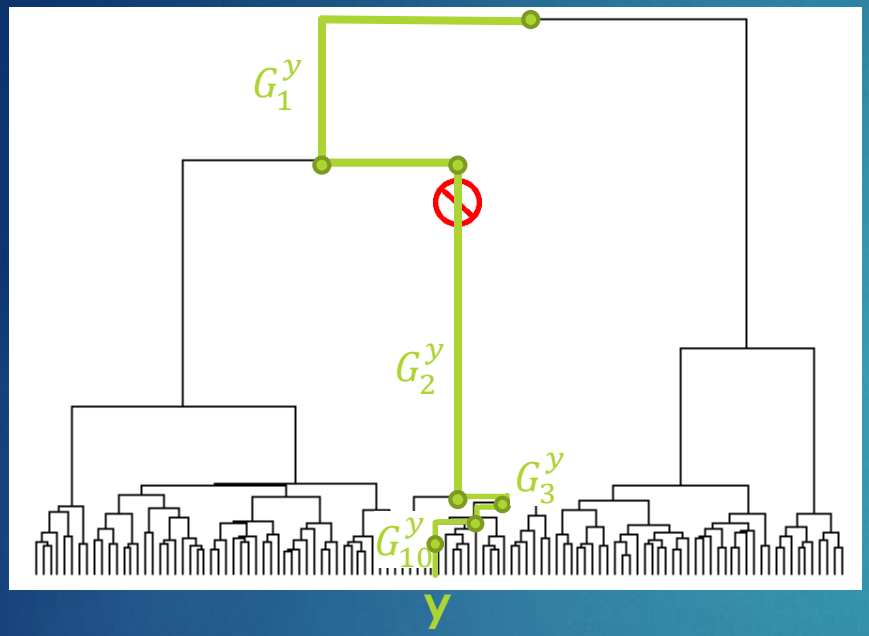


Quel niveau q
↔ Quel groupe de variables
↔ Quel variable latente de groupe considérer pour expliquer **y** ?

L'objectif est que le groupe des variable **X** associées à **y**, G_q^y , soit **le plus large possible** mais aussi **le plus unidimensionnel possible**.

1ère étape

discussion



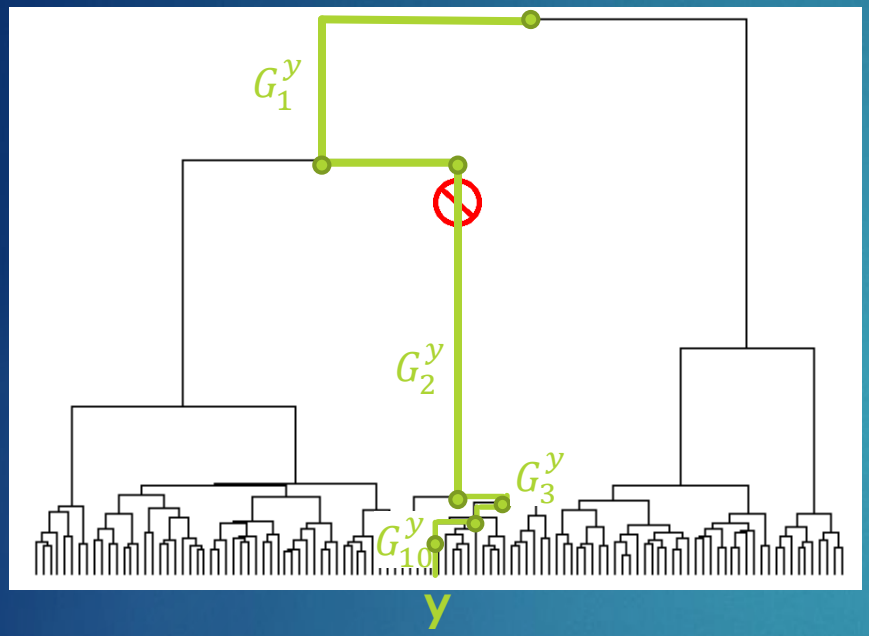
L'objectif: choisir G_q^y

Plusieurs critères ont été considérés :

- Variation locale du critère CLV (Chen & Vigneau, 2016).
- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.
- Test de Bartlett de sphéricité sur les résidus des variables **X** de G_q^y non expliqués par la variable latente du groupe, c^q.
- Minimiser les erreurs de prédiction (LOO) en utilisant la variable latente c^q.

1ère étape

discussion



L'objectif: choisir G_q^y

Plusieurs critères ont été considérés :

- Variation locale du critère CLV (G[illegible] & Vigneau, 2014)

  Difficulté de déterminer un seuil d'arrêt.*

- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.

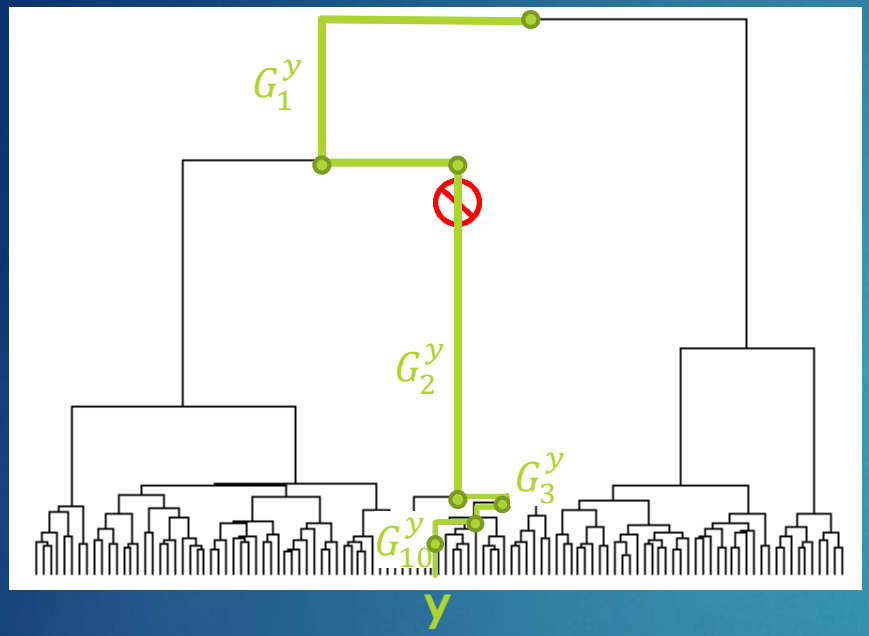
- Test de Bartlett de sphéricité sur les résidus des variables **X** de G_q^y non expliqués par la variable latente du groupe, c^q.

- Minimiser les erreurs de prédiction (LOO) en utilisant la variable latente c^q.

* Rq sur la base de l'étude de cas ci-après

1ère étape

discussion



L'objectif: choisir G_q^y

Plusieurs critères ont été considérés :

- Variation locale du critère CLV (Chen & Vigneau, 2016).

- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.

- Test de Bartlett de sphéricité sur les résidus des variables **X** de G^y [illegible] du groupe, c^q.

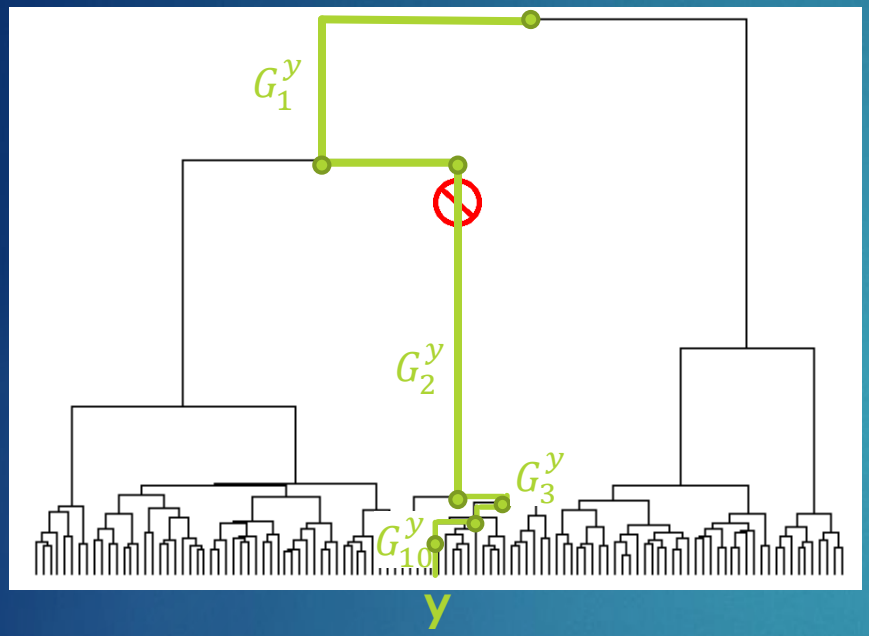
Tendance à choisir des groupes de très petite taille.*

- Minimiser les erreurs de prédiction (LOO) en utilisant la variable latente c^q.

* Rq sur la base de l'étude de cas ci-après

Proposition : 1ère étape

discussion



L'objectif: choisir G_q^y

Plusieurs critères ont été considérés :

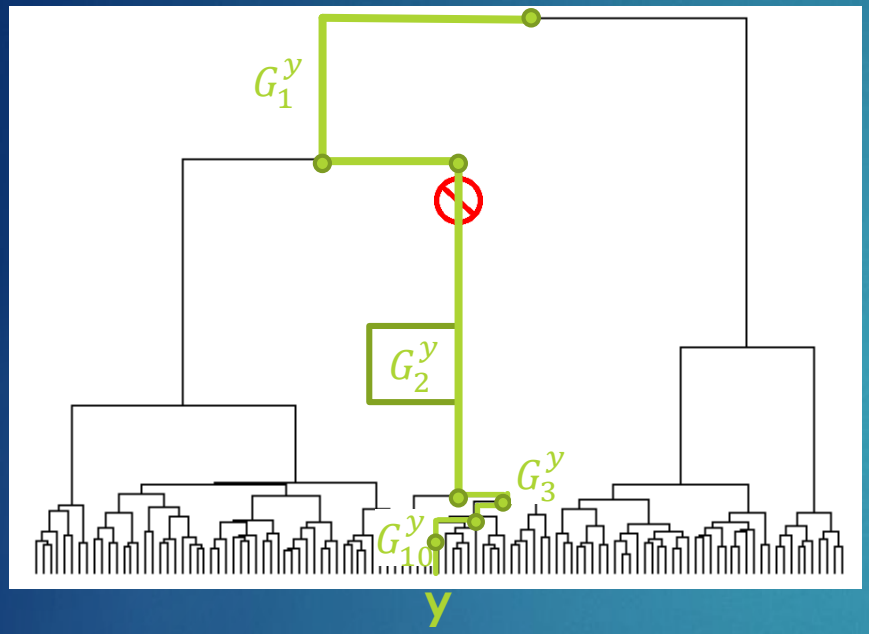
- Variation locale du critère CLV (Chen & Vigneau, 2016).
- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.
- Test de Bartlett de sphéricité sur les résidus des variables **X** de G_q^y non expliqués par la variable latente du groupe, c^q.
- Minimiser les erreurs de prédiction (LOO) en utilisant la variable

Allongement des temps de calcul,
Tendance à choisir des groupes de très grande taille.*

*Réf sur la base de l'étude de cas ci-après

1ère étape

discussion



L'objectif: choisir G_q^y

Plusieurs critères ont été considérés :

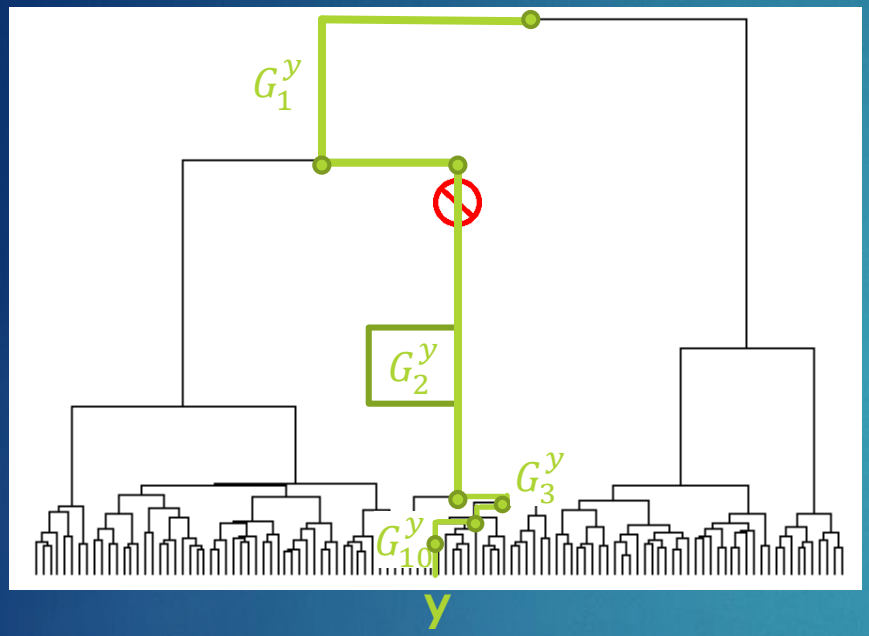
- Variation locale du critère CLV (Chen & Vigneau, 2016).

- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.

- Test de Bartlett de sphéricité sur les résidus des variables **X** de G_q^y non expliqués par la variable latente du groupe, c^q.

- Minimiser les erreurs de prédiction (LOO) en utilisant la variable latente c^q.

1ère étape



L'objectif: choisir G_q^y

- Règle de Kaiser-Guttman (KG) adaptée par Karlis et al, 2003.

- $\mathbf{X}_q$ la matrice de données restreinte aux variables **X** associées à **y** dans G_q^y .
- $\mathbf{X}_q$ est de taille (n, p_q)
- **R** la matrice de corrélation de $\mathbf{X}_q$.
- λ_1 et λ_2 les première et deuxième valeurs propres de **R**.
- On définit un seuil $l = 1 + 2\sqrt{\frac{p_{q-1}}{n-1}}$

Si $\lambda_1 > l$ et $\lambda_2 \leq l$ alors on considère que $\mathbf{X}_q$ est unidimensionnel.

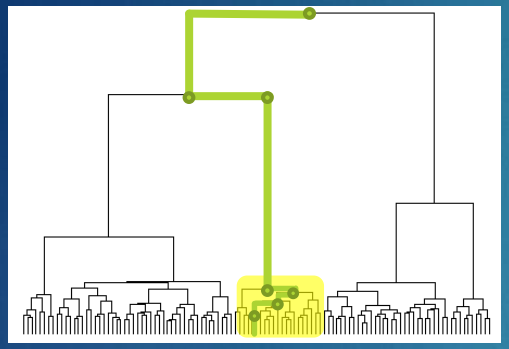
Karlis, D., Saporta, G. et Spinakis, A. (2003). A simple rule for the selection of Principal Components. *Communications in Statistics, Theory and methods*, 32, 643-666.

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

CLV sur $[\, X^{(0)} \mid y \,]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(0)T} X^{(0)}) / p^{(0)}$



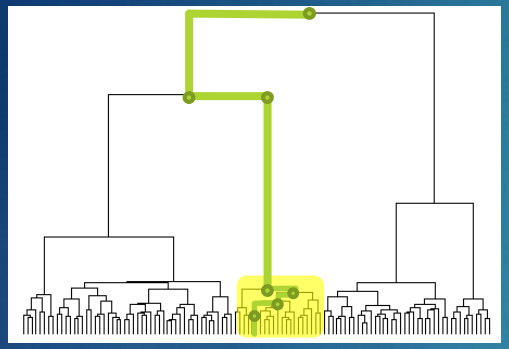
$LV^{(1)}$

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

CLV sur $[\, X^{(0)} \mid y \,]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(0)T} X^{(0)})/p^{(0)}$



$LV^{(1)}$

$\alpha_1^{(1)}$

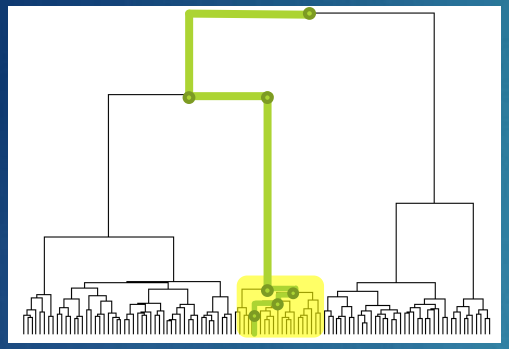
y

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

CLV sur [$X^{(0)}$ | y]

avec $trace(y^T y) = trace(X^{(0)T} X^{(0)})/p^{(0)}$

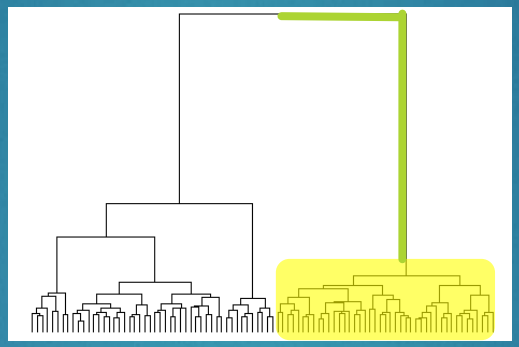


$LV^{(1)}$

$X^{(1)} \leftarrow X^{(0)} \setminus G^{y(1)}$

CLV sur [$X^{(1)}$ | y]

avec $trace(y^T y) = trace(X^{(1)T} X^{(1)})/p^{(1)}$



$LV^{(2)}$

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

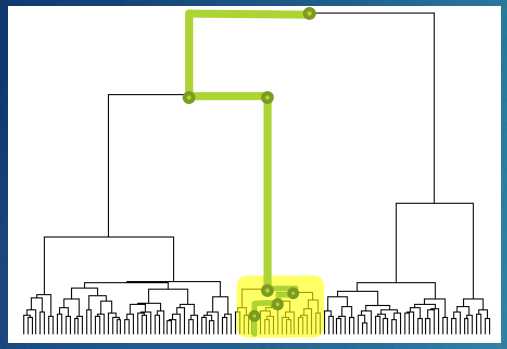
CLV sur $[\ X^{(0)}\ |\ y\ ]$

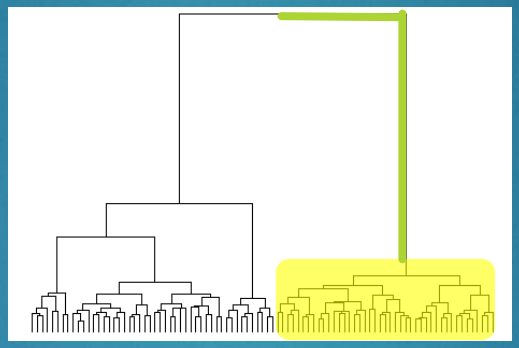
avec $trace(y^T y) = trace(X^{(0)T} X^{(0)})/p^{(0)}$

$X^{(1)} \leftarrow X^{(0)} \setminus G^{y(1)}$

CLV sur $[\ X^{(1)}\ |\ y\ ]$

avec $trace(y^T y) = trace(X^{(1)T} X^{(1)})/p^{(1)}$





$LV^{(1)}$ **$LV^{(2)}$**

$\alpha_1^{(2)}$ $\alpha_2^{(2)}$

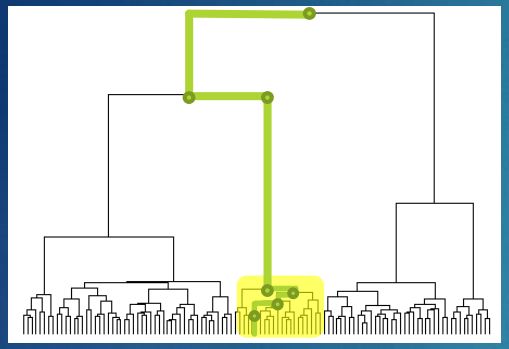
y

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

CLV sur $[\ X^{(0)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(0)T} X^{(0)})/p^{(0)}$

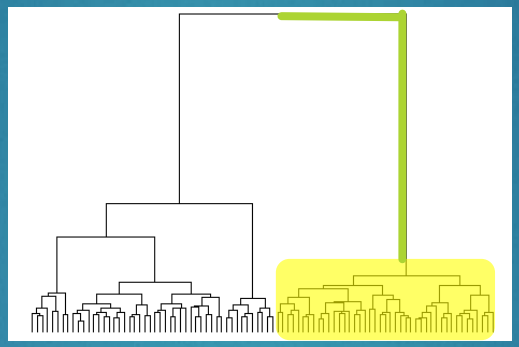


$LV^{(1)}$

$X^{(1)} \leftarrow X^{(0)} \setminus G^{y(1)}$

CLV sur $[\ X^{(1)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(1)T} X^{(1)})/p^{(1)}$

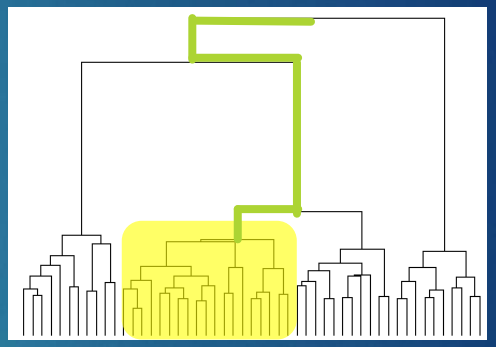


$LV^{(2)}$

$X^{(2)} \leftarrow X^{(1)} \setminus G^{y(2)}$

CLV sur $[\ X^{(2)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(2)T} X^{(2)})/p^{(2)}$



$LV^{(3)}$

Itérations : classification/ extraction

$X^{(0)} \leftarrow X$

CLV sur $[\ X^{(0)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(0)T} X^{(0)})/p^{(0)}$

$X^{(1)} \leftarrow X^{(0)} \setminus G^{y(1)}$

CLV sur $[\ X^{(1)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(1)T} X^{(1)})/p^{(1)}$

$X^{(2)} \leftarrow X^{(1)} \setminus G^{y(2)}$

CLV sur $[\ X^{(2)}\ |\ y\ ]$

avec $\text{trace}(y^T y) = \text{trace}(X^{(2)T} X^{(2)})/p^{(2)}$

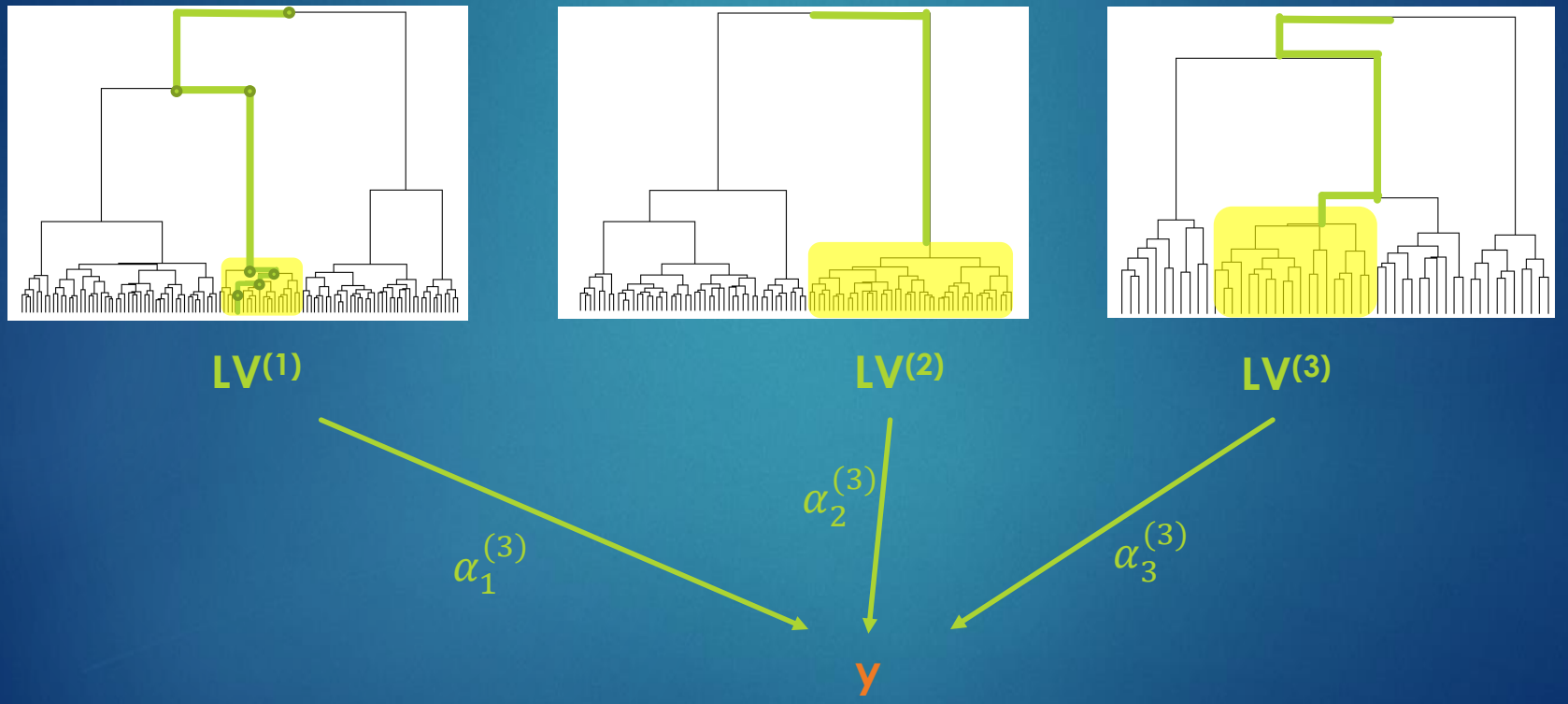


Illustration: identification du taux d'adultération de jus d'orange (*Citrus sinensis*) avec du jus de mandarine (*Citrus reticulata*).

Citrus sinensis

20 pur jus d'origines différentes

100	0
90	10
80	20
70	30
60	40
50	50
40	60
0	100

Citrus reticulata

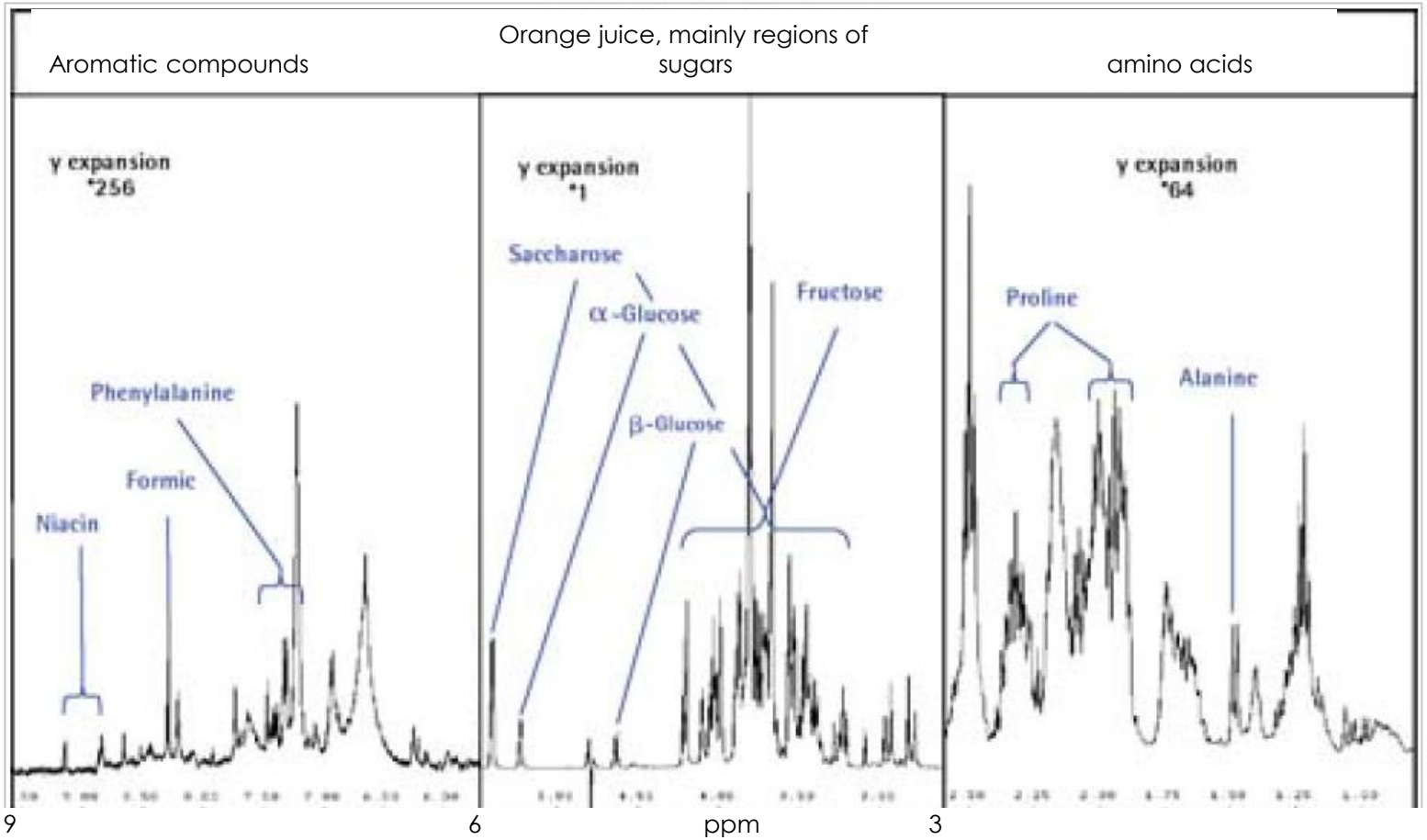
10 pur jus d'origines différentes

150 mélanges

Vigneau E., Thomas F. (2012). Model calibration and feature selection for orange juice authentication by 1H NMR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 117, 22–30.

Illustration : spectres RMN

Spectre RMN-^{1}H typique d'un jus d'orange concentré



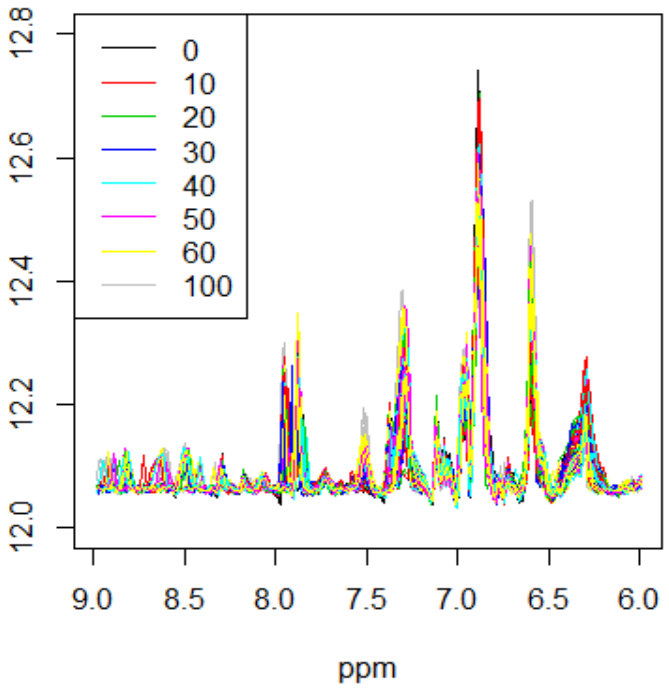
D'après Rinke et al., 2007, Quality control

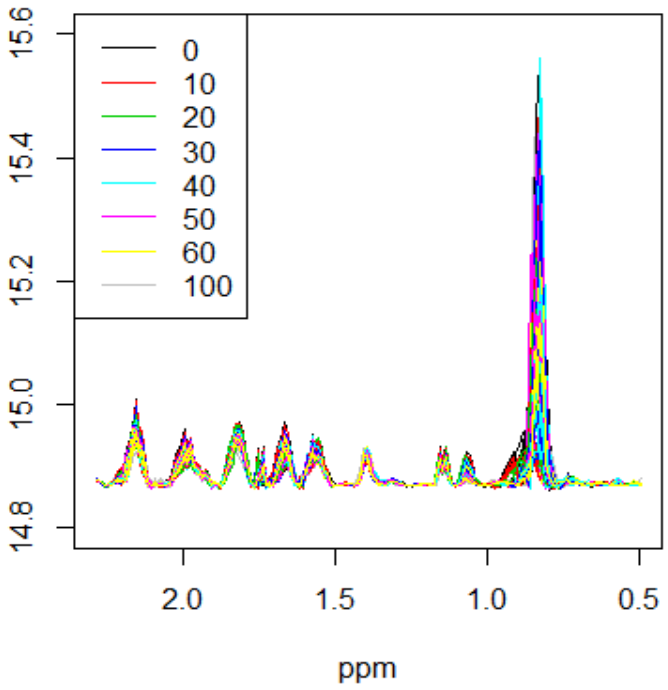
Illustration : spectres RMN

150 spectres RMN

colorés en fonction du pourcentage de jus de mandarine dans le mélange

z_1: zone des composés aromatiques z_2: zone des acides aminés





300 variables spectrales

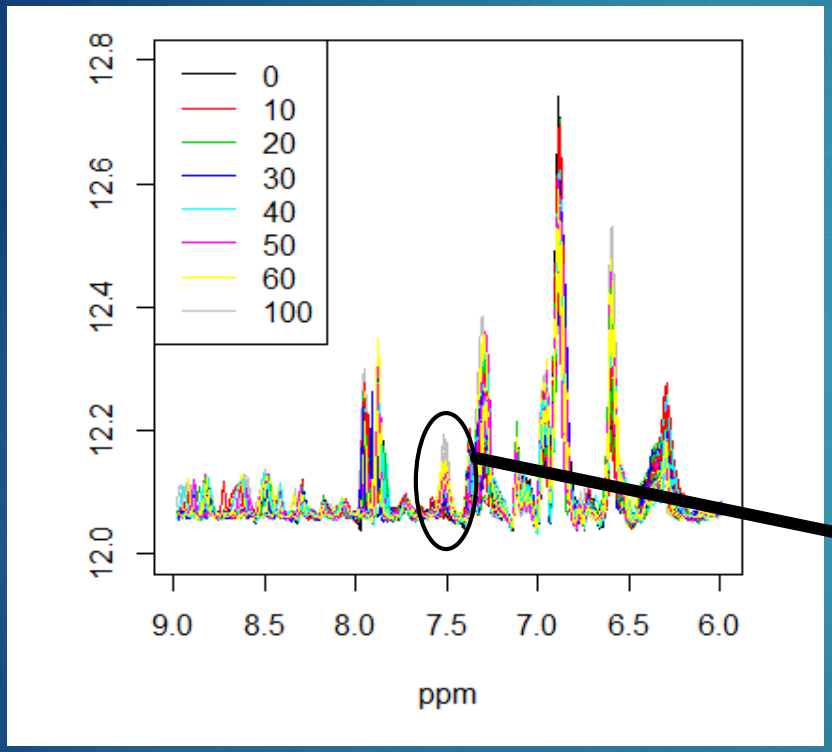
180 variables spectrales

Illustration : spectres RMN

150 spectres RMN

colorés en fonction du pourcentage de jus de mandarine dans le mélange

z_1: zone des composés aromatiques z_2: zone des acides aminés



300 variables spectrales

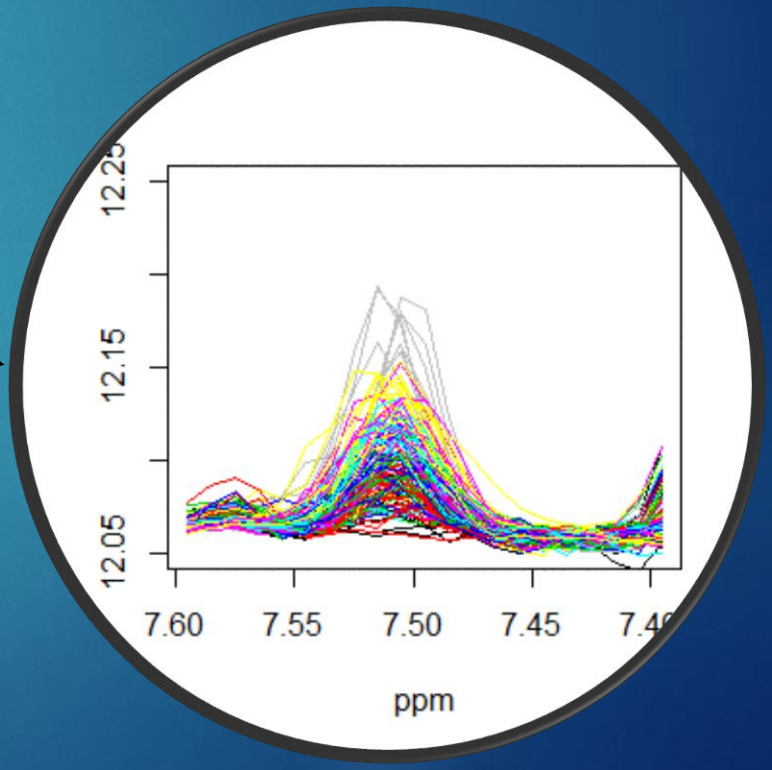


Illustration : prédiction

- **Prédiction** de **y** (150, 1) % jus mandarine dans le mélange
  en fonction de **X** (150, 480) données spectrales (standardisées)

- Evaluation de la qualité prédictive des modèles par **Cross-Validation** :
  10 segments, de 15 mélanges, respectant la structure du dispositif expérimental
  => racine carré de la moyenne des carrés des erreurs (**$RMSE_{CV}$**)

Illustration : prédiction

modèle des moindres carrés intégrant les A premières variables latentes (A=0,…,60) de **l'approche CLV supervisée.**

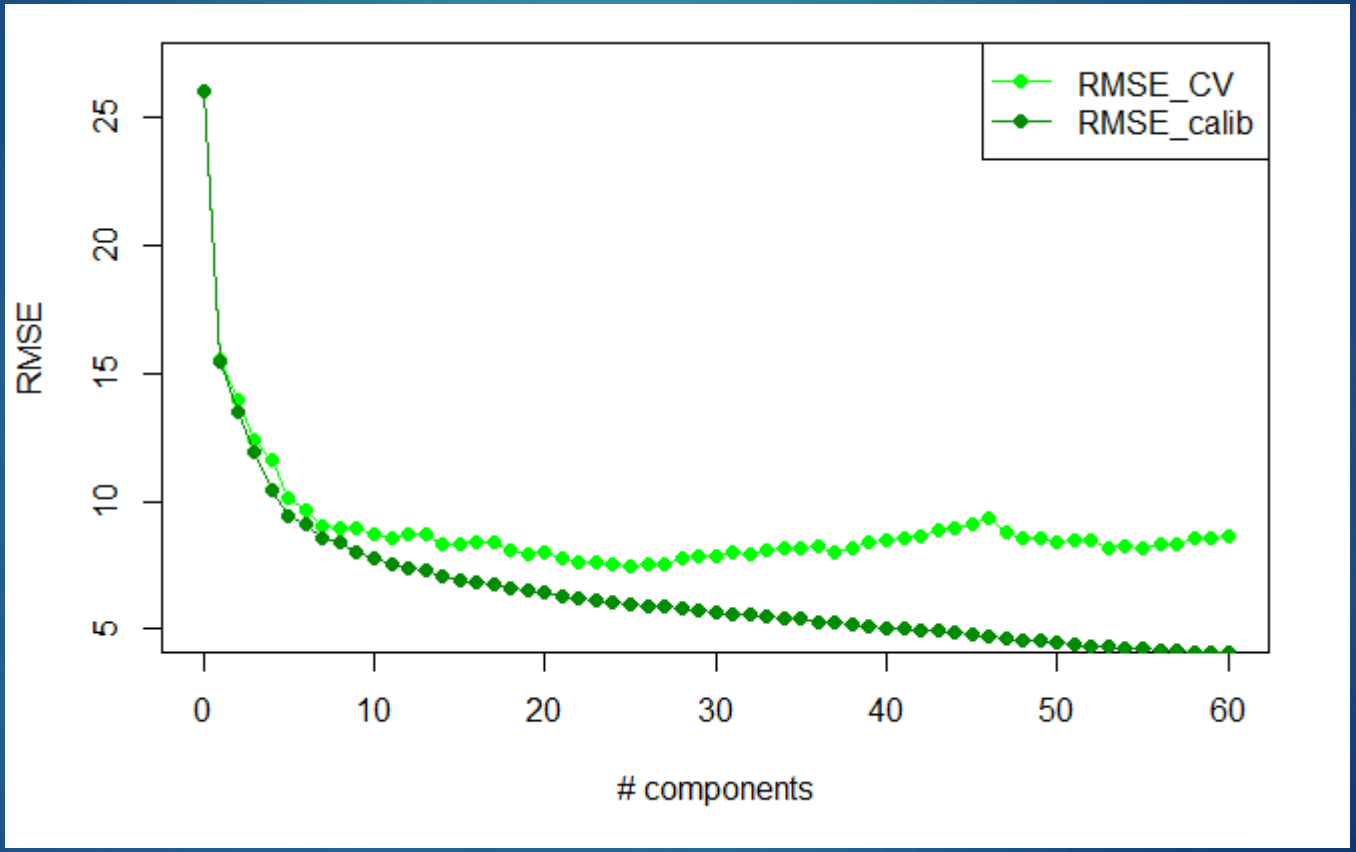


Illustration : prédiction

$RMSE_{CV}$ avec

approche CLV supervisée
régression PLS
régression PLS sparse (ξ=0.5)
+ elasticnet (α=0.5)

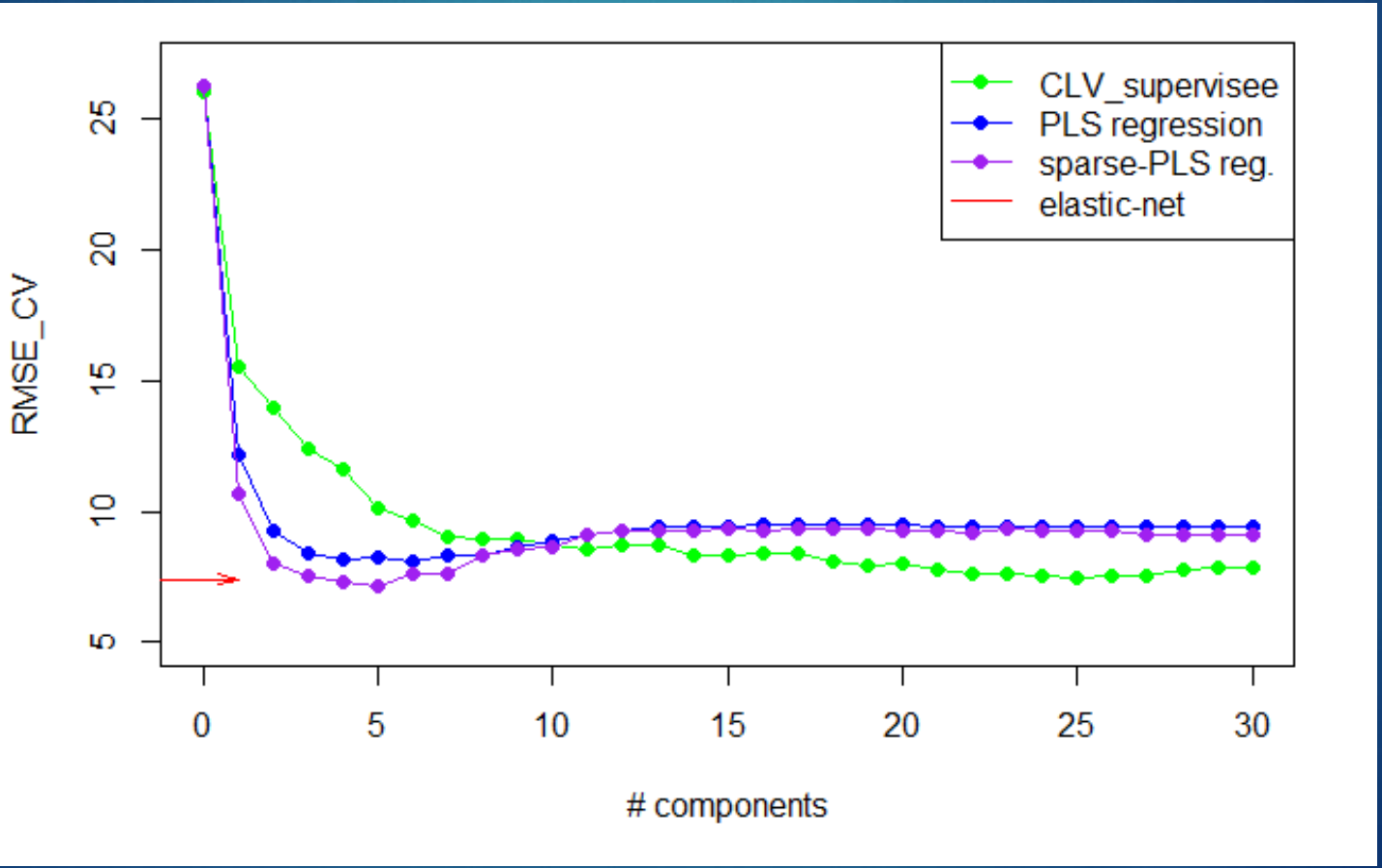
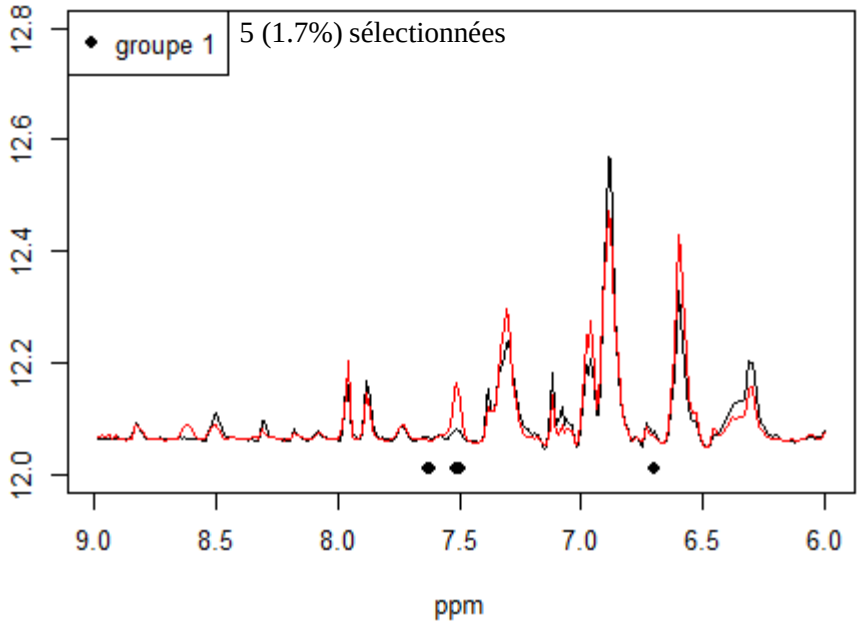
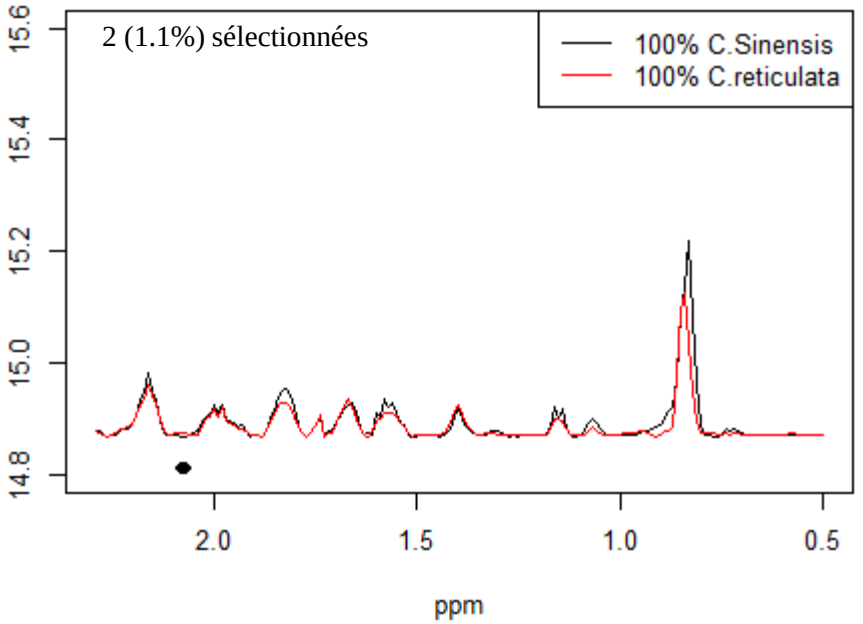


Illustration : interprétation

CLV supervisée. 1ere étape

Profils moyens et variables sélectionnées dans un groupe (jeu de données complet)





Fréquence de sélection sur l'ensemble des 10 boucles de CV

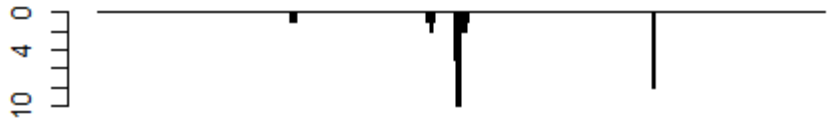
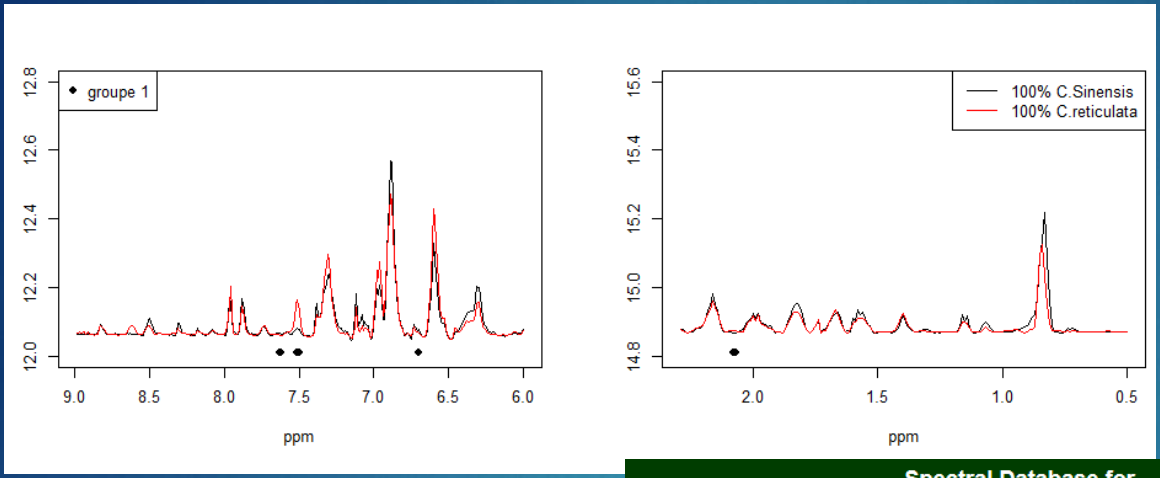


Illustration : interprétation



Tentative d'identification

Spectral Database for Organic Compounds SDBS

Japanese | Introduction | Disclaimer | HELP | Contact | What's New | RI

SDBS Information

SDBS No.: 12185

Compound Name:
4-amino-3-methylbenzoic acid

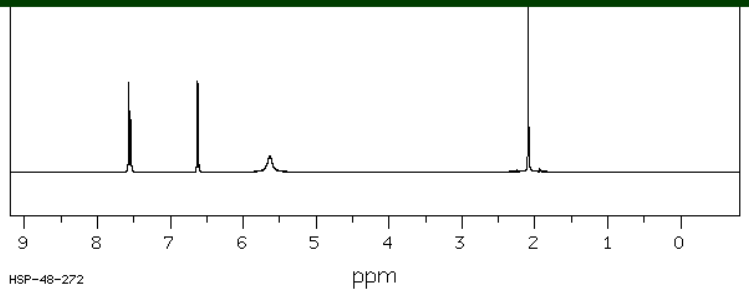
Molecular Formula: $C_8H_9NO_2$

Molecular Weight: 151.2

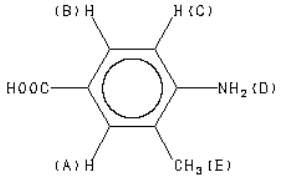
CAS Registry No.:
2486-70-6

Derivatives:
display in a separate page
SDBS Structures Web (on trial since 2013-05-14)

Spectral Code:
^{13}C NMR : in DMSO-d_6
^{1}H NMR : 400 MHz in DMSO-d_6
IR : nujol mull
IR : KBr disc



HSP-48-272

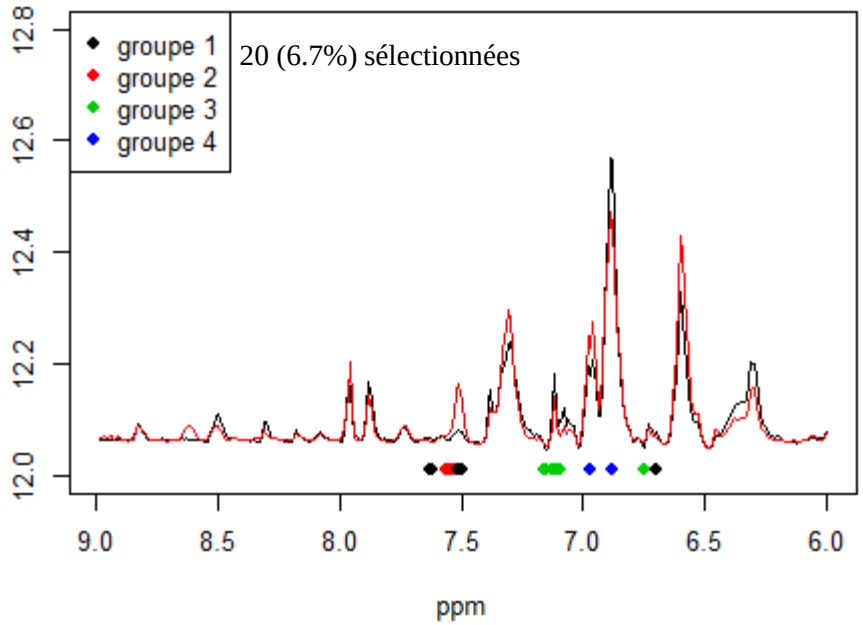


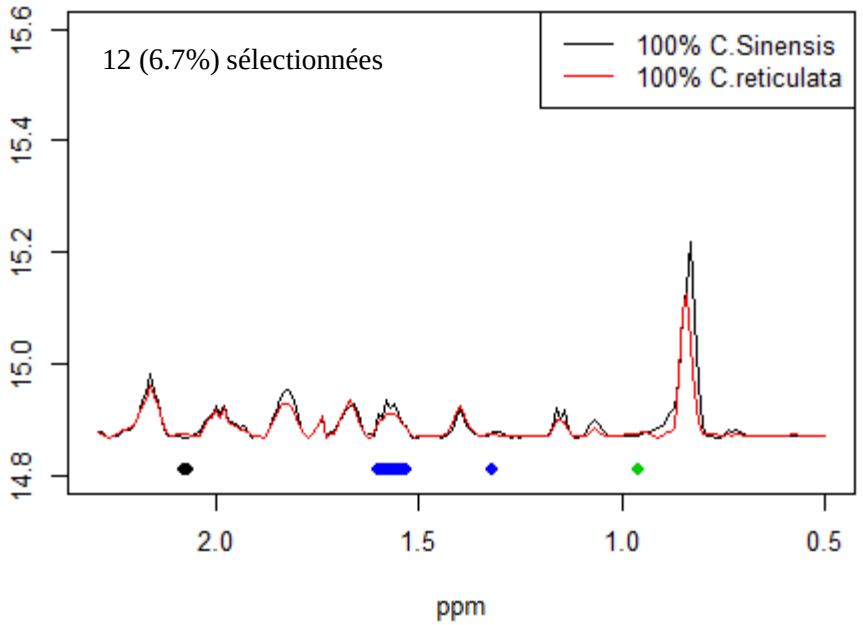
Assign.	Shift (ppm)
A	7.571
B	7.548
C	6.624
D	5.64
E	2.089

Illustration : interprétation

CLV supervisée. Étapes 1à 4

Profils moyens et variables sélectionnées dans un groupe (jeu de données complet)

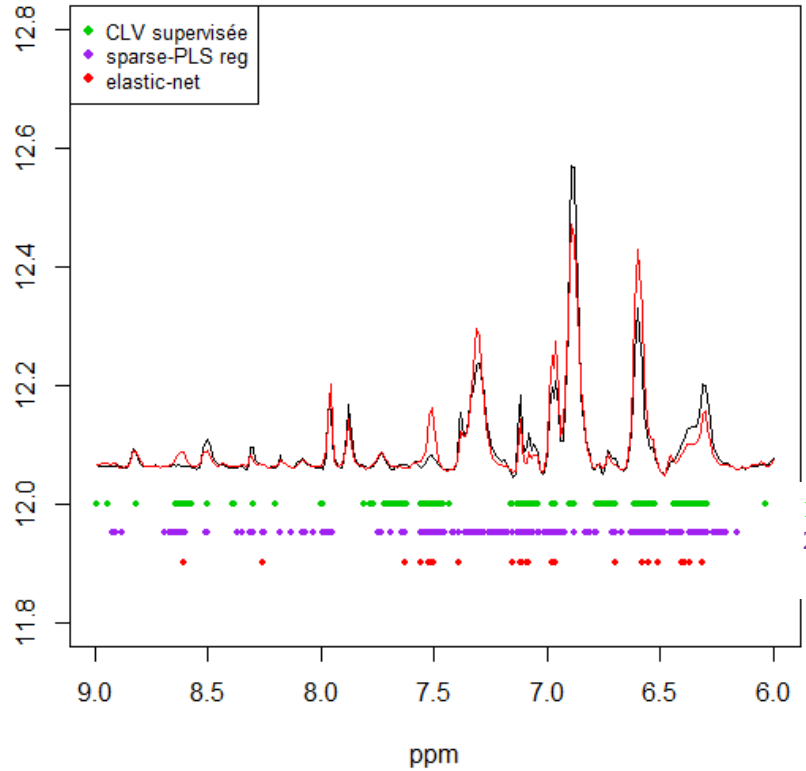




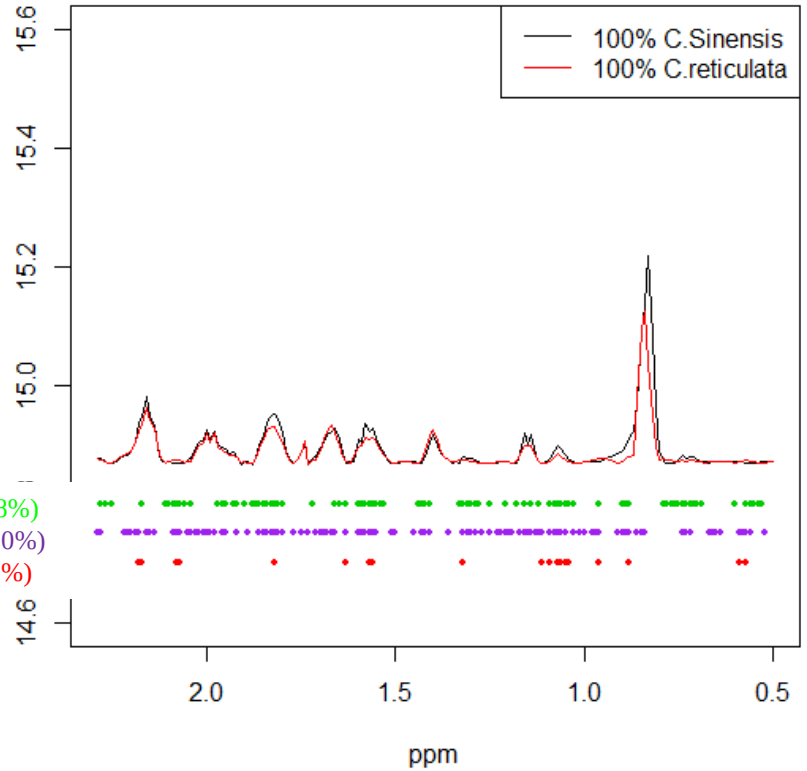
Fréquence de sélection sur l'ensemble des 10 boucles de CV

Illustration : interprétation

Variables retenues avec approche CLV supervisée (25 var. latentes)
régression PLS sparse (ξ=0.5, 4 dimensions)
elasticnet (α=0.5)



180 var (38%)
239 var. (50%)
43 var. (9%)



Pour conclure

Méthode de classification de variables autour de variables latentes, **CLV**,

- mise à profit dans une optique prédictive.
- Sur la base de l'étude de cas, présente une qualité prédictive proche d'autres approches.
- Un assez grand nombre de variables latentes, par forcément orthogonales, peuvent être intégrées dans le modèle.
- Chacune de ces variables latentes est censée représenter un groupe de variables prédictrices, fortement redondantes,
- pour une interprétabilité améliorée.

Merci de votre attention