Classification de variables avec possibilité de mettre à l'écart des variables atypiques ou de bruit.

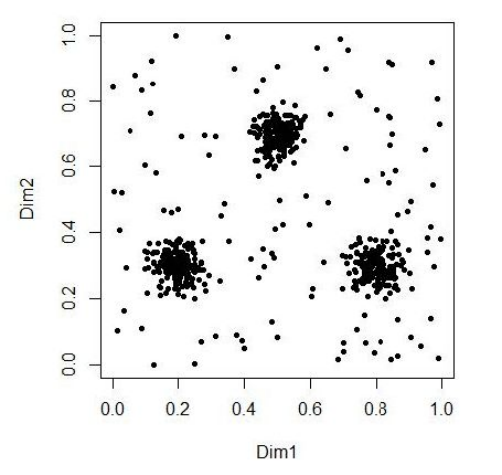
Implémentation dans le package ClustVarLV

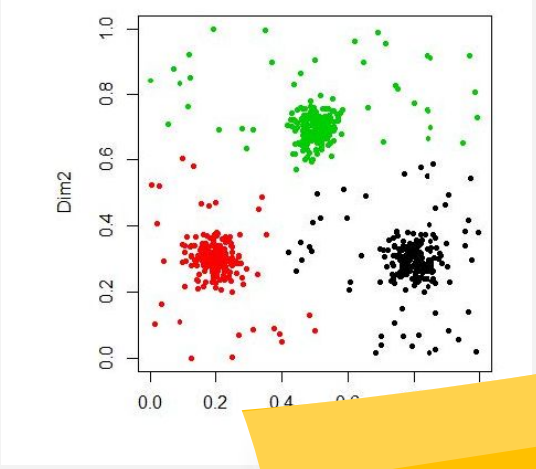
Evelyne Vigneau, El Mostafa Qannari

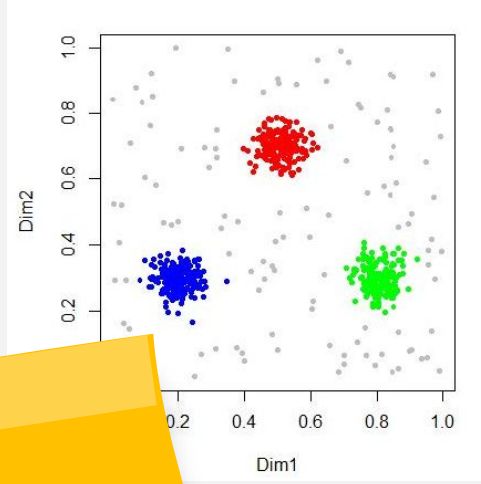
Plan

- Problématique
- La méthode CLV et modifications proposées pour prendre en compte les variables atypiques
- Implémentation dans le package ClustVarLV
- Exemple 1 : *Segmentation d'un panel de consommateurs.*
- Exemple 2 : *Analyse métabolomique pour la détection de traitements analobisants chez les bovins*
- Conclusion

Classification et observations atypiques





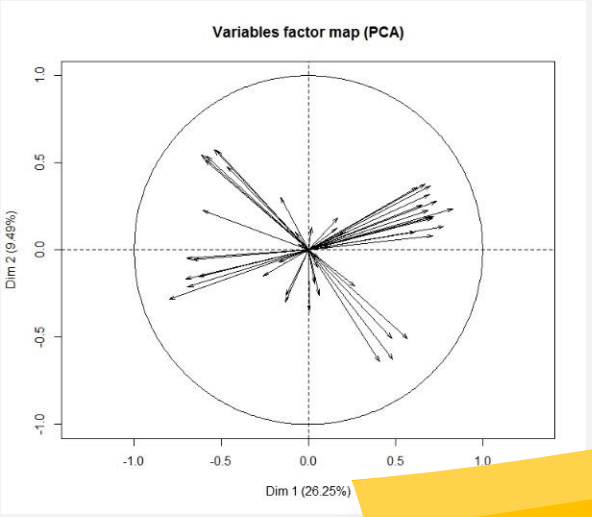


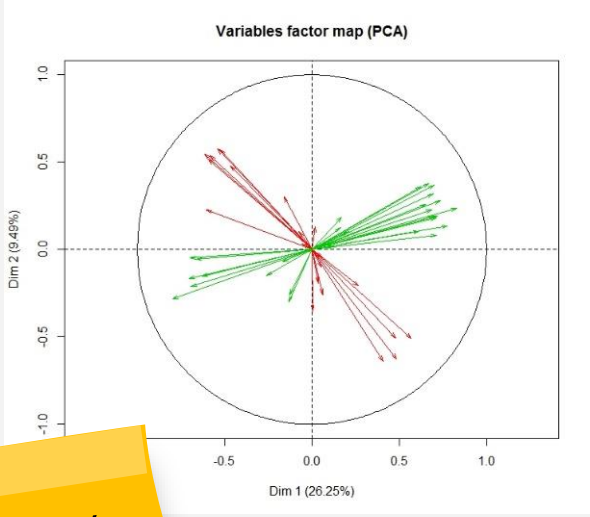
Classification

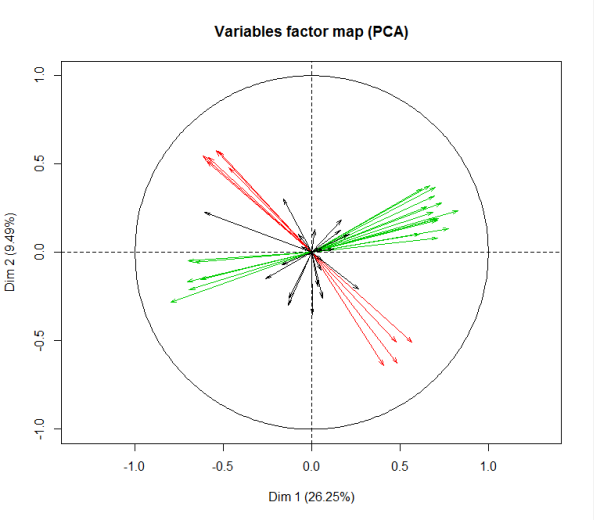
- *floue* [1]
- *robuste* [2]
- *classe supplémentaire = "noise cluster"* [3]

[1] Berget, I., Mevik, B. H., & Næs, T. (2008). New modifications and applications of fuzzy C-means methodology. *Computational Statistics & Data Analysis*, **52**, 2403–2418.
[2] García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advanced Data Analysis and Classification*, **4**, 89–109.
[3] Davé, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, **12**, 657–664.

Classification de variables.
Variables atypiques ou de bruit







Méthode CLV
"Classification de variables autour de variables latentes" [4]

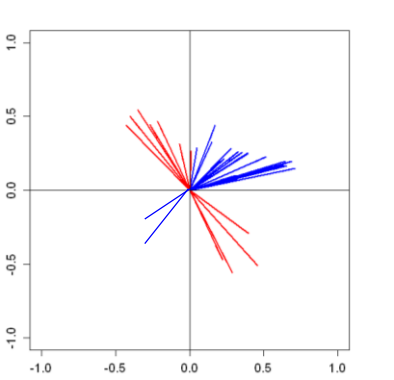
?

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

La méthode CLV : type de groupes

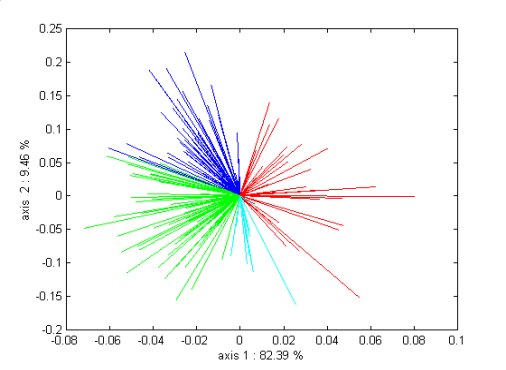
groupes directionnels

Fortes corrélations positives ou négatives ⇨ accord



groupes locaux

Fortes corrélations positives ⇨ accord
Fortes corrélations négatives ⇨ désaccord



maximisation de :

$$\mathrm{T} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}^2(\boldsymbol{x}_j, \boldsymbol{c}_k)$$

$$\mathrm{S} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}(\boldsymbol{x}_j, \boldsymbol{c}_k)$$

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

La méthode CLV : algorithme

Optimisation alternée

0 Initialisation : par coupure du dendrogramme produit par une approche hiérarchique **ou** aléatoire avec répétitions

1 Etape d'estimation (e.g. groupes directionnels)

$\boldsymbol{c_k}$ $(k = 1, \dots, K)$; première composante principale normée de $\mathbf{X_k}$ (matrice formée par les variables appartenant au groupe G_k).

2 Etape d'affectation (e.g. groupes directionnels)

Chaque variable $\mathbf{x_j}$ $(j = 1, \dots, p)$ est affectée dans le groupe G_{k*} pour lequel on a $cor^2(\mathbf{x_j}, \mathbf{c_{k*}})$ maximale.

CLV et stratégies pour prendre en compte les variables atypiques

Stratégie « K+1 »

mettre de côté, dans un groupe additionnel ("noise cluster"), les variables qui sont trop peu corrélées avec les variables latentes des groupes en formation [3].

Stratégie « sparse LV »

définir des variables latentes "sparse" en adoptant une démarche similaire à celle de l'ACP Sparse [5] / approche de classification "robuste" .

[3] Davé, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, **12**, 657–664.
[5] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265–286.

CLV et stratégies pour prendre en compte les variables atypiques

Stratégie « K+1 »

mettre de côté, dans un groupe additionnel ("noise cluster"), les variables qui sont trop peu corrélées avec les variables latentes des groupes en formation [3].

e.g. groupes directionnels, maximisation de :

$$\mathrm{T}_{\mathrm{new}} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}^2(\boldsymbol{x}_j, \boldsymbol{c}_k) \quad + \quad \sum_{j=1}^{p} (1-\delta_{kj}) \rho^2 \operatorname{var}(\boldsymbol{x}_j)$$

paramètre de seuillage, compris entre 0 et 1

stratégie " K+1 "

Modification du critère => Modification de l'algorithme

⓪ **Initialisation :** par coupure du dendrogramme produit par une approche hiérarchique **ou** aléatoire avec répétitions

❶ **Etape d'estimation** (e.g. groupes directionnels)

$\boldsymbol{c_k}$ $(k = 1, \dots, K)$; première composante principale normée de $\mathbf{X_k}$ (matrice formée par les variables appartenant au groupe G_k).

❷ **Etape d'affectation** (e.g. groupes directionnels)

Si $cor^2(\mathbf{x_j}, \mathbf{c_k}) < \rho^2,\ \forall\, k$, alors $\delta_{kj} = 0, \forall\, k \Longrightarrow \mathbf{x_j}$ est affectée au "groupe" supplémentaire

Sinon la variable $\mathbf{x_j}$ $(j = 1, \dots, p)$ est affectée dans le groupe G_{k*} pour lequel on a $cor^2(\mathbf{x_j}, \mathbf{c_{k*}})$ maximale.

Stratégie "Sparse LV"

Stratégie « K+1 »

mettre de côté, dans un groupe additionnel ("noise cluster"), les variables qui sont trop peu corrélées avec les variables latentes des groupes en formation [3].

Stratégie « sparse LV »

définir des variables latentes "sparse" en adoptant une démarche similaire à celle de l'ACP Sparse [5].

e.g. groupes directionnels, procédure itérative de définition des loadings avec une étape de seuillage "doux", en fonction du paramètre ρ

Stratégie "Sparse LV"

Optimisation alternée

⓪ Initialisation : par coupure du dendrogramme produit par une approche hiérarchique **ou** aléatoire avec répétitions

❶ Etape d'estimation (e.g. groupes directionnels)

$\boldsymbol{c_k}$ $(k = 1, \dots, K)$ **:** première composante principale sparse, normée, de $\mathbf{X_k}$

❷ Etape d'affectation (e.g. groupes directionnels)

Chaque variable $\mathbf{x_j}$ $(j = 1, \dots, p)$ est affectée dans le groupe G_{k*} pour lequel on a $cov^2(\mathbf{x_j}, \mathbf{c_{k*}})$ maximale.

Implémentation dans le package ClustVarLV

Algorithme de partitionnement (*k-means*).

Approche "classique"

```
> CLV_kmeans(X, method="directional", clust= K)
```

ou "local"
choix du type de groupes

matrice de données
(n x p)
p : nombre de
variables à grouper

- si clust est un entier, K : nb de groupes de la partition avec initialisation au hasard, nstart=100 fois (par défaut)

- si clust est un vecteur de p entiers$\in\{1,...,K\}$: partition initiale

Implémentation dans le package ClustVarLV

Algorithme de partitionnement (*k-means*)

Approche "K+1"

```
> CLV_kmeans(x, method="directional",clust=K, strategy="kplusone",rho=0.3)
```

rho=0.3 par défaut

Approche "sparse LV"

```
> CLV_kmeans(x, method="directional",clust=K, strategy="sparselv",rho=0.3)
```

Appréciation de différentes variétés de carottes. Segmentation d'un panel de consommateurs [6]

- 10 lots de carottes de variétés/régions de production différentes.
- 299 consommateurs (Paris, Angers, Avignon).
- Réponse : score d'appréciation globale (de 0 à 10) pour carottes râpées.

Profil moyen **sur l'ensemble du panel** *apprécié* *peu apprécié*

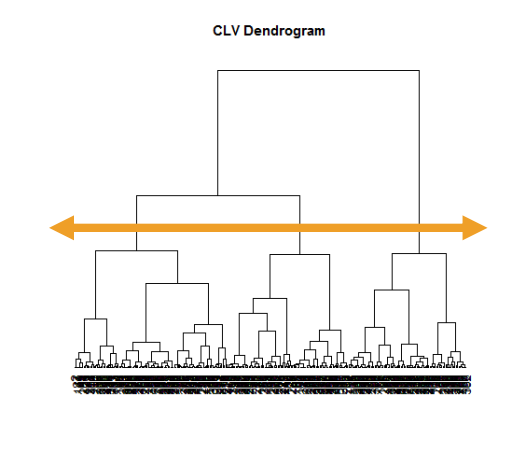
AMST	ARTE M	BOLER	BONFI	DELUC	DORD1	DORD2	MAES1	MAES2	PHOEN	*mean*	*range*
5.1	6.2	6.2	5.1	6.3	5.9	6.3	4.9	6.7	7.0	6.0	2.1
peu sucré, peu juteuse			peu sucré, amer				pâle	sucré et parfumé			

[6] Vigneau E., Qannari E.M., Navez B., Cottet V. (2015). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. *Food Quality and Preference* DOI: http://dx.doi.org/10.1016/j.foodqual.2015.02.008.

Appréciation de différentes variétés de carottes. Segmentation d'un panel de consommateurs

- CLV "classique"

```
> resclv <- CLV (Xhedo, method="local", sX=FALSE)
```



- K=3 segments considérés

```
> partini=resclv[[3]]$clusters[1,]
```

Appréciation de différentes variétés de carottes. Segmentation d'un panel de consommateurs

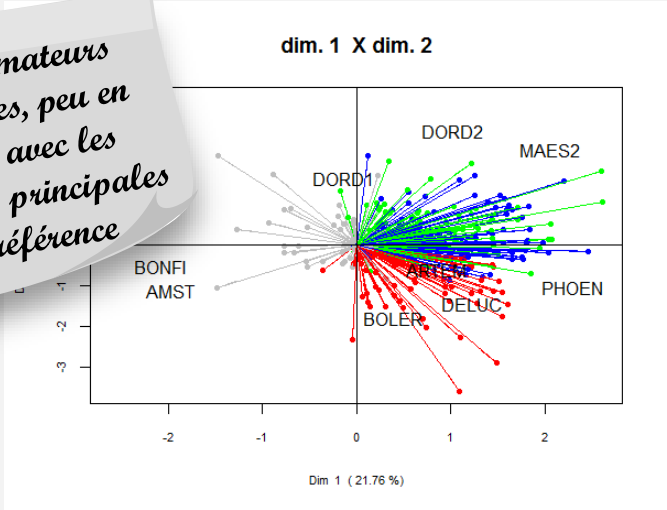
- CLV avec la stratégie “K+1”
- K=3 segments et ρ=0.3 (fixé)

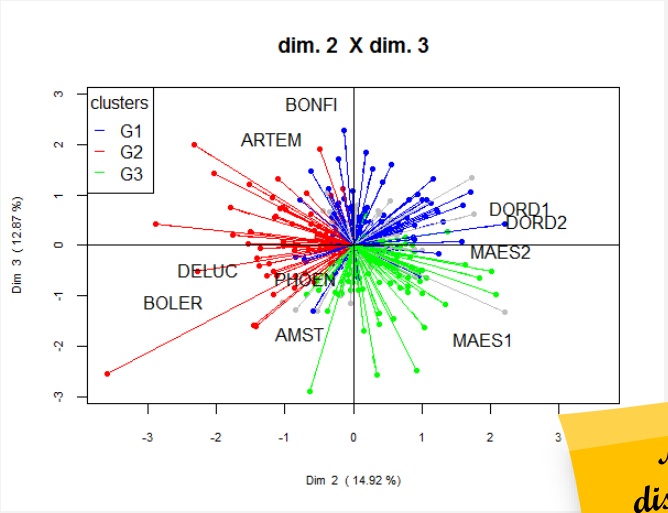
```
> resclvkp1 <- CLV_kmeans(Xhedo, method="local", sX=FALSE,
                          clust=partini,
                          strategy="kplusone", rho=0.3)
```

Appréciation de différentes variétés de carottes. Segmentation d'un panel de consommateurs

- CLV avec la stratégie “K+1”
- K=3 segments et ρ=0.3(fixé)

Consommateurs atypiques, peu en accord avec les directions principales de préférence





Meilleure discrimination des produits au sein des 3 segments

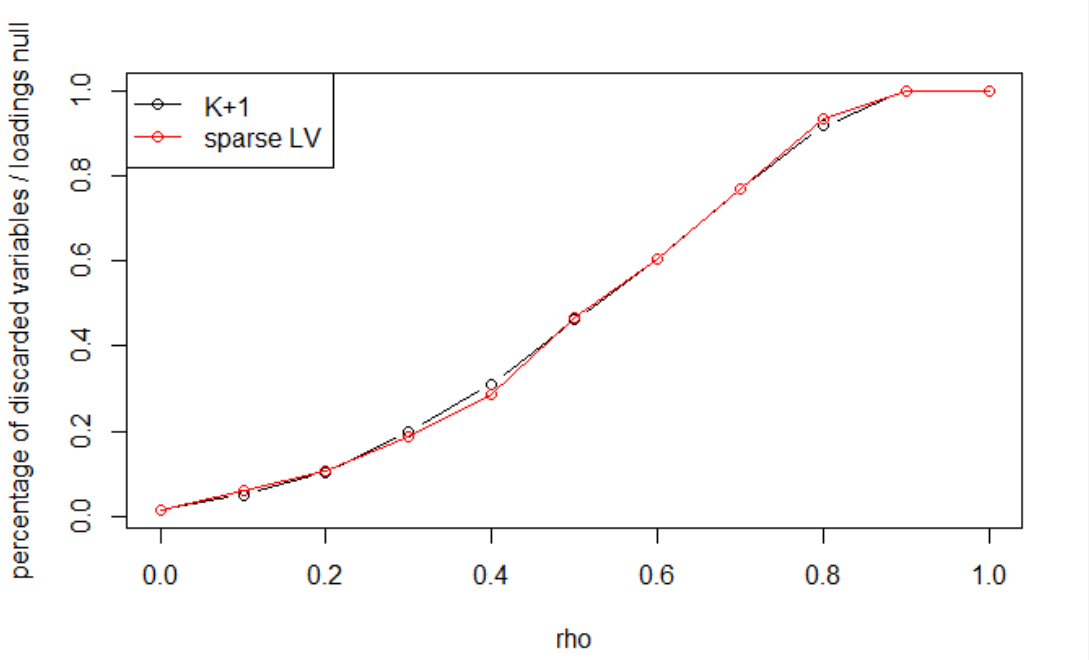
Seg	size	AMST	ARTEM	BOLER	BONFI	DELUC	DORD1	DORD2	MAES1	MAES2	PHOEN	[illegible]	[illegible]
G1'	97	3.5	7.0	5.5	5.4	6.4	5.8	7.0	4.9	6.9	7.7	6.0	4.2
G2'	62	6.0	7.1	7.5	5.5	7.4	5.0	4.5	3.5	5.8	7.0	5.9	3.9
G3'	81	5.6	5.0	6.8	3.5	5.8	6.3	7.1	5.0	7.8	6.6	6.0	4.3
Gpe suppl.	59	6.2	5.7	5.4	6.3	5.4	6.3	5.9	6.2	6.1	6.2	6.0	0.9

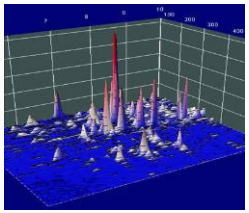
Appréciation de différentes variétés de carottes. Segmentation d'un panel de consommateurs

en fonction du seuil de corrélation ρ

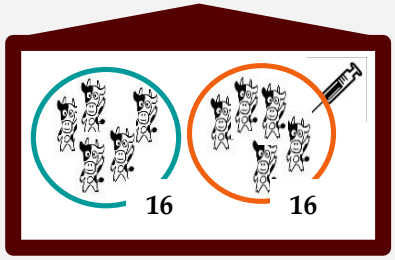
Stratégie “K+1”: Proportion de variables mis de côté

Stratégie "Sparse LV" : proportion de variables ayant un loading nul





Empreintes métabolomiques / détection de traitements analobisants chez les bovins [7]



Contrôle (C) Traitement (T)

J42

LABERCA

Chromatographie en phase liquide couplée avec une spectrométrie de masse haute réslution (RPLC-HRMS)

4515 métabolites

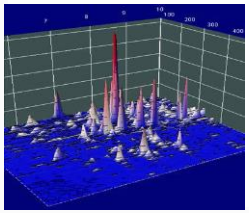
Contrôles

Traités

32 échantillons

Prétraitement : transformation log Pareto scaling => X

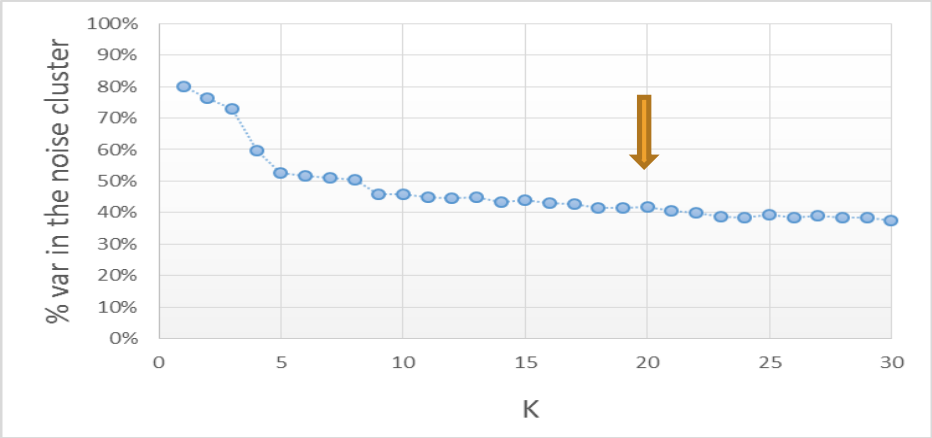
[7] Jacob, C. C., Dervilly-Pinel, G., Biancotto, G., Monteau F. & Le Bizec, B. Global urine fingerprinting by LC-ESI(+)-HRMS for better characterization of metabolic pathway disruption upon anabolic practices in bovine. *Metabolomics*, DOI 10.1007/s11306-014-0685-3, 2014

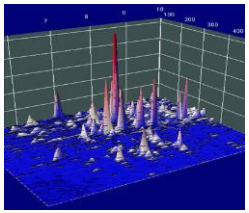


Empreintes métabolomiques / détection de traitements analobisants chez les bovins

```
> resK<- CLV_kmeans(X, method="directional", sX=FALSE,
                        clust=K,
                        strategy="kplusone", rho=0.5)
```

- Pour un nombre de groupes variant de K=1 à K=30
  % de métabolites placés dans le "noise cluster"

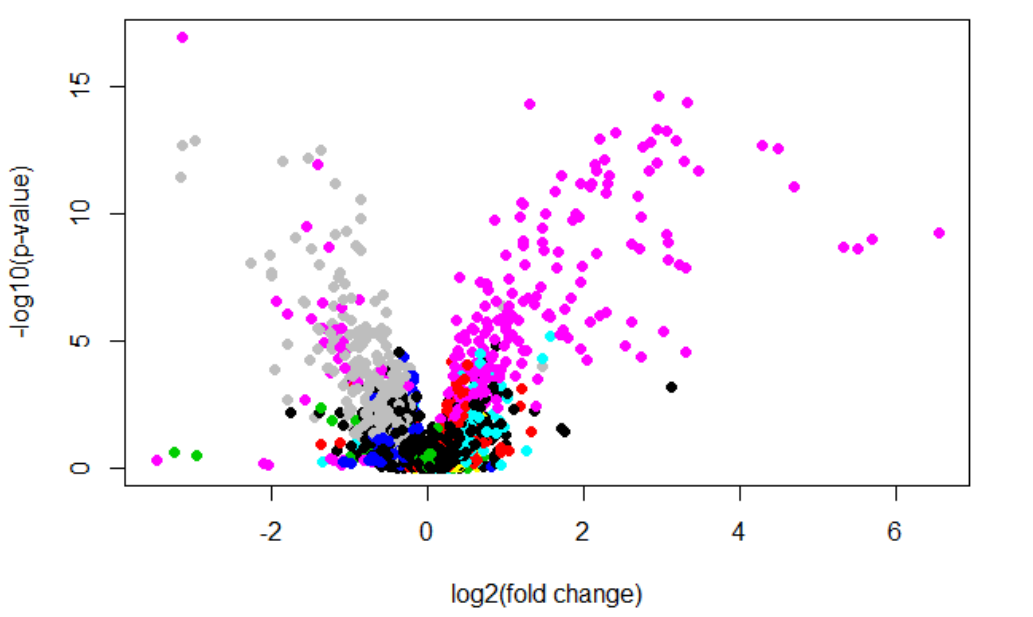




Empreintes métabolomiques / détection de traitements analobisants chez les bovins

"Volcano plot" et identification des groupes de métabolites (K=20)

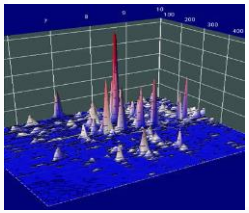
p-value de l'effet "Traitement", pour chaque métabolite



Groupe 13
(225 métabolites)
Groupe 15
(269 métabolites)

« noise cluster »
(1886 métabolites)

Ratio de changement : moy. obs. « T » / moy. obs. « C », pour chaque métabolite



Empreintes métabolomiques / détection de traitements analobisants chez les bovins

	nb métabolites
1. X_1=X, matrice complète	4575
2. X_2 = X \ « noise cluster » (CLV, *K*=20)	2629
3. X_3 = G13 et G15 uniquement (CLV, *K*=20)	494

Qualité de prédiction
du statut "traité"/"non traité" des observations

		X_1	X_2	X_3
	nb métabolites	4575	2629	494
Régression PLS-DA RMSEP (LOO)	1 comp.	0.190	0.149	0.131
	2 comp.	0.136	0.137	0.110

Conclusion

Classification de variables CLV

stratégie "K+1"	stratégie "sparse LV"
• Mettre de côté les variables "atypiques", • Définir des groupes de variables plus "compacts".	Définir des variables latentes de groupes "robustes".
Filtrage des variables et classification simultanées	Mise en évidence de structures simples.

Merci
pour
votre
attention