SUJET :

DÉVELOPPEMENT DE MÉTHODES CHIMIOMÉTRIQUES POUR LE TRAITEMENT DE DONNÉES MASSIVES

Doctorant : Maxime Metz (INRAE, DigitAg)
Directeurs de thèse : Matthieu Lesnoff (CIRAD), Jean-Michel Roger (INRAE)
Encadrants : Florent Masseglia (INRIA), Reza Akbarinia (INRIA), Florent Abdelghafour (INRAE)

Plan de la présentation

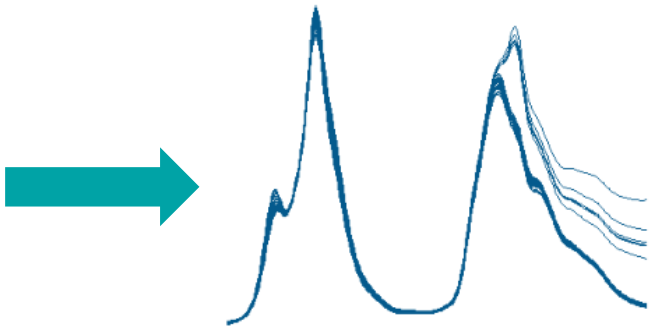
- Introduction
- Section 1 : Comment mettre à profit des paradigmes du “big-data” pour améliorer les modèles PLS locaux actuels ?
- Section 2 : Comment estimer la pertinence d’un individu par rapport à un modèle PLS ?
- Section 3 : Comment associer les paradigmes de la chimiométrie et du big-data ?
- Conclusion générale, perspectives et questionnements

Introduction

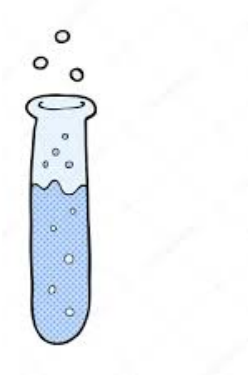
L’agriculture numérique

L'agriculture numérique se définit comme la convergence de **l'agriculture** et des technologies de l'**information** (capteurs, réseaux intelligents, **outils de la science de la données**) pour **améliorer** la productivité et répondre aux **attentes environnementales** et **sociétales**.

L’agriculture numérique et la chimiométrie



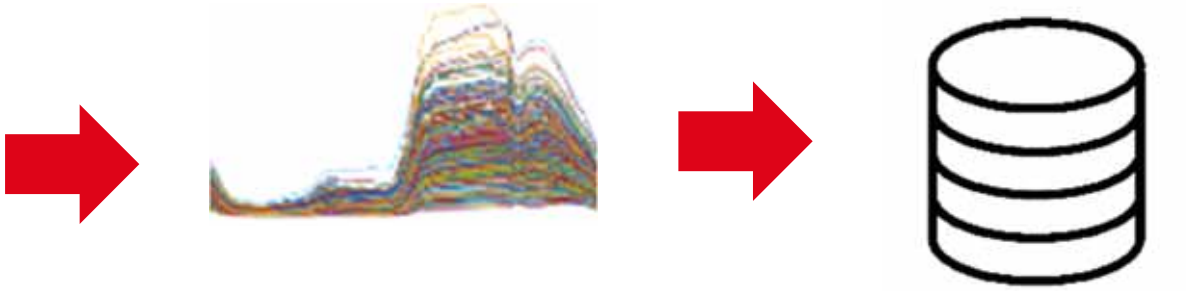
Chimiométrie



[Geraudie, 2009]

[Chawade, 2019]

Les données massives et la chimiométrie



Les données massives et la chimiométrie

La quantité de données… pose problème

De simples opérations deviennent problématiques

- Prétraitements
- Méthodes linéaires

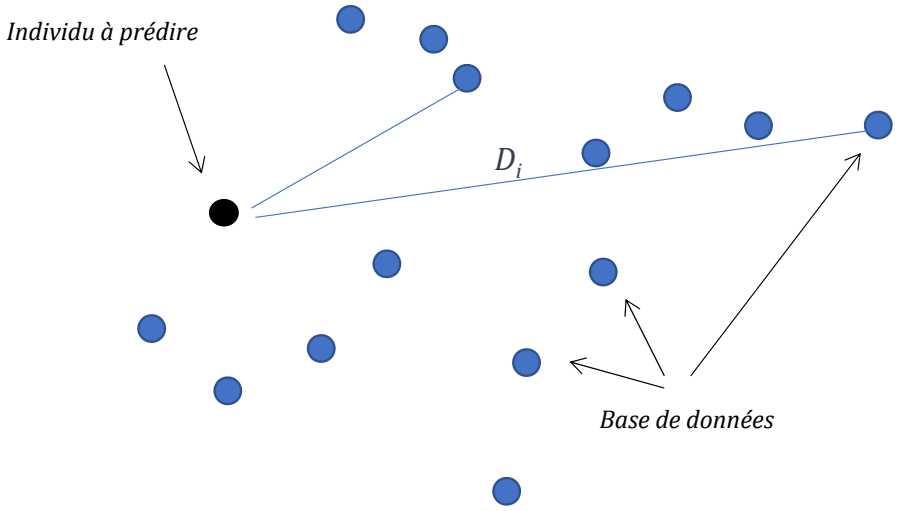
- Temps de calcul
- Non-linéarité

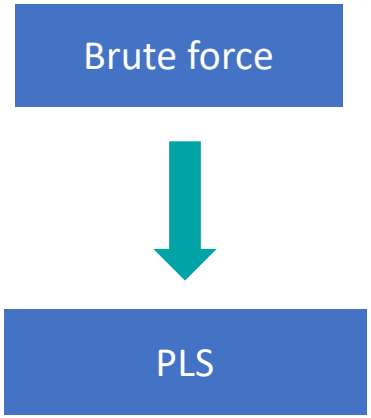
[Szymańska, 2018]
[Dardenne et al., 2000]
[Bertran et al., 1999]

Les outils de la chimiométrie doivent désormais répondre à ces problématiques

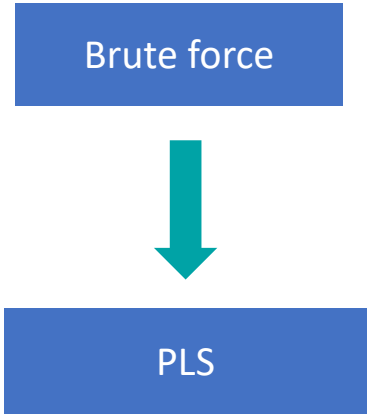
Les méthodes PLS locales et les données massives

Un ensemble d'outils de la chimiométrie : Les méthodes PLS locales





Les méthodes PLS locales et les données massives



Distances / similarités :

- Malahanobis
- Euclidiennes
- Avec réduction de dimension
- ……

[Zhang, 2020]
[Shen, 2019]
[Hazama, 2015]

[Fearn, 2003]

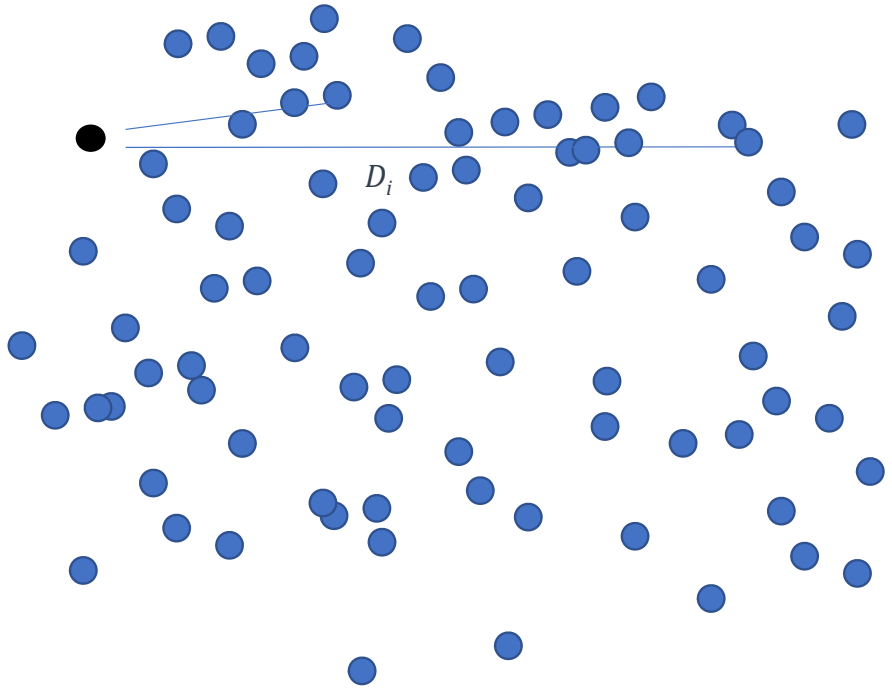
Les **distances** développées en **chimiométrie** permettent d'obtenir des modèles avec de meilleures capacités **prédictives**

Les méthodes PLS locales et les données massives

- Le volume des données
- La pertinence des voisins
- Les connaissances de la chimiométrie pour les données massives

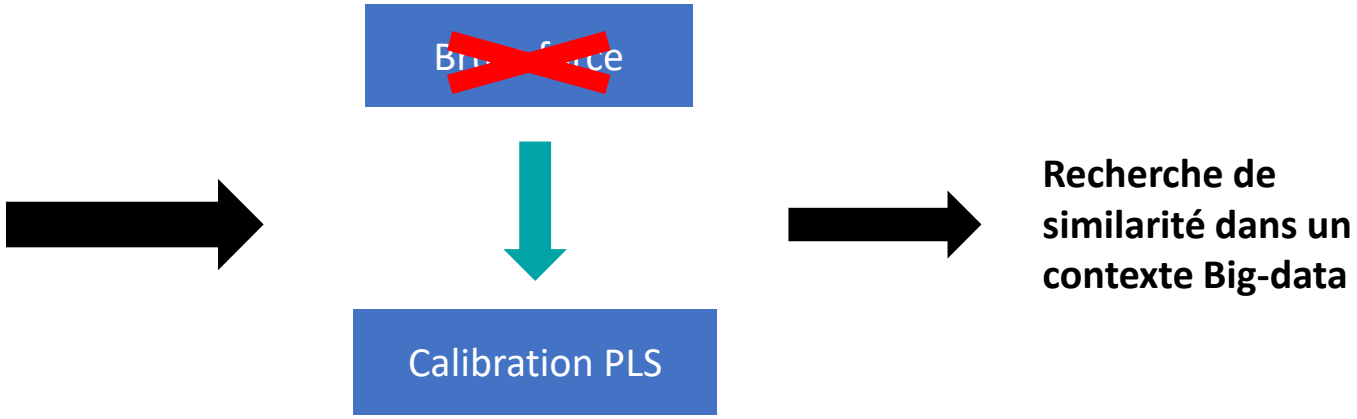
Les méthodes PLS locales et les données massives

Le volume des données



Les méthodes PLS locales et les données massives

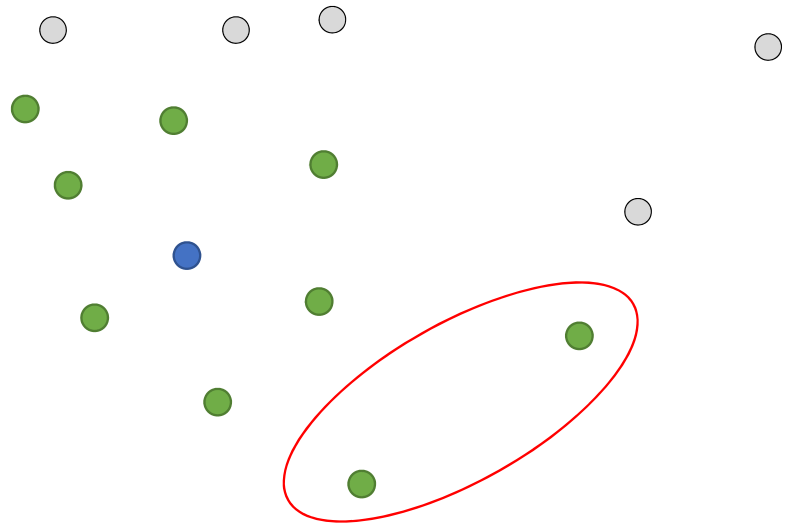
Le volume des données



Comment mettre à profit des paradigmes du “big-data” pour améliorer les modèles PLS locaux actuels ?

Les méthodes PLS locales et les données massives

La pertinence des voisins

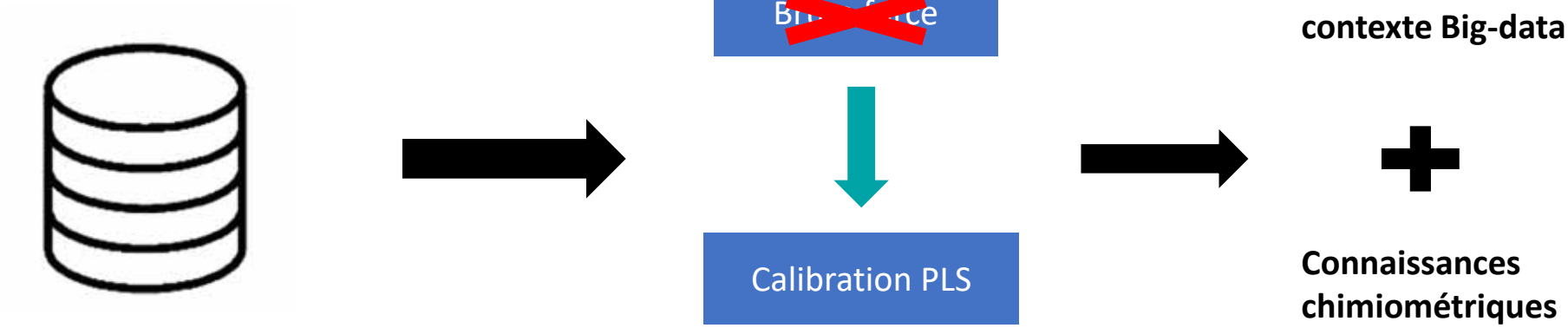


Les **voisins** les plus **proches** ne sont **pas** forcément les plus **pertinents**

Comment estimer la pertinence d'un individu par rapport à un modèle PLS ?

Les méthodes PLS locales et les données massives

Les connaissances de la chimiométrie pour les données massives



Comment associer les paradigmes de la chimiométrie et du “big-data” ?

Questions scientifiques

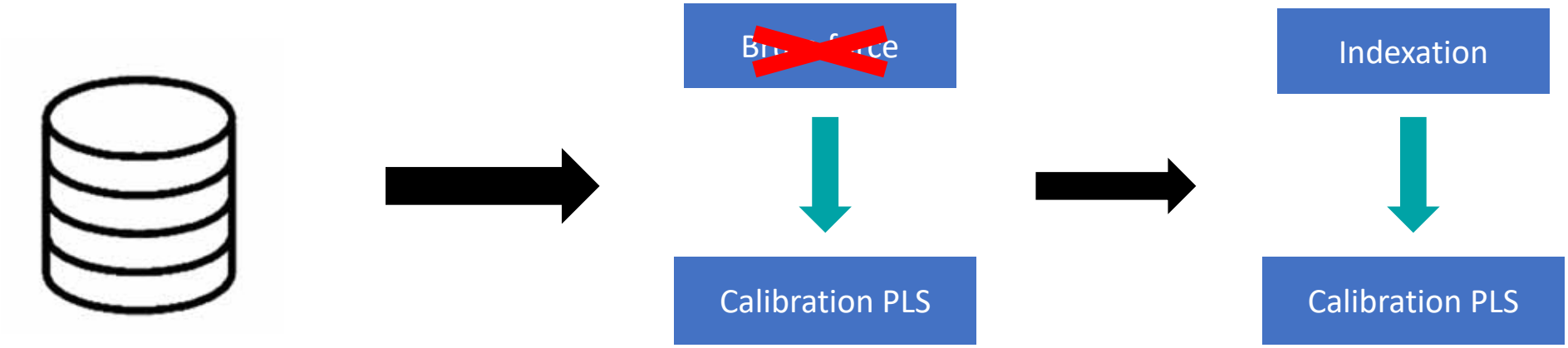
Comment mettre à profit des paradigmes du **“big-data”** pour améliorer les modèles PLS locaux actuels ?

Comment estimer la **pertinence** d’un individu par rapport à un modèle PLS ?

Comment associer les paradigmes de la **chimiométrie** et du **“big-data”** ?

Section 1 : Comment mettre à profit des paradigmes du “big-data” pour améliorer les modèles PLS locaux actuels ?

1.1. L’indexation



1.1. L’indexation

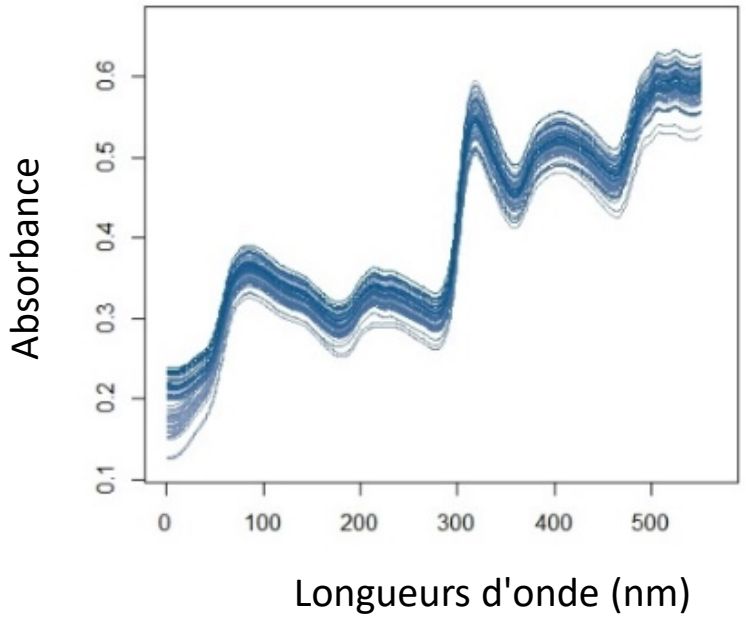
L’indexation a pour objectif d’organiser un ensemble d’individus pour faciliter ultérieurement la recherche de voisins

Bibliothèque Mazarine

1.2. L’indexation des données massives spectroscopiques

Quelles méthodes d’indexation utilisées pour les données massives spectroscopiques ?

- Beaucoup d’individus
- **Beaucoup de variables**



1.3. La méthode parSketch

Méthode de recherche de voisins massivement parallélisable développée pour les données de grande dimension.

[Levchenko, 2018]

Les étapes de parSketch :

1.3. La méthode parSketch

Réduction de dimension

V : nb de vecteurs aléatoires

ACP → **parSketch** → $T_* = X P_*'$

Scores

Base de données

Loadings

Sketches

Base de données

Vecteurs aléatoires (-1,1)

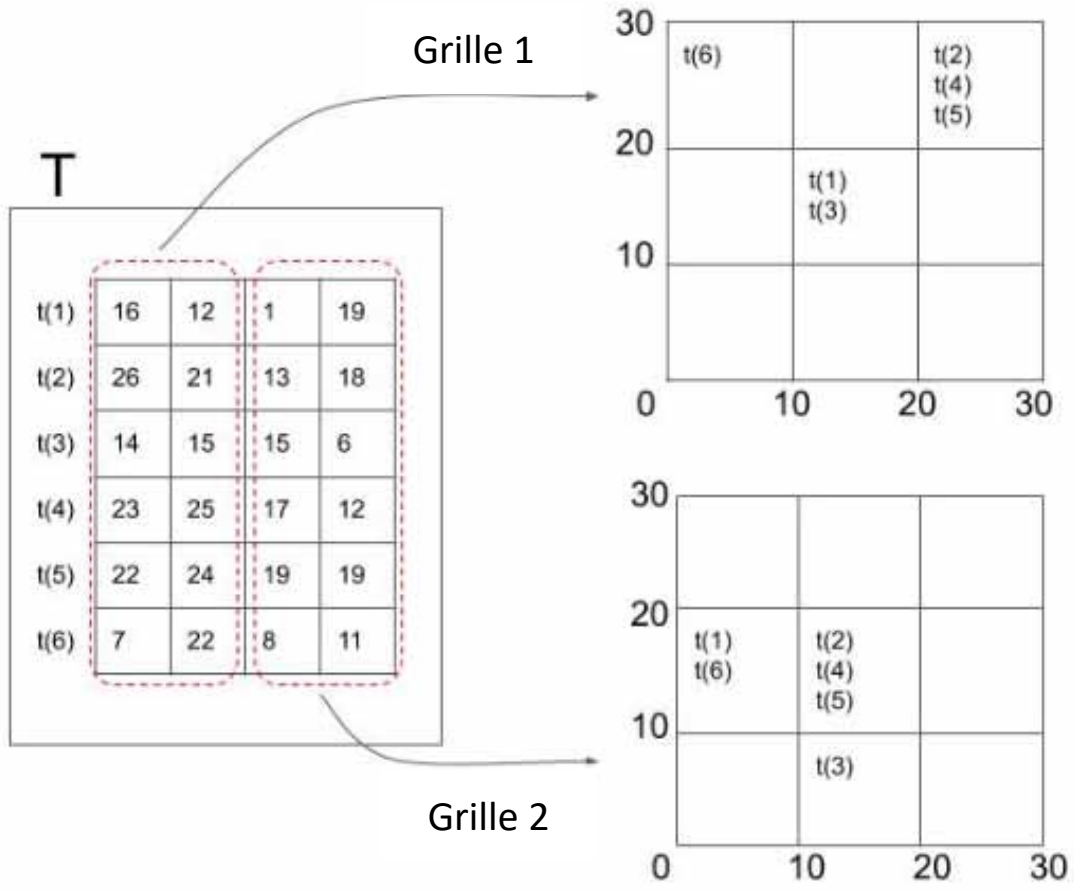
Pourquoi cette technique de réduction de dimension ?

- Approche les distances euclidiennes *[Johnson, 1984]*
- Opération rapide
- Pas d'*a priori*

↳ **Création de grilles**

1.3. La méthode parSketch

1.3. La méthode parSketch

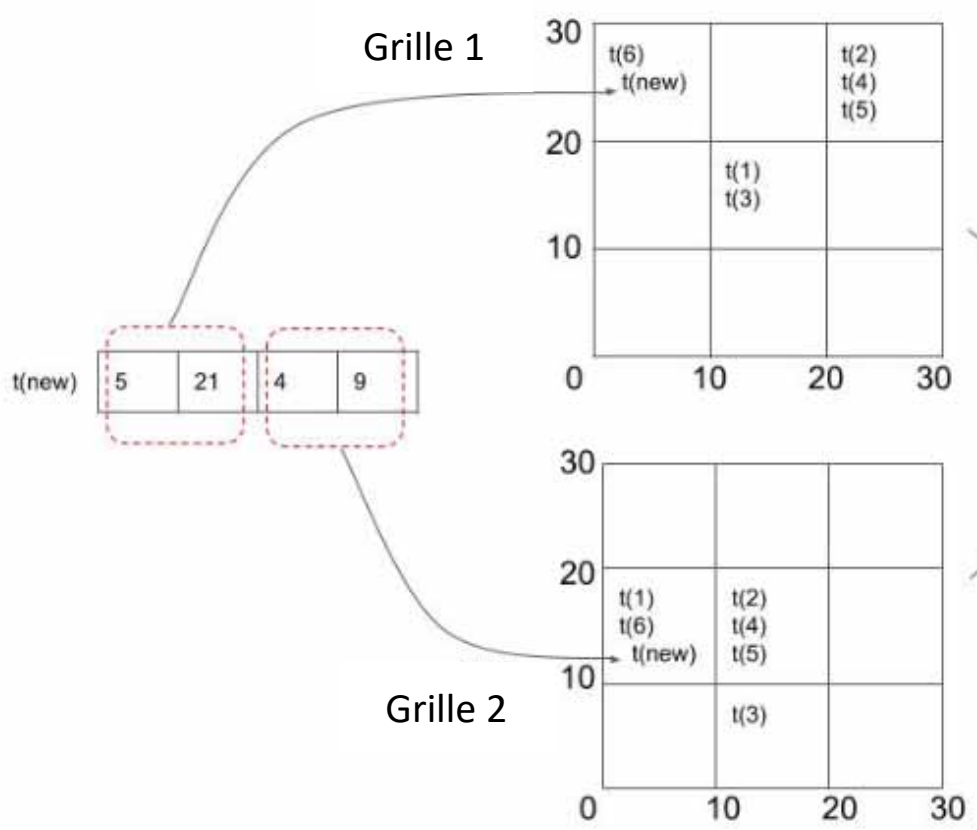


S : nb segments

1.3. La méthode parSketch

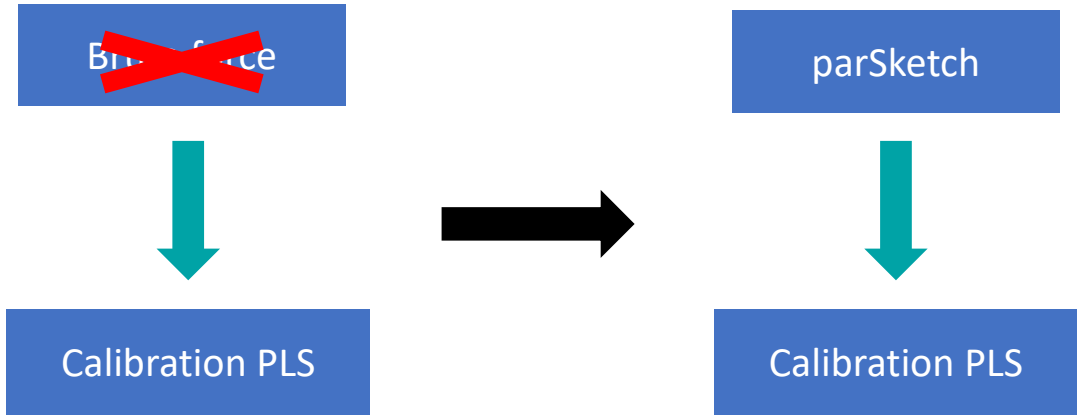
1.3. La méthode parSketch

x_{new} P_* ⟶



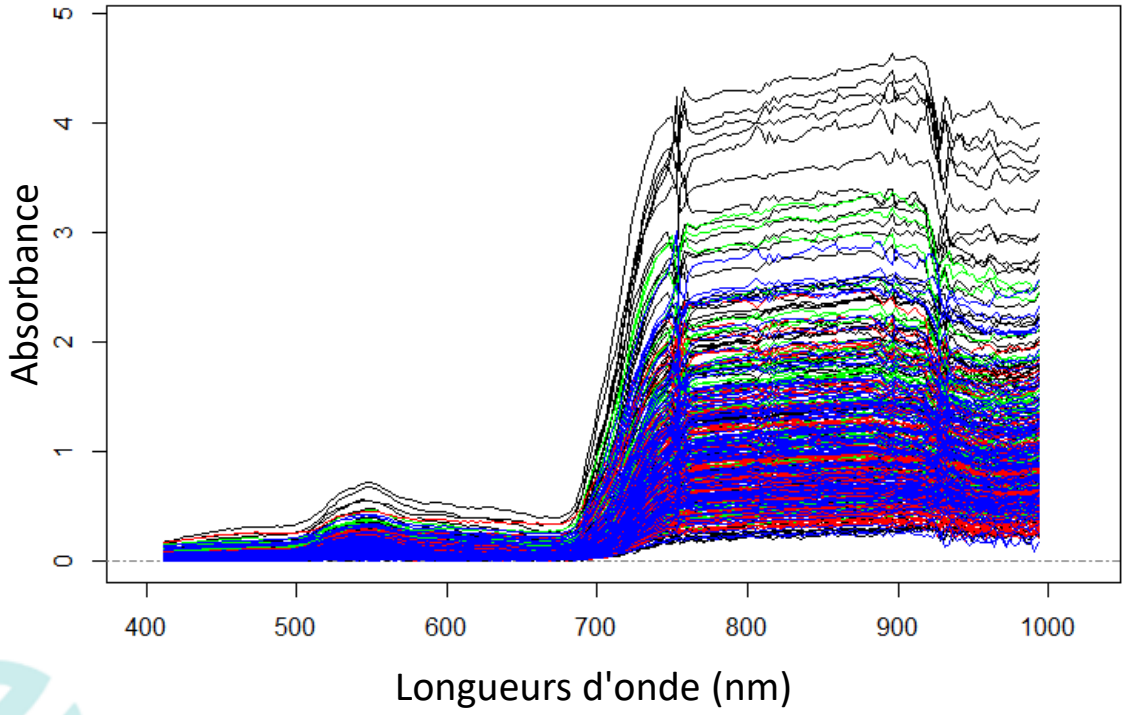
m : % min de cellules en commun

1.4. Une application de parSketch-PLS



1.4. Une application de parSketch-PLS

Problème de classification sur un jeu de données à l'interface des données massives/standards



- Spectres de feuilles
- 4 génotypes (4 images)*
- 360 000 spectres
- 256 variables

* Merci **Martin Ecarnot** ! (AGAP)

1.4. Une application de parSketch-PLS

Les 3 stratégies utilisées :

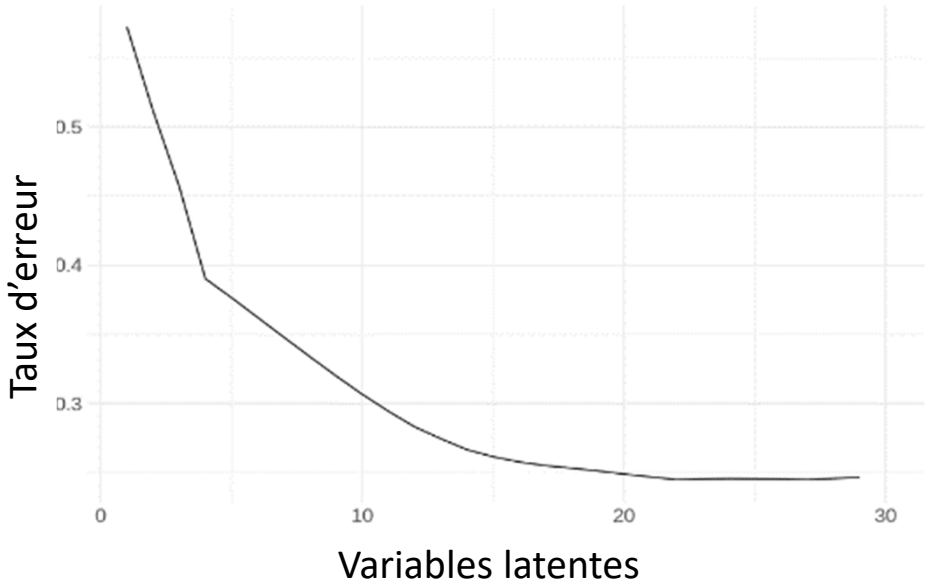
PLS-DA globale

BF-PLS-DA

parSketch – PLS-DA

1.4. Une application de parSketch-PLS

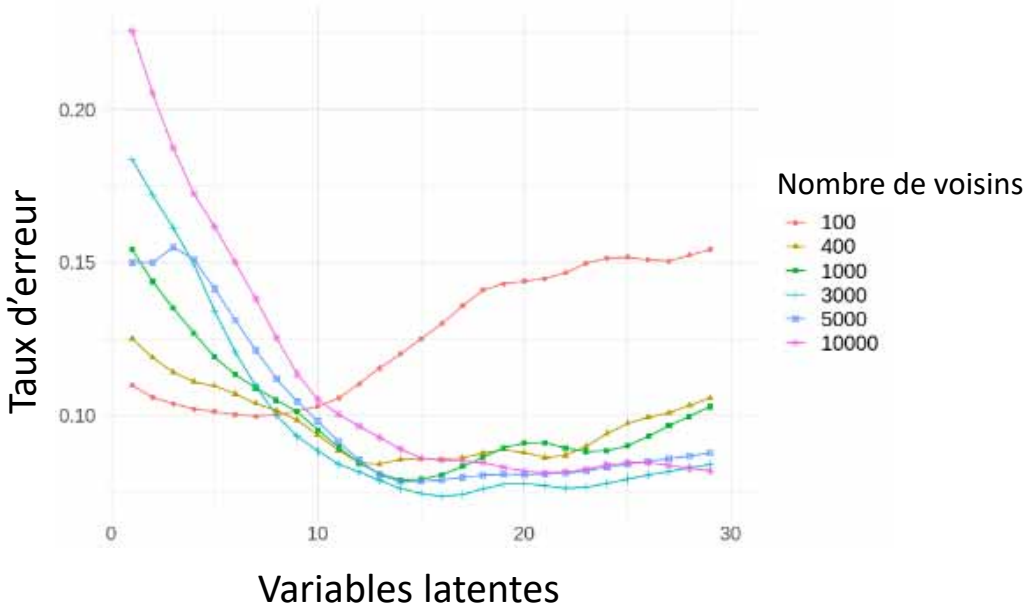
PLS-DA globale :



Meilleur résultat :

Err min : 24 %
20 variables latentes
Temps : ~ 10 min

BF-PLS-DA :



Meilleur résultat :

Err min : 6-7 %
15 variables latentes
Temps : ~4 h

1.4. Une application de parSketch-PLS

parSketch – PLS-DA :

3 paramètres :

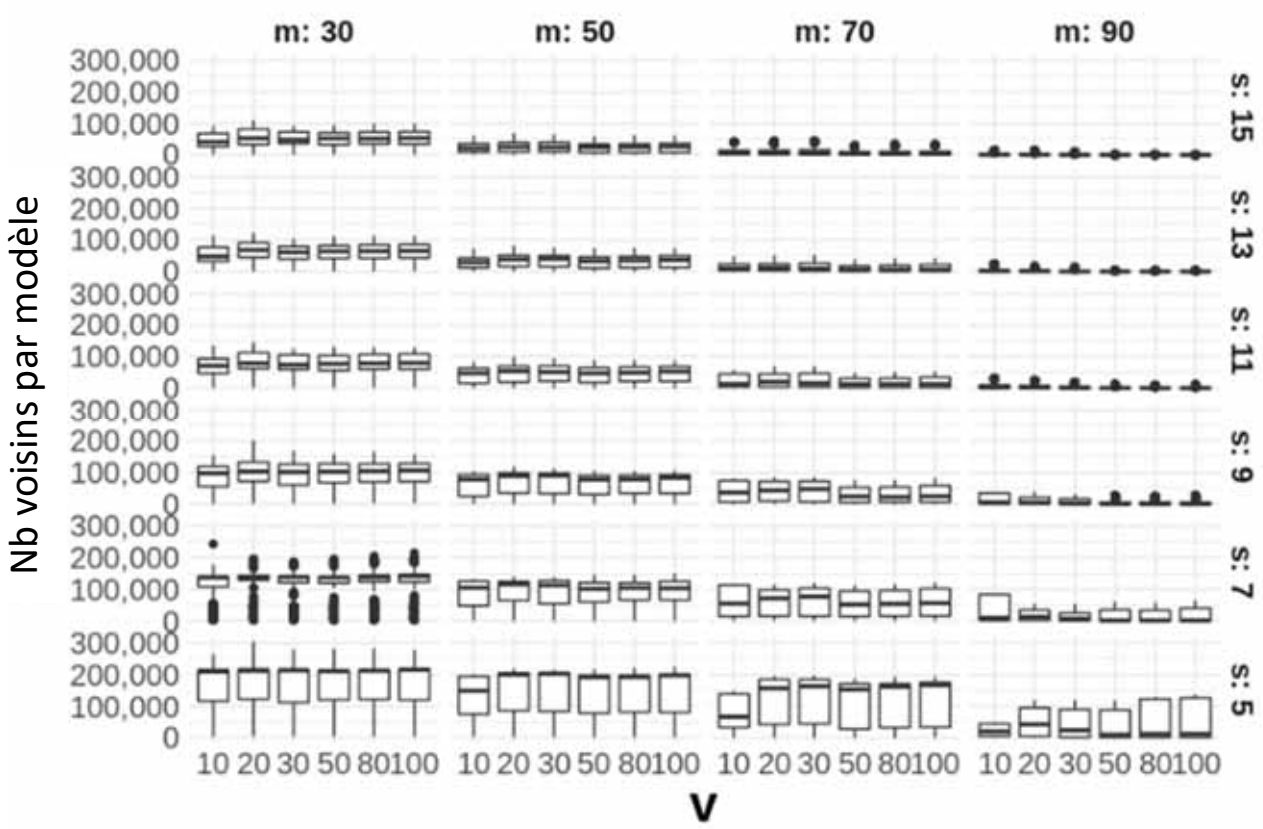
V : 10-100 vecteurs aléatoires

S : 5-15 segments

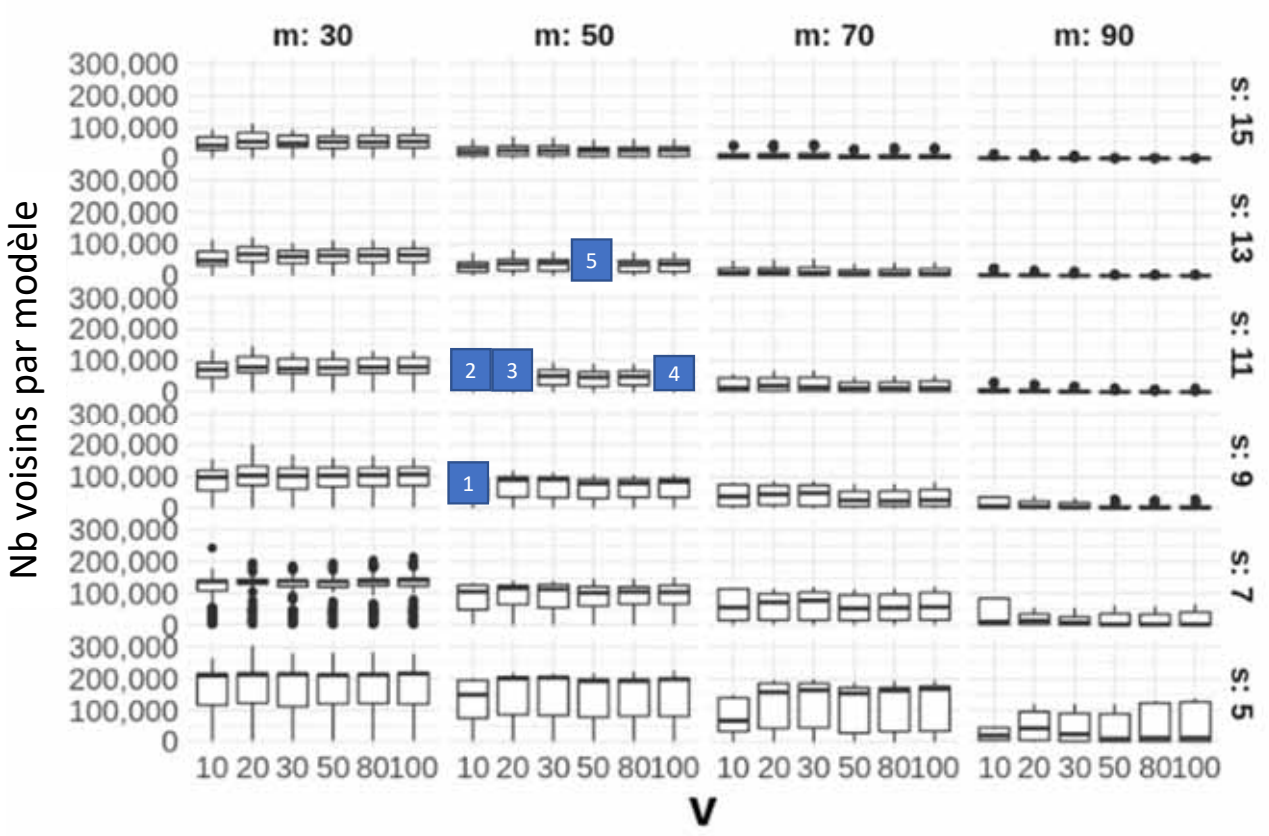
m : 30-90 %

1.4. Une application de parSketch-PLS

Quels sont les impacts des paramètres sur le voisinage ?

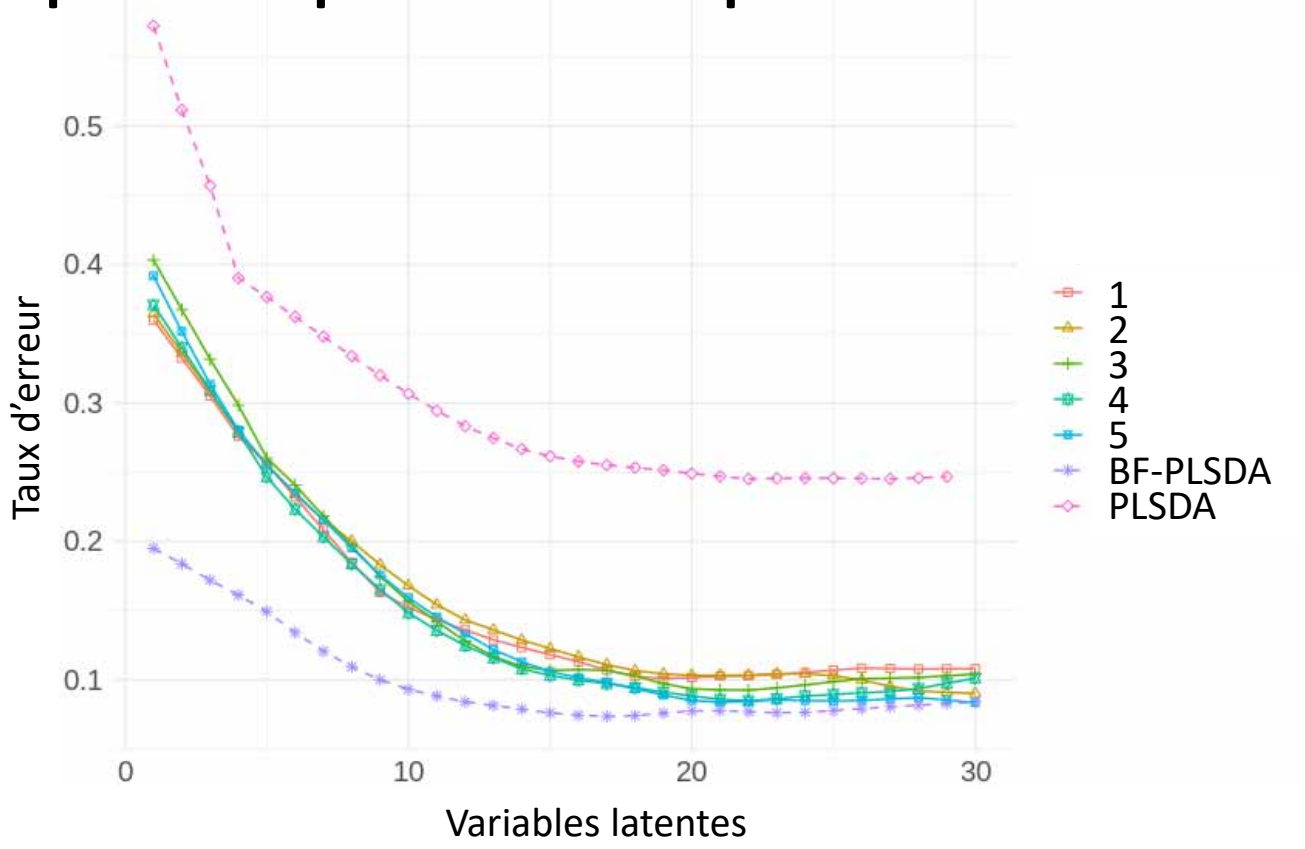


1.4. Une application de parSketch-PLS



1.4. Une application de parSketch-PLS

Quels sont les impacts des paramètres de parSketch sur l'erreur de prédiction ?



1.5. Conclusions et perspectives

Conclusions

- parsketch renvoie un **voisinage très grand**
- parSketch impose une **distance**
- L'intérêt d'utiliser des outils de traitement des **données massives** a été mis en évidence
- **parSketch-PLS** offre une **alternative** aux méthodes usuelles

Perspectives

- parSketch est utilisable dans n'importe quelle méthode utilisant des **sous-ensembles**
- Amélioration de parSketch-PLS en **filtrant** les voisins (BF, …)

1.6. Contributions

ELSEVIER

Chemometrics and Intelligent Laboratory Systems

Volume 203, 15 August 2020, 104076

A “big-data” algorithm for KNN-PLS

Maxime Metz [a, b], Matthieu Lesnoff [b, c, d], Florent Abdelghafour [a, b], Reza Akbarinia [e], Florent Masseglia [e], Jean-Michel Roger [a, b]

ELSEVIER

Biosystems Engineering

Volume 210, October 2021, Pages 69-77

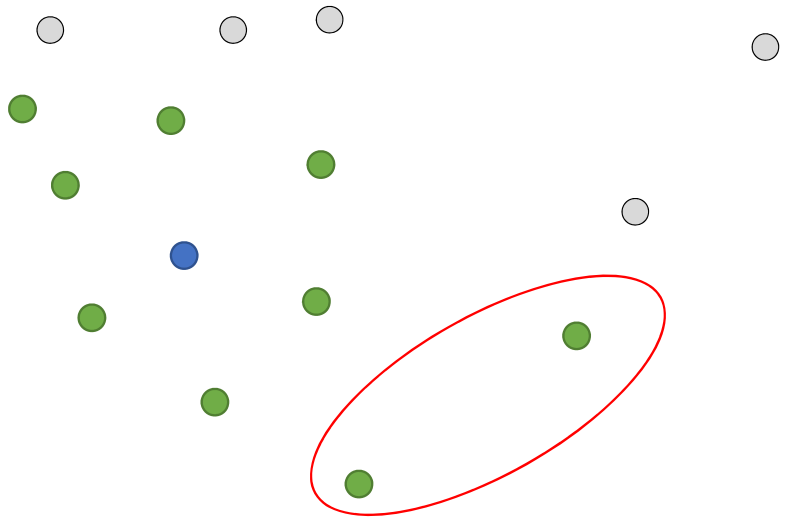
Research Paper

Massive spectral data analysis for plant breeding using parSketch-PLSDA method: Discrimination of sunflower genotypes

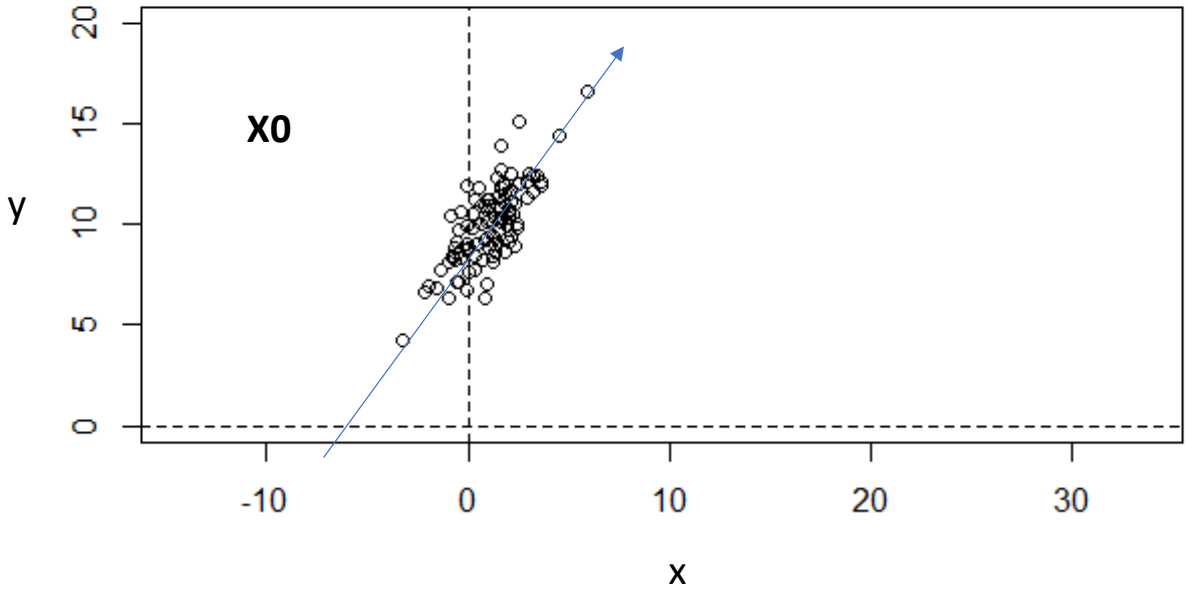
Maxime Ryckewaert [a, b], Maxime Metz [a, b], Daphné Héran [a], Pierre George [c], Bruno Grèzes-Besset [c], Reza Akbarinia [d], Jean-Michel Roger [a, b], Ryad Bendoula [a]

Section 2 : Comment estimer la pertinence d’un individu par rapport à un modèle PLS ?

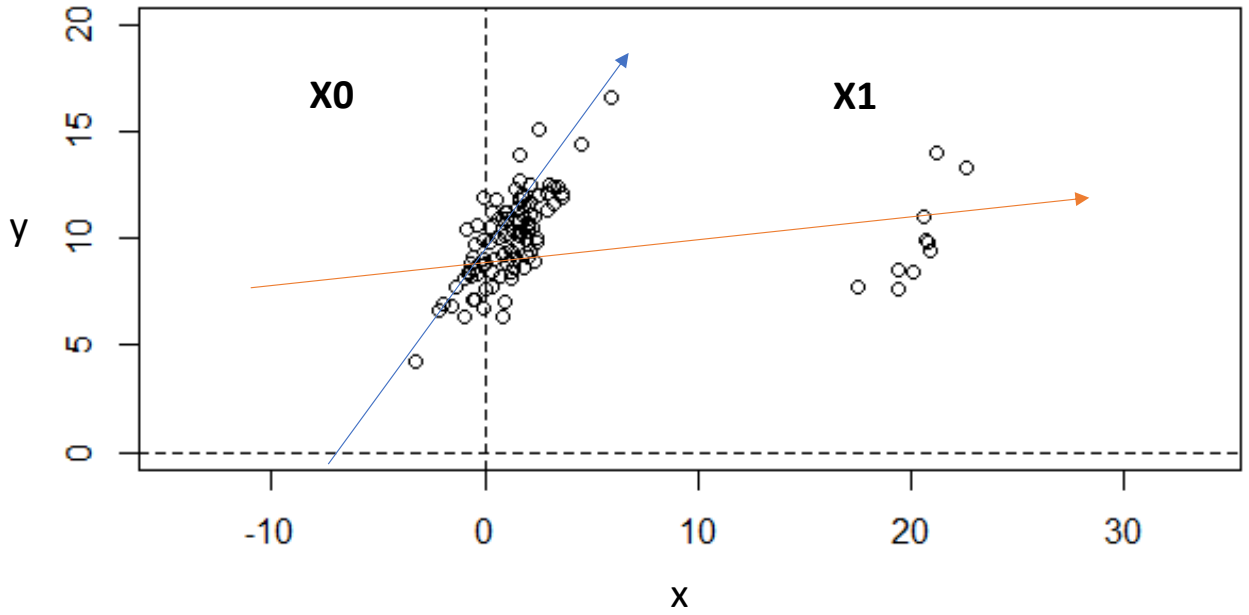
2.1. La robustesse



2.1. La robustesse



2.1. La robustesse



2.1. La robustesse

Hypothèse principale :

- Plus d'individus dans X0 que dans X1

Principale difficulté des méthodes robustes :

Trouver une bonne mesure pour mettre en évidence les valeurs aberrantes

Où est notre aberrant ?

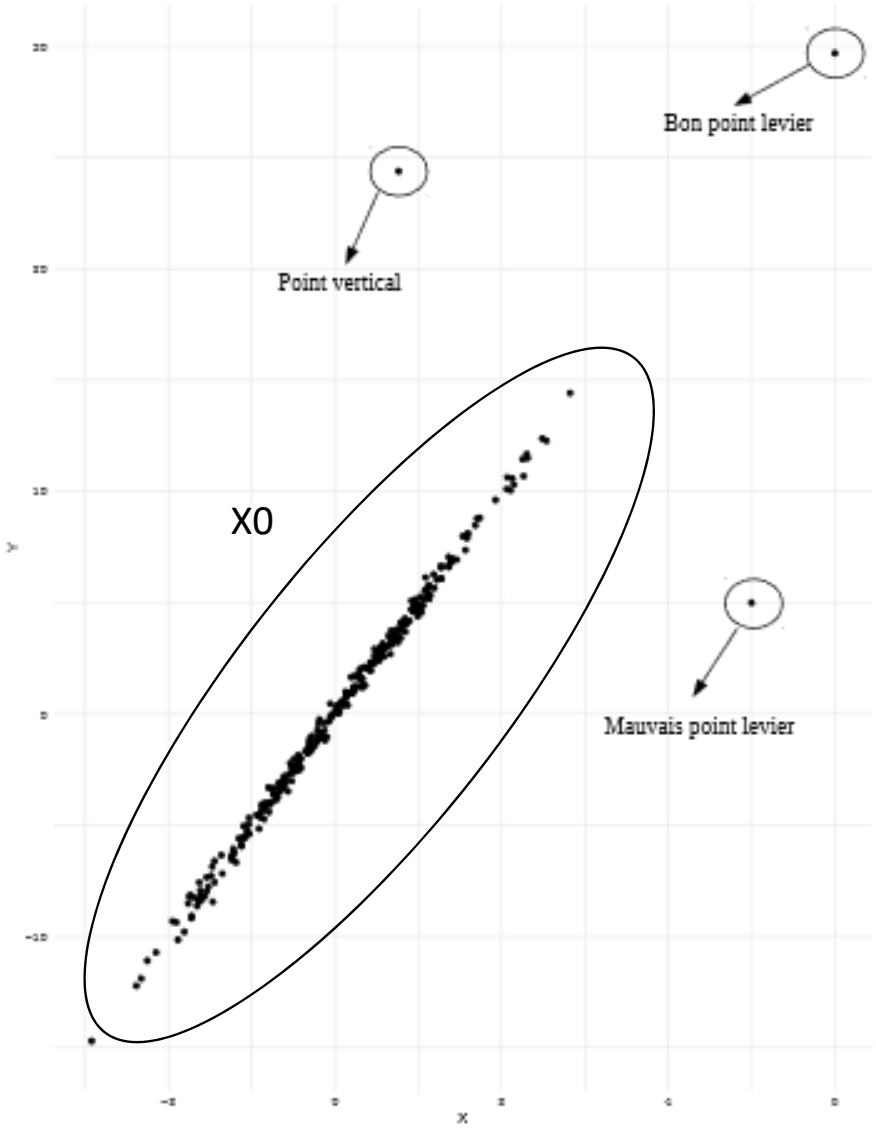
2.1. La robustesse

Point vertical : loin selon la droite calibrée sur X0. Influence modérée sur la calibration du modèle.

Point levier : loin du centre des données selon X.

Bon point levier : contribue à la bonne calibration d'un modèle sur X0.

Mauvais point levier : à la fois vertical et levier, influence fortement la calibration.



2.2. Les méthodes PLS robustes

Wakelinc et Macfie, « A Robust PLS Procedure »

Cummins , « Iteratively reweighted partial least squares: A performance analysis by monte carlo simulation »

Pell, « Multiple Outlier Detection for Multivariate Calibration Using Robust Statistical Techniques »

Griep et al., « Comparison of Semirobust and Robust Partial Least Squares Procedures »

Serneels et al., « Partial Robust M-Regression »

Gil et Romera, « On Robust Partial Least Squares (PLS) Methods »

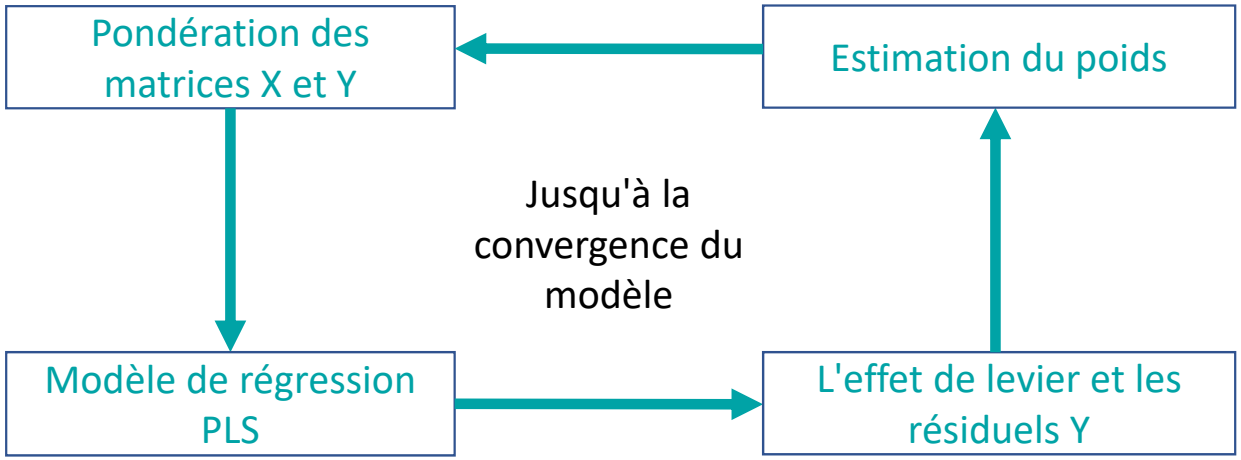
Møller, Frese, et Bro, « Robust Methods for Multivariate Data Analysis »

Hubert et Branden, « Robust Methods for Partial Least Squares Regression »

2.2. Les méthodes PLS robustes

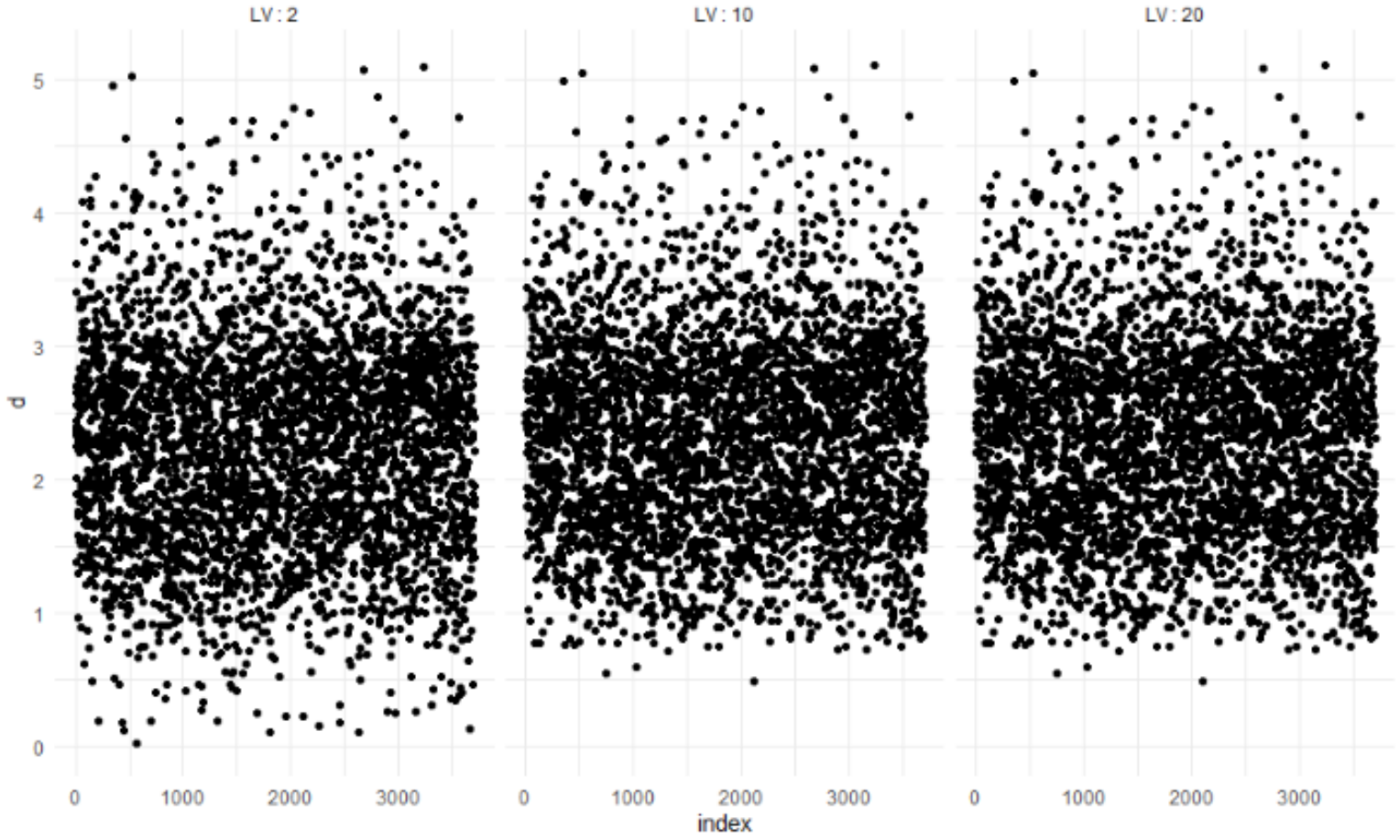
PRM (Partial Robust M-regression) : *[Serneels, 2005]*

Calcul d'un modèle PLS avec x variables latentes définies puis pondération en fonction des résidus Y et effet de levier.



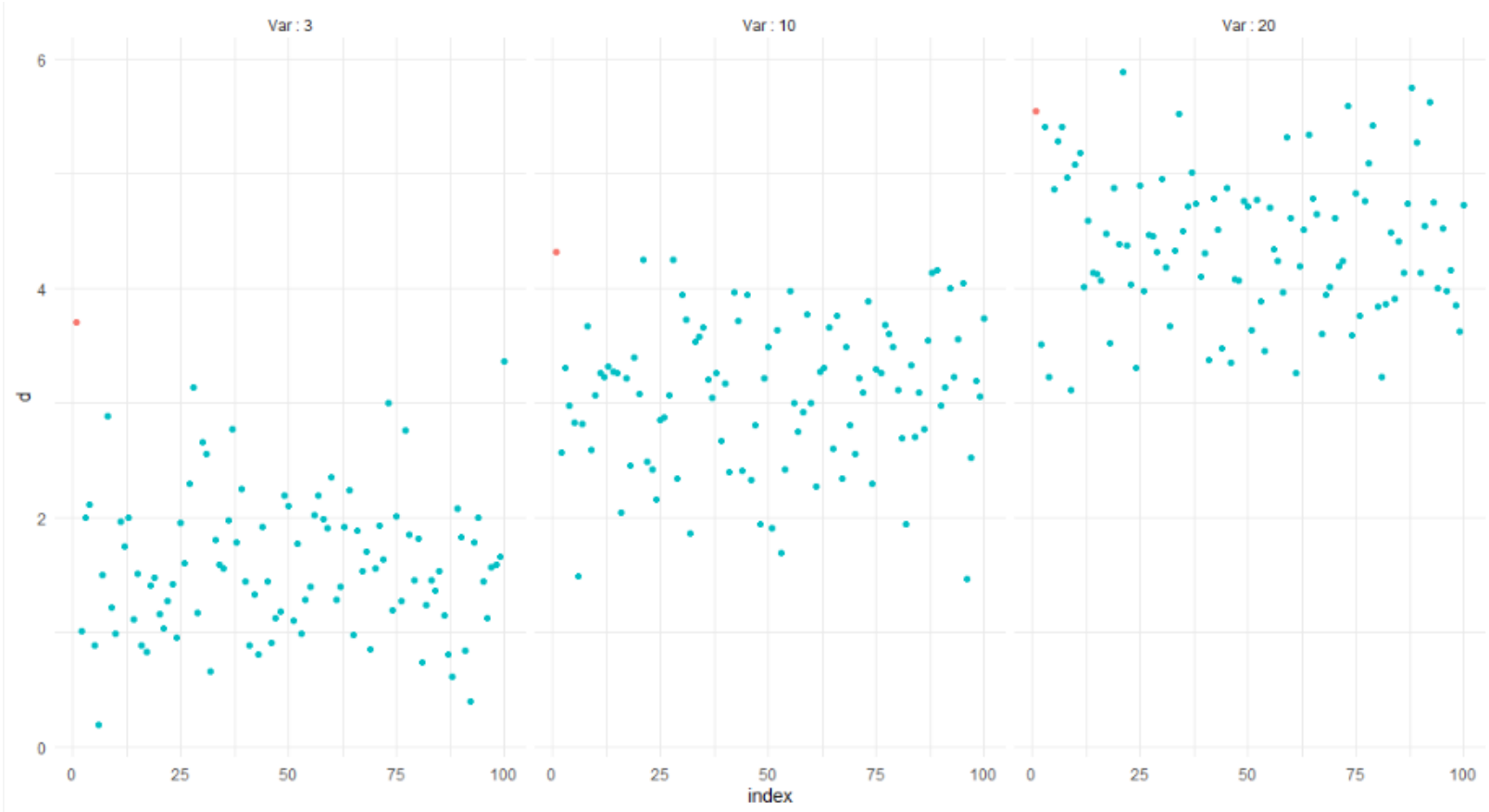
2.2. Les méthodes PLS robustes

Distance Euclidienne au centre des scores PLS

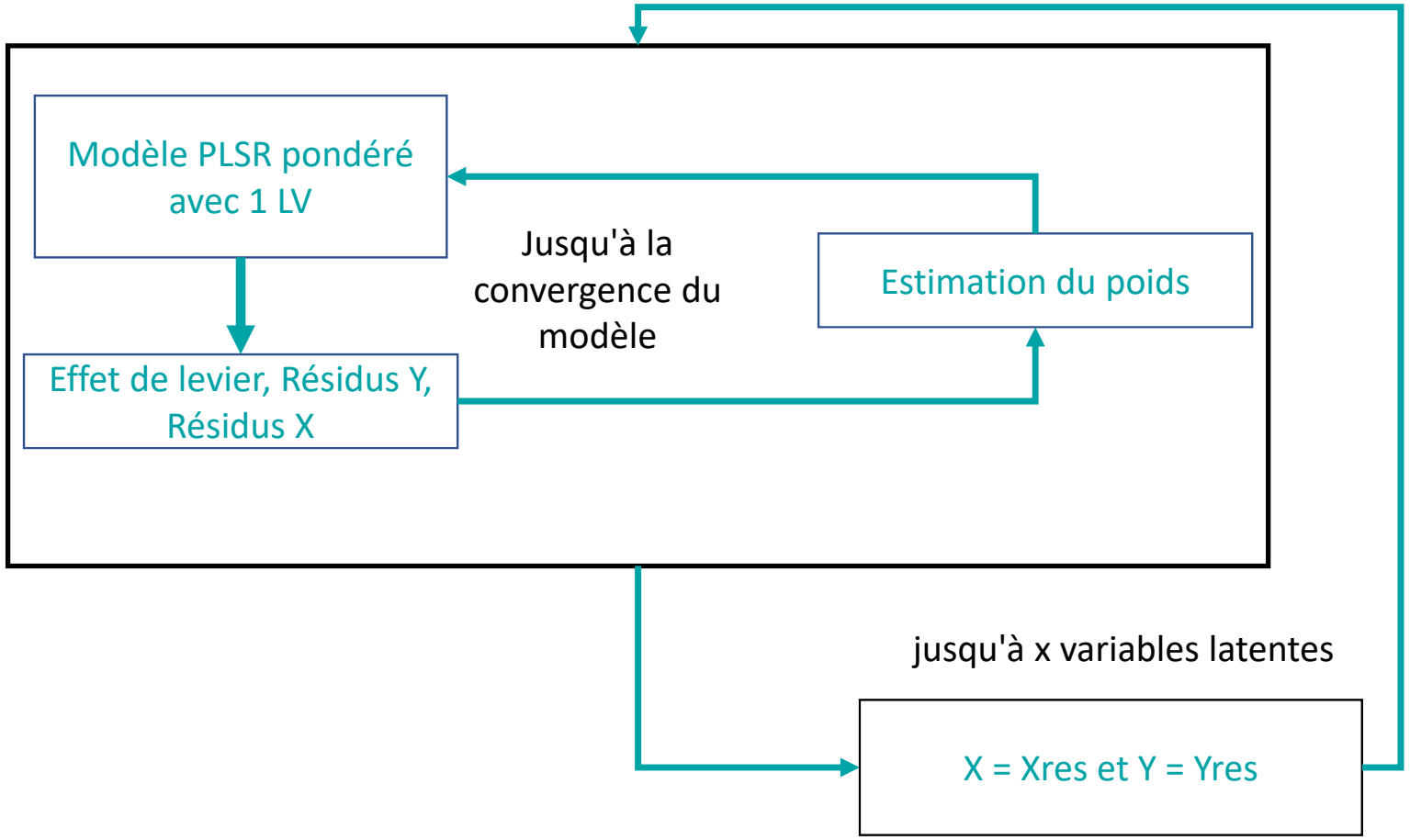


2.2. Les méthodes PLS robustes

Distance de Mahalanobis



2.3. Roboost-PLSR



2.4. Application de Roboost-PLSR

4 méthodes différentes :

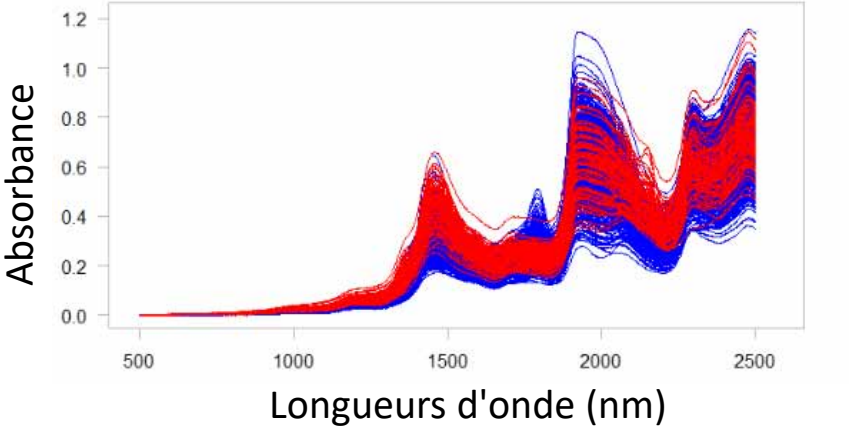
PLSR avec des valeurs aberrantes
PLSR sans valeurs aberrantes → Référence
PRM avec des valeurs aberrantes
RoBoost-PLSR avec des valeurs aberrantes

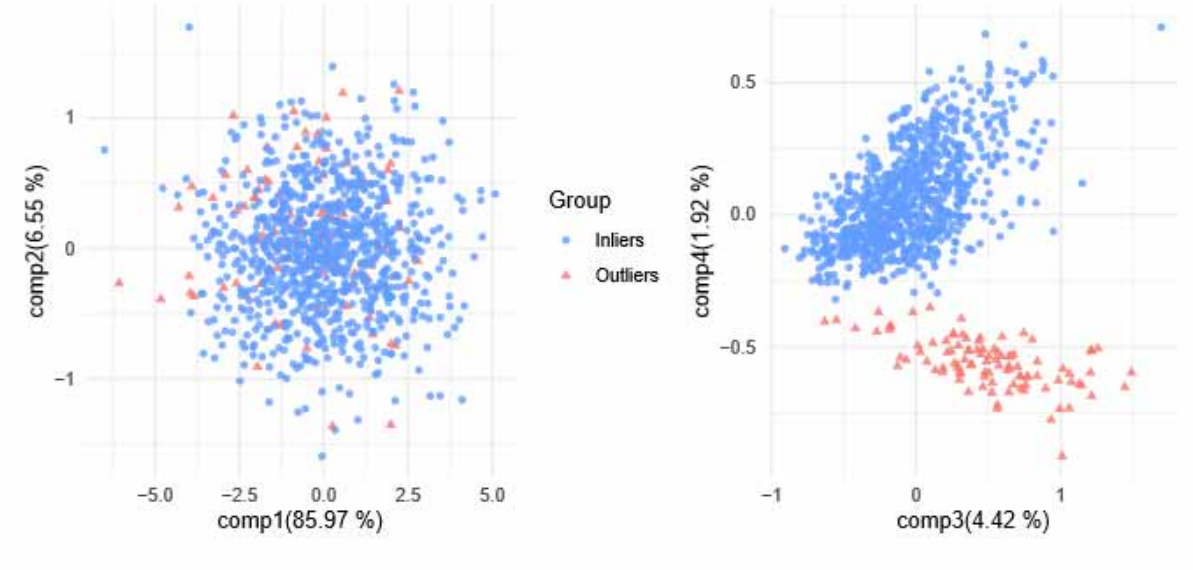
1 jeu de données :

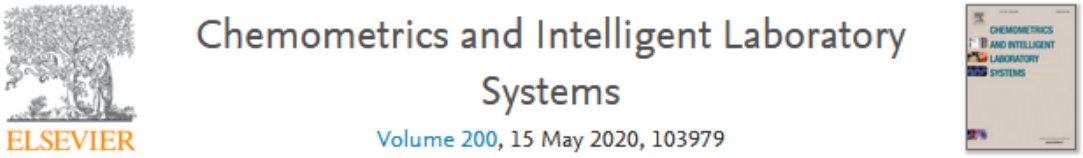
Données simulées

2.4. Applications de Roboost-PLSR

Données simulées

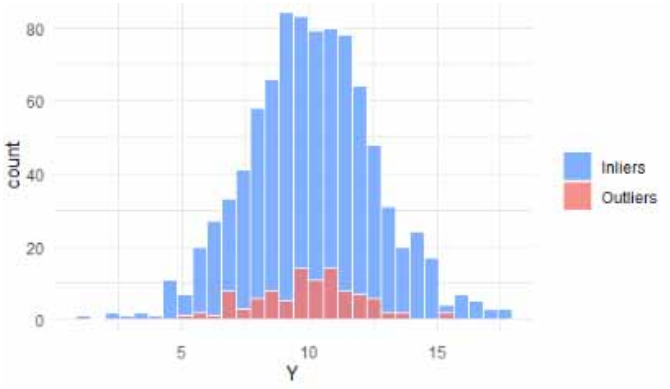






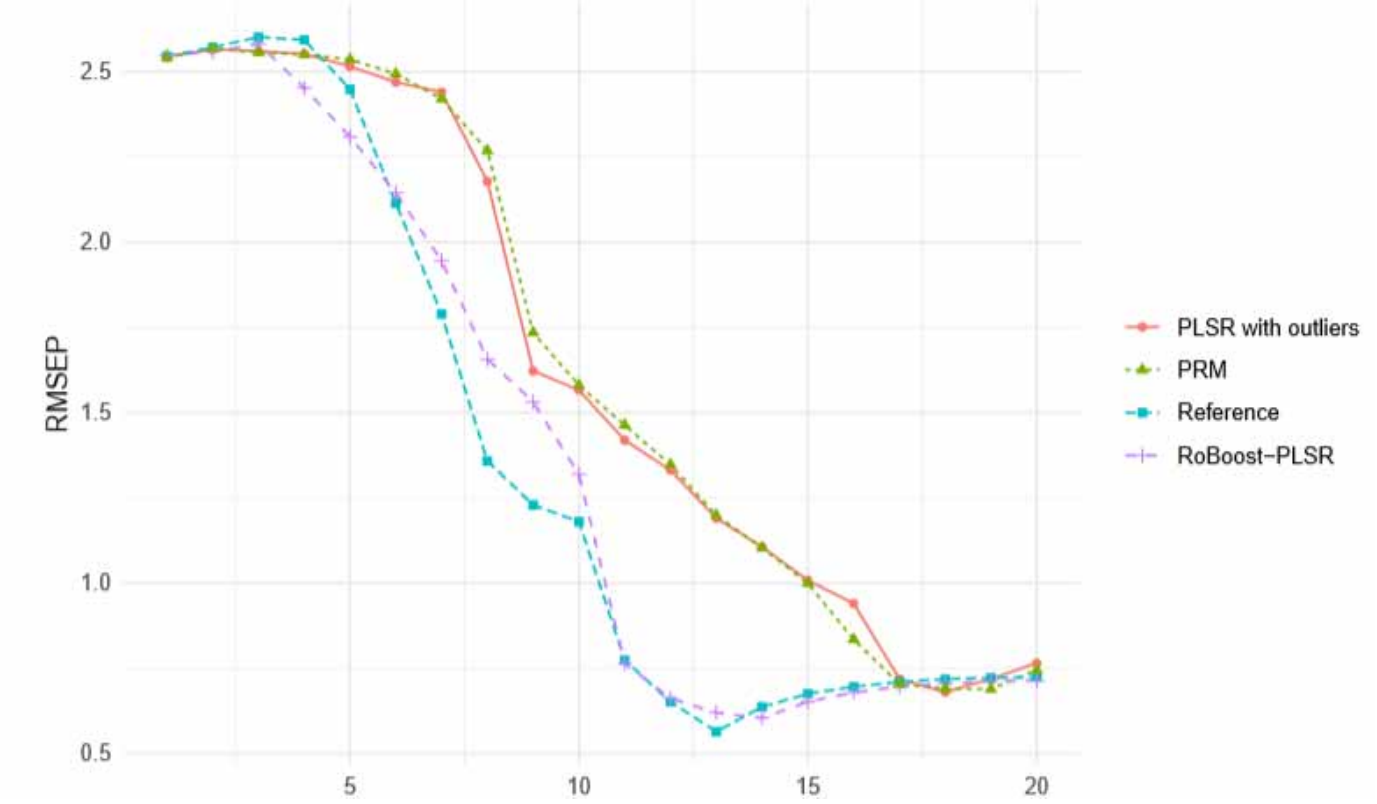
A note on spectral data simulation

Maxime Metz [a, b], Alessandra Biancolillo [a], Matthieu Lesnoff [b, c], Jean-Michel Roger [a, b]



2.4. Applications de Roboost-PLSR

Données simulées



2.5. Conclusions et perspectives

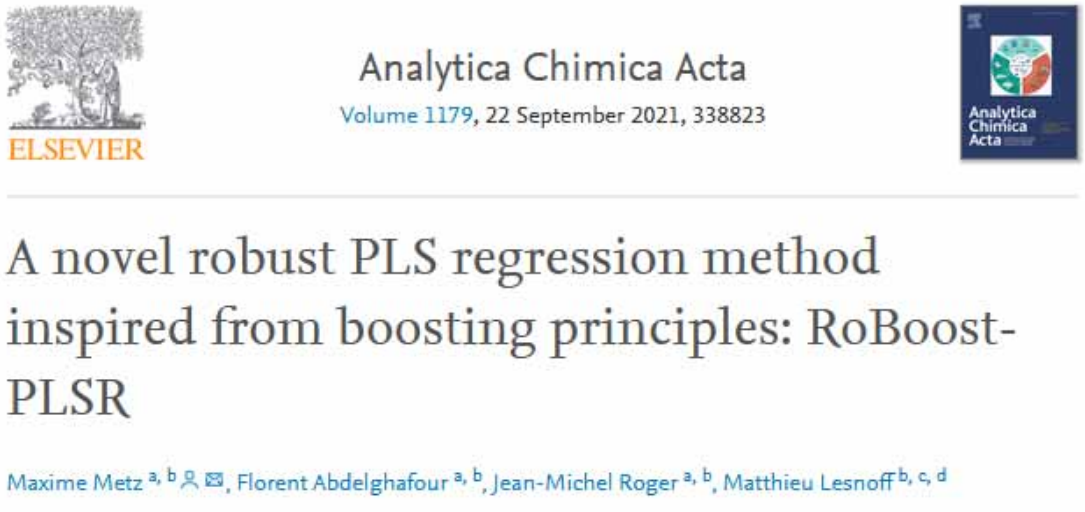
Conclusions

- RoBoost-PLSR permet de **réduire/éliminer** l'effet des outliers sur la calibration
- RoBoost-PLSR permet de faciliter la **cross-validation**

Perspectives

- Etudier les fonctions de **poids**
- Etudier les **coefficients de régression**
- Etudier la présence de données aberrantes dans un cas **multi-tableaux**
- Développer une méthode robuste pour la **classification**

2.6. Contributions

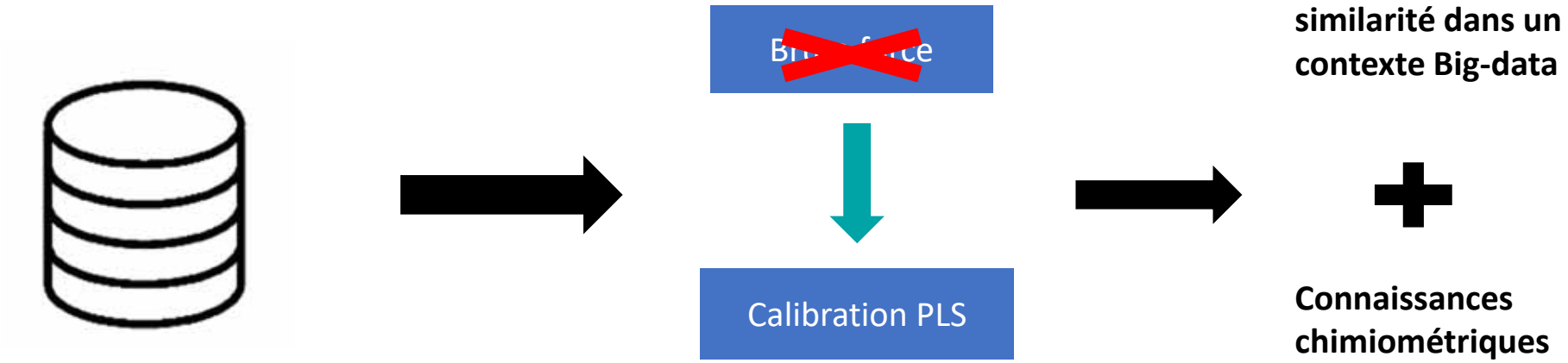


Aldrig Courand, Maxime Metz, Daphné Héran, Carolen Feilhes, Fanny Prezman,Eric Serrano, Ryad Bendoula, and Maxime Ryckewaert. Evaluation of a regression method (roboost-plsr) to predict biochemical variables for agronomic applications : case study of grape berry maturity monitoring. Chemometrics and Intelligent Laboratory Systems, XXXX(under revision)

Maxime Metz, Maxime Ryckewaert, Silvia Mas Garcia, Pierre Dardenne, Ryad Bendoula, Matthieu Lesnoff, and Jean-Michel Roger. Roboost-pls2-r : An extension of roboost-plsr method for multi-responses. Chemometrics and Intelligent Laboratory Systems,XXXX (under revision)

Section 3 : Comment associer les paradigmes de la chimiométrie et du “big-data” ?

3.1. Introduction



3.1. Introduction

Indexation orientée

Permet d'indexer en fonction d'une distance chimiométrique

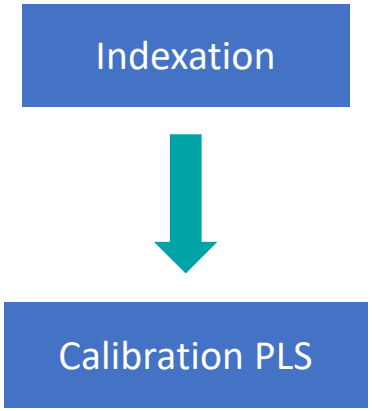
Filtrage du voisinage

Permet de limiter l'impact de voisin non pertinent pour la calibration du modèle PLS

3.2. Indexation orientée

Indexation orientée

Permet d'indexer en fonction d'une distance chimiométrique

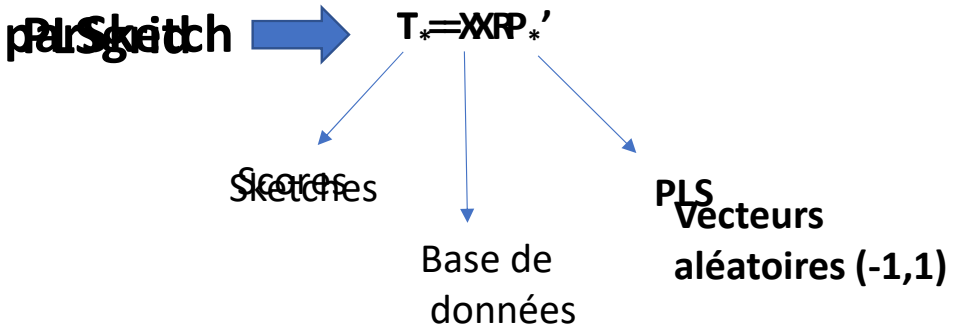


Distances / similarités :

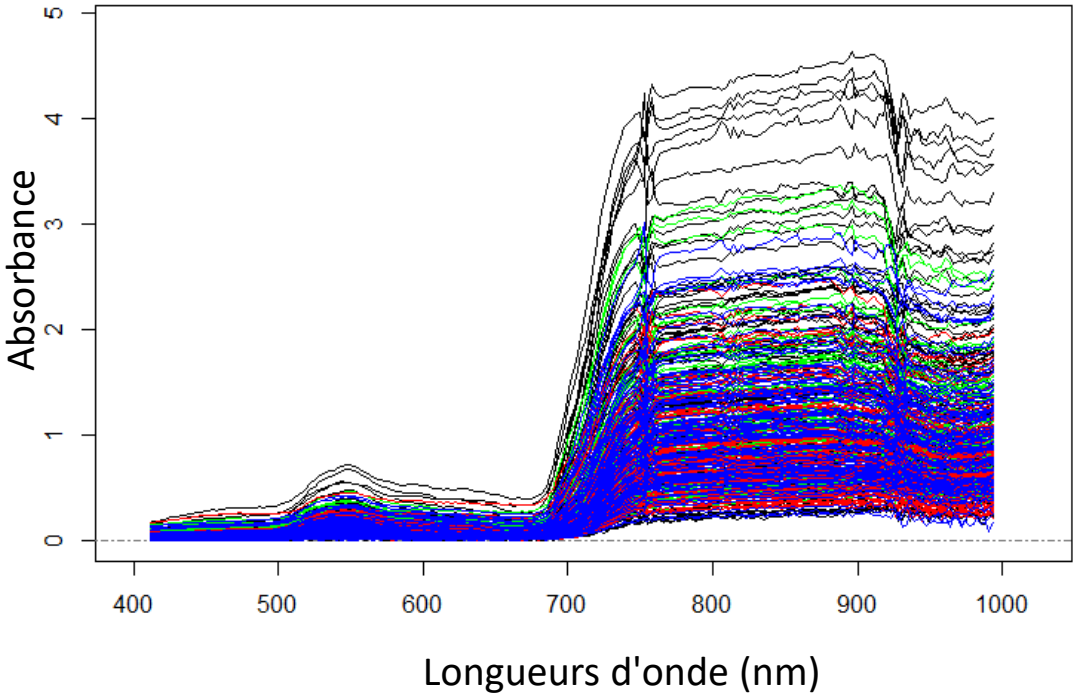
- Malahanobis
- Euclidiennes
- Avec réduction de dimension
- …….

3.3. PLSgrid


Réduction de dimension
Création des grilles
Recherche des voisins

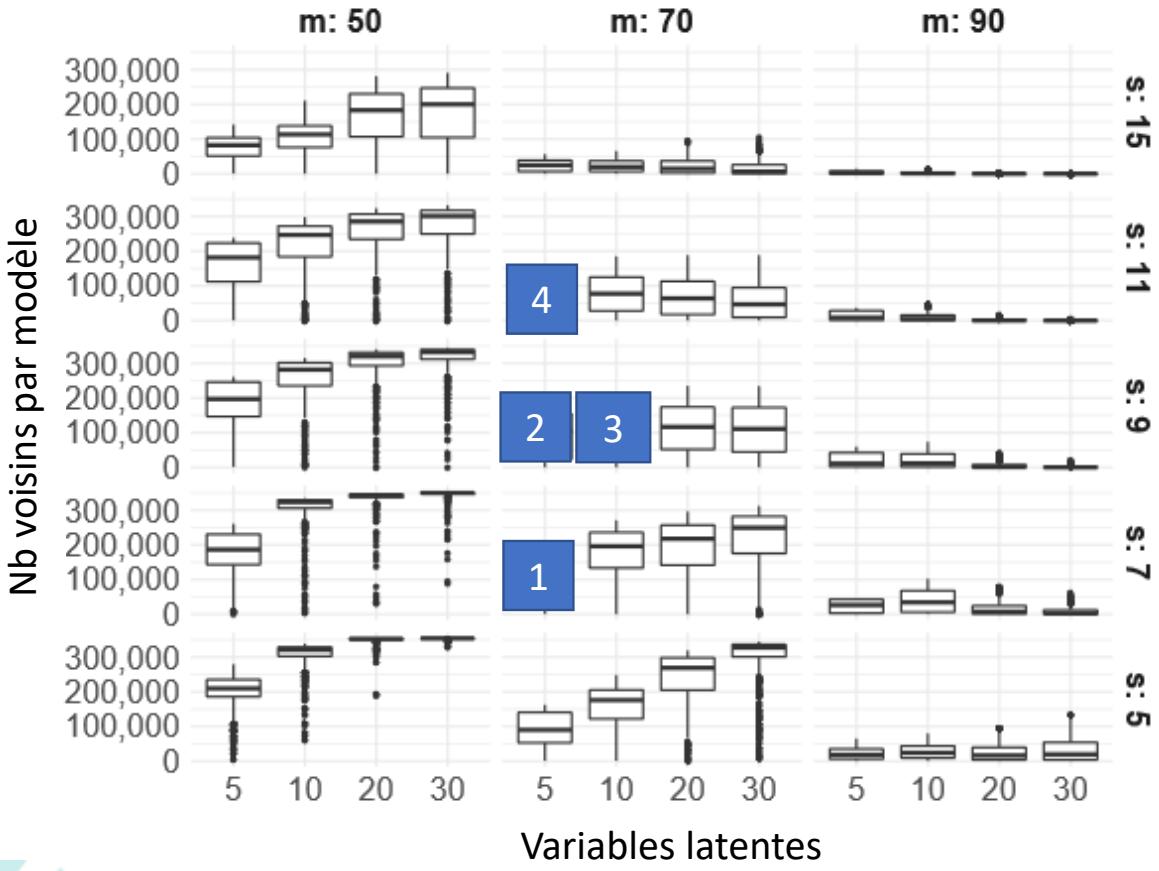

PLSSketch
T* = XP*'
Scores
Sketches
Base de données
PLS
Vecteurs aléatoires (-1,1)

3.4. Application de PLSgrid



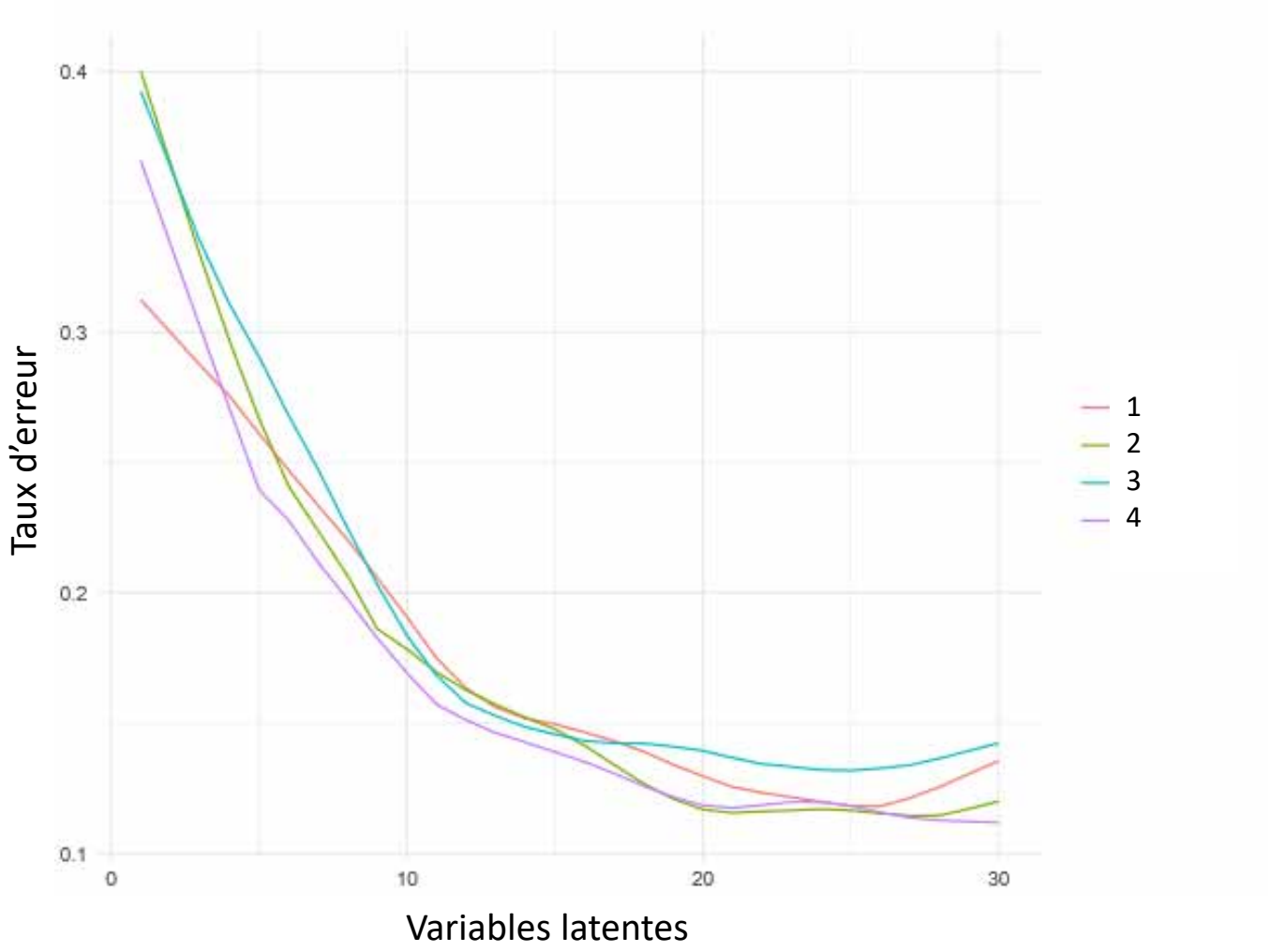
- Spectres de feuilles
- 4 génotypes (4 images)
- 360 000 spectres
- 256 variables

3.4. Application de PLSgrid



S : nb segments
m : % min de cellules en commun

3.4. Application de PLSgrid



3.5. Conclusion et perspectives

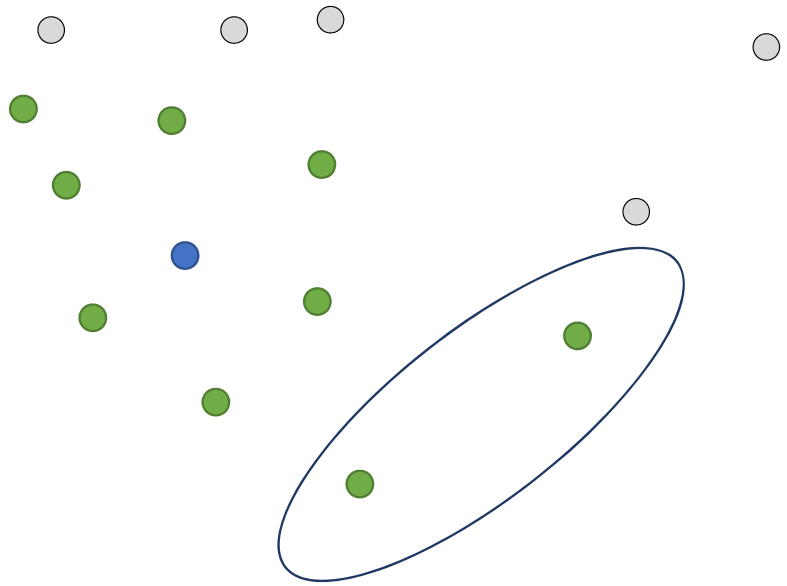
Conclusions

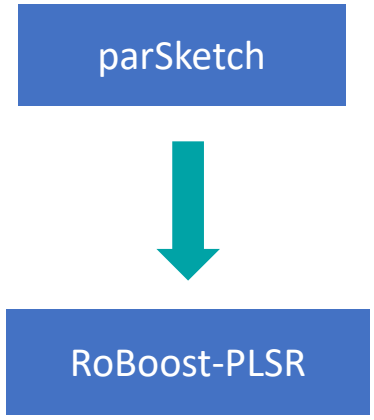
- PLSgrid-PLSDA **approche** les résultats de parSketch-PLSDA
- PLS est **une alternative** potentielle à la projection sur vecteurs aléatoires
- Par cette stratégie, il est possible de représenter **une autre distance** que les distances Euclidiennes

Perspectives

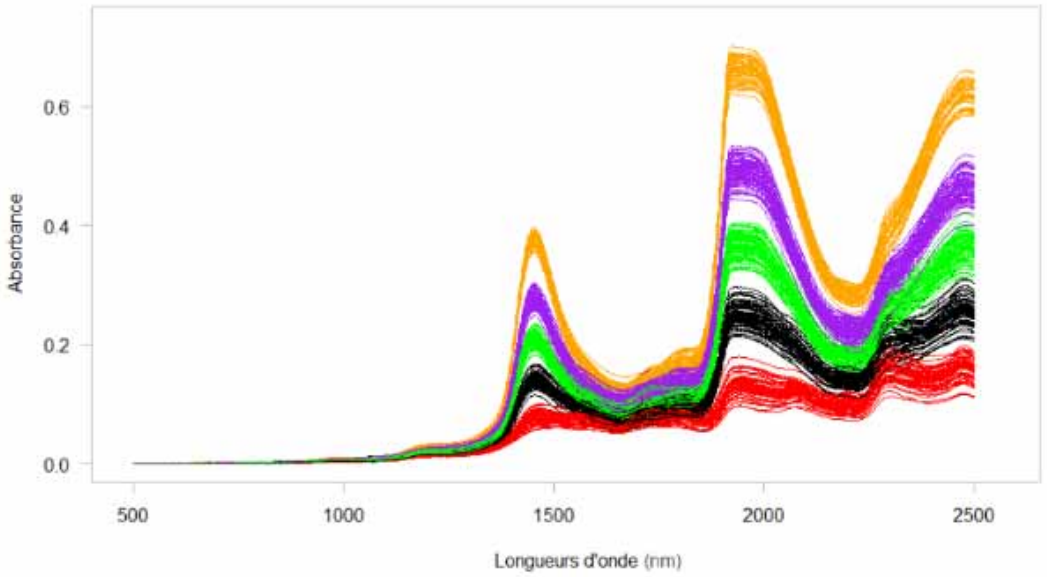
- La PLS pourrait être combinée avec **d'autres méthodes d'indexation**
- Il serait intéressant de combiner la méthode PLSgrid-PLSDA avec la méthode **brute-force**
- Il serait intéressant d'utiliser des outils d'approximation du voisinage permettant d'approcher une plus grande diversité de **distances**

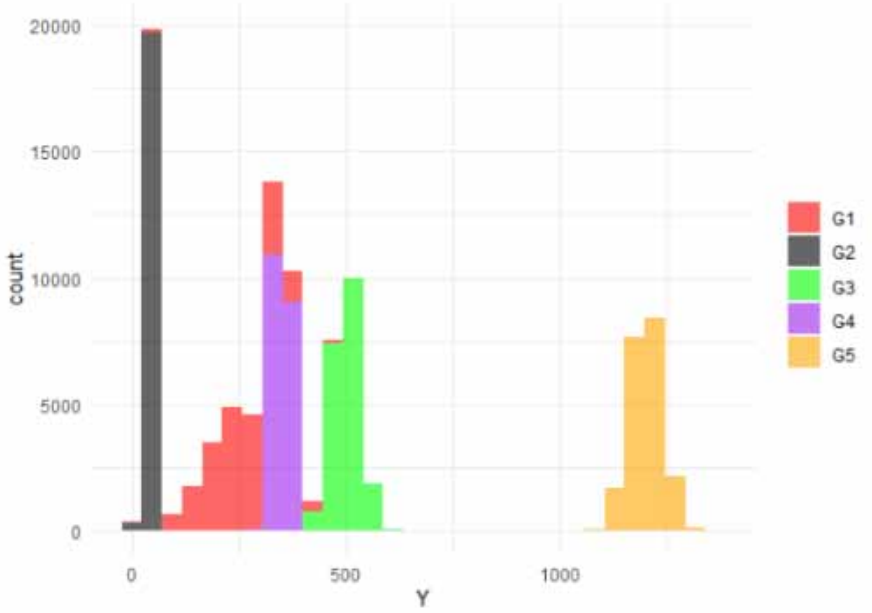
3.6. Filtrage du voisinage





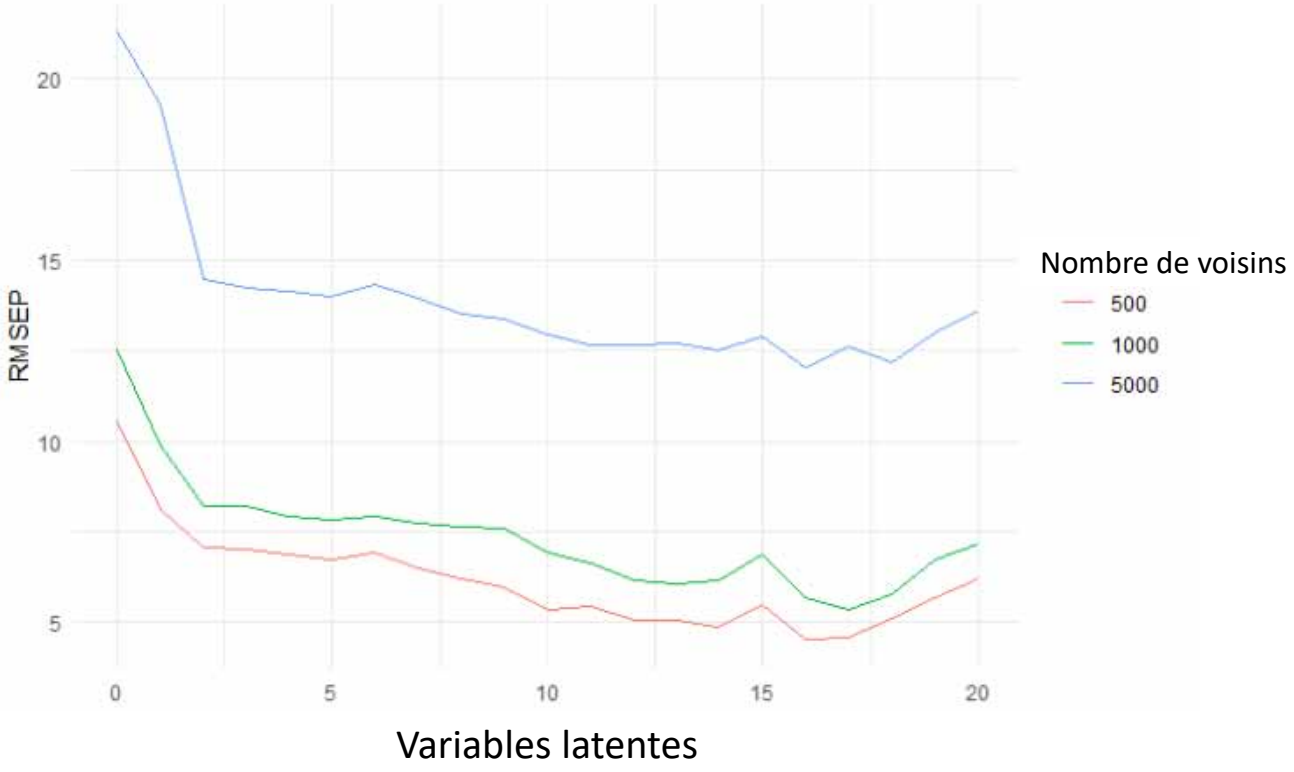
3.7. Application de parSketch-RoBoost-PLSR



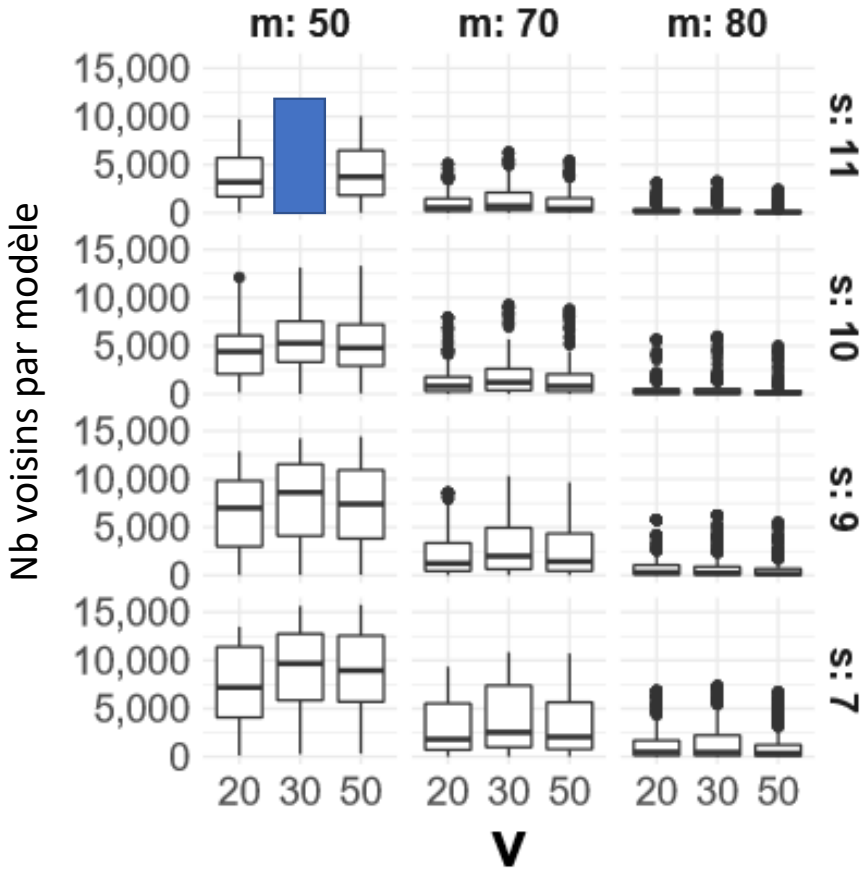


3.7. Application de parSketch-RoBoost-PLSR

BF-PLSR

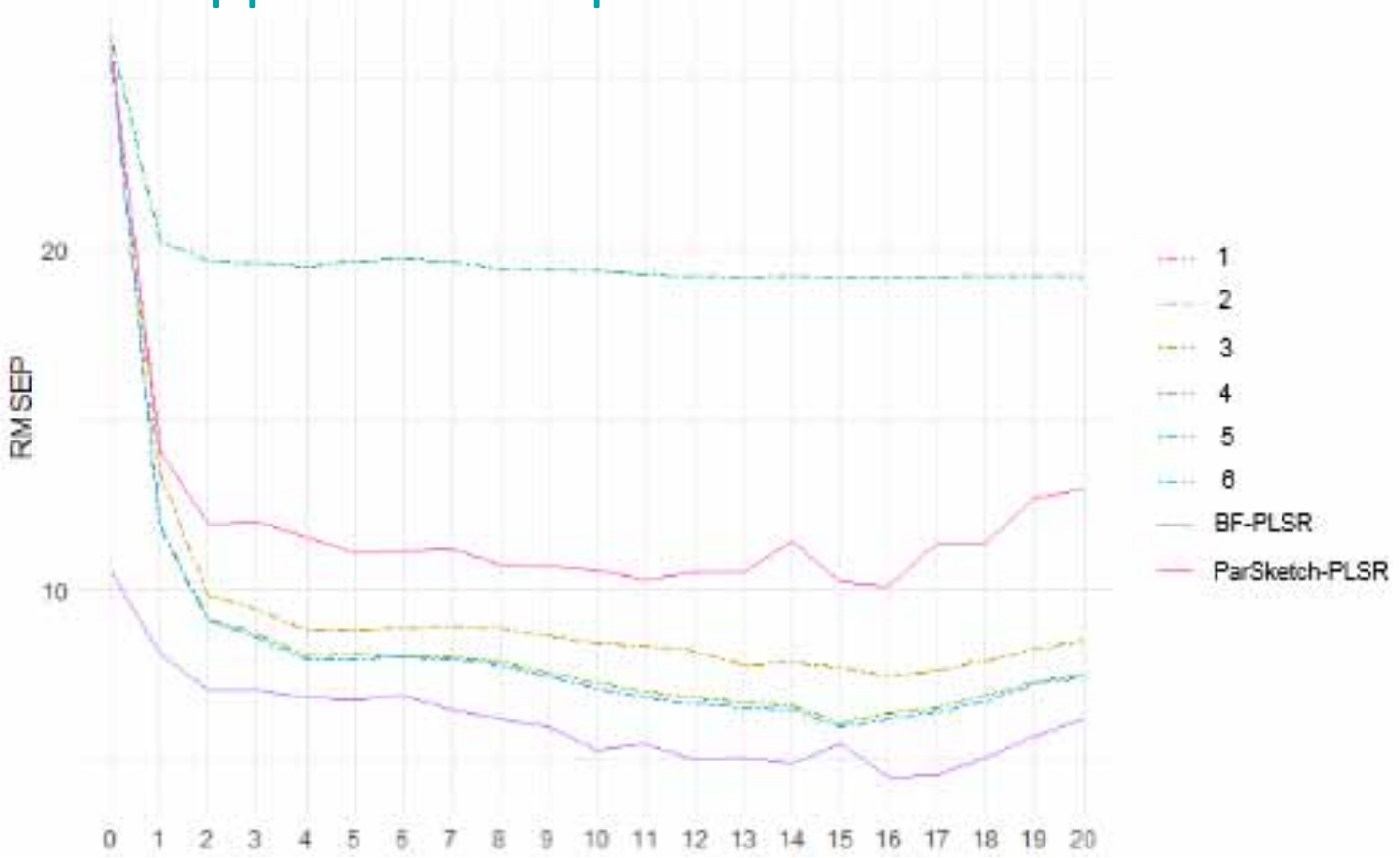


3.7. Application de parSketch-RoBoost-PLSR



S : nb segments
m : % min de cellules en commun
V : nb vecteurs aléatoires

3.7. Application de parSketch-RoBoost-PLSR



Combinaison	α	β	γ
1	Inf	6	Inf
2	Inf	4	Inf
3	Inf	8	Inf
4	Inf	6	6
5	6	6	Inf
6	4	6	Inf

3.8. Conclusion et perspectives

Conclusions

- Certains voisins peuvent être **non pertinents** bien que la **distance** choisie soit **adaptée**
- Les méthodes robustes peuvent être utilisées pour **améliorer les capacités prédictives**
- RoBoost-PLSR a un **coût calculatoire** élevé

Perspectives

- Sous-**échantillonner** les voisinages renvoyés par parSketch
- Développer une **approche massivement parallélisable** de RoBoost-PLSR
- Combiner les approches d'indexation telles qu'**iSAX** avec RoBoost-PLSR

Conclusion générale, perspectives et questionnements

Conclusion générale

Dans cette thèse, l'intérêt de combiner les **connaissances** issues du traitement de **données massives** avec les connaissances de **la chimiométrie** a été mis en évidence.

Les connaissances métiers de **la chimiométrie** permettent de développer des outils adaptés aux problématiques du traitement des **données chimiques** et les outils du **big-data** permettent de traiter les **données chimiques massives**.

Dans de futurs travaux, il sera donc intéressant de joindre ces deux domaines afin de développer **des outils big-data pour le traitement de données chimiques**.

Perspectives générales et questionnements

Perspectives

- Evaluer **d'autres outils** utilisés dans le domaine du traitement des données massives
- Développer les outils de **chimiométrie** existants pour les données massives (GPU et calcul matriciel)

Questionnements

- De grands ensembles de données n'impliquent pas toujours de meilleures performances : **smart data.**
- L'augmentation de la quantité de données, l'utilisation d'outils complexes et de moyens de calcul onéreux posent la question de l'**impact de nos stratégies de modélisation sur l'environnement.**

Merci à toute l'équipe !!!

ChemHouse