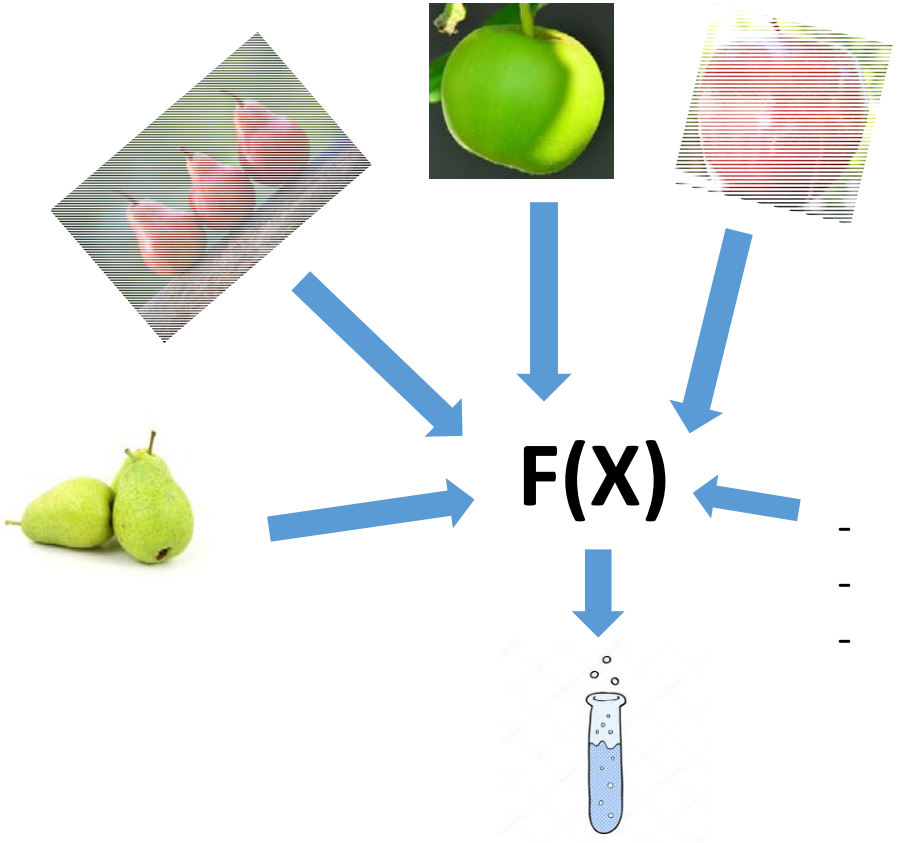
SUJET:

UN ALGORITHME ‘BIG-DATA’ POUR LA LOCAL-PLS

Présenté par: Maxime Metz
Supervisé par: Matthieu Lesnoff, Jean-Michel Roger

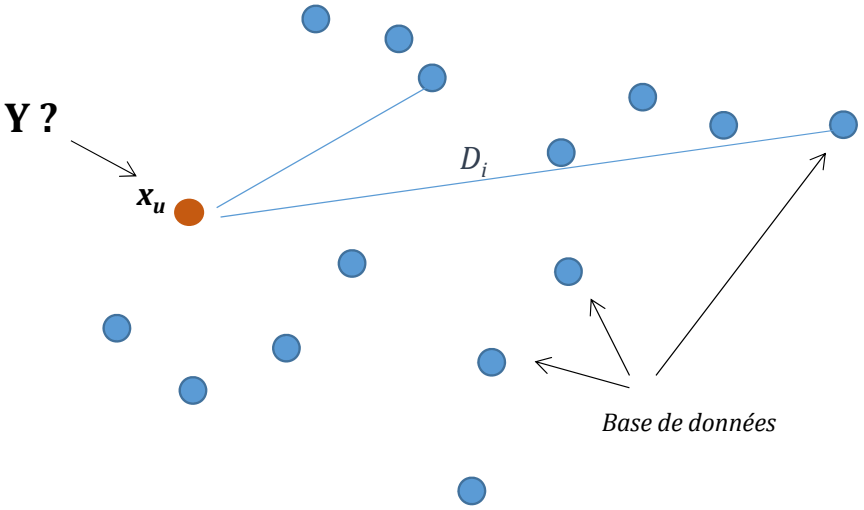
CONTEXTE



- **Spectro**
- **Acquisition**
- **….**

Méthodes locales

Les méthodes locales :



Idées de base :

- Un point à prédire un modèle
- Les points ont une importance différente dans le modèle calculé

Deux cas :

- Localement pondérée
- Sélection de voisins

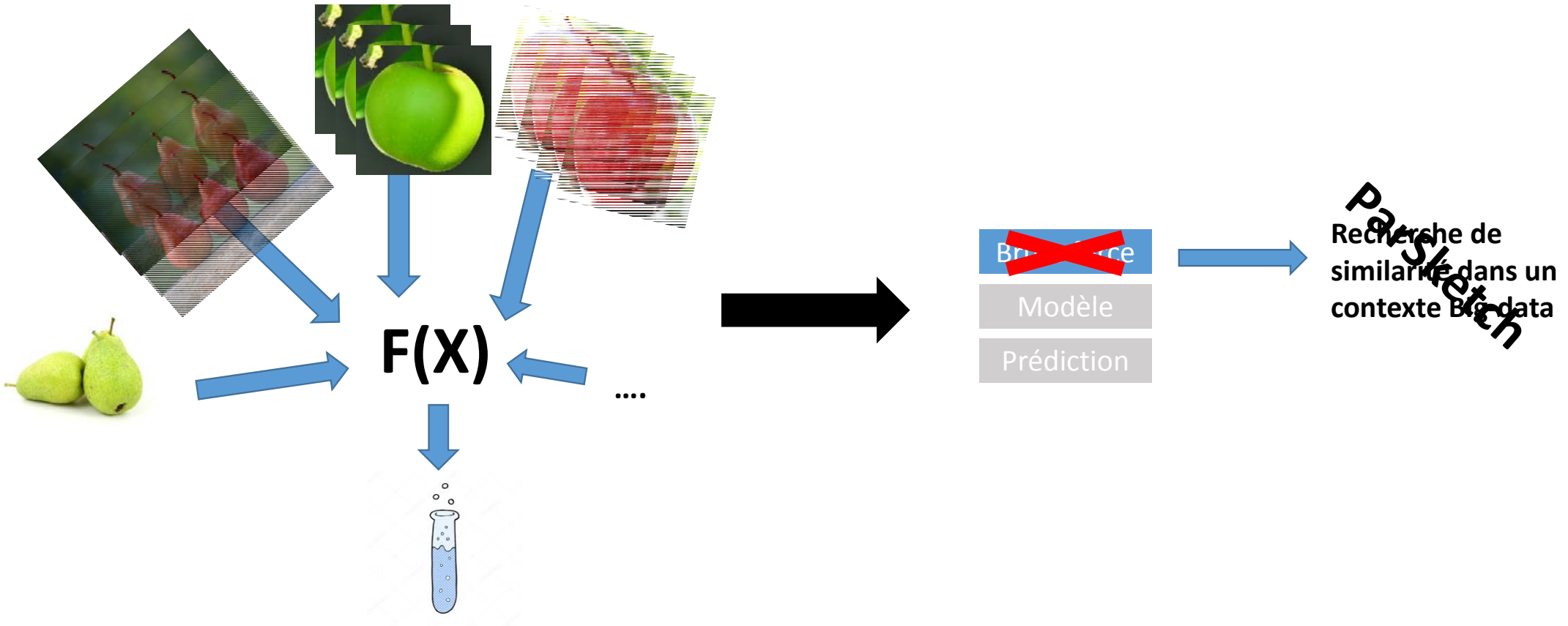
Une procédure générale :

Brute force

Modèle

Prédiction

Les méthodes locales pour les données massives :

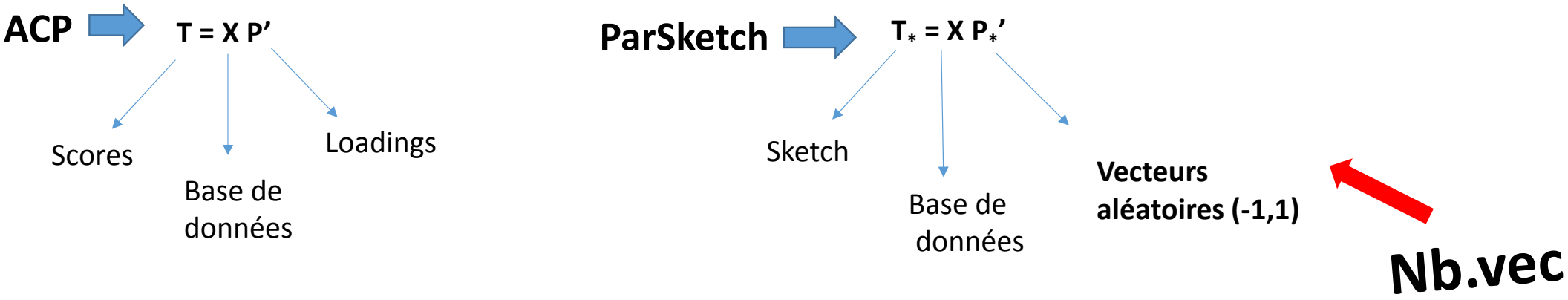


Intégration de ParSketch

ParSketch :

Intégration de ParSketch

Réduction de dimensions



Pourquoi cette technique de réduction de dimensions ?

- Approche les distances euclidiennes
- Opération rapide
- Les variables obtenues représentent la même chose

↳ **Méthode « grid »**

Intégration de ParSketch

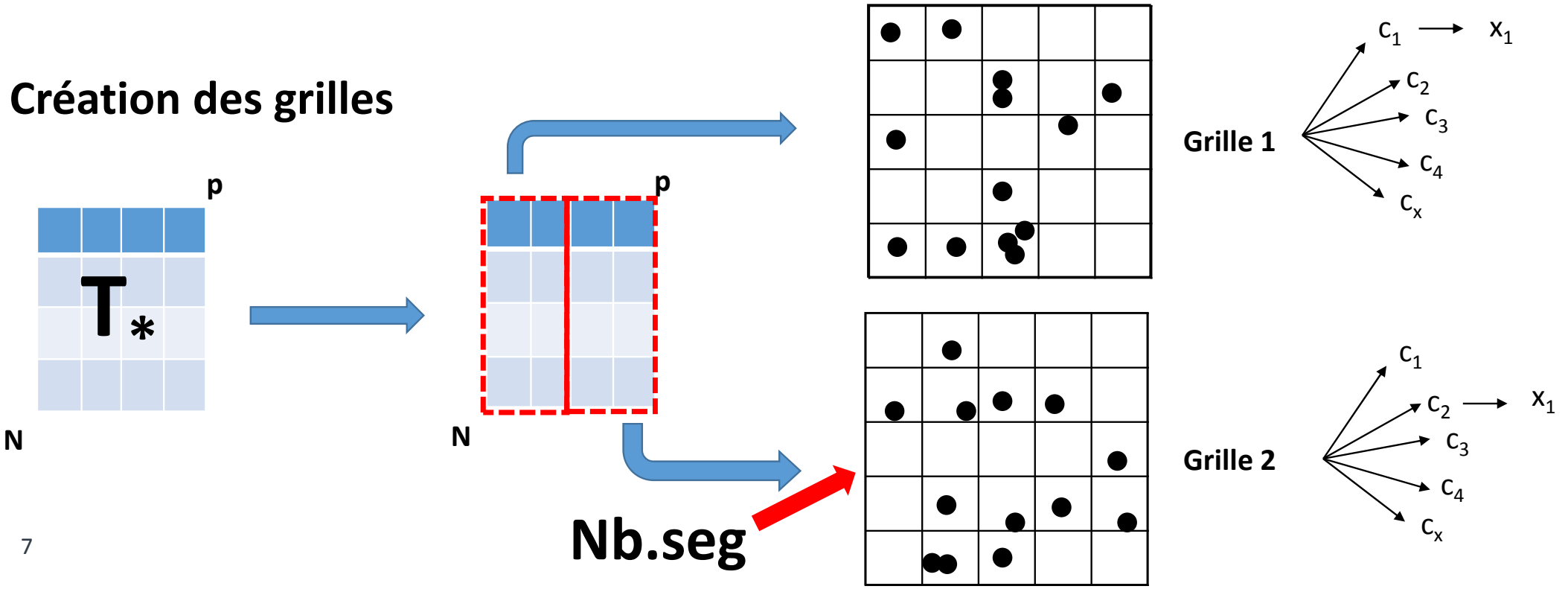
ParSketch :

Réduction de dimensions → **Création des grilles** → Recherche des voisins

Pourquoi créer des grilles ?

- **Permet une recherche O(g) et pas O(n)**

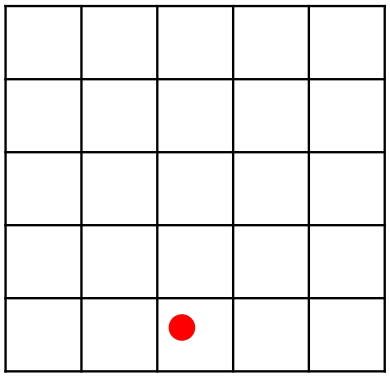
Création des grilles

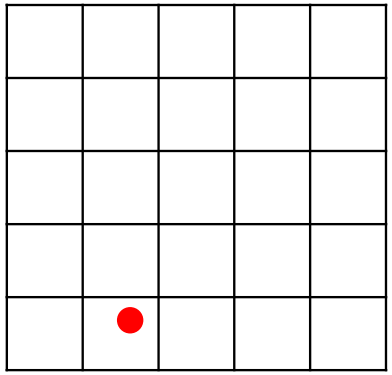


Intégration de ParSketch

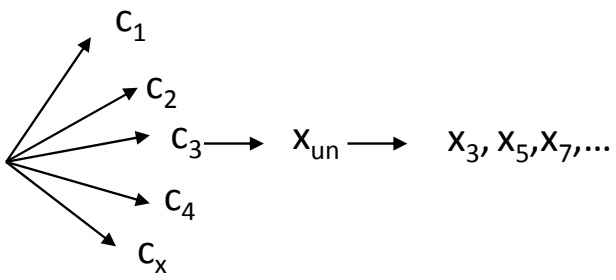
ParSketch :

Réduction de dimensions → Création des grilles → Recherche des voisins



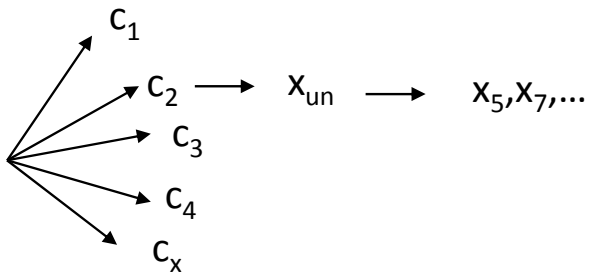


Grille 1



x_3	1
x_5	2
x_7	2

Grille 2



Min.grid

Application :

Problème de classification
Interface données massives/standards
Observation de l’impact des paramètres de Parsketch sur les voisins renvoyés.
Observation de la position de Parsketch par rapport à une situation globale et locale (sélection)

Données

Méthodes

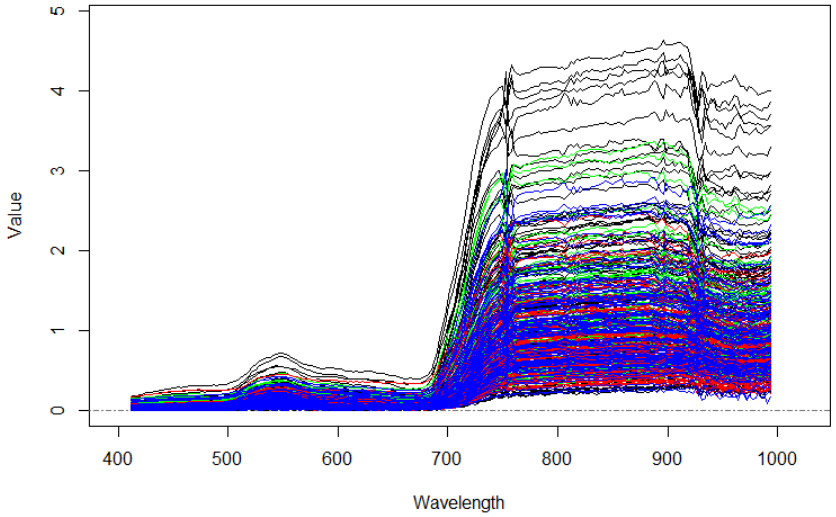
Résultats

Conclusions / perspectives

Application

1

Données



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

Objectifs

Discriminer les 4 classes (360 000 spectres/400 points tests)

2

Méthodes

Les 3 stratégies utilisées :

PLS-DA globale

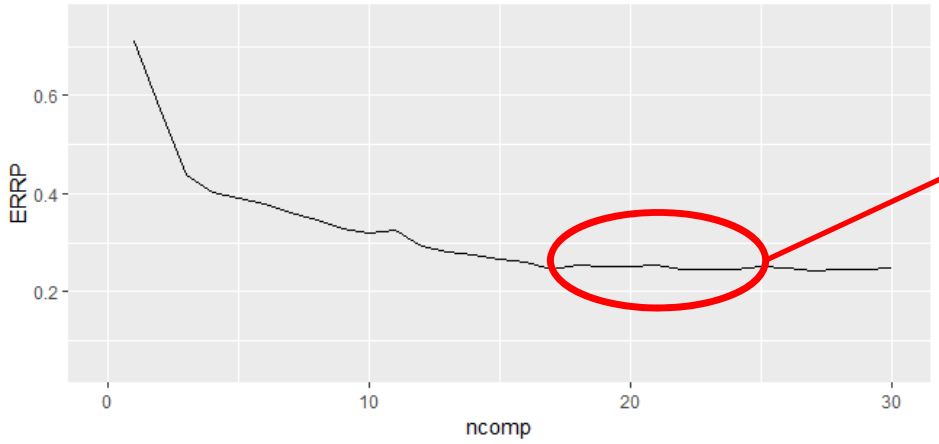
Knn(Brut force) -PLS-DA

ParSketch – PLS-DA

3

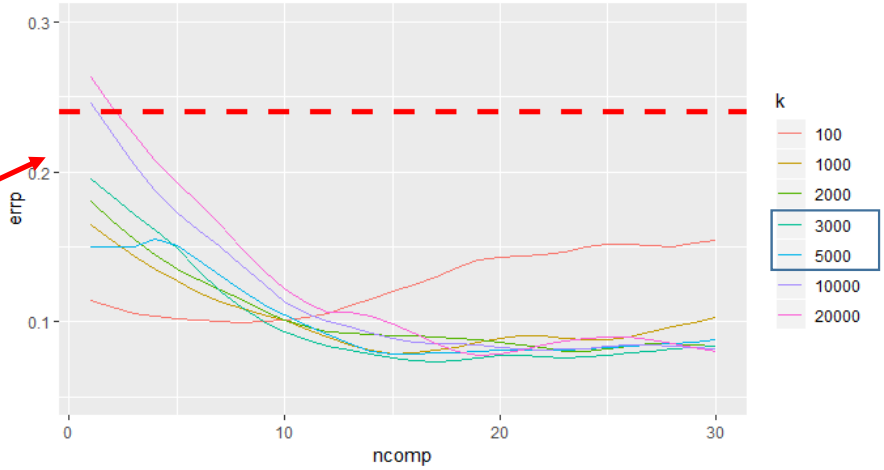
Résultats

PLS-DA :



Meilleur résultat :
Err min : 24 %

Knn(Brut force) PLS-DA :



Meilleur résultat :
Err min : 6-7 %
Temps : ~4 h

Observation :
Temps de calcul longs / bonnes prédictions

3

Résultats

ParSketch – PLS-DA :

3 paramètres :

Nb.vec : 10-100 vecteurs aléatoires

Nb.seg : 5-13 segments

Min.grid: 30-90 %

2 questions principales :

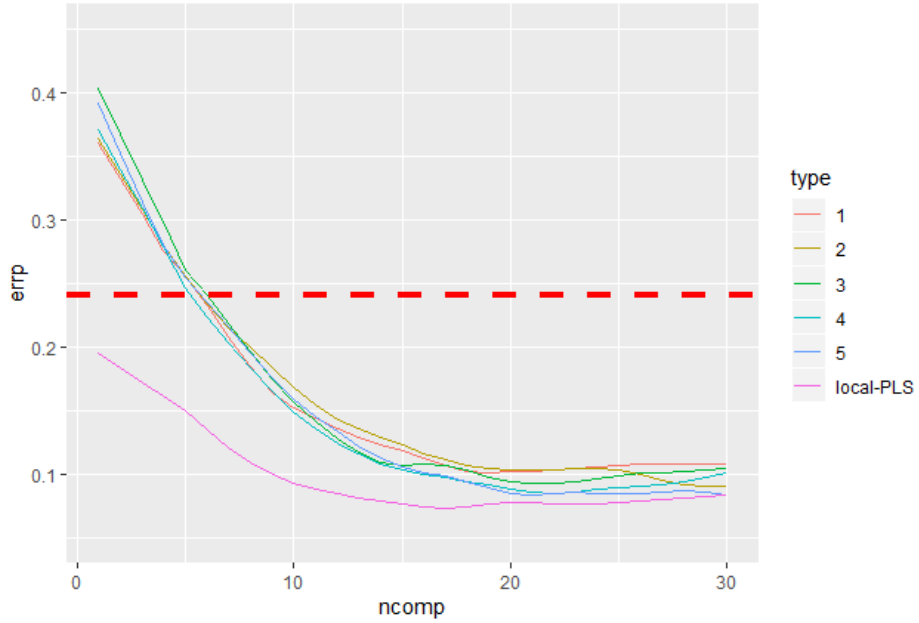
Quels sont les impacts des paramètres de Parsketch sur le voisinage renvoyé ?
Quels sont les impacts des paramètres de Parsketch sur l’erreur de prédiction ?

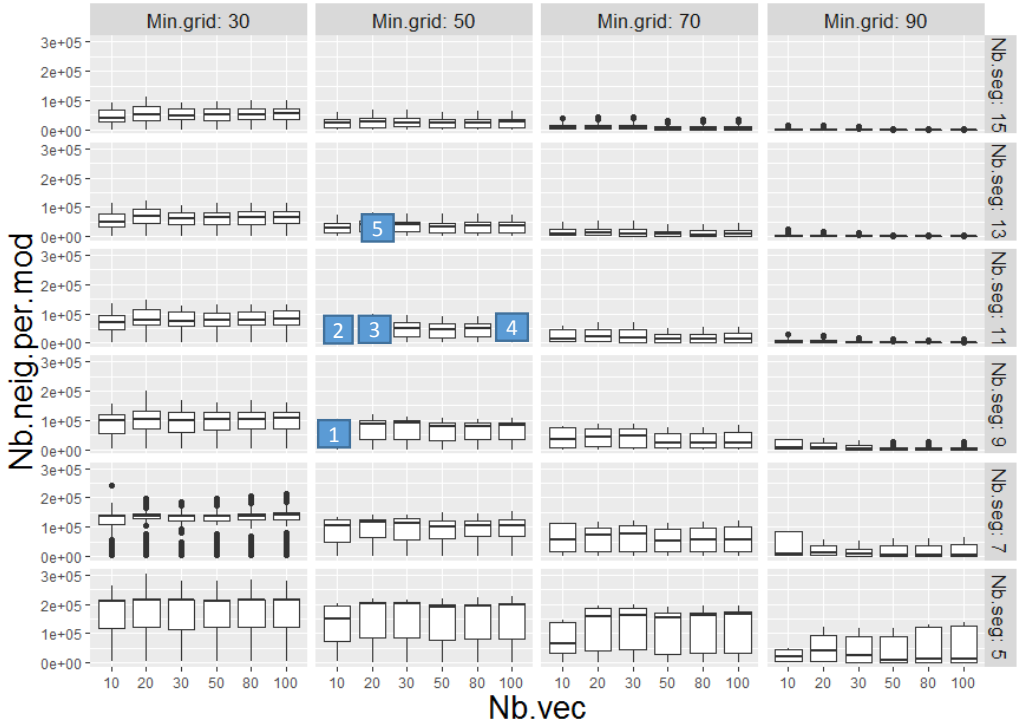
Etude de la méthode

3

Résultats

Quels sont les impacts des paramètres de sketch sur le voisinage renvoyé ?





Quels sont les impacts des paramètres de sketch sur l’erreur de prédiction ?

4

Conclusions / perspectives

Conclusions

- ParSketch offre une alternative aux méthodes locales (avec sélection).
- ParSketch impose une métrique (distance euclidienne). Il est donc possible que Parsketch ne réponde pas à tous les problèmes.

Perspectives

- ParSketch est utilisable dans n’importe quelle méthode utilisant des sous-ensembles
- Combiner Parsketch et une étape de pondération

MERCI !!

Présenté par: Maxime Metz
Supervisé par: Matthieu Lesnoff, Jean-Michel Roger