Introduction

Alain Franc

INRA BioGeCo & INRIA Equipe Pleiade

Saint-Pierre d'Oléron
ANF "Données massives"
2017

Cadre général

- Un ensemble de n objets
- n très grand
- décrits chacuns par p variables
- p très grand

Réduction de la dimension

On cherche une description fidèle de cet ensemble de n objets avec $r \ll p$ variables

Plusieurs objectifs

- visualiser les données : $\mathbb{R}^2$ ou $\mathbb{R}^3$, coordonnées parallèles, ...
- trouver des variables qui résument l'information
- approcher des estimations et inférences dans des modèles statistiques
- compresser les données (coût de stockage, de transfert, ...)
- avec une complexité de calcul raisonnable (temps, mémoire, ...)

Plusieurs objectifs

- visualiser les données : $\mathbb{R}^2$ ou $\mathbb{R}^3$, coordonnées parallèles, ...
- trouver des variables qui résument l'information
- approcher des estimations et inférences dans des modèles statistiques
- compresser les données (coût de stockage, de transfert, ...)
- avec une complexité de calcul raisonnable (temps, mémoire, ...)

Une géométrie non intuitive

En grande dimension, des phénomènes étranges se passent ...

- Le volume d'un cube est immense, il contient une très longue diagonale ($\ell = \sqrt{n}$)
- alors que le volume de la sphère se ratatine
- et bien d'autres choses ...

Mille méthodes ...

issu de https://en.wikipedia.org/wiki/Dimensionality_reduction

- PCA
- kernel PCA
- graph based kernel PCA
- linear discriminant analysis
- general discriminant analysis
- ...

Domaines d'application

- classifications (empreintes digitales, reconnaissance faciale, ...)
- traitement d'image
- classification et recherche dans les textes
  individus : les textes ; variables : les mots ; valeur : fréquence du mot
- bioinformatique : classification et recherche dans des mots très longs d'un alphabet de 4 ou 21 lettres
- liens avec la représentation par graphes (communautés sur graphes) : réseaux de neurones, de protéines, de gènes, métabolomique, etc ...
- réseaux sociaux et d'énergie (mails, portables, etc ...)
- visualisation des données
- POD (Principal Orthogonal Decomposition) : extraire des structures cohérentes dans un écoulement

Domaines d'application

- classifications (empreintes digitales, reconnaissance faciale, ...)
- traitement d'image
- classification et recherche dans les textes
  individus : les textes ; variables : les mots ; valeur : fréquence du mot
- bioinformatique : classification et recherche dans des mots très longs d'un alphabet de 4 ou 21 lettres
- liens avec la représentation par graphes (communautés sur graphes) : réseaux de neurones, de protéines, de gènes, métabolomique, etc ...
- réseaux sociaux et d'énergie (mails, portables, etc ...)
- visualisation des données
- POD (Principal Orthogonal Decomposition) : extraire des structures cohérentes dans un écoulement

Idée générale

Extraire d'un jeu de données une structure cohérente, le reste étant du bruit.
"core/periphery models" en structure de graphes

Focalisation

Histoire

Bien des méthodes sont classiques et ont une longue histoire :

- ACP : Pearson (1901) ; Hotelling (1933)
- MDS : Torgerson (1952)
- POD : Karhunen (1946) ; Loève (1955)

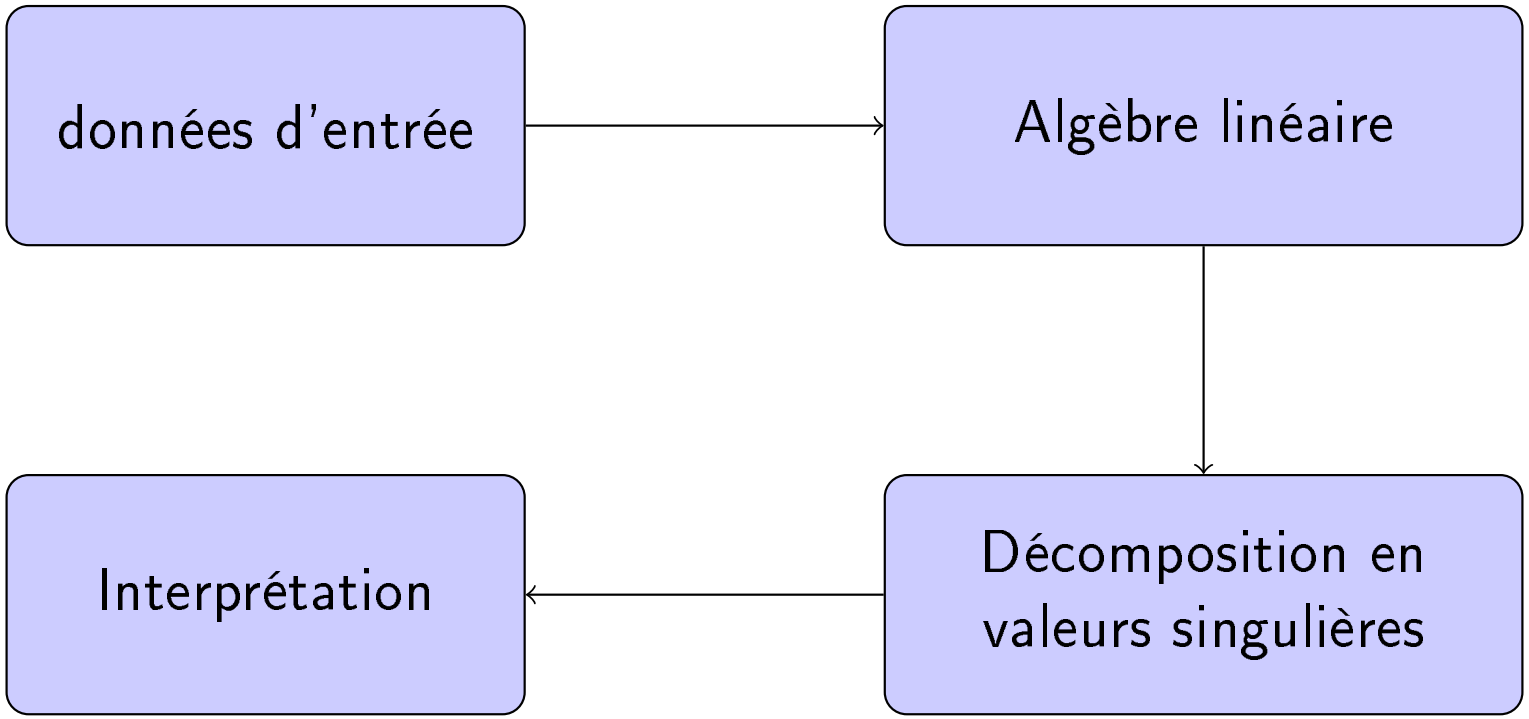
Techniquement

Ces méthodes se ramènent à

- la recherche de valeurs et vecteurss propres (EVD)
- la décomposition en valeurs singulières (SVD)

qui sont de complexité $\mathcal{O}(n^3)$

Démarche générale



Déroulé de la formation