Apprentissage non supervisé

APM_4STA3_TA - M1 Mathématiques Appliquées

Partie I

Christine Keribin

Laboratoire de Mathématiques
Université Paris-Saclay

2024-2025

université PARIS-SACLAY
FACULTÉ DES SCIENCES D'ORSAY

ENSTA IP PARIS

Sommaire

Apprentissage

- Modification d'un comportement sur la base d'une expérience [1]
  $\hookrightarrow$ série d'expériences pour augmenter la réussite à la tâche demandée
- apprentissage machine : capacité à apprendre une procédure qui n'est pas directement programmée [2]
  $\hookrightarrow$ utiliser les données d'entrée et les réponses du système afin de produire une procédure apprise faisant le (meilleur) lien entre les deux

algorithme d'apprentissage + données $\rightarrow$ modèle appris

$\hookrightarrow$ méthode d'apprentissage $\neq$ modèle appris

1. Fabien Benureau (2015)
2. Arthur Samuel (1959)

Apprentissage

Définition (Tom Mitchell (1997) `http://www.cs.cmu.edu/~tom/`)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

$\hookrightarrow$ problème d'optimisation

Apprentissage

Définition (Tom Mitchell (1997) `http://www.cs.cmu.edu/~tom/`)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

$\hookrightarrow$ problème d'optimisation

- Apprentissage machine (*machine learning*), apprentissage automatique,...
- Modéliser un phénomène à partir d'exemples

Apprentissage

Modéliser un phénomène à partir d'exemples pour résoudre des problèmes

- qu'on ne sait pas résoudre directement
- dont on ne sait pas formaliser la résolution algorithmique
- qu'on sait résoudre mais d'exécution très coûteuse

Convertir une expérience en expertise et connaissance

Machine Learning : exemples typiques

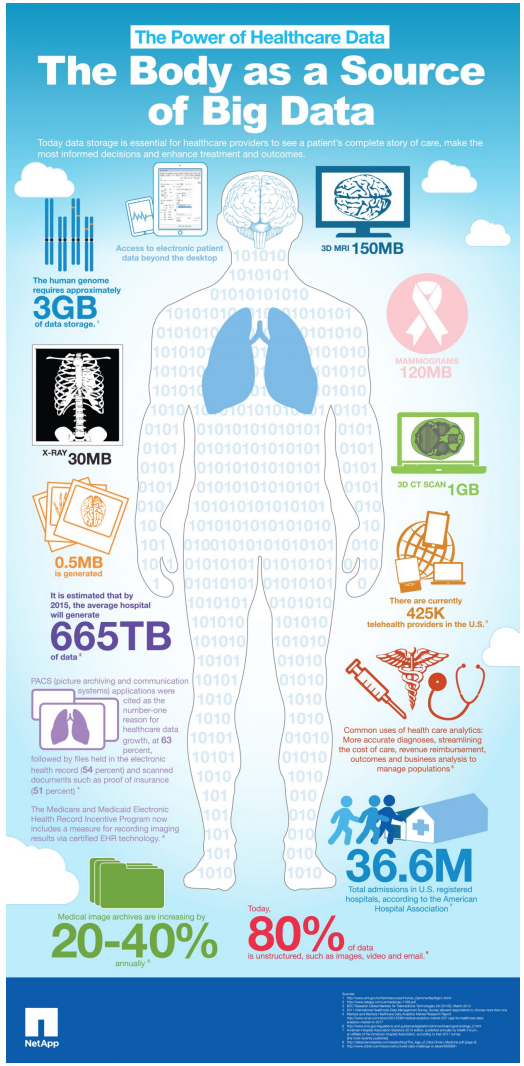
- filtrage de spam
- reconnaissance de caractères, d'objets, de visages, etc
- traitement automatique du langage
- analyse de scène
- système de recommandation
- moteurs de recherche
- détection de communautés
- ...

Machine Learning

Un composant dans des applications de plus en plus nombreuses

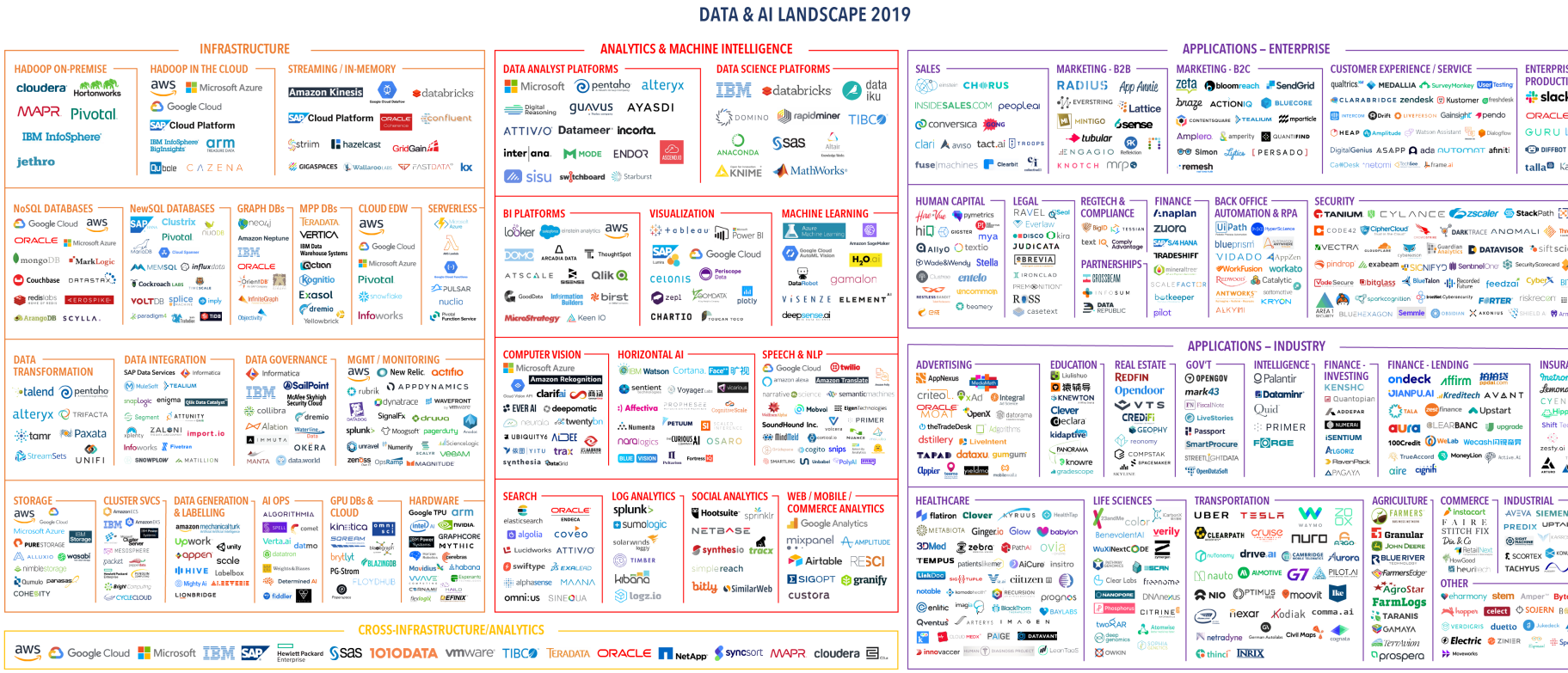
- publicité
- santé, bioinfo, analyse ADN
- cybersécurité
- cités intelligentes
- véhicule autonomes
- sport
- ...

↪ Big Data : Volume, Variété, Vitesse, Véracité (qualité), Visualisation, Valeur...



https://datafloq.com/read/body-source-big-data-infographic/

Panorama des data et de l'IA [3]



© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap) mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

3. Matt Turck 2019

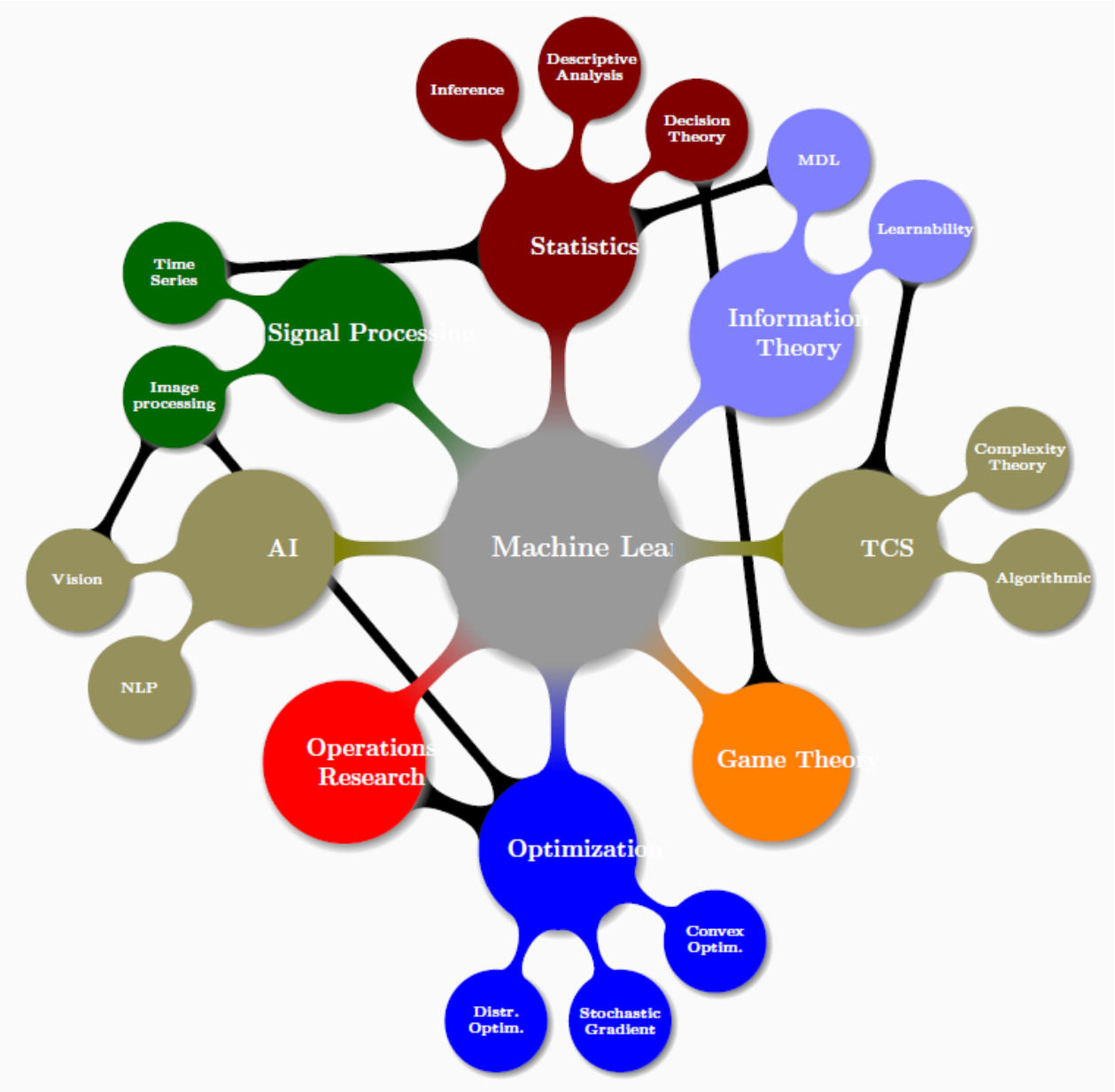
Modèles de mélange

Data science ou statistique ?

Statistique / Statistical learning / machine learning / IA

- apprentissage statistique : apprentissage automatique (ML) à l'aide d'outils et de garanties inspirés des statistiques
- ML : les modèles (s'il y en a) sont instrumentaux
- IA : imiter les capacités cognitives des humains
- Data science : étude de l'extraction généralisable de connaissance à partir de données
  $\hookrightarrow$ data scientist/ML engineer : mélange de compétences en Math/Statistics/ML/optim, informatique (data management, programmation, visualisation, interface), Business et communication

Eco-système de l'apprentissage



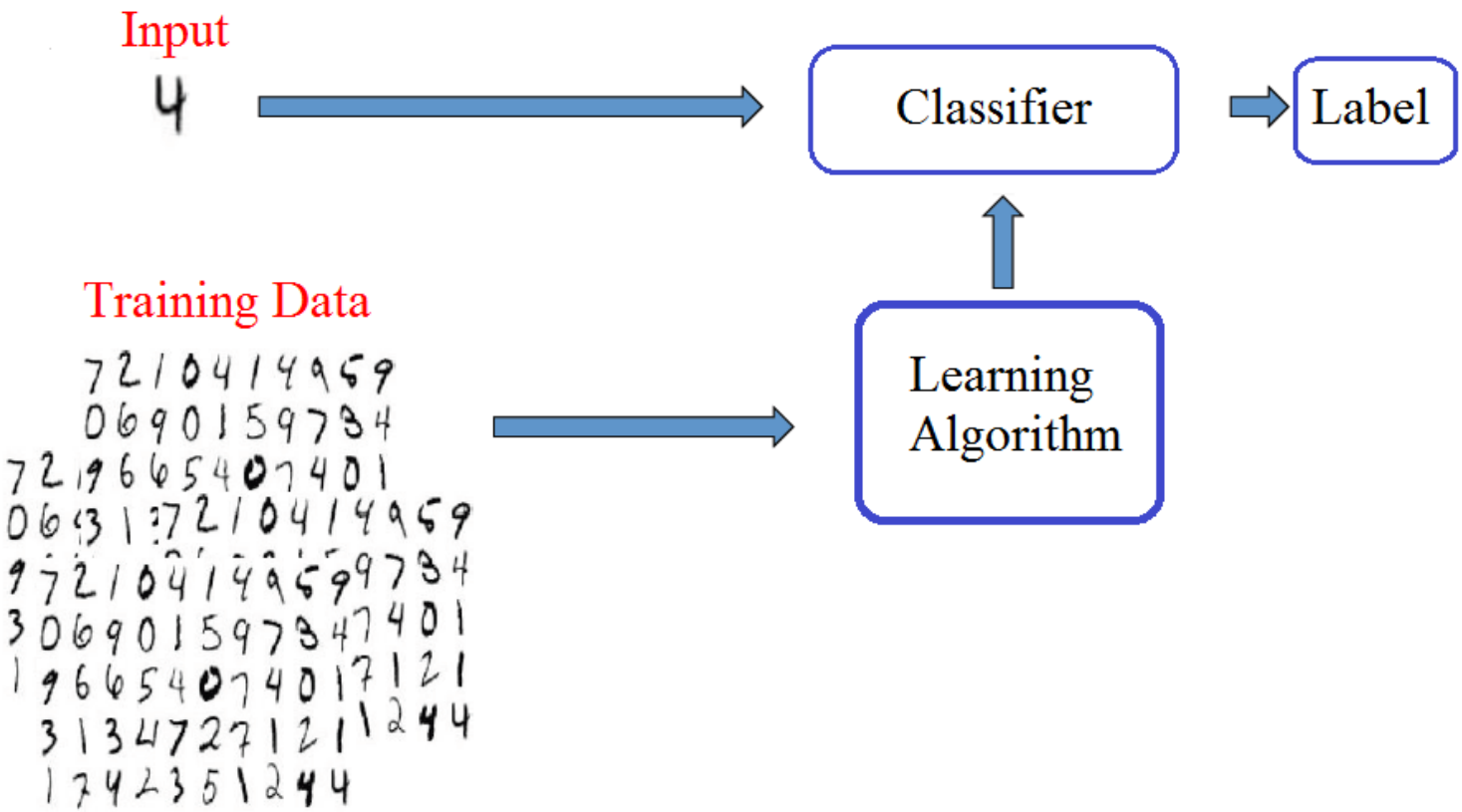
Crédit : Aurélien Garivier

Apprentissage(s)

- supervisé : construire une règle de décision à partir de couples (X, Y) connus pour estimer ou prédire Y^{new} à partir X^{new}
- non supervisé : étudier un ensemble d'observations multivariées, dont aucune variable ne joue un rôle particulier
- semi supervisé : apprendre des étiquettes à partir d'un jeu de données partiellement étiqueté
- par renforcement : le système maximise la récompense de ses actions

Apprentissage supervisé

Construire un modèle statistique à partir de couples (X, Y) connus pour estimer ou prédire Y^{new} à partir X^{new}



Apprentissage supervisé

- Données d'apprentissage :
  $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbf{P}$)
- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$ mesurable
- Fonction de perte ou de coût : $\ell(f(\mathbf{X}), Y)$ mesure avec quelle qualité $f(\mathbf{X})$ "prédit" Y, d'où le risque :

$$\mathcal{R}(f) = \mathbb{E}\ell(Y, f(\mathbf{X}))$$

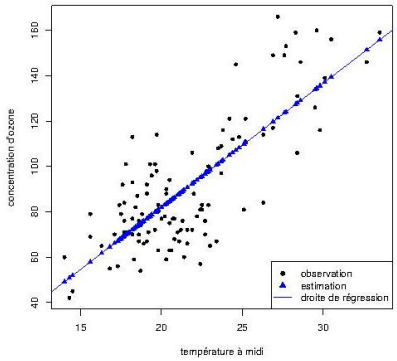
$\hookrightarrow$ souvent $\ell(f(\mathbf{X}), Y) = \|f(\mathbf{X}) - Y\|^2$ ou $\ell(f(\mathbf{X}), Y) = \mathbf{1}_{Y \neq f(\mathbf{X})}$

But : apprendre une règle (régresseur/classifieur) $\widehat{f} \in \mathcal{F}$ pour prédire (X^{new}, Y^{new}) à partir des données d'apprentissage $\mathcal{D}_n$ avec un risque $\mathcal{R}(\widehat{f})$ petit en moyenne ou avec grande probabilité par rapport à $\mathcal{D}_n$.

Régression linéaire

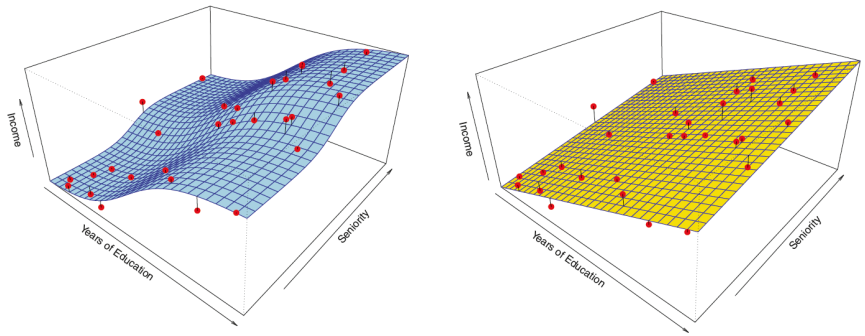
$Y_{i|X_i=x_i} = x_i\theta + \varepsilon_i$ où $\varepsilon_{|X=x} \sim_{i.i.d.} \mathcal{L}(0, \sigma^2 I_n)$.

- simple



Source : Cornillon et Matzner-Løber (Springer, 2007)

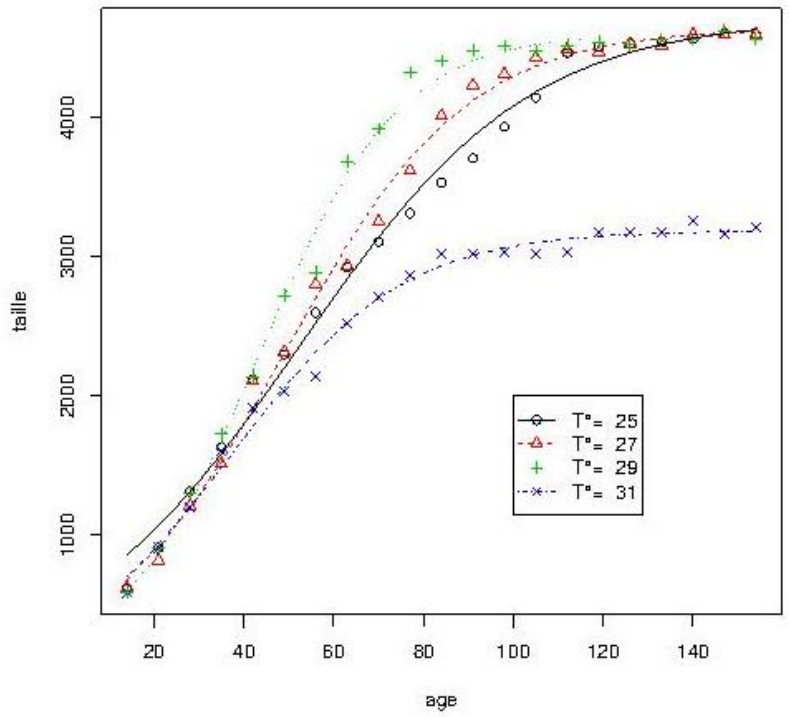
- multiple



Source : An Introduction to Statistical Learning : James, Witten, Hastie, Tibshirani (Springer)

Régression non linéaire

$$Y_{i|X_i=x_i} = m(x_i, \theta) + \varepsilon_i \text{ où } \varepsilon_{|X=x} \sim_{i.i.d.} \mathcal{L}(0, \sigma^2 I_n)$$



Dans l'exemple $m(x) = k/(1 + \frac{k-n_0}{n_0} e^{-rx})$

Classification. Ex : Credit Scoring

Modèles de décision binaire permettant d'aider à assurer un crédit à la consommation :
prédire $\mathbb{P}(Y = \text{non remboursement}|x)$

$$Y_{i|X_i=x_i} \sim \mathcal{B}(\pi(x_i))$$

avec, par exemple, $\text{logit}\,\pi(x_i) = \mathrm{x}_i\theta$ et x_i un vecteur de l'ordre de 100 à 200 covariables qualitatives ou quantitatives

- Quel est l'effet d'une covariable sur la réponse (augmente ou diminue la probabilité) ?
- Quelle est la forme de $\pi(x)$?
- Comment estimer $\pi(x)$?
- Toutes les covariables sont-elles utiles ?
- A quelle classe affecter une nouvelle observation ?

Bien d'autres applications...

Méthodes d'apprentissage supervisé

- linéaires, linéaires généralisées, non linéaires
- GAM
- classification : analyse discriminante, Bayes Naïf, plus proches voisins, SVM, ...
- arbres de décision
- méthodes d'ensemble (bagging, boosting, forêts aléatoires)
- réseaux de neurones
- ... du statistical learning au machine learning ...

Apprentissage supervisé : problématiques communes

- Compromis biais variance entre la qualité d'ajustement et la précision de la prédiction
  $\hookrightarrow$ comparaison de modèle
  $\hookrightarrow$ choix de modèle, sélection de variables
- Mesurer la performance du modèle
  $\hookrightarrow$ validation croisée, bootstrap

Apprentissage non supervisé

Pas de variable prioritaire, toutes les variables sont considérées au même niveau

- Données d'apprentissage :
  $\mathcal{D} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ (i.i.d. $\sim \mathbb{P}$)
- Tâche : ? ? ? pas de tâche évidente !
- Mesure de performance : ? ? ?

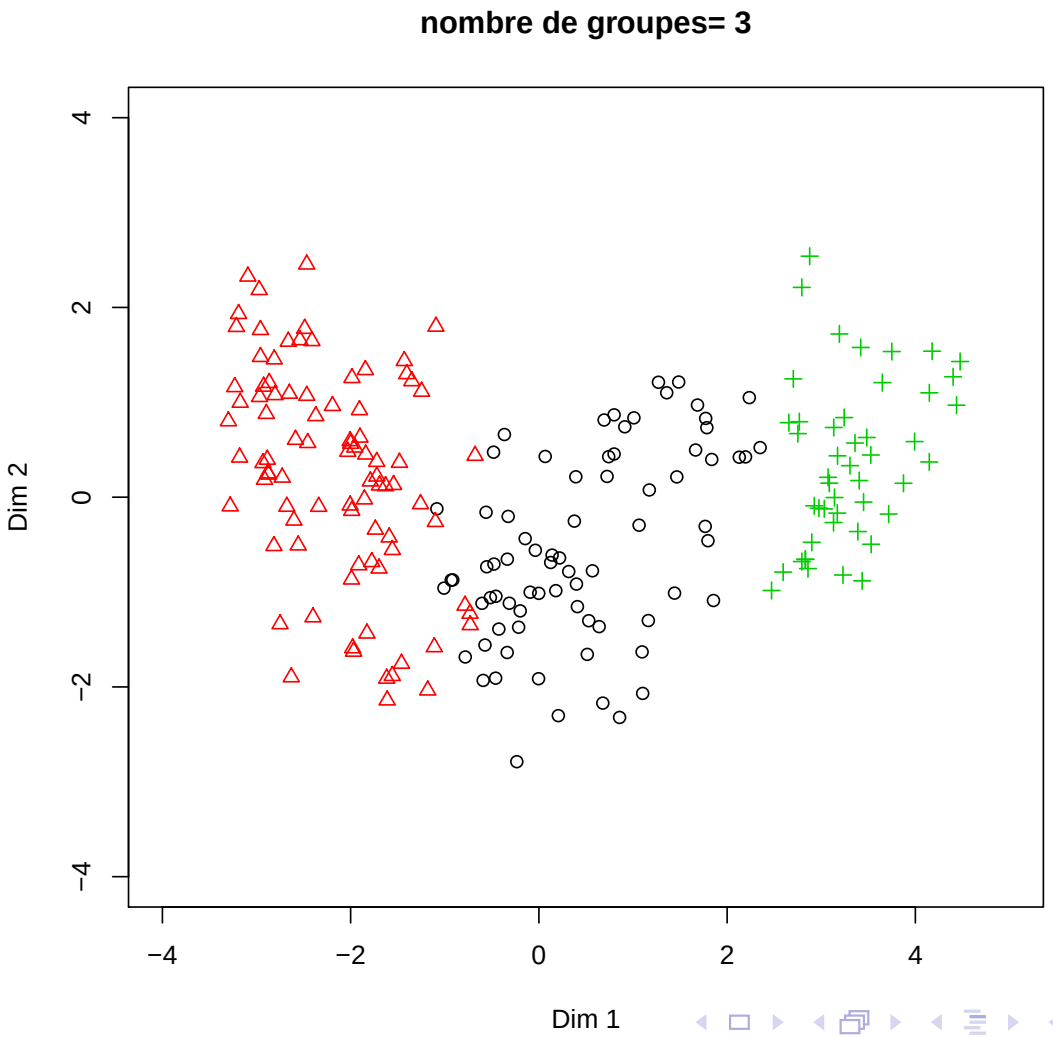
On ne cherche pas à prédire une variable, mais à découvrir une structure cachée dans les données ou les résumer

Apprentissage non-supervisé

- Réduction de dimension : construire une relation avec des données dans un espace de plus petite dimension sans trop les déformer.
  - ↪ Analyse exploratoire multidimensionnelle, méthodes factorielles : ACP, AFC, ACM
  - ↪ visualisation et interprétation, étape préliminaire d'étapes supervisées ultérieures
- Classification non supervisée (ou clustering) : construire des groupes de données homogènes.
  - ↪ K-moyennes, classification hiérarchique, modèles de mélange, ...

Apprentissage non supervisé

```
> head(seeds)
   area perimeter compact length width  asym lgroove
1 15.26     14.84  0.8710  5.763 3.312 2.221   5.220
2 14.88     14.57  0.8811  5.554 3.333 1.018   4.956
3 14.29     14.09  0.9050  5.291 3.337 2.699   4.825
4 13.84     13.94  0.8955  5.324 3.379 2.259   4.805
5 16.14     14.99  0.9034  5.658 3.562 1.355   5.175
6 14.38     14.21  0.8951  5.386 3.312 2.462   4.956
```



Contenu STA203

- Apprentissage non supervisé :
  - ↪ méthodes factorielles
  - ↪ classification non supervisée ou *clustering*
- Apprentissage supervisé
  - ↪ Inférence en régression (logistique)
  - ↪ Classification
  - ↪ Choix de modèle et sélection de variables
  - ↪ Mesure de la performance d'un modèle
  - ↪ Méthodes de régularisation (ridge, lasso)
  - ↪ SVM
- Applications pratiques avec le logiciel R
  - ↪ Comparaison de méthodes

Prérequis : bases de la statistique inférentielle ; modèle linéaire ; les bases du langage R

Documents sur e-campus !

Objectifs STA203

Être capable de dérouler une méthodologie d'apprentissage statistique, supervisée ou non supervisée, sur un jeu de données multivarié

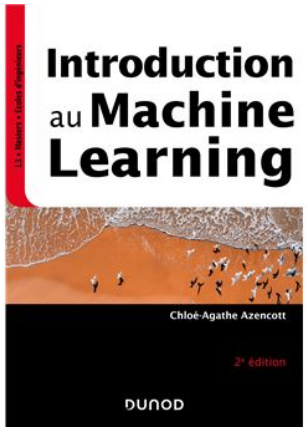
- Définir le machine learning, en identifier les principaux composants et reconnaître les principales méthodes
- Connaître les principes théoriques des méthodes étudiées
- Savoir les utiliser à bon escient
- Mettre en place une stratégie de choix de modèle
- Être capable d'appréhender une nouvelle méthode
- Utiliser un logiciel statistique pour mettre en œuvre les analyses et interpréter les résultats

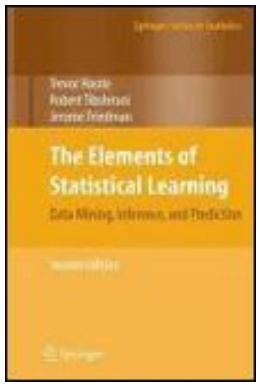
Evaluation STA203

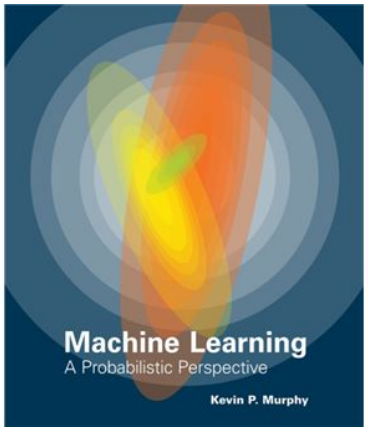
- non supervisé, un EX noté sur 20 (poly et calculette autorisés) (15 avril)
- supervisé : un PR en binôme noté sur 20
  Énoncé donné le 6 mai, à rendre pour le 25 mai

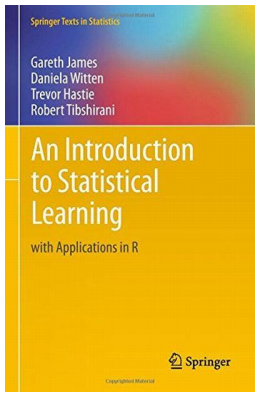
Note finale = 0.4*EX+0.6*PR

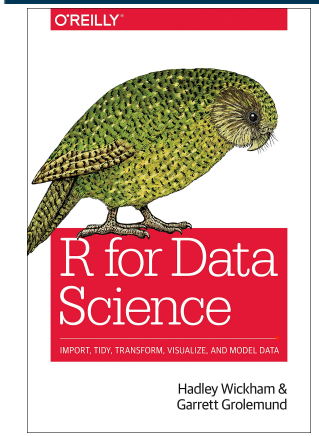
Quelques références générales











Sommaire

Références


J. Pagès.
Statistique générales pour utilisateurs.
Presses Universitaires de Rennes, Rennes, 2005.


Husson F., Lê S., Pagès J..
Analyse de données avec R .
Presses Universitaires de Rennes, Rennes, 2009.


Lebart L., Morineau A., Piron M.
Statistique exploratoire multidimensionnelle
Dunod, 2004

Statistique Exploratoire Multidimensionnelle

Soit X un tableau (`data.frame`)

- lignes : individus $i = 1, \ldots, I$
- colonnes : variables $k = 1, \ldots, K$

$$x_{ik} = x_i[k] = X_k[i]$$

Objectif : outils pour explorer et visualiser de façon simplifiée des (grands) tableaux, tout en conservant le maximum d'information : interprétation et réduction de dimension

- Exploratoire $\leftrightarrow$ inférentielle
- Multidimensionnelle $\leftrightarrow$ unidimensionnelle

$\hookrightarrow$: analyse de données, analyse factorielle

Statistique Exploratoire Multidimensionnelle

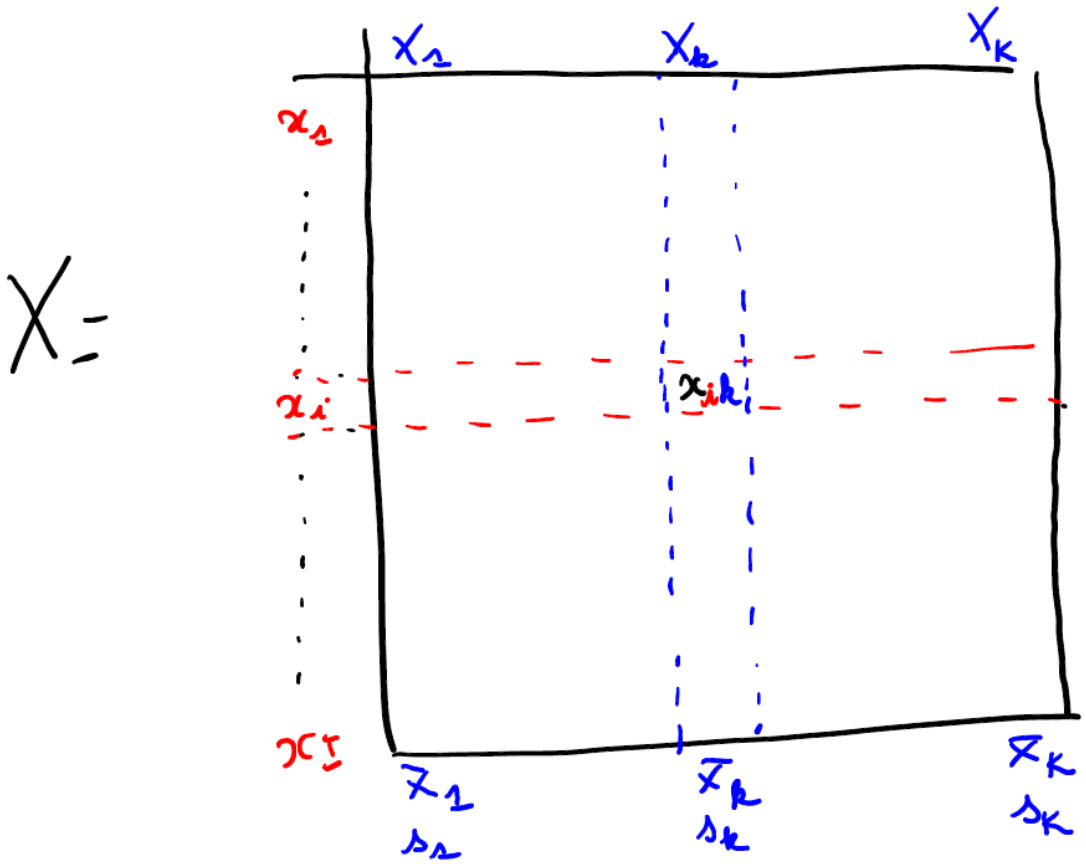
Une même base mathématique : définition de distance, diagonalisation de matrices, projection. Les techniques diffèrent suivant les types de variables considérées

- ACP : variables quantitatives uniquement
- AFC : deux variables qualitatives à I et K niveaux respectivement
- ACM : strictement plus de deux variables qualitatives.

Sommaire

ACP

- Toutes les variables sont quantitatives, mais de nature (potentiellement) très différente. Exemples !



Objectifs ACP

- Dépasser les études univariée et bivariée

$$\bar{X}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}, \quad s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{X}_k)^2}$$

$$cor(X_k, X_{k'}) = \frac{\sum_i (x_{ik} - \bar{X}_k)(x_{ik'} - \bar{X}_{k'})/I}{s_k s_{k'}}$$

- Analyse exploratoire, pas de variable particulière à expliquer en fonction des autres.
  - ↪ dans quel repère représenter les données ?
  - ↪ comparaison des individus en fonction de leur profil de réponse sur l'ensemble des variables
  - ↪ comparaison des variables entre elles
  - ↪ mettre en rapport les individus et les variables

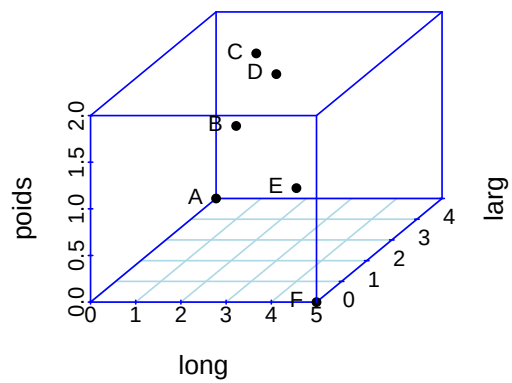
Exemple ACP

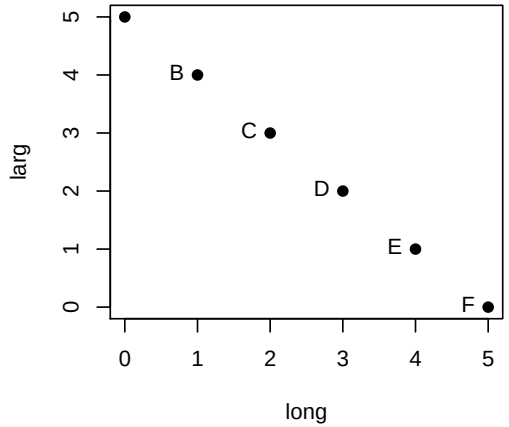
- Exemple [4]

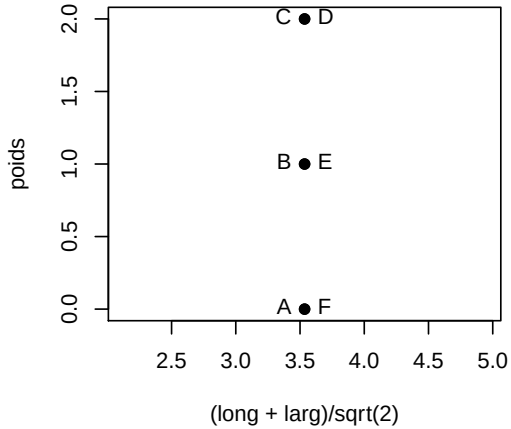
nom	*Long.*	*Larg.*	*Poids*
A	0	5	0
B	1	4	1
C	2	3	2
D	3	2	2
E	4	1	1
F	5	0	0
moy	2.5	2.5	1
e.t.	1.708	1.708	0.816

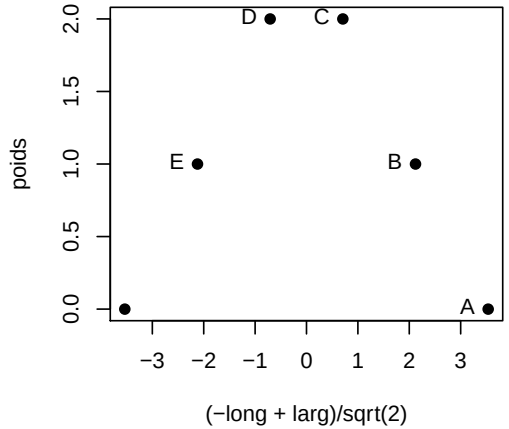
4. Pagès (2005)

Exemple ACP : dans quel plan visualiser ?

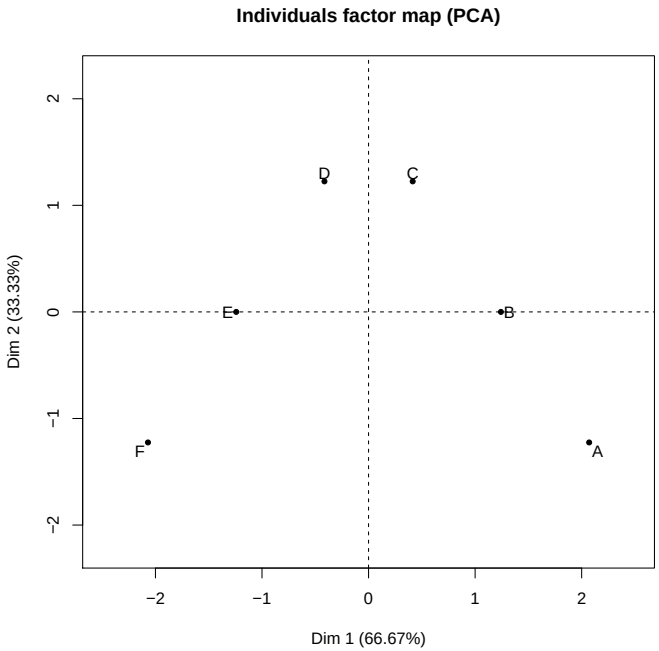


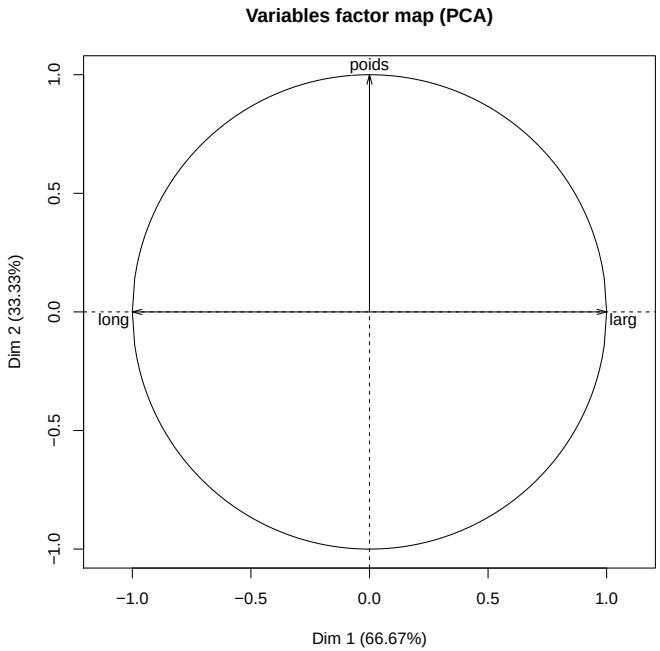




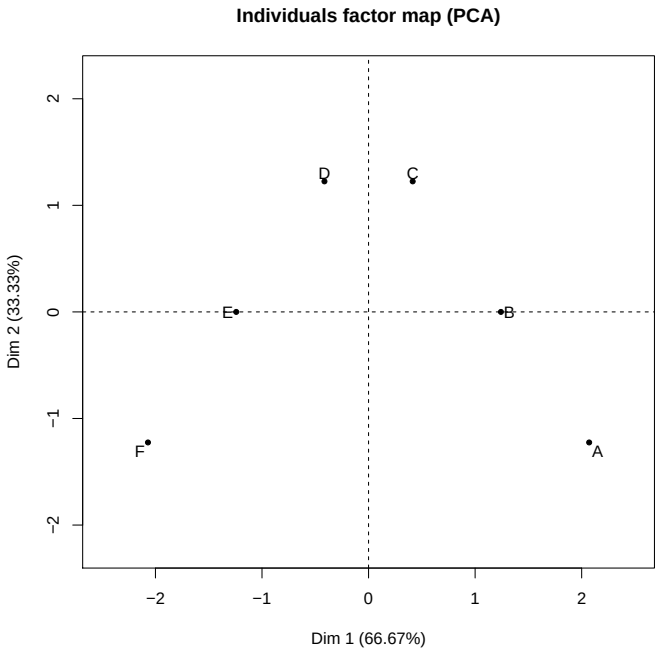


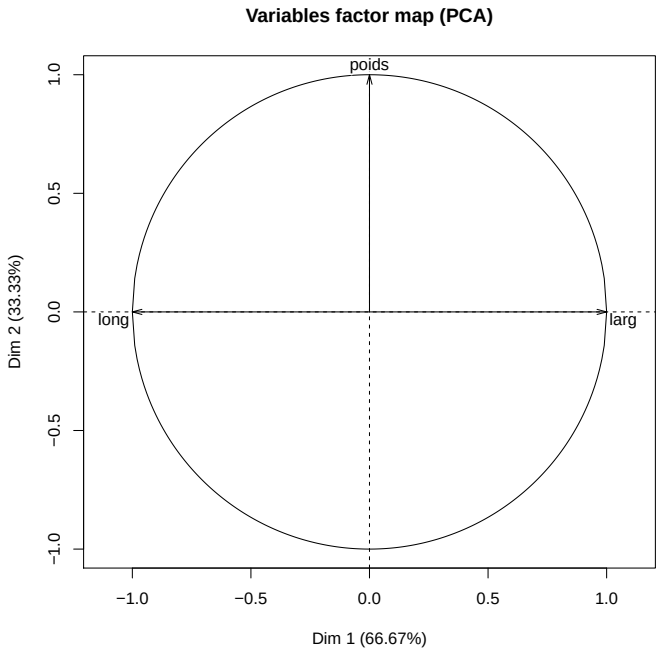
Exemple ACP : et la visualisation des variables ?





Exemple ACP : et la visualisation des variables ?





Teaser : l'ACP permet de trouver le plan qui déforme le moins le nuage projeté, et de faire une interprétation croisée entre les individus et les variables

Deux transformations préliminaires

Le vecteur des moyennes $(\bar{X}_k) \in R^K$ est le centre de gravité G_I du nuage où chaque individu est affecté du poids $1/I$

- Le centrage $x_{ik} - \bar{X}_k$ déplace G_I à l'origine O du repère. Ne modifie pas la forme du nuage, et sera toujours effectuée.
- La réduction $(x_{ik} - \bar{X}_k)/s_k$ modifie la forme du nuage en harmonisant sa variabilité dans toutes les directions de base (ACP normée). C'est indispensable quand les variables ne s'expriment pas dans les mêmes unités, souhaitable pour donner le même poids à toutes les variables.
- Tous les individus ont a priori le même poids $\frac{1}{I}$. Généralisation à un poids quelconque.

Dans la suite, on suppose l'ACP normée et équipondérée.

Nuage des individus N^I dans $\mathbb{R}^K$

- Considérer simultanément les valeurs des K variables d'un individu i
  - $\hookrightarrow$ un individu = une ligne : un point dans l'espace R^K
  - $\hookrightarrow$ choix de la distance euclidienne entre les individus i et i'

$$d^2(i,i') = \sum_k (x_{ik} - x_{i'k})^2 = ||x_i - x_{i'}||^2$$

- Analyser la variabilité entre les individus = étudier l'ensemble des distances interindividuelles

Nuage des individus N^I dans $\mathbb{R}^K$

- Considérer simultanément les valeurs des K variables d'un individu i
  - $\hookrightarrow$ un individu = une ligne : un point dans l'espace R^K
  - $\hookrightarrow$ choix de la distance euclidienne entre les individus i et i'

$$d^2(i,i') = \sum_k (x_{ik} - x_{i'k})^2 = ||x_i - x_{i'}||^2$$

- Analyser la variabilité entre les individus = étudier l'ensemble des distances interindividuelles
  $\hookrightarrow$ lié à l'inertie du nuage N_I, c'est à dire sa forme.

$$Inertie = \sum_i \frac{1}{I} ||x_i||^2$$

Objectifs

- Réduction de dimension : Définir une direction (un plan, un sous-espace de dimension donnée) qui permet de déformer le moins la projection du nuage parmi toutes les directions (plans, sous-espace de dimension finie)
  ↪ approximation
- Définir des profils type qu'on tentera d'interpréter en liaison avec des caractéristiques individuelles profondes (facteurs), interprétées à partir des variables
  ↪ apprentissage de représentations

Ajustement du nuage des individus

- u : vecteur unitaire de la direction cherchée
- $M_i = (x_{ik})_{k=1,\dots,K}$ un point du nuage N_I de R^K
- H_i la projection de M_i sur u dans $\mathbb{R}^K$: la distance signée $OH_i = X[i,]u = \sum_{k=1}^{K} x_{ik} u_k$;

$$\begin{pmatrix} OH_1 \\ \vdots \\ OH_I \end{pmatrix} = \begin{pmatrix} x_1 u \\ \vdots \\ x_I u \end{pmatrix} = Xu$$

Objectif

Trouver u normé rendant maximum le critère

$$Inertie(u) = \frac{1}{I} \sum_i OH_i^2 = \frac{1}{I} u' X' X u = u' C u$$

où $1/I$ est le poids de chaque individu et C la matrice de corrélation. $\hookrightarrow$ La projection du nuage sur cette droite est donc d'inertie maximum (de déformation minimum)

Ajustement du nuage des individus suivant un axe

Proposition

Le vecteur u solution est un vecteur propre normé associé à la plus grande valeur propre de C

Ajustement du nuage des individus suivant un axe

Proposition

Le vecteur u solution est un vecteur propre normé associé à la plus grande valeur propre de C

- u maximise l'inertie ou "variance expliquée".
- L'axe de projection u permet de déformer le moins possible la forme du nuage.

Remarque : différent de la régression linéaire !

Ajustement du nuage des individus suivant un plan

Objectif : chercher le plan maximisant l'inertie projetée.

Plan d'inertie maximum :

- u_1 : définit le meilleur axe (vecteur propre de plus grande valeur propre)
- u_2 : orthogonal à u_1 et qui exprime le plus de variabilité
  $\hookrightarrow$ u_2 est un vecteur propre normé associé à la 2ème plus grande valeur propre.

Reconstruction par ACP

Les axes de l'ACP sont obtenus par diagonalisation de la matrice de corrélation.

- valeurs propres = inertie projetée :
  $\lambda_1 \geq \cdots \geq \lambda_j \geq \ldots \geq \lambda_s$

  Pourcentage d'inertie associé à l'axe u_j : $\dfrac{\lambda_j}{\sum_j \lambda_j}$.
- Si l'ACP est normée : Inertie de $N_I = K$
- composantes principales : $F_j = Xu_j = \begin{pmatrix} x_1 u_j \\ \vdots \\ x_I u_j \end{pmatrix}$,

  coord. des I individus sur l'axe j
- Reconstruction : $X = \sum_{j=1}^{K} F_j u_j'$

Aide à l'interprétation

- Pourcentage d'inertie associé à un axe
  - ↪ qualité globale de représentation du nuage (variabilité exprimée).
  - ↪ importance relative des axes, choix du nombre de directions
- Contribution d'un individu à un axe : dans l'inertie associée à un axe, la part de chaque individu

$$\frac{\frac{1}{I}(OH_i^j)^2}{\lambda_j} \times 100$$

Individus remarquables : loin de l'origine $(OM_i)^2$ grand.

- Qualité de représentation d'un individu (*cos carré*).

$$\frac{(OH_i^j)^2}{(OM_i)^2} = \cos^2(\theta_i^j)$$