COMMENT GERER DES BASES DE DONNEES IMPORTANTES ET/OU UN NOMBRE DE VARIABLES IMPORTANT?

Modèles à composantes multiples ou multidimensionnels

Septembre 2022
Diane Haering

UNIVERSITÉ RENNES 2
UFR STAPS

Rappels

- Statistiques descriptives
- Statistiques analytiques
- Statistiques prédictives → modèles

Rappels

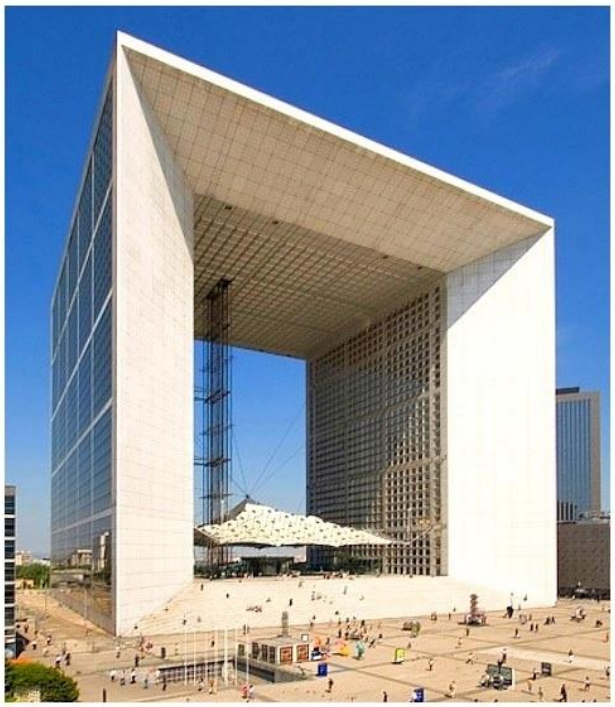
- Statistiques descriptives
  - Objets: dénombrement, centralité, dispersion
  - Outils: comptage, moyenne, écart type…
- Statistiques analytiques
  - Objets: comparaison, relations, facteurs, classification
  - Outils: t-test, anova, corrélation, ACP, machine learning
- Statistiques prédictives
  - Objets: Interpolation, extrapolation
  - Outils: régression, machine learning

Modèle linéaire multiple

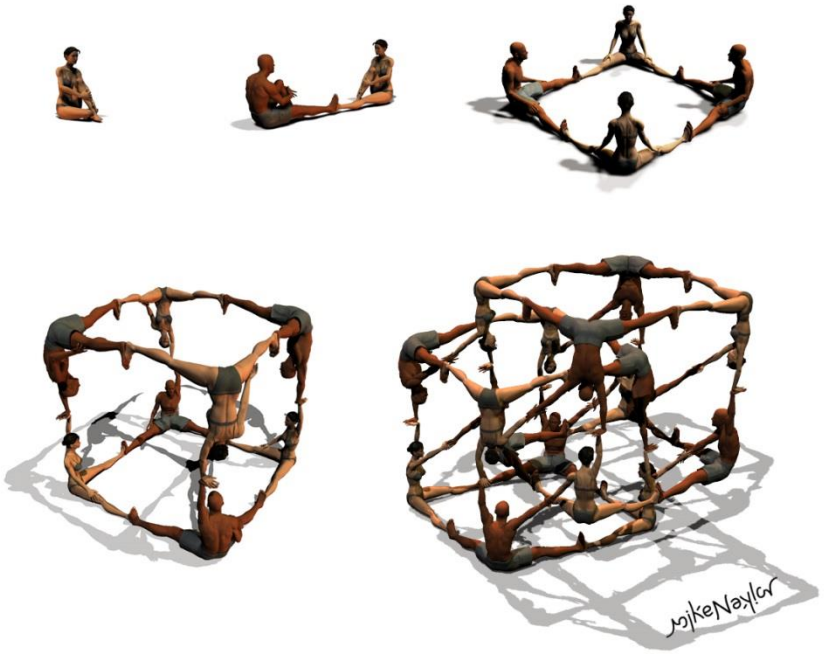
- Modèle linéaire 2D
  - 1 variable dépendante
  - 1 variable indépendante
- Modèle linéaire N-D
  - 1 variable dépendante
  - (N-1) variables indépendantes
  - $y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_{N-1} x_{N-1}$

Modèle linéaire multiple

- Modèle linéaire N-D
  - Calcul très simple
  - Mais difficulté de représentation (pour N>3)

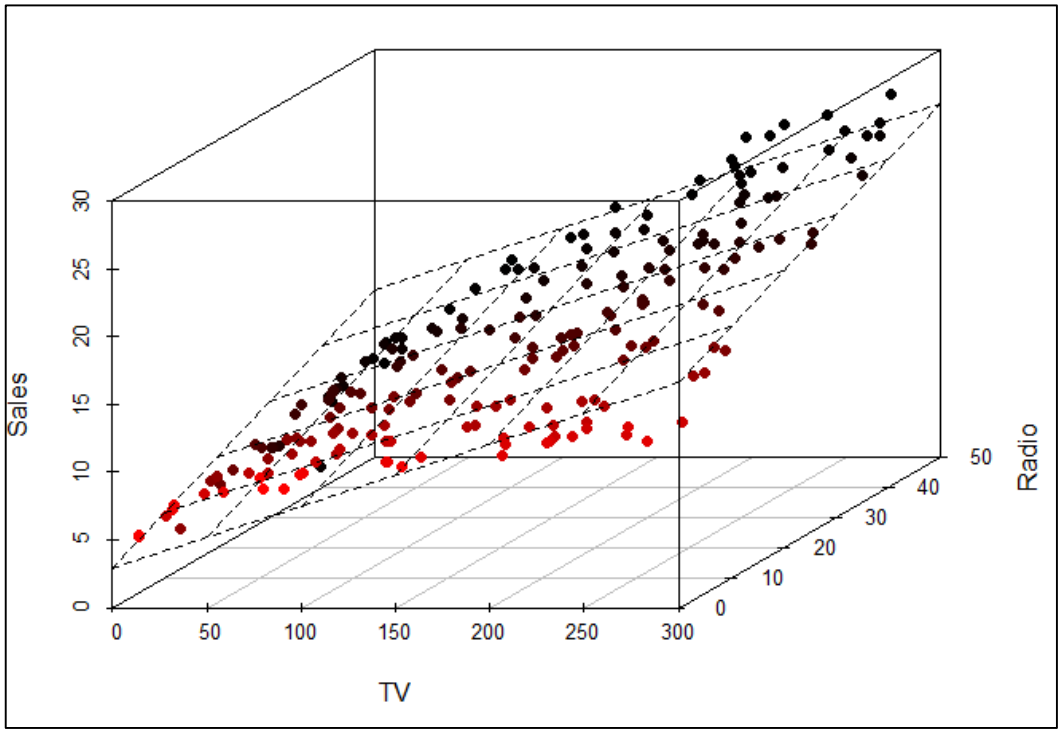


Arche de la défense:
ombre en 3D d'un cube 4D



Modèle linéaire multiple

- Modèle linéaire N-D : pour quoi faire?
  - Une variable dépendante peut être liée à plusieurs variables indépendantes



Ventes en fonction de la diffusion publicitaire à la télé et à la radio

Modèle linéaire multiple

- Beaucoup de variables ➔ comment trier?

- Exemple, dans 60 villes industrielles:
  - Mortalité → variables dépendante
  - 15 variables indépendantes
    - Météorologiques (précipitations, températures,…)
    - Sociologiques (âge moyen, études, revenus,…)
    - Niveaux de pollution

Modèle linéaire multiple

- Trier ?
  - Toutes les variables sont-elles utiles?
  - Le sont-elles au même niveau?

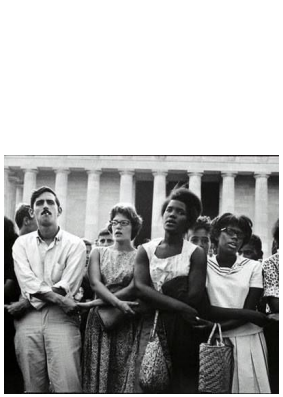
- Méthodes usuelles
  - Régression pas-à-pas
  - Meilleur sous-ensemble

RÉGRESSION PAS-A-PAS

Ascendante ou descendante

Modèle linéaire multiple

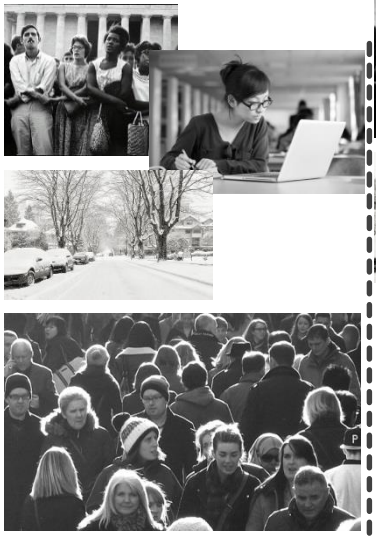
- Régression pas-à-pas
  - Ascendante ou descendante
  - Ajout ou retrait successif d'une variable
  - Critère → meilleure statistique (R^2 par exemple)
- Exemple: ascendante sur la mortalité



N = 1, R^2 = 0,41

N = 2, R^2 = 0,56

N = 3, R^2 = 0,64



N = 4, R^2 = 0,70

N = 5, R^2 = 0,75

Modèle linéaire multiple

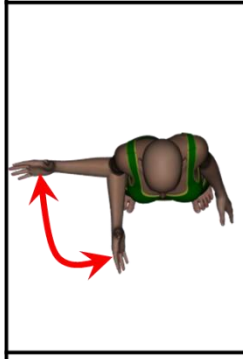
- Meilleur sous-ensemble
  - Pour N de 1 au nombre de variables
  - Recherche du meilleur sous-ensemble de N variables (par exemple au sens de R^2)

- Peut différer de la régression pas-à-pas
  - Une variable dans le meilleur sous-ensemble N peut ne plus être dans le N+1ème

  - Mortalité: pollution au NO dans le 6ème mais plus dans les suivants jusqu'au 14ème

Modèle linéaire multiple

- Mobilité de l'épaule
  - Mesures mono-axiales

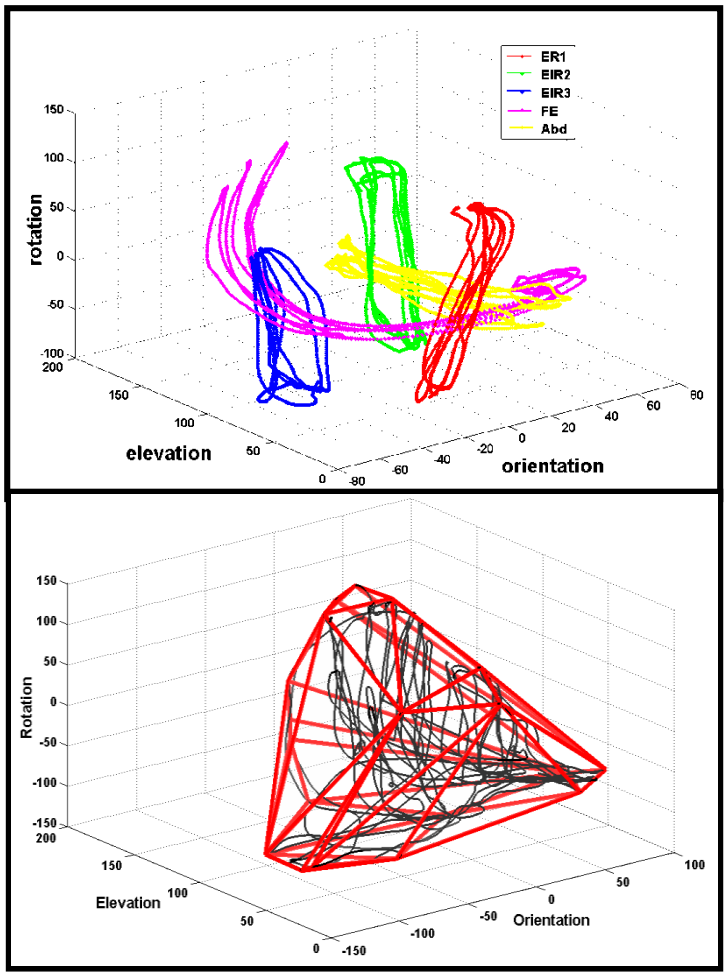
ER1 External rotation from neutral	**EIR2** Ext./ internal rotation in mid-abduct.	**EIR3** Ext./ internal rotation in mid-flexion	**Abd** Abduction from neutral	**FE** Flexion/ extension amplitude
				
				

Modèle linéaire multiple

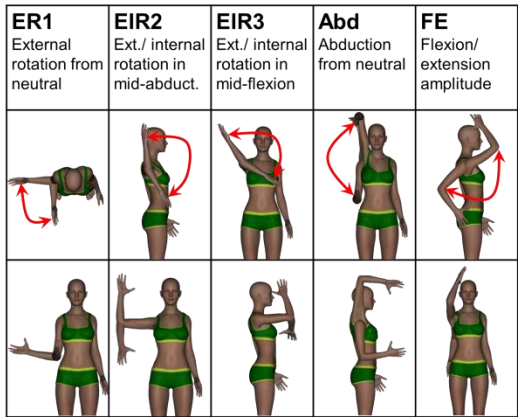
- Mobilité de l’épaule
  - Espace atteignable → volume articulaire

Modèle linéaire multiple

- Mobilité de l'épaule
  - Mise en relation → volume = f(amplitudes)



ER1	EIR2	EIR3	Abd	FE
External rotation from neutral	Ext./ internal rotation in mid-abduct.	Ext./ internal rotation in mid-flexion	Abduction from neutral	Flexion/ extension amplitude



Mobilité épaule.xlsx

ANALYSE EN COMPOSANTES PRINCIPALES

Analyse en Composantes Principales

- Différent de modèle linéaire multiple
  - Ne cherche pas à « expliquer » une variable par les (N-1) autres

- Objectifs
  - Identifier les individus proches et éloignés de l'individu
  - Concept clé de ressemblance : dans le but de former des groupes d'individus proches les uns des autres et éloignes des autres groupes
  - Quelles sont les (groupes de) variables qui expliquent le plus la variabilité inter-individu?

  ➢ Classification selon les liens entre individus et entre les variables

Analyse en Composantes Principales

- Exemple:

ORIGINAL ARTICLE

WILEY

Cross-country skiing movement factorization to explore relationships between skiing economy and athletes' skills

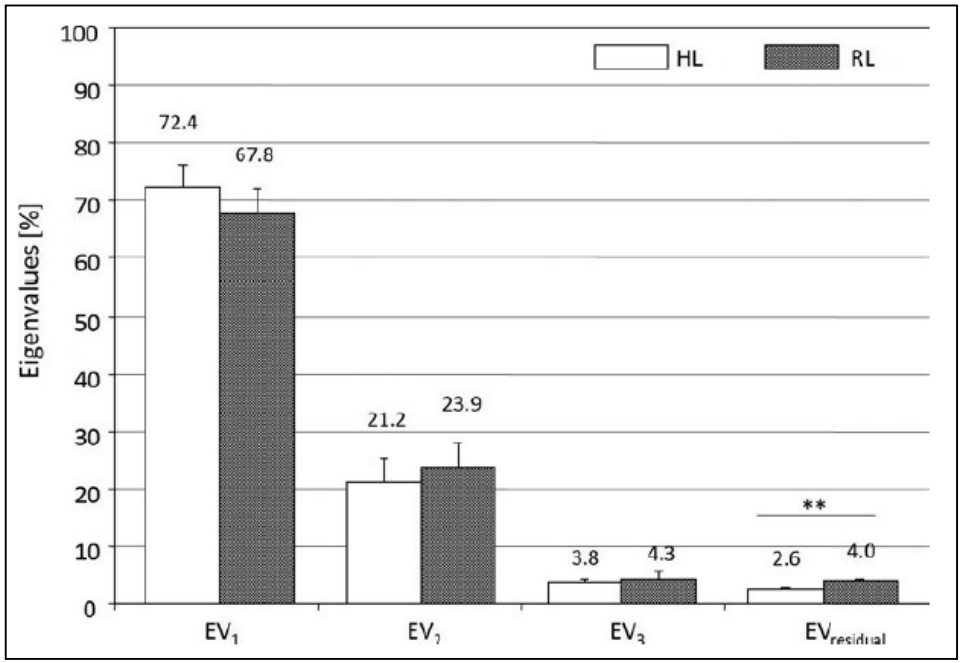
B. Pellegrini[1,2] | C. Zoppirolli[1,2] | G. Boccia[1,2,3] |
L. Bortolan[1,2] | F. Schena[1,2]

Groupe de Haut-Niveau

Groupe de Niveau Régional

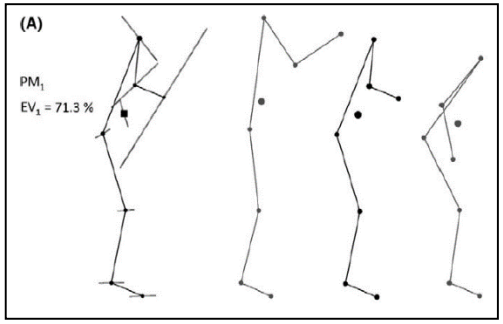
Analyse en Composantes Principales

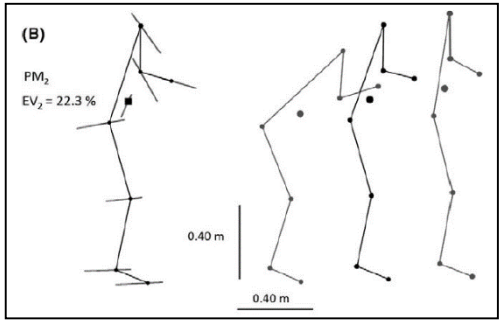
- ACP sur les postures

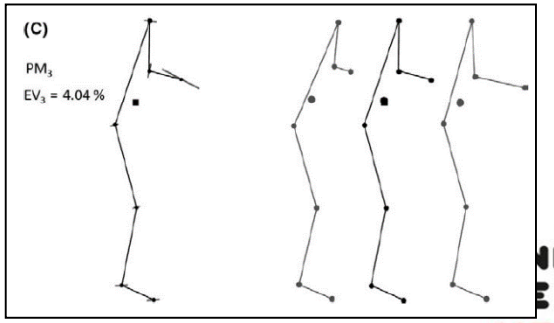


45 paramètres cinématiques mesurés

96 à 97,4% de l'information sur les 3 premières CP







Conditions d'application

- **Taille suffisante de l'échantillon**
  - **Nombre de variables** : au moins 5-10 fois plus d'observations que les variables initiales pour une ACP..
  - **Variabilité des données** : Si les données sont très variables, il est nécessaire d'avoir un échantillon plus important.
  - ➢ Analyse de puissance pour estimer la taille d'échantillon requise en fonction de la taille de l'effet
- **Absence de données manquantes**
- **Normalité des données** (Shapiro-Wilk ou Kolmogorov-Smirnov)
  - si données non normales → tests non-parmétriques pour l'ACP
- **Pas de corrélation des variables 2 à 2** (tableau de corrélation)

Vérification de la corrélation des variables

Tableau de corrélation entre les variables du test Oxford Cognitive Screen, réalisé chez 1973 patients post AVC

TABLE 3 | Heatmap correlation matrix for the OCS scores (Pic Nam, picture naming; Sem, semantics; Read, reading; Num. Wr., number writing; Calc, calculation; Ori, orientation; SR, sentence recall; EM, episodic memory; IM, imitation; VF, visual field; Canc., cancelation; O AS, object asymmetry; S AS, space asymmetry; TR, trails).

	Language			Number Cognition		Memory			Pra-xis	Attention				
	Pic Nam	**Sem**	**Read**	**NumWr**	**Calc**	**Ori**	**SR**	**EM**	**IM**	**VF**	**Canc**	**O As**	**S As**	**TR**
Pic Nam	1	0.30	0.43	0.37	0.29	0.23	0.36	0.35	0.33	0.19	0.28	−0.05	−0.09	−0.10
Sem	0.28	1	0.35	0.28	0.27	0.17	0.16	0.24	0.28	0.26	0.23	0.01	−0.08	−0.05
Read	0.45	0.32	1	0.47	0.41	0.25	0.33	0.31	0.26	0.28	0.29	−0.02	−0.08	−0.08
NumWr	0.42	0.26	0.50	1	0.42	0.27	0.26	0.29	0.30	0.24	0.31	−0.08	−0.13	−0.14
Calc	0.36	0.24	0.43	0.47	1	0.32	0.28	0.25	0.25	0.17	0.29	−0.07	−0.14	−0.14
Ori	0.27	0.18	0.24	0.31	0.33	1	0.29	0.28	0.25	0.20	0.29	−0.12	−0.15	−0.09
SR	0.38	0.15	0.31	0.30	0.32	0.29	1	0.40	0.18	0.06	0.12	0.01	−0.02	−0.11
EM	0.41	0.25	0.33	0.36	0.31	0.31	0.42	1	0.25	0.14	0.24	−0.04	−0.03	−0.05
IM	0.39	0.27	0.29	0.37	0.30	0.26	0.22	0.32	1	0.27	0.38	−0.10	−0.16	−0.14
VF	0.20	0.25	0.30	0.24	0.18	0.20	0.08	0.17	0.28	1	0.40	−0.14	−0.22	−0.03
Canc	0.38	0.23	0.33	0.39	0.37	0.32	0.19	0.34	0.44	0.40	1	−0.21	−0.42	−0.20
O As.	−0.12	−0.01	−0.08	−0.16	−0.12	−0.12	−0.02	−0.11	−0.16	−0.16	−0.29	1	0.25	0.16
S As.	−0.17	−0.11	−0.15	−0.19	−0.21	−0.17	−0.07	−0.13	−0.22	−0.25	−0.49	0.31	1	0.09
Trails	−0.21	−0.06	−0.15	−0.23	−0.24	−0.15	−0.19	−0.17	−0.24	−0.08	−0.31	0.22	0.18	1

Above the diagonal, partial correlations corrected for demographical factors (age and education), below the diagonal, not corrected correlations.

Iosa, M., Demeyere, N., Abbruzzese, L., Zoccolotti, P., & Mancuso, M. (2022). Principal component analysis of oxford cognitive screen in patients with stroke. *Frontiers in neurology*, *13*, 779679.

Préparation des données

- S'assurer que les données sont **quantitatives**. Dans la pratique, on considère souvent les variables ordinales comme des quantitatives.
  - Par exemple, « Pas du tout satisfait »=1 ; « plutôt pas satisfait »=2; « moyennement satisfait »=4; « plutôt satisfait »=4; « Tout à fait satisfait »=5 …

- Remarque : on s'autorise une certaine liberté d'interprétation qui n'a pas de fondement statistique:
  - 4 « plutôt satisfait » est supérieur à 2 « plutôt pas satisfait », mais rien ne justifie le fait que « plutôt satisfait » traduise une satisfaction deux fois plus importante que « plutôt pas satisfait ».

Préparation des données

- Elles se présentent dans un tableau (matrice) à n lignes et p colonnes que l'on notera X.

- Chaque ligne représente un individu, chaque colonne représente une variable.

$$X \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & X_{32} & x_{33} & \cdots & x_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

- La matrice X peut être analysée a travers ses lignes (les individus) ou a travers ses colonnes (les variables) ce qui induit plusieurs types de questions.

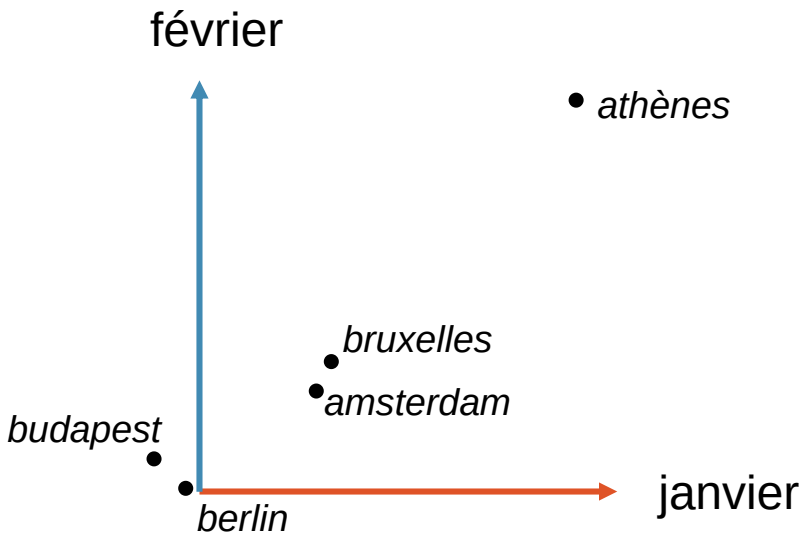
Températures moyennes relevées dans 35 grandes villes Européennes

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5	11.4	7.0	4.4
Athènes	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8	19.2	14.6	11.0
Berlin	-0.2	0.1	4.4	8.2	13.8	16.0	18.3	18.0	14.4	10.0	4.2	1.2
Bruxelles	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15.0	11.1	6.7	4.4
Budapest	-1.1	0.8	5.5	11.6	17.0	20.2	22.0	21.3	16.9	11.3	5.1	0.7

- Qui sont les individus? Quelles sont les variables?

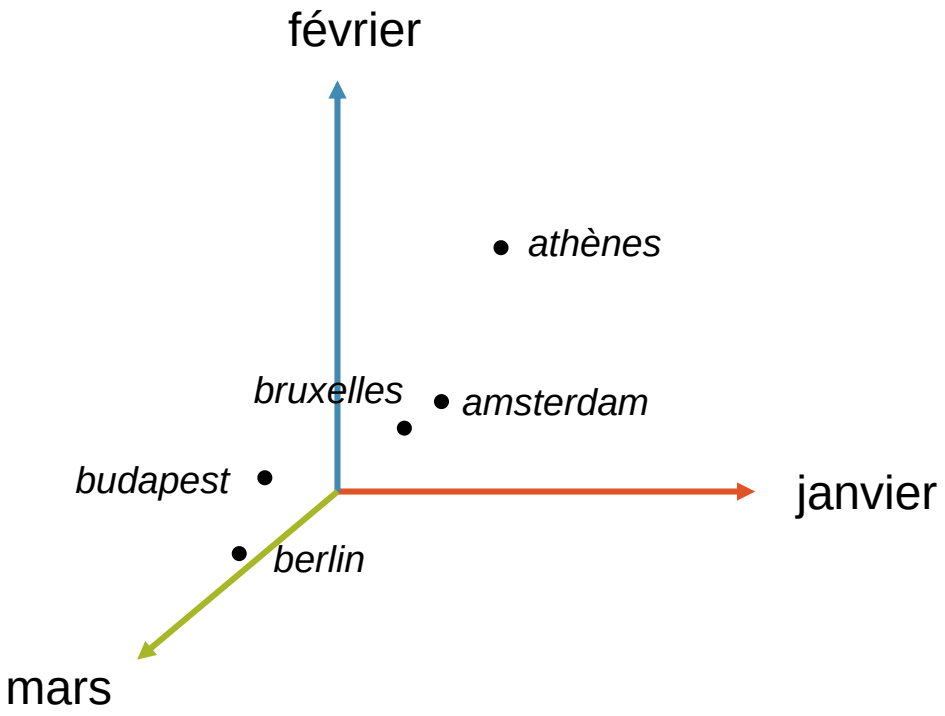
Représentation en nuage à 2 dimensions

	Janvier	Février
Amsterdam	2.9	2.5
Athènes	9.1	9.7
Berlin	-0.2	0.1
Bruxelles	3.3	3.3
Budapest	-1.1	0.8



Représentation en nuage à 3 dimensions

	Janvier	Février	Mars
Amsterdam	2.9	2.5	5.7
Athènes	9.1	9.7	11.7
Berlin	-0.2	0.1	4.4
Bruxelles	3.3	3.3	6.7
Budapest	-1.1	0.8	5.5



Représentation en nuage à p dimension

- Au-delà de la 3ème dimension, la représentation spatiale usuelle n'est plus possible

Calcul de la distance entre les individus

- La ressemblance entre les individus correspond à la distance (numérique et géométrique) entre ces individus selon l'ensemble des variables

Espace à 2 variables $$d_{n1,n2}{}^2 = (n1_{p1} - n2_{p1})^2 + (n1_{p2} - n2_{p2})^2$$

Espace à 3 variables $$d_{n1,n2}{}^2 = (n1_{p1} - n2_{p1})^2 + (n1_{p2} - n2_{p2})^2 + (n1_{p3} - n2_{p3})^2$$

Espace à p variables $$d_{n1,n2}{}^2 = \sum_{p=1}^{p} (n1_p - n2_p)^2$$

Exemple

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5	11.4	7.0	4.4
Athènes	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8	19.2	14.6	11.0

$$d_{Amsterdam,Athènes}{}^2 = (2{,}9-9{,}1)^2+(2{,}5-9{,}7)^2+(5{,}7-11{,}7)^2\ldots$$
$$+(8{,}2-15{,}4)^2+(12{,}5-20{,}1)^2+(14{,}8-24{,}5)^2\ldots$$
$$+(17{,}1-27{,}4)^2+(17{,}1-27{,}2)^2+(14{,}5-23{,}8)^2\ldots$$
$$+(11{,}4-19{,}2)^2+(7{,}0-14{,}6)^2+(4{,}4-11{,}0)^2$$
$$d_{Amsterdam,Athènes}{}^2 = 786{,}72$$

Alors que: $d_{Amsterdam,Berlin}{}^2 = 42{,}49$

Mesure de l'inertie

- Une mesure de l'information portée par le nuage : la somme des distances inter-individus

- Si les points sont tous proches les uns des autres, cette quantité sera faible alors que des points très éloignés des autres auront tendance a l'augmenter.

- Un objectif de l'ACP sera de décomposer une quantité dérivant de cette somme: l'**inertie.**

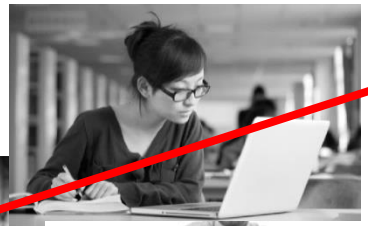
- On cherchera en particulier à faire apparaître :
  - des (groupes d')individus
  - les directions de l'espace: (groupes de) variables

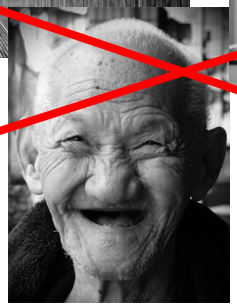
y contribuant le plus.

Analyse

- Exemple, mortalité en villes industrielles (variable dépendante)
- individus → 60 villes
- variables (indépendantes) → 15
  - Météorologiques (précipitations, températures,…)
  - Sociologiques (âge moyen, études, revenus,…)
  - Niveaux de pollution

$= f($





$)$

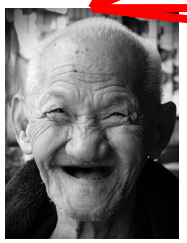
UNIVERSITE RENNES UFR STA

Analyse

- Exemple, mortalité en villes industrielles (variable dépendante)
- individus → 60 villes
- variables (indépendantes) → 15
  - Météorologiques (précipitations, tem
  - Sociologiques (âge moyen, études, r
  - Niveaux de pollution

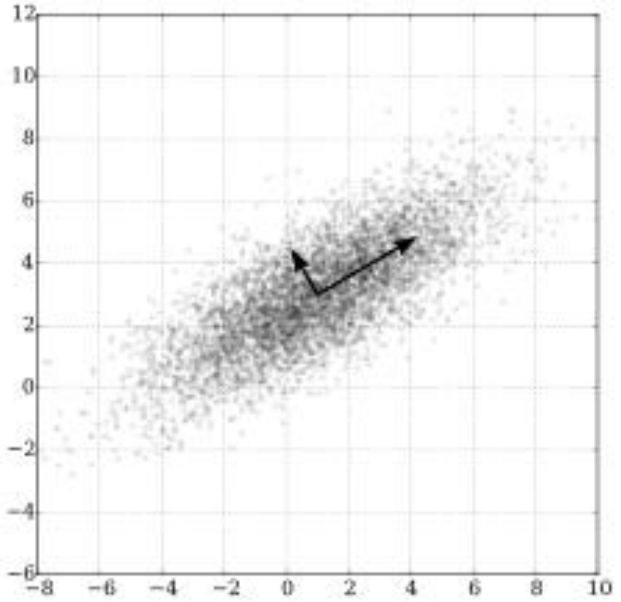







Calcul des axes principaux

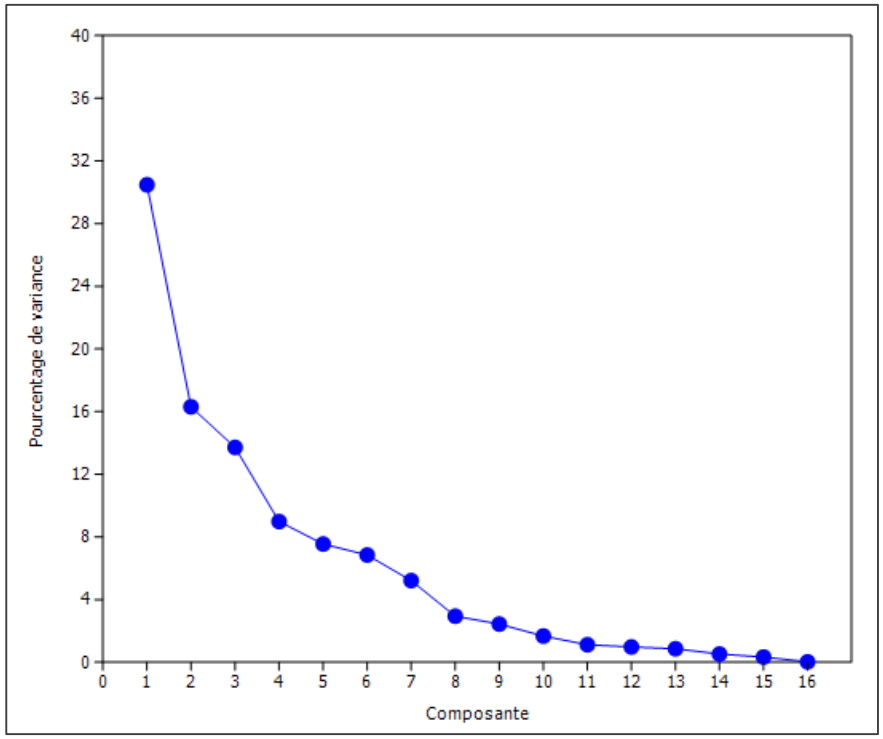
- Calculer les axes du nuage de points N-D qui possèdent le plus d'inertie



- Les classer par ordre décroissant d'importance
  - ➔ R^2 = pourcentage d'explication de la variance
- Travailler avec les p premières composantes
  (l'analyse visuelle se fait en général en fonction des 2 premières composantes)

Analyse

- Exemple, dans 60 villes industrielles:
  - Mortalité → variables dépendante
  - 15 variables indépendantes

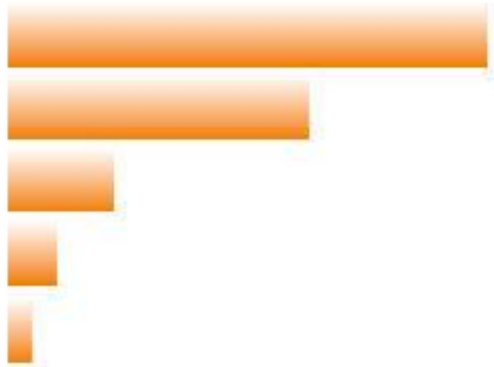


ACP sur les données « mortalité »
Inertie relative de chaque composante

Interprétation des résultats

1. Nombre d'axes de l'analyse

Valeur propre	% d'inertie	Somme
2,0424	53,8%	53,8%
1,2509	29,7%	83,5%
0,9003	9,8%	93,3%
0,5375	4,3%	97,2%
0,2690	2,4%	100,0%



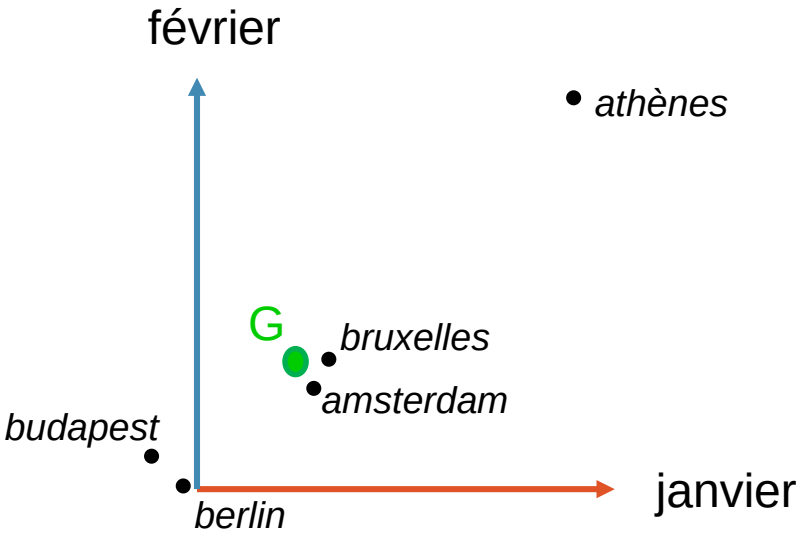
- Sur le schéma précédent on remarque qu'en conservant les deux premiers axes on va expliquer 83,5% de l'inertie totale du nuage de point.

Centrage des données

- On choisit en général de représenter le centre du repère au centre de gravité des données.
- Les coordonnées de ce point sont les moyennes de chacune des variables. Les coordonnées des autres points peuvent être recalculées par rapport à ce centre.

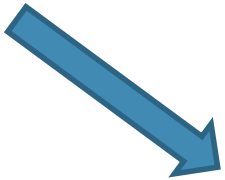
	Janvier	Février
Amsterdam	0.1	-0.8
Athènes	6.3	6.4
Berlin	-3.0	-3.2
Bruxelles	0.5	0
Budapest	-3.9	-2.5
G	0	0

$$x\_centré_{np} = x_{np} - \overline{x_p}$$

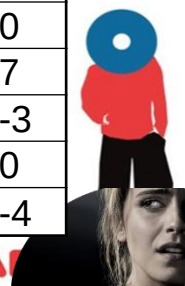


Centrage des données

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	2,9	3	5,7	8,2	13	15	17	17	15	11	7	4,4
Athènes	9,1	10	12	15	20	25	27	27	24	19	15	11
Berlin	-0,2	0	4,4	8,2	14	16	18	18	14	10	4,2	1,2
Bruxelles	3,3	3	6,7	8,9	13	16	18	18	15	11	6,7	4,4
Budapest	-1,1	1	5,5	12	17	20	22	21	17	11	5,1	0,7
moyenne	**2,8**	**3,3**	**6,8**	**10,5**	**15,2**	**18,2**	**20,5**	**20,3**	**16,9**	**12,6**	**7,5**	**4,3**



	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	0,1	-1	-1,1	-2,3	-2,7	-3,4	-3,4	-3,2	-2,4	-1,2	-0,5	0
Athènes	6,3	6	4,9	4,9	4,9	6,3	6,9	6,9	6,9	6,6	7,1	7
Berlin	-3	-3	-2,4	-2,3	-1,4	-2,2	-2,2	-2,3	-2,5	-2,6	-3,3	-3
Bruxelles	0,5	0	-0,1	-1,6	-2,4	-2,6	-2,7	-2,5	-1,9	-1,5	-0,8	0
Budapest	-4	-2	-1,3	1,1	1,8	2	1,5	1	-0	-1,3	-2,4	-4



Réduction des données

- Si les variables sont exprimées selon des unités ou ordres de grandeurs différents, les coordonnées des individus seront inégalement dispersées selon chaque axe

- Cela donnera artificiellement plus d'importance aux variables dont les ordres de grandeur sont plus petits.

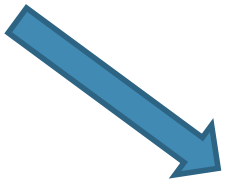
  Exemple: une variabilité de 1m est égale à une variabilité de 1000mm

- Réduire les données, c'est diviser les observations centrées par l'écart type.

$$x\_centré\_réduit_{np} = \frac{x_{np} - \overline{x_p}}{\sigma_p}$$

Réduction des données

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	0,1	-1	-1,1	-2,3	-2,7	-3,4	-3,4	-3,2	-2,4	-1,2	-0,5	0
Athènes	6,3	6	4,9	4,9	4,9	6,3	6,9	6,9	6,9	6,6	7,1	7
Berlin	-3	-3	-2,4	-2,3	-1,4	-2,2	-2,2	-2,3	-2,5	-2,6	-3,3	-3
Bruxelles	0,5	0	-0,1	-1,6	-2,4	-2,6	-2,7	-2,5	-1,9	-1,5	-0,8	0
Budapest	-4	-2	-1,3	1,1	1,8	2	1,5	1	-0	-1,3	-2,4	-4



DONNEES CENTREES REDUITES	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amsterdam	0,03	-0,23	-0,43	-0,82	-0,94	-0,94	-0,89	-0,85	-0,68	-0,36	-0,14	0,02
Athènes	1,76	1,88	1,92	1,78	1,67	1,72	1,79	1,84	1,93	1,98	1,92	1,81
Berlin	-0,84	-0,93	-0,94	-0,82	-0,50	-0,61	-0,58	-0,61	-0,71	-0,78	-0,90	-0,85
Bruxelles	0,14	0,01	-0,04	-0,56	-0,84	-0,72	-0,71	-0,66	-0,54	-0,45	-0,22	0,02
Budapest	-1,09	-0,73	-0,51	0,41	0,61	0,54	0,39	0,27	-0,01	-0,39	-0,66	-0,99

Matrice des covariances entre les variables

1. Calculer la covariance entre chaque paire de variables en utilisant la formule suivante :

$$\mathbf{cov(Xi, Xj) = \sum(xi - \mu i)(xj - \mu j) / (n - 1)}$$

- Exemple:

Cov(Jan,Fév)=(0,03*(-0,23)+1,76*1,88+(-0,84)*(-0,93)+0,14*0,01+-1,09*(-0,73))/(5-1)= 4,4

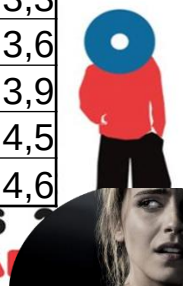
DONNEES CENTREES REDUITES	Janvier	Février
Amsterdam	0,03	-0,23
Athènes	1,76	1,88
Berlin	-0,84	-0,93
Bruxelles	0,14	0,01
Budapest	-1,09	-0,73

Matrice des covariances entre les variables

1. Créer la matrice des covariances: les lignes et les colonnes représentent les variables qui sont ainsi confrontées 2 à 2.

- Exemple:

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Janvier	4,5	4,4	4,2	2,9	2,3	2,5	2,8	3,0	3,5	4,0	4,3	4,5
Février	4,2	4,3	4,2	3,4	2,9	3,1	3,3	3,4	3,8	4,1	4,2	4,2
Mars	3,0	3,2	3,2	2,8	2,4	2,6	2,7	2,8	3,0	3,1	3,2	3,1
Avril	2,3	2,8	3,0	3,5	3,4	3,4	3,4	3,4	3,4	3,1	2,9	2,4
Mai	1,9	2,5	2,8	3,5	3,6	3,6	3,6	3,6	3,4	3,0	2,6	2,1
Juin	2,6	3,3	3,7	4,5	4,6	4,6	4,6	4,5	4,4	3,9	3,5	2,8
Juillet	3,0	3,7	4,0	4,8	4,8	4,8	4,8	4,8	4,7	4,3	3,8	3,2
Août	3,1	3,8	4,1	4,7	4,6	4,6	4,7	4,7	4,6	4,3	3,9	3,3
Septembre	3,5	4,0	4,2	4,3	4,2	4,2	4,3	4,4	4,4	4,3	4,1	3,6
Octobre	3,7	4,0	4,1	3,8	3,5	3,6	3,7	3,9	4,1	4,2	4,1	3,9
Novembre	4,4	4,6	4,5	3,8	3,3	3,5	3,7	3,9	4,2	4,5	4,6	4,5
Décembre	4,6	4,5	4,4	3,2	2,6	2,8	3,1	3,3	3,8	4,2	4,5	4,6



Matrice des covariances entre les variables

- Notez que la matrice de covariance peut être calculée à partir de la matrice de variance (diagonale) en utilisant la formule suivante :

$$\mathbf{cov(Xi, Xj) = \sigma i * \sigma j * \rho(Xi, Xj)}$$

avec :

  - σi et σj sont les écarts types (ou standard deviations) des variables Xi et Xj respectivement
  - ρ(Xi, Xj) est la corrélation entre les variables Xi et Xj

- Cependant, il est généralement préférable de calculer directement la covariance en utilisant la formule précédente, car elle est plus précise et moins susceptible d'erreurs.

Valeurs propres de la matrice de covariance

- La matrice de covariance est semi-définie positive → valeurs propres sont non négatives. Cette propriété est importante pour l'ACP, car elle permet de diagonaliser la matrice et de trouver les composantes principales.

- **Méthodes de calcul**
  - La méthode de Gauss, qui consiste à appliquer des permutations et des réductions à la matrice pour obtenir une forme diagonale.
  - La méthode de décomposition spectrale, qui utilise la théorie des valeurs propres pour diagonaliser la matrice de covariance.

  → Calculateur en ligne, python, matlab…

- **Interprétation**

Les valeurs propres représentent les variances d'échantillon des composantes principales. La somme des valeurs propres correspond à la dispersion totale des individus considérés. Dans l'ACP, les valeurs propres sont utilisées pour déterminer la proportion de variance expliquée par chaque composante principale.

Vecteurs propres et valeurs propres

Exemple :

…

$$\circ\ v\approx\begin{pmatrix}-543{,}818\\-76{,}377\\-250{,}410\\809{,}032\\-40{,}791\\92{,}622\\-426{,}673\\-535{,}580\\-291{,}268\\311{,}699\\891{,}616\\1\end{pmatrix}, \text{ valeur propre } \lambda_5\approx 0{,}034$$

$$\circ\ v\approx\begin{pmatrix}1{,}091\\0{,}660\\0{,}312\\-0{,}475\\-0{,}786\\-0{,}871\\-0{,}770\\-0{,}621\\-0{,}253\\0{,}192\\0{,}609\\1\end{pmatrix}, \text{ valeur propre } \lambda_7\approx 6{,}123$$

$$\circ\ v\approx\begin{pmatrix}-0{,}016\\-1{,}947\\-2{,}623\\-2{,}231\\1{,}006\\-0{,}775\\0{,}508\\0{,}545\\0{,}638\\2{,}523\\1{,}376\\1\end{pmatrix}, \text{ valeur propre } \lambda_6\approx 0{,}105$$

$$\circ\ v\approx\begin{pmatrix}0{,}940\\0{,}989\\0{,}766\\0{,}811\\0{,}793\\1{,}028\\1{,}109\\1{,}108\\1{,}087\\1{,}021\\1{,}087\\1\end{pmatrix}, \text{ valeur propre } \lambda_8\approx 44{,}639$$

…

UNIVERSITÉ
RENNES
UFR STA

Exercice

- Une étude sur des fournisseurs de matériel informatique a conduit à apprécier le service, la qualité et le prix de quatre fournisseurs. Pour cela un expert a noté ces entreprises avec des notes allant de -3 à 3. Les résultats sont consignés ci-dessous (on considère que la distribution des donnée est normale):

Entreprise	Service	Qualité	Prix
E1	-2	3	-1
E2	-1	1	0
E3	2	-1	-1
E4	1	-3	2

1) Calculer le vecteur moyen des individus. Qu'en conclure?
2) Calculer la matrice de corrélation.
3) Calculer le tableau des covariances.
4) Calculer les vecteurs propres et les valeurs propres à l'aide d'un calculateur en ligne.

Axes principaux

- **Projection des données sur les vecteurs propres : l**es données sont projetées sur les vecteurs propres en utilisant les formules de projection vectorielle.

- **Obtention des composantes principales (CP)**: Les composantes principales sont obtenues en prenant les produits scalaires entre les données projetées et les vecteurs propres.

Axes principaux et valeurs propres

Résultats pour la décomposition des items du test Oxford Cognitive Screen, réalisé chez 1973 patients post AVC

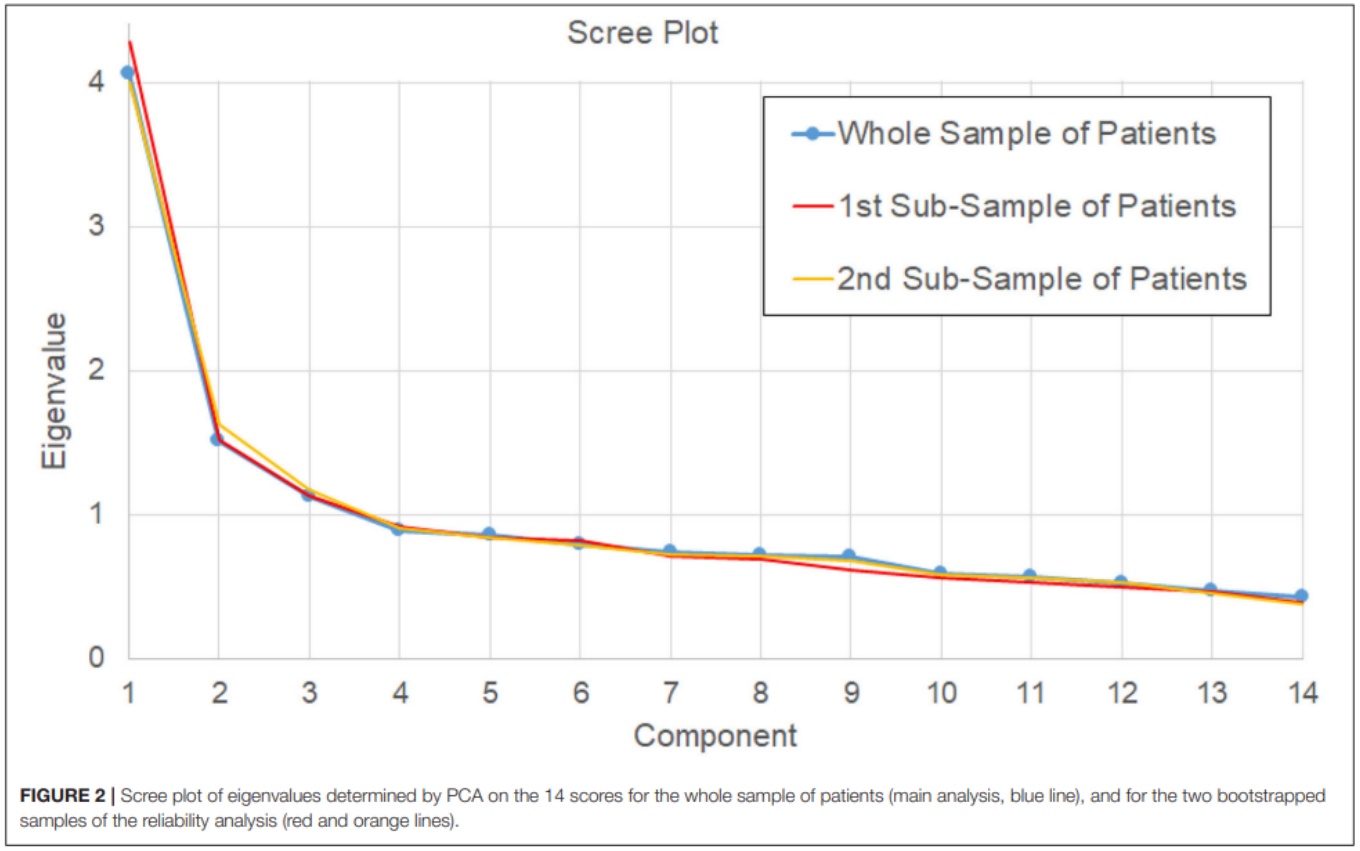


FIGURE 2 | Scree plot of eigenvalues determined by PCA on the 14 scores for the whole sample of patients (main analysis, blue line), and for the two bootstrapped samples of the reliability analysis (red and orange lines).

Iosa, M., Demeyere, N., Abbruzzese, L., Zoccolotti, P., & Mancuso, M. (2022). Principal component analysis of oxford cognitive screen in patients with stroke. *Frontiers in neurology*, *13*, 779679.

Axes principaux et valeurs propres

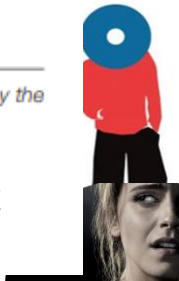
Résultats pour la décomposition des items du test Oxford Cognitive Screen, réalisé chez 1973 patients post AVC

TABLE 4 | The pattern matrix from the principal component analysis on the patients' sample (in bold the higher value for each task, forming clear aggregation of subtasks with absolute values > 0.4).

OCS Subtask	Components						Communality	95% CI main load
	1	2	3	4	5	6		
Sentence Reading	**0.771**	0.006	0.123	0.093	0.128	−0.159	0.699	0.66–0.76
Number Writing	**0.713**	−0.051	−0.083	0.074	0.032	0.010	0.611	0.64–0.78
Calculation	**0.761**	0.013	−0.129	−0.055	−0.102	0.250	0.678	0.78-0.85
Cancelation	0.115	**−0.430**	−0.166	0.019	0.383	0.241	0.642	0.34-0.64
Object Asymmetry	0.004	**0.852**	0.055	−0.132	0.178	0.211	0.723	0.46–1.00
Space Asymmetry	−0.024	**0.676**	−0.021	0.101	−0.121	−0.201	0.592	0.60–0.96
Trails	−0.083	0.053	**0.921**	0.082	0.000	0.035	0.860	0.91–0.91
Sentence Recall	0.137	0.060	−0.043	**0.721**	−0.161	0.148	0.640	0.65–0.86
Episodic Memory	−0.061	−0.080	0.088	**0.808**	0.111	0.090	0.681	0.80–0.82
Picture naming	0.278	−0.077	−0.056	**0.514**	0.214	−0.155	0.590	0.31–0.73
Semantics	0.175	0.166	0.048	0.078	**0.666**	−0.135	0.556	0.63–0.74
Visual Field	0.107	−0.228	0.202	−0.151	**0.609**	0.149	0.581	0.49–0.70
Imitation	−0.110	0.016	−0.342	0.214	**0.629**	0.076	0.615	0.54–0.68
Orientation	0.092	0.035	0.024	0.230	−0.006	**0.813**	0.792	0.69–0.93

The last two columns report the results of the communality table on the whole sample of patients and the 95% confidence interval of the main load for each subtest obtained by the reliability analysis of the two subsamples of patients.

Iosa, M., Demeyere, N., Abbruzzese, L., Zoccolotti, P., & Mancuso, M. (2022). Principal component analysis of oxford cognitive screen in patients with stroke. *Frontiers in neurology*, *13*, 779679.



Axes principaux et valeurs propres

Résultats pour la décomposition des items du test Oxford Cognitive Screen, réalisé chez 1973 patients post AVC

FIGURE 3 | On the left the original structure of OCS with five domains and on the right the six components identified by the principal component analysis, with arrows reported for values >0.25 according to the legend.

Iosa, M., Demeyere, N., Abbruzzese, L., Zoccolotti, P., & Mancuso, M. (2022). Principal component analysis of oxford cognitive screen in patients with stroke. *Frontiers in neurology*, *13*, 779679.

Analyse

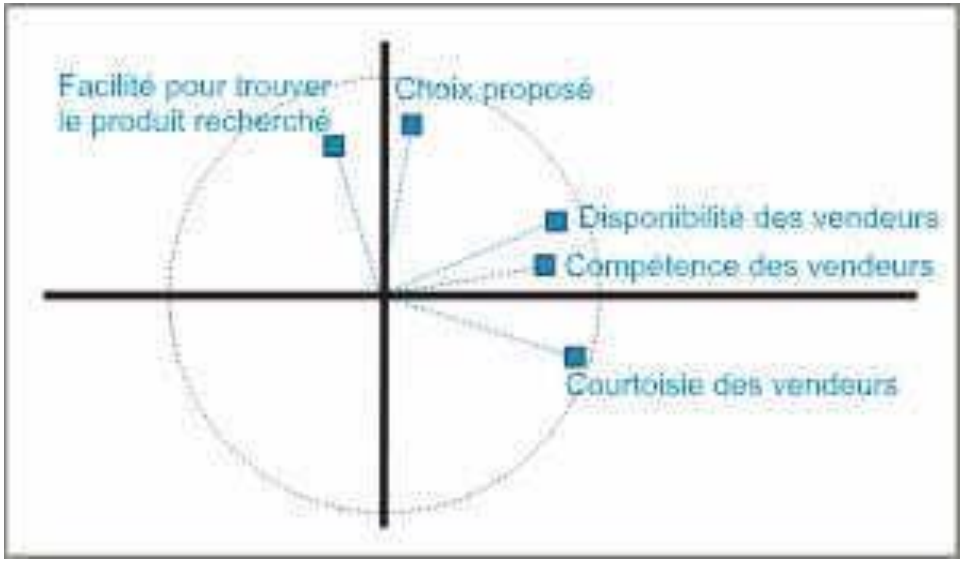
- Identifier les colonnes de paramètres de la matrice X à des vecteurs va permettre de mesurer des corrélations entre variables

- Ces corrélations peuvent être interprétées comme des mesures du degré de colinéarité entre les vecteurs:
  - Si l'angle est proche de 0, la corrélation est proche de 1.
  - Si l'angle est proche de 90°, la corrélation est proche de 0.
  - Si l'angle est proche de 180°,la corrélation est proche de -1.

- L'opération de centrage-réduction des variables permet de normer les vecteurs de paramètres

Interprétation des résultats

2. Signification des axes principaux, ou du plan principal



- Les points “choix”, “disponibilité” et “courtoisie” proches du cercle de corrélation sont bien représentés sur le mapping.

- L’angle entre “compétence” et “disponibilité” indique que ces 2 variables sont assez bien corrélées. L’angle entre “compétence” et “choix” indique que ces deux variables sont indépendantes.

- Le point “compétence” proche de l’axe 1 est très bien représenté par cet axe. En revanche, il est peu représenté par l’axe 2.
- Selon l’axe 2, le point “choix” est mieux corrélé que le point “facilité”.

➢ L’axe 1 correspond plutôt à l’appréciation des vendeurs (accueil)

➢ L’axe 2 correspond plutôt à l’appréciation du magasin (produits)

Interprétation des résultats

2. Signification des axes principaux

- Nîmes et Strasbourg semblent proches mais ils sont peut être opposés sur l'axe 3.

- En synthétisant les informations des 5 variables analysées:
  - Il y a des efforts à faire en matière d'accueil dans les magasins de Nice, Marseille, Amiens et Toulon. Ce dernier est également très peu apprécié en matière de produits.
  - Les magasins de Paris, de Lyon et de Marseille sont appréciés de la clientèle pour leurs produits.
  - Lyon se distingue selon les 2 axes et peut être considéré comme le meilleur magasin

- Ces conclusions sont à confirmer par l'examen des tableaux de corrélations et de coordonnées des individus, fournis par le logiciel d'analyse.

Kollias, I., Hatzitaki, V., Papaiakovou, G., & Giatsis, G. (2001). Using principal components analysis to identify individual differences in vertical jump performance. *Research Quarterly for Exercise and Sport*, *72*(1), 63-67.

- Quel est l'objectif de l'article? Pourquoi l'ACP est-elle utilisée?
- Est-ce que toutes les conditions d'application ont été explicitement vérifiées?
- Toutes les étapes de l'analyse ont bien été réalisées?
- Combien de composantes sont retenues? Quels groupes de variables représentent-elles? Peut-on leur donner un nom? Combien de variabilité expliquent-elles?

Kollias, I., Hatzitaki, V., Papaiakovou, G., & Giatsis, G. (2001). Using principal components analysis to identify individual differences in vertical jump performance. *Research Quarterly for Exercise and Sport*, *72*(1), 63-67.

- Que montrent la répartition des individus selon les axes principaux?

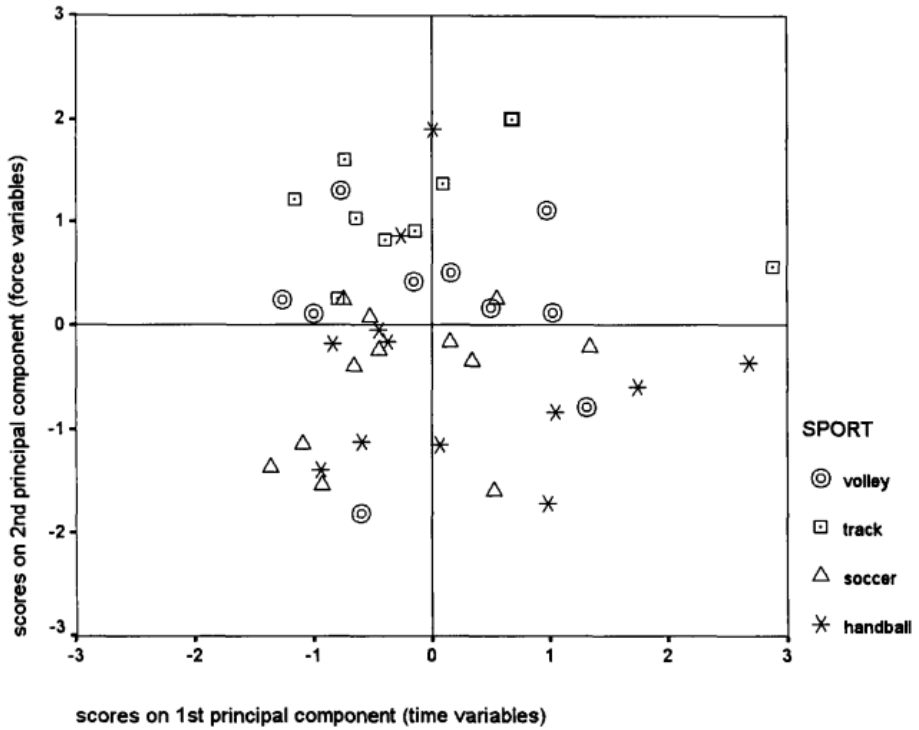
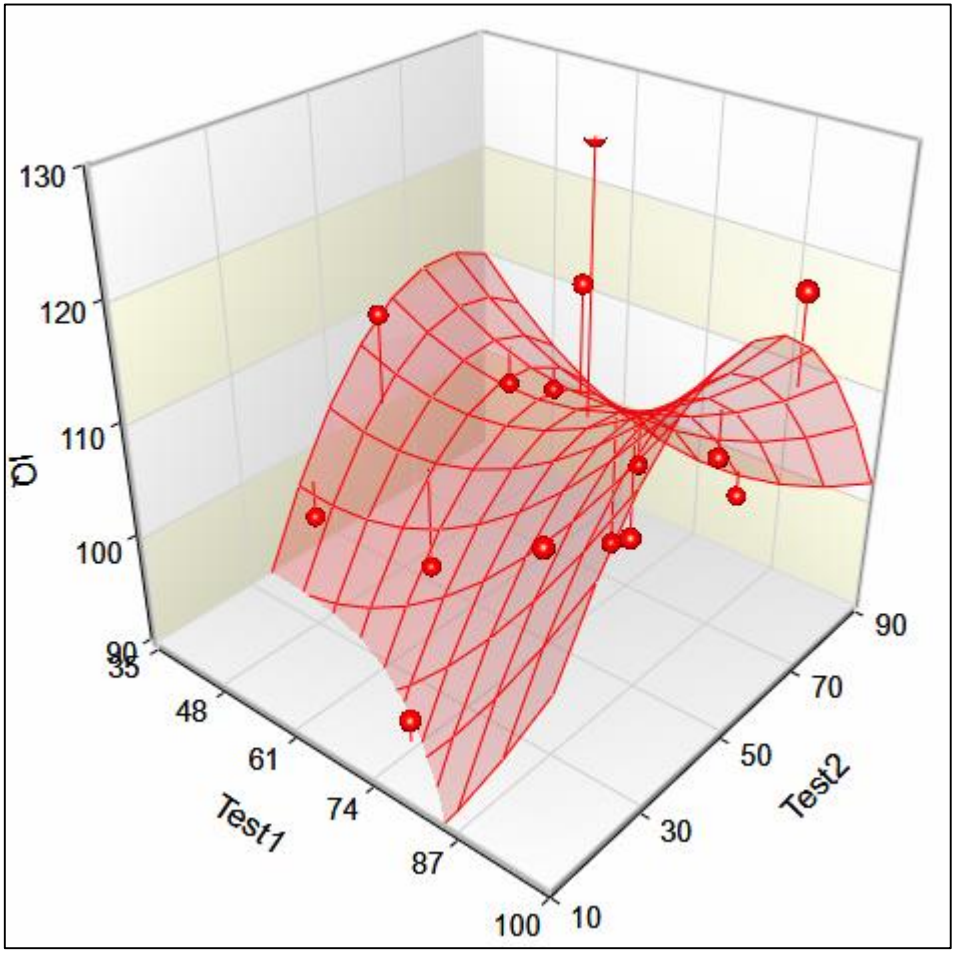


Figure 1. Factor scores for each athlete on the two rotated principal components. The x-axis represents the first principal component identified with the time variables (TIME, TFMAX and RFDMAX); the y-axis shows the regression scores on the second principal component identified with the force variables (RFMAX, RPMAX).

Limite importante

- Ces méthodes sont linéaires



Comment rendre compte de ce type de relation?

Limite importante

- Ces méthodes sont linéaires

- Cependant,
  - Dans beaucoup de domaines, on observes des liaisons linéaires ou assimilables
  - Le fait de considérer des liaisons multiples apporte une grande richesse d'analyse et d'interprétations par rapport à des modèles de régression
  - Etablir des liaisons linéaires dans un premier temps, n'empêche pas d'explorer d'autres aspects ensuite

INTRODUCTION AU « MACHINE LEARNING »

Contexte

« Informations, données, bases de données, “data”, “big data”… Quand les données se multiplient, se stockent et se transmettent facilement, il y a aussi de vrais enjeux de société qui émergent et qui touchent tout le monde : moi, nous, vous… »

- exemple: étude américaine visant à détecter les comportements schizophrènes parmi un large panel d’utilisateurs du réseau social Twitter, à partir de toutes les données postées sur ce réseau social.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525233/

- Des humains ne pourraient pas forcément déceler une information de ces données: nombre d'amis, mots les plus utilisés, liens avec d'autres personnes ou sujets d'intérêt, choix d'images…, mais une machine qui calcule le peut.

https://openclassrooms.com/fr/courses/3930076-manipuler-linformation/3930696-vos-donnees-pour-quoi-faire

Historique

https://www.cours-gratuit.com/cours-statistique/cours-d-introduction-a-la-statistique-big-data

- 1940-70 – hOctets La Statistique expérimentale est une question, (biologique), associée à une **hypothèse testée expérimentalement** avec n=30 individus sur p<10 variables.

- 1970s – kO Les **premiers outils informatiques** se généralisant, l'analyse des données (Multivariate statistics) explore, sans modèle, des données plus volumineuses.

- 1980s – MO En **Intelligence Artificielle**, les systèmes experts expirent, supplantés par l'apprentissage (machine learning) des réseaux de neurones. La Statistique aborde des modèles non paramétriques ou fonctionnels.

- 1990s – GO 1er changement de paradigme. **Les données ne sont plus planifiées**, elles sont préalablement acquises en continu et stockées pour les objectifs usuels (i.e. comptables) de l'entreprise. Le Data Mining aide à la décision à travers de logiciels qui regroupent gestions de données, techniques exploratoires et modélisation statistique.

- 2000s –TO 2ème changement de paradigme. **Le nombre p de variables explose** (de l'ordre de 10^4 à 10^6), notamment avec les biotechnologies omiques où p >> n. L'objectif de qualité de prévision l'emporte sur la réalité du modèle devenu "boîte noire". Face à la (sur-)dimension, Apprentissage Machine et Statistique s'unissent en Apprentissage Statistique.

- 2010s – PO 3ème changement de paradigme. Dans les applications industrielles, le e-commerce, la géolocalisation... **le nombre n d'individus explose**, les bases de données se structurent en nuages (cloud), les moyens de calculs se groupent (cluster), mais la puissance brute ne suffit plus à la voracité (greed) des algorithmes. L'optimisation est nécessaire pour limiter le temps de calcul ou le volume/flux de données considéré. La décision devient adaptative ou séquentielle.

Big Data

- Données d'un volume supérieurs à plusieurs terabytes.

  (10^{12}bits ou 1000 Gb)

- Caractéristiques:
  - Grand Volume,
  - Grande Vitesse,
  - Grande diversité (en termes de type, format,…),
  - Grande Véracité.

- La donnée massive est la matière première du machine learning. Il consiste à instaurer un nouveau rapport à la machine, à passer d'un rapport de programmation (l'homme programme une machine) à un rapport d'enseignement.

- Le machine learning (et plus généralement l'IA) est indubitablement une des applications les plus prometteuses du Big Data.

Planifier l'acquisition de Big Data

- Choisir le bon protocole
- Réfléchir en amont aux usages
- Sécuriser

- « voici le triptyque pour gérer les quantités de données considérables provenant des objets connectés qui devront être traitées et analysées en temps réel. »

https://www.usinenouvelle.com/article/des-donnees-pour-quoi-faire.N1850187

Data Mining

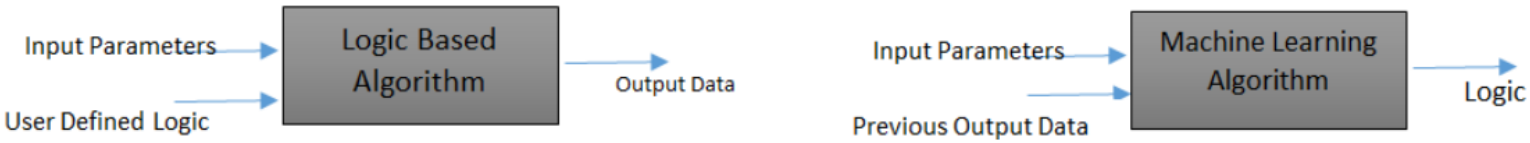
- Le data mining ou la fouille de données est présenté comme la recherche d'informations pertinentes (des "pépites" d'information) pour l'aide à la décision et la prévision. Il met en œuvre des techniques statistiques et d'apprentissage machine en tenant compte de la spécificité de grandes à très grandes dimensions des données.

Machine Learning

- Pourquoi: Lorsque les modèles déterministes analytiques font défaut car les phénomènes physique, biologique, …sont trop complexes ou trop bruités.

- Objectifs:
  - Utiliser les données existantes
  - Automatiser des tâches chronophages
  - Développer des algorithmes sans les programmer explicitement
  - Définir des modèles prédictifs

- Moyen: Ensemble d'approches (statistiques) permettant de décrire au mieux le comportement d'un jeu de données à partir d'une série d'observations



- Catégories: **Supervised, Unsurpervised,**

Apprentissage Supervisé

- Prend des caractéristiques identifiées (features) en données d'entrée et ressort des données labélisées ou classifées.

- Il est utilisé pour effectuer des régression et classifications:
  - **Régression:** prédire des nombres réels (variable continue)

  Exemple: prédire le prix de vente d'une maison en fonction de la taille, l'année de construction, le nombre de pièce, la localisation …

  - Méthodes
    - Simple Linear Regression Model: analyses the statistical relationship between two quantitative variables. This technique is mostly used in financial fields, real estate, etc.
    - Lasso Regression: Least Absolute Selection Shrinkage is used when there is a need for a subset of the predictor to minimize the prediction error in a continuous variable.
    - Logistic Regression: It is carried out in cases of fraud detection, clinical trials, etc. wherever the output is binary.
    - Support Vector Regression: SVR is a bit different from SVM. In simple regression, the aim is to minimize the error, while in SVR, we adjust the error within a threshold.
    - Multivariate Regression Algorithm: This technique is used in the case of multiple predictor variables. It can be operated with matrix operations and Python's Numpy library.
    - Multiple Regression Algorithm: works with multiple quantitative variables in both linear and non-linear regression algorithms.

Apprentissage Supervisé

- Prend des caractéristiques identifiées (features) en données d'entrée et ressort des données labélisées ou classifées.

- Il est utilisé pour effectuer des régression et classifications:
  - **Classification:** affilier à une ou plusieurs catégories dans un système binaire ou multi-classes

  Exemple binaire: automatiser le processus d'évaluation d'accord de prêt (categories: oui, non) à partir de paramètres comme l'âge, le revenu, l'éducation, la ville, etc

  - Méthodes
    - Logistic Regression,
    - Decision Tree,
    - Random Forest,
    - Multilayer Perception, etc.

Apprentissage non-supervisé

- Les techiques d'apprentissage non supervisé n'ont pas pour objectif de labelliser les données en sortie. Elles sont utilisées pour regrouper (**clustering**) les données à partir de différentes caractéristiques d'entrée.

- Exemple: Répartir un groupe de 100 personnes en sous-groups de 5 à partir de données d'intérêt, de loisirs, de réseaux sociaux, …

- Méthodes:
  - Méthodes basées densité: les clusters sont formés de regions denses selon leur similitudes et differences par rapport aux regions peu denses.
  - Méthodes hierarchiques agglomeratives (ascendante) ou divisives (déscendante): Les clusters sont organisés sous forme d'arbres: de nouveaux clusters sont formés à partir des cluster précédents.
  - Méthodes de partitions: les objects sont partitionnés à partir de k-clusters et chaque méthode forme un cluster unique.
  - Méthodes basées "Gris": les données sont combinées dans une grille…

Avantages/Inconvénients

- **Advantages:**
  - **Automate time-consuming tasks:** ML-based applications have automated several tasks like low-level decision making, data entry, tele-calling, and loan approval processes.
  - **Cost saving:** Once the algorithm is developed and put into production, it can cause significant cost savings as human labor and decision-making are minimal.
  - **Turnaround time:** For a lot of applications, total time is of paramount importance. ML has reduced time in domains such as auto insurance claims, where users upload pictures, and the insurance amount gets calculated. It has also helped e-commerce companies in handling returns of inventory sold.
  - **Data-driven decision making:** Not only corporates but many governments are relying on ML to decide which projects to invest in and how to utilize existing resources optimally.

- **Disadvantages:**
  - **ML algorithms can be biased:** Often, input data to the ML algorithm is biased to a specific gender, Country, Caste, etc. This results in ML algorithms propagating unwanted bias into the decision-making process. This has been observed in some applications which deployed ML-like school/college admission process and social media recommendations.
  - **Require large data to achieve acceptable accuracy:** While people can learn easily for small datasets, for some applications, introduction to machine learning requires huge amounts of data to achieve sufficient accuracy.
  - **ML technique trained on the current dataset may not be well suited for the future** as input distribution may change significantly over time. One of the countermeasures to overcome this is to re-train the model periodically.

Exemples généraux

- Identifier les facteurs de risque d'un certain type de cancer, en fonction de variables cliniques et démographiques: rechercher des gènes potentiellement impliqués dans une maladie à partir de données de biopuces ou plus généralement des biomarqueurs pour un **diagnostic précoce**,
- Identifier des chiffres manuscrits sur un code postal à partir d'une **image** digitalisée,
- Prévoir le prix d'un stock dans 6 mois à partir de mesures de performance de l'entreprise et de **données économiques**,
- **Prévision de la consommation** électrique pour un client EDF en fonction de variables climatiques et de caractéristiques spécifiques à ce client,
- **Maintenance préventive** à partir de relevés d'incidents,
- Construire un modèle de substitution à un code numérique complexe qui permet de prédire une **carte de concentration** d'un polluant dans un sol un an après un rejet accidentel en fonction de la carte initiale et des caractéristiques du sol (porosité, perméabilité...).

Applications (exemples)

- A vous de trouver application pour laquelle le machine learning serait utile dans votre pratique professionnelle (ou autour)?

Sources

- Agence Universitaire de la Francophonie – MOOC – L'analyse en composante principale en pratique (en ligne)
- Universite de Rennes 2 Statistiques des données M1-GEO (en ligne)

- Statistiques Générales pour les Utilisateurs. 1 Méthodologie & 2 Exercices corrigés. J. Pagès (B.U.)

- cours-gratuit.com
- https://www.educba.com/introduction-to-machine-learning/