

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

*Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique*

*Ecole Supérieure d'Informatique
De Sidi Bel Abbès*



مدرسة العليا للإعلام الآلي
بسبدي بلعباس

Analyse En composantes principales

Présenté par Dr.Nabil KESKES.

Année 2018-2019.

PLAN

- Introduction
- Nature de données étudiées
- Démarches de la méthode
- Algorithme General
- Conclusion

*Ecole Supérieure d'Informatique
De Sidi Bel Abbés*



مدرسة العليا للإعلام الآلي
بمبدي بلعباس

1. Introduction

1.1 Definition

Analyse en composantes principales ,notée en abrégé par ses initiales A.C.P est l'une des méthodes **factorielle** de données **multidimensionnelles** les plus courantes. Elle donne une description des unités statistique et des variables observées fondée sur l' étude des coefficients de corrélation linéaire.

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

*Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique*

Ecole Supérieure d'Informatique

De Sidi Bel Abbés



مدرسة العليا للإعلام الآلي
بسيدي بلعباس

1.2 Objectifs

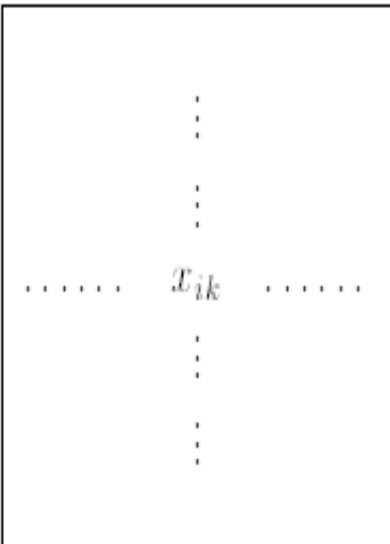
- ✓ Former des groupes homogènes d'unités statistiques.
- ✓ Analyses les liaisons entre les variables.

Ces deux points relèvent de la statistique descriptive: on cherche à mettre en évidence les propriétés fondamentales des données, à l'aide de paramètres numérique et de graphique

2. Nature de données

2.1 Données initiales

Les données pour l'ACP sont généralement présentées sous la forme d'un tableau rectangulaire des individus en lignes et des variables **quantitatives** en colonnes.

		VARIABLES				
		1	k	K
INDIVIDUS	1					
	\vdots					
	\vdots					
	i					
	\vdots					
	\vdots					
	I					

le terme général a la i^{eme} ligne et la k^{eme} colonne est noté x_{ik} et c'est l'observation de la variable k sur l'individu i .

Figure 2.1 : Tableau de données

2.2 Exemples

Analyse sensorielle: note du **descripteur** k pour **le produit** i .

Ecologie: concentration du **polluant** k dans **la rivière** i .

Economie: valeur de l'**indicateur** k pour **l'année** i .

Biologie: **mesure** k pour l'**animal** i .

Marketing : valeur **d'indice de satisfaction** k pour **la marque** i .

2.3 Exemple

	Poids	Taille	Age	Note
1	45	1.50	13	14
2	50	1.60	13	16
3	50	1.65	13	15
4	60	1,75	15	9
5	60	1,70	14	10
6	60	1,70	14	7
7	70	1,60	14	8
8	65	1.60	13	13
9	60	1.55	15	17
10	65	1.70	14	11

tableau 2.2 : Exemple de données

Remarques

- ✓ Les données traitées par l'ACP doivent être quantitative (homogène ou hétérogène, discrète ou continue).
- ✓ l'ACP ne donnera des résultats intéressants que sur les tableaux suffisamment grand (nb des unités statistique >15 et le nb de variables >4).
- ✓ On peut introduire dans le tableau des données que l'on appelle supplémentaire (passives) pour faciliter l'interprétation des résultats.

3. Démarches de la méthode

3.1 Notion de distance

✓ Il est évident que deux unités statistiques se ressemblent si les variables observées prennent des valeurs voisines sur ces deux unités statistique

4	60	1,75	15	9
5	60	1,70	14	10
6	60	1,70	14	7

$$d^2(4,5) = (60-60)^2 + (1.75-1.70)^2 + (15-14)^2 + (9-10)^2 = 2.00025$$

$$d^2(4,6) = (60-60)^2 + (1.75-1.70)^2 + (15-14)^2 + (9-7)^2 = 5.00000$$

$$d^2(5,6) = (60-60)^2 + (1.70-1.70)^2 + (14-14)^2 + (10-7)^2 = 9.00000$$

Exprimons maintenant la taille en centimètres

$$d^2(4,5) = (60-60)^2 + (175-170)^2 + (15-14)^2 + (9-10)^2 = 27$$

$$d^2(4,6) = (60-60)^2 + (175-170)^2 + (15-14)^2 + (9-7)^2 = 30$$

$$\mathbf{d^2(5,6) = (60-60)^2 + (170-170)^2 + (14-14)^2 + (10-7)^2 = 9.00000}$$

Lorsque la taille est exprimée en mètre ,l' élève 6 est donc plus proche de l' élève 4 que 5,lorsque elle est exprimée en centimètre , c'est l'inverse.

La distance ainsi définie dépend donc des unités de mesure choisie .Pour stabiliser la distance, le procédé habituel consiste a **centrer et a réduire** les variable

Calculons les distances entre les élèves 4,5,6

$$d^2(4,5)= (60-60)^2 / 55.25 +(1.75-1.70)^2 / 0.005525 +(15-14)^2 / 0.56 +(9-10)^2 / 11 =2.328$$

$$d^2(4,6)= (60-60)^2 / 55.25 +(1.75-1.70)^2 / 0.005525 +(15-14)^2 / 0.56 +(9-7)^2 / 11 =2.601$$

$$d^2(5,6)= (60-60)^2 / 55.25 +(1.70-1.70)^2 / 0.005525 +(14-14)^2 / 0.56 +(10-7)^2 / 11 =0.819$$

La distance ne dépend plus maintenant des unités mesures dans lesquelles sont exprimées les variables. Donc on peut affirmer que l' élève 6 est plus proche de l' élève 5 que l' élève 4 **indépendamment** des unités de mesure.

3.2 Description de la methode

On peut calculer ,par la formule précédente, toutes les distances entre les unités statistique: dans le cas de 10 unités statistique , cela donne $10 * 9 / 2$ distances ,soit 45.
dans le cas 100 unités statistiques, on obtient $100 * 99 / 2$,soit 4950,

⇒ ce qui est impossible d'analyser directement .

L'ACP, pour décrire le mieux possible les donnée fournit un système d'axe orthonormé conservant l'ensemble de ces distance .

⇒ Les axes possédant cette priorités sont les droites les plus proches des observations suivant le critère de moindre Carré .

3.2.1 Definitions

✓ Les droites les plus proches des unités statistique sont appelées axes principaux.

Leurs vecteurs directeurs sont appelés **vecteurs principaux**

⇒ Les vecteurs principaux ,qui engendrent les axes principaux , sont les vecteurs propres , de la matrice de corrélation associés aux valeurs propres.

Exemple

On donne dans cet exemple les corrélations entre les variables initiales ,les valeurs propres, et les vecteurs principaux obtenus par L'ACP.

	poids	taille	age	note
poids	1.0000000	0.3665158	0.4854043	-0.5678917
taille	0.3665158	1.0000000	0.3955146	-0.6287373
age	0.4854043	0.3955146	1.0000000	-0.3223292
note	-0.5678917	-0.6287373	-0.3223292	1.0000000

\$values

[1] 2.3908988 0.7503114 0.5844061 0.2743837

\$vectors

	[,1]	[,2]	[,3]	[,4]
[1,]	-0.5079730	0.3065293	0.6593007	0.4618721
[2,]	-0.5038351	-0.4647035	-0.5253477	0.5041931
[3,]	-0.4453032	0.7057876	-0.4712381	-0.2854882
[4,]	0.5383481	0.4381258	-0.2593591	0.6715355

Remarque

✓ Ces vecteurs sont unitaire et orthogonaux

$$\|U1\|^2 = -0.5080^2 + -0.5038^2 + -0.4453^2 + 0.5383^2 \\ = 1$$

$$U1 \cdot U2 = -0.5079730 * 0.3065293 + -0.5038351 * -0.4647035 + -0.4453032 * 0.705787 + 0.5383481 * 0.438125 \\ = 0$$

✓ on appelle **composante principale** la liste des coordonnées des unités statistiques sur l'axe principal engendré par le vecteur principal

Exemple

✓ Nous calculons les coordonnées de l'unité statistique 1 sur les deux premiers axes principaux

poids: $x_1(1)=45$ $m_1=58.5$, $\sigma=7.43303$

$$X'_1(1) = (45 - 58.5) / 7.43303 = -1.816$$

taille: $x_2(1)=1.50$ $m_2=1.635$, $\sigma=0.07433$

$$X'_2(1) = (1.50 - 1.635) / 0.07433 = -1.816$$

Age: $x_3(1)=13$ $m_3=13.8$, $\sigma=0.74833$

$$X'_3(1) = (13 - 13.8) / 0.74833 = -1.069$$

Note: $x_4(1)=14$ $m_4=12$, $\sigma=3.31662$

$$X'_4(1) = (14 - 12) / 3.31662 = 0.603$$

$$c1(1) = -0.5079730 * -1.816 + -0.5038351 * -1.816 + -0.4453032 * -1.069 + 0.5383481 * 0.603 = -2.638$$

$$c2(1) = 0.3065293 * -1.816 + -0.4647035 * -1.816 + 0.7057876 * -1.069 + 0.4381258 * 0.603 = -0.203$$

L'exemple suivant donne la liste des coordonnées des unités statistique sur les axes principaux obtenus par L'ACP.

	Dim.1	Dim.2	Dim.3	Dim.4
1	-2.6383478	-0.20304035	-0.10408766	-1.04443266
2	-1.9434521	-0.35783346	0.31559230	0.34952125
3	-1.4422179	-0.80252677	0.59077968	0.48620265
4	2.0830449	0.07837172	1.20080516	-0.19195868
5	0.9867493	-0.42008337	0.29589940	0.05285961
6	1.4737035	-0.81638306	0.06130029	-0.55456712
7	1.3169524	0.35329157	-1.45426142	-0.40902793
8	-0.4313990	-0.13555130	-1.24948767	0.67416087
9	-0.5711642	2.38554456	0.41285443	0.07121803
10	1.1661308	-0.08178953	-0.06939452	0.56602398

Remarque

Chaque composante principale définit une nouvelle variable .ces composantes principale possèdent les propriétés suivantes:

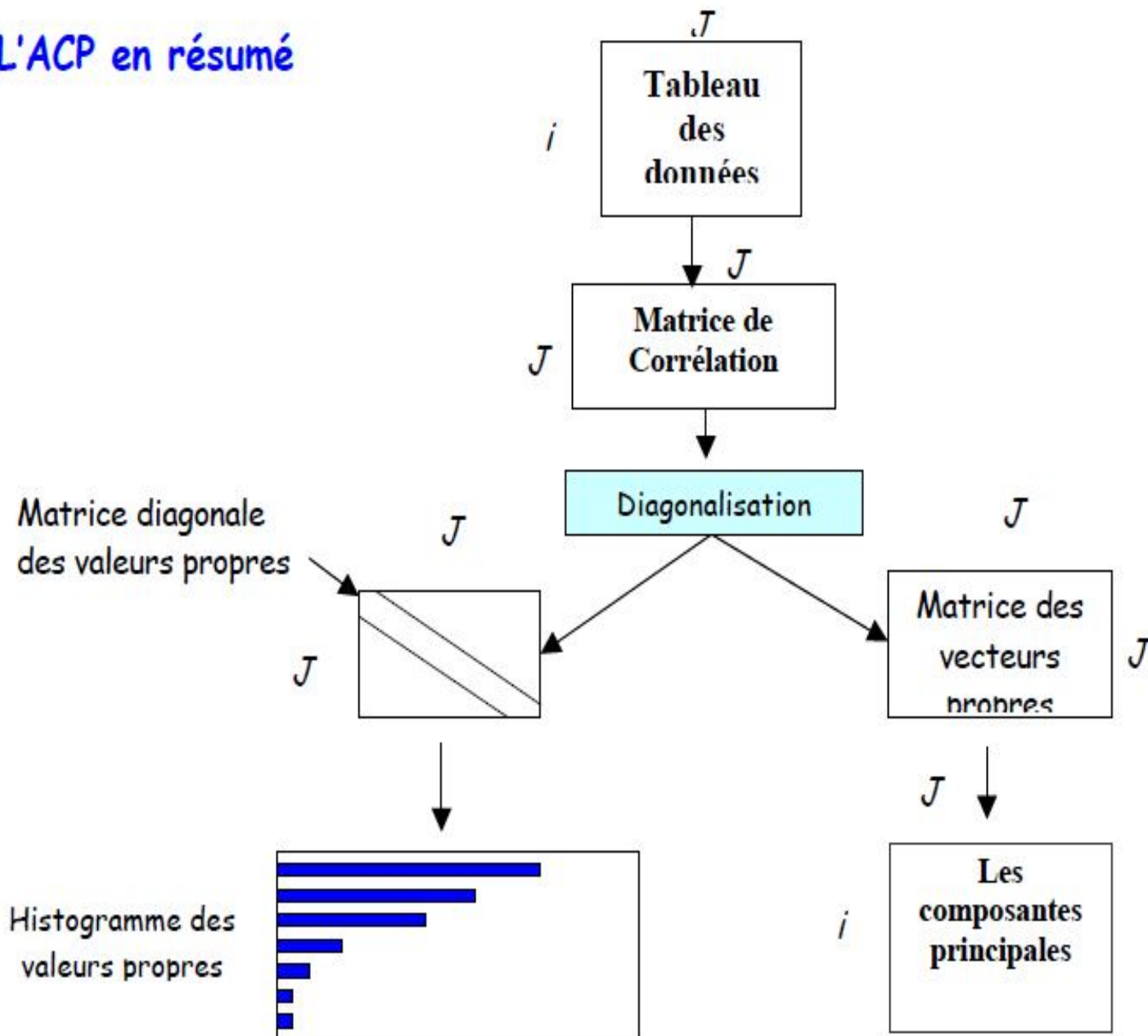
- ✓ les composantes principales sont centrées.
- ✓ Les composantes principales sont non corrélées deux a deux (les coefficients de corrélation sont nuls)

Remarque

Lorsque tous les axes sont considérés , toutes les distances sont exactement reconstruites.

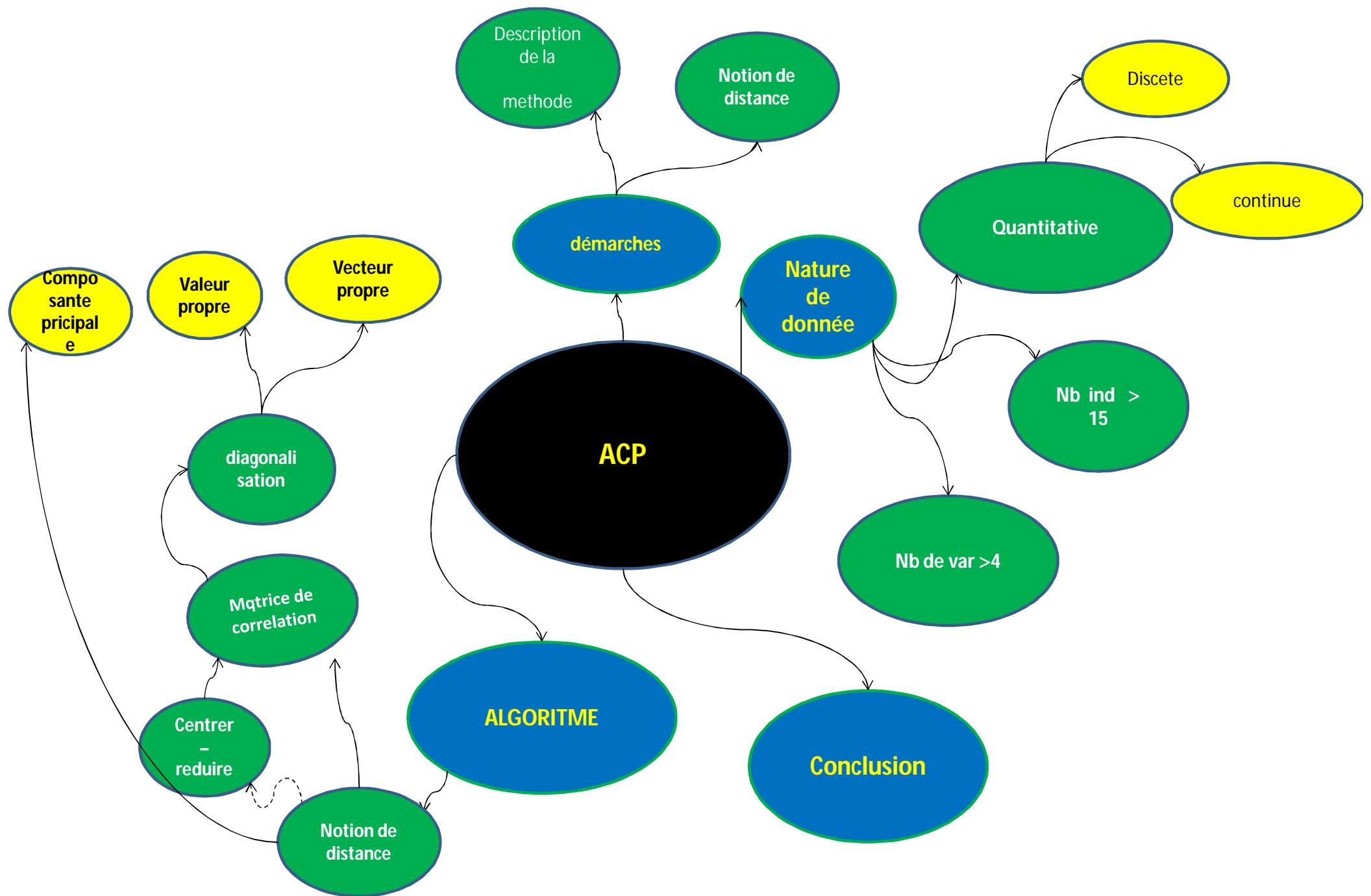
	Axe 1	Axe 1 et 2	Axe 1 ,2,3	Axe 1 ,2,3,4
$d^2(4,5) =$	1.201	1.449	2.268	2.328
$d^2(4,6) =$	0.371	1.170	2.470	2.601
$d^2(5,6) =$	0.237	0.394	0.449	0.819

L'ACP en résumé



4. Conclusion

Les composantes principales sont très grande utilités dans **l'interprétation**: ce sont elles qui permettent d'explicitier les relations entre les variables initiales et de justifier la formation des groupes homogènes d'unités statistique



Une Carte Mentale de l'exposé