

# DATA MINING

# LANALYSE

ACP
ANALYSE EN COMPOSANTES PRINCIPALES

# DES DONNÉES PAYS/PROTÉINES



Pays	VR	VB	Oeufs	Lait		on Céréa:	les Stai	rchNoix	FL
Albanie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5		4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Cheko.	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-	E 8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-	0 11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavi	e 4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

## INTRODUCTION



Mapping de R<sup>n</sup> dans R<sup>k</sup> : avec k<=n.

Projection dans un **espace 2D** d'un problème a n dimensions.

Système d'axes indépendants.

Réduction de la dimensionnalité d'un problème.

Minimisation de la Perte d'information.

Décomposition en valeurs propres.

nouvel espace est une combinaison linéaire de l'espace d'origine.

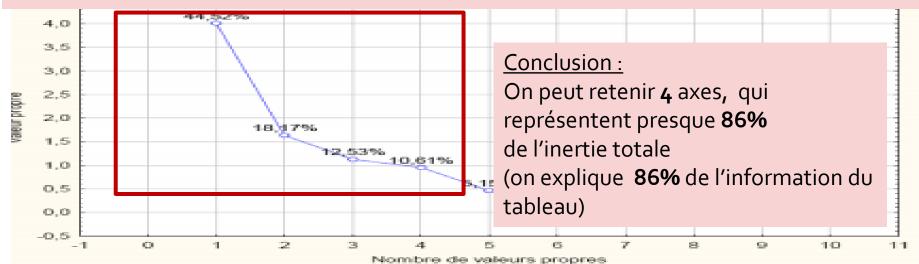
#### **CHOIX DU NOMBRE D'AXES A RETENIR**



	Val. Propres (matrice de corrél.) & stat. associées (proteines: Variables actives seules								
Valeur numéro	Val Propre	% Total variance	Cumul Val Propre	Cumul %					
Calem Burners									
1	4,006438	44,51597	4,006438	44,5160					
2	1,634999	18,16666	5,641437	62,6826					
3	1,127920	12,53244	6,769357	75,2151					
4	0,954664	10,60738	7,724020	85,8224					
5	0,463838	5,15376	8,187859	90,9762					
6	0,325131	3,61257	8,512990	94,5888					
7	0,271606	3,01785	8,784596	97,6066					
8	0,116292	1,29213	8,900888	98,8988					
9	0,099112	1,10124	9,000000	100,0000					

Deux critères empiriques pour sélectionner le nombre d'axes :

- Critère de Kaiser: on ne retient que les axes associés à des valeurs propre supérieures à 1
- Critère du coude : sur l'évolution des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement



#### PROJECTION DU NUAGE DE POINTS



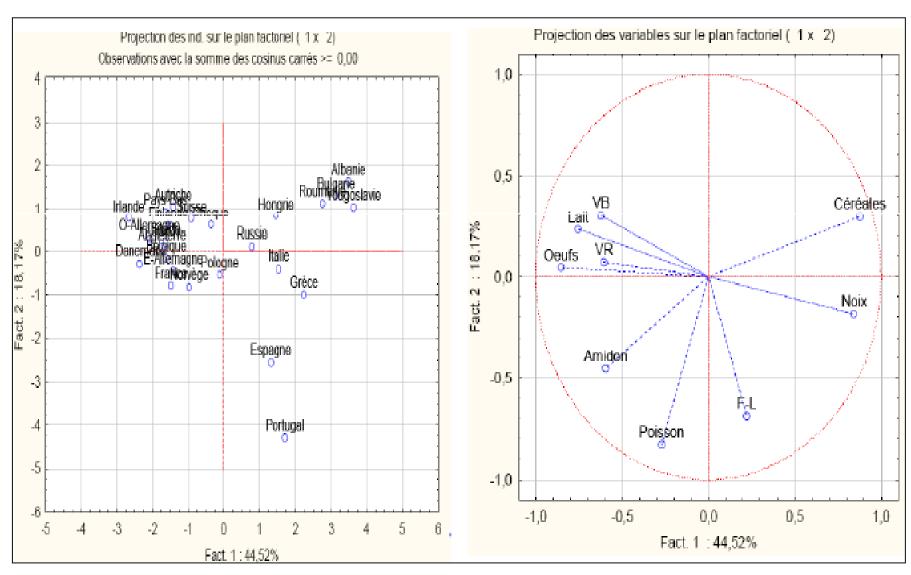
Chaque nuage de points (variables et individus) est construit en projection sur les plans factoriels

un plan factoriel est un repère du plan défini par deux des q axes factoriels retenus.

L'examen des plans factoriels permettra de visualiser les corrélations entre les variables et d'identifier les groupes d'individus ayant pris des valeurs proches sur certaines variables.

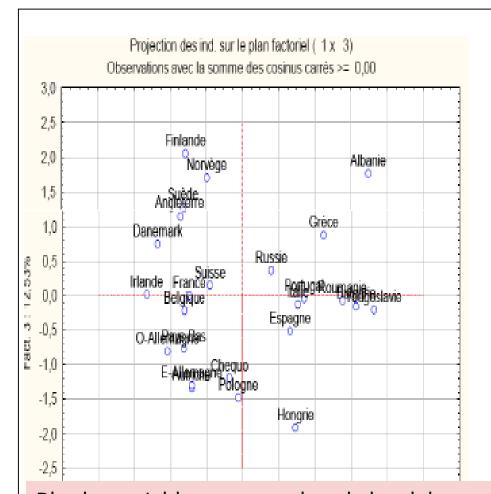
# PLAN FACTORIEL 1 : 1X 2 CARTE DES INDIVIDUS CERCLE DE CORRÉLATION

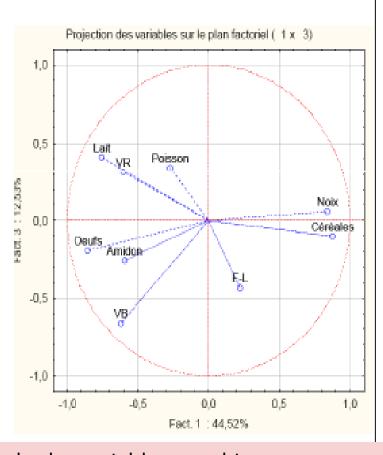




# PLAN FACTORIEL 2 : 1X 3 CARTE DES INDIVIDUS CERCLE DE CORRÉLATION







Plus les variables sont proches du bord du cercle et plus les variables sont bien représentées par le plan factoriel, c'est-a-dire que la variable est bien corrélée avec les deux facteurs constituant ce plan.

### **ETUDE DE PROXIMITÉ ENTRE LES POINTS**



Regarder les graphiques et analyser plus finement les proximités entre points.

Un point est dit bien représenté sur un axe ou un plan factoriel s'il est proche de sa projection sur l'axe ou le plan. S'il est éloigné, on dit qu'il est mal représenté.

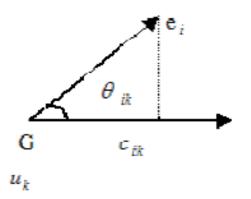
Indicateur = angle formé entre le point et sa projection sur l'axe :

au plus il est proche de 90 degrés, au moins le point est bien représenté

L'analyse se fera à l'aide des individus et variables contribuant le plus à l'axe : si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

## QUALITÉ DE REPRÉSENTATION DES INDIVIDUS





$$qlt_k(e_t) = \cos^2(\theta_{tk}) = \frac{c_{tk}^2}{\|e_t\|^2}$$

$$||e_i||^2 = \sum_{k=1}^p c_{ik}^2$$

- Lorsque l'angle est proche de 0, c'est-à-dire que l'individu est bien représenté, le cosinus est proche de 1.
- Dans le cas inverse, l'angle est proche de 90° et le cosinus est proche de 0.

## PROXIMITÉ ENTRE LES POINTS



- La proximité dans l'espace entre deux individus bien représentés traduit la ressemblance de ces deux individus de point de vue des valeurs prises par les variables.
- Lorsque la qualité de représentation de deux individus est bonne, leur proximité observée retrace leur proximité réelle (dans l'espace).
- La proximité entre deux variables sur un axe donne, si les deux variables sont bien représentées sur l'axe (proches de l'axe et du bord du cercle), une approximation de leur corrélation :
- Deux variables proches sont corrélées positivement
- Deux variables qui s'opposent sont corrélées négativement
- Deux variables orthogonales sont non corrélées.

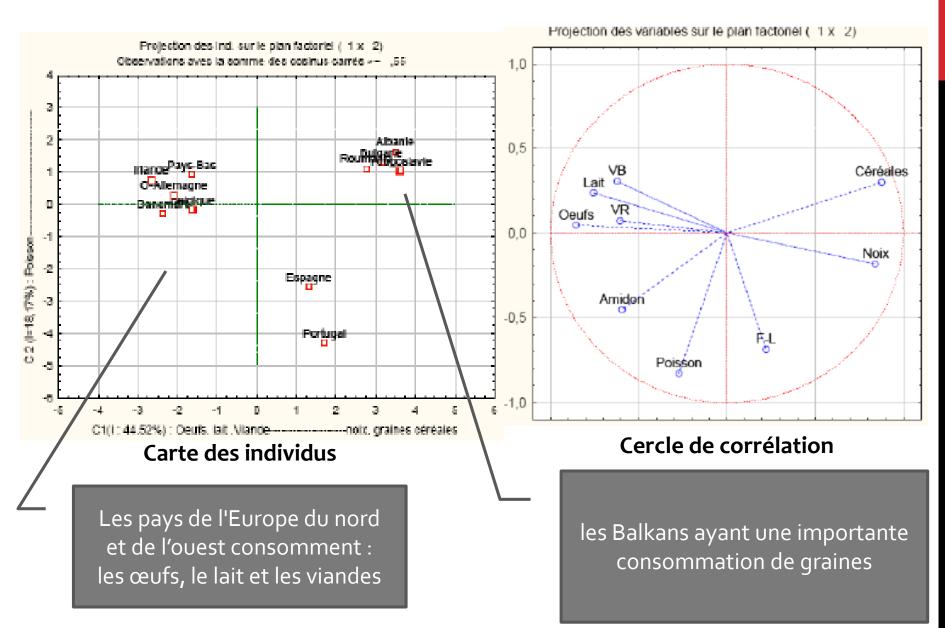


# QUALITÉ DE REPRÉSENTATION : EXEMPLE

	Cosinus c	Cosinus carrés, basées sur les corrélations (proteines2)					
Individus	Fact. 1	Fact. 2	Fact 3	Fact. 4	fact1"2	fact1*3	fact1"4
Albanie	0.612154	0.133966	0.156314	0.002658	0.746119	0.768467	0.614811
Autriche	0,337993	0,181048			0,51904054		
Belgique	0,709293	0.006858	0.012640	0.073101	0.71615085	0.72193315	0.78239405
Bulgarle	0,740647	0.127641	0.001726		0.86828752		0.74410514
Chequo	0.045579	0.120522	0.474998	0.071495	0.16620146	0.5205766	0.11707367
Danemark	0,662846	0.009654	0.067049			0.72989438	
E-Allemagne	0,325524	0.032633	0.273047		0.35815734		0,53319811
Finlande	0.237490		0.407898	0.194516	0.27198467		0.43200585
France	0,253759	0,070692	0.000000		0.32445026		0,69290548
Grêce	0.426677	0.085239	0.066260	0.273854	0.51191654	0.49293733	0.70053178
Hongrie	0,276110	0,086543	0,476279	0,006143	0,362653	0,75238942	0,28225333
Irlande	0,776498	0,063840	0.000043	0,020686	0.84033882	0,7765417	0,7971845
Italle	0,445901	0,030143	0,003010	0,282968	0,47604352	0,44891104	0,72886841
Pays-Bas	0,530279	0,163694	0,115628	0,003132	0,69397267	0,64590662	0,53341092
Norvège	0,157443	0,111983	0,481236	0,214475	0,26942693	0,63867957	0,37191797
Pologne	0,003157	0,060106	0,462355	0,044634	0,06326324	0,46551268	0,04779152
Portugal	0,127331	0,804916	0,000083	0,034938	0,93224748	0,12741472	0,16226949
Roumanie	0,801256	0,131962	0,000518	0,039877	0,93321825	0,8017736	0,84113285
Espagne	0,172268	0,652748	0,026580	0,012917	0,82501618	0,19884861	0,18518502
Suède	0,454371	0,007322	0,279076	0,091741	0,46169263	0,73344706	0,54611181
Sulsse	0,206235	0,139771	0,005896	0,339446	0,34600535	0,21213034	0,54568045
Angleterre	0,333229	0,000977	0,147021	0,332586	0,33420631	0,4802495	0,66581465
Russle	0,128014	0,002565	0,028565	0,179837	0,13057881	0,15657902	0,30785084
O-Allemagne	0,795640	0,015663	0,117306	0,002148	0,81130324	0,91294668	0,79778838
Yougoslavie	0,860599	0,070645	0,002784	0,044252	0,93124418	0,86338274	0,90485107
_							1

#### **ÉTUDE DES POINTS BIEN REPRÉSENTÉS 1 X 2**

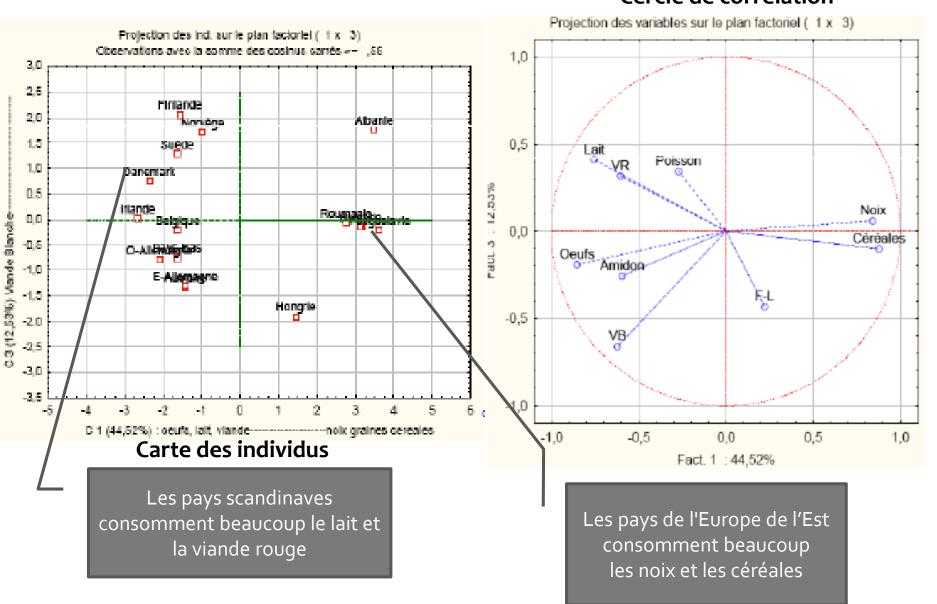




#### **ÉTUDE DES POINTS BIEN REPRÉSENTÉS 1 X 3**



#### Cercle de corrélation



## INTÉRÊTS DE L'ACP



## Visualisation

 Représentation assez fidèle des individus d'une population en 2 dimensions via de nouvelles variables

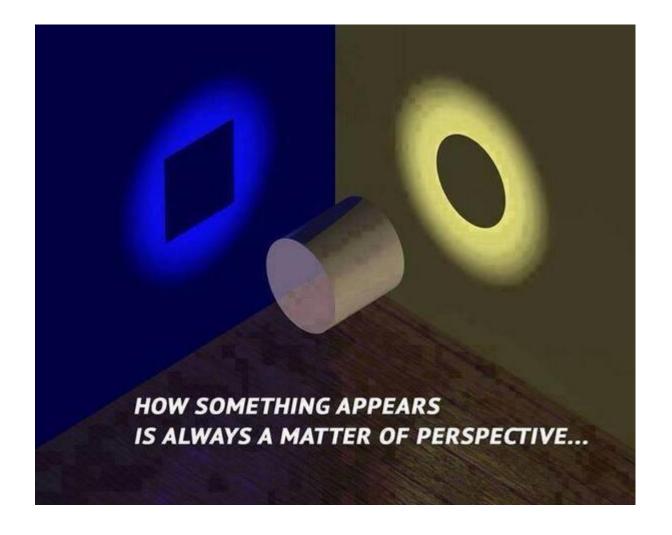
# Pertinence de Variables

 Détection des facteurs les plus pertinents dans une dynamique observée

Relation
V / V, I/I, ou V/I

 Mesure du taux de dépendance entre les variables et les individus





Difficulté à mettre en évidence les relations globales existant entre les variables dès que p>3

# DES DONNÉES PAYS/PROTÉINES



Pays	VD	T/D	Court	e Tait	Doie	eon <sup>c</sup> óróa	log Cto	rch Noiv	PT.
Albanie	10.1	1.4	0.5	8.9	0.2	12.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Choko	0.7	11.4	2.0	10 5	2.0	24.3	5.0	11	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-E	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0		3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2		2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9		4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2		5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1		3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4		4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-O	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavie	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

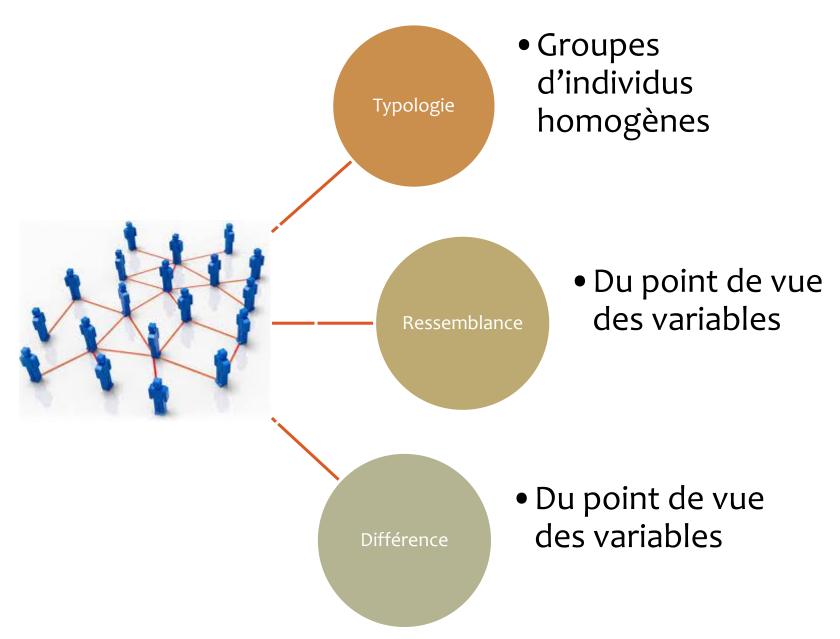
# DES DONNÉES PAYS/PROTÉINES



Pays	WD.	WD.	Ooufe	Lait	Doice	on Cáráal	ac Ctar	n Noiv	PI.
Albanie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Cheko.	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-E	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-0	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavie	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

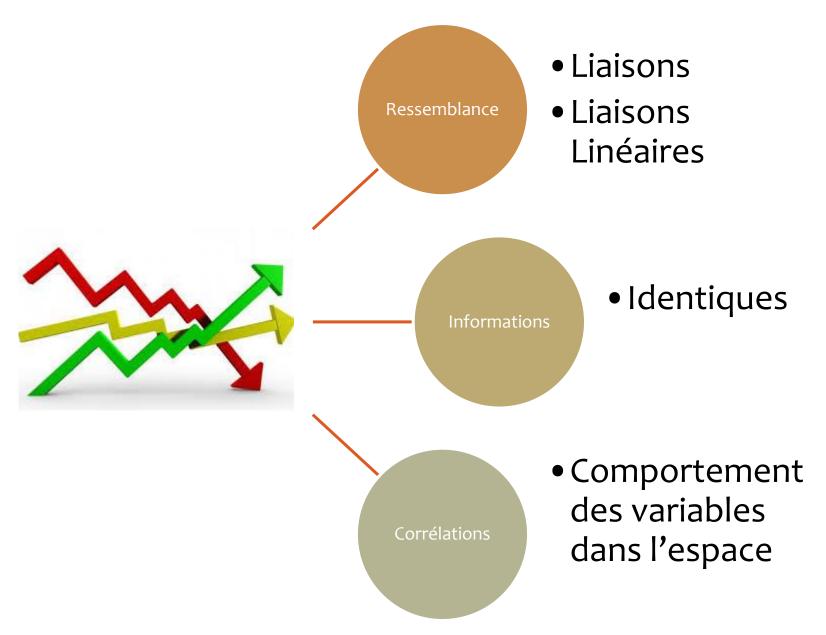
### **OBJECTIFS - ÉTUDES DES INDIVIDUS**





### **OBJECTIFS – ÉTUDES DES VARIABLES**





#### **OBJECTIFS DES DEUX ÉTUDES**



Caractérisations

Des classes d'individus par les variables

Besoin de Procédure automatique Compréhension

Individus spécifiques / liaisons entre variables

À l'aide de visualisations graphiques

- Condenser l'information de l'entrepôt de données
  - de manière à retirer les relations vraiment caractéristiques
    - en limitant la perte d'information.



## **CENTRAGE ET RÉDUCTION**



# Centrer les données ne modifie pas la forme du nuage :

c'est translater!

• On peut toujours procéder par un centrage de l'entrepôt

Réduire les données :

Indispensable si les unités de mesure sont différentes!

$$\bullet \ x_{ij} = \frac{x_{ij} - \overline{x_k}}{S_k}$$

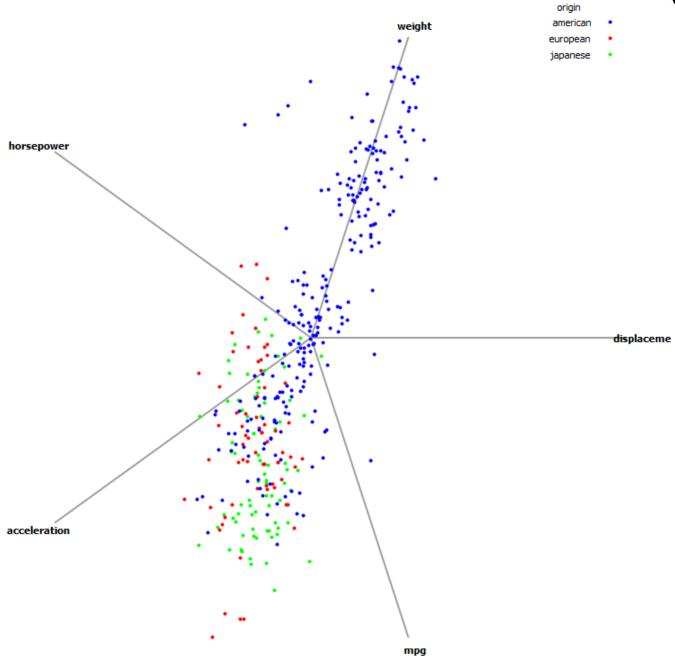
• Permet de comparer les valeurs prises par les variables



#### **ENTREPÔT DE DONNÉES – PLUSIEURS PROJECTIONS POSSIBLES**

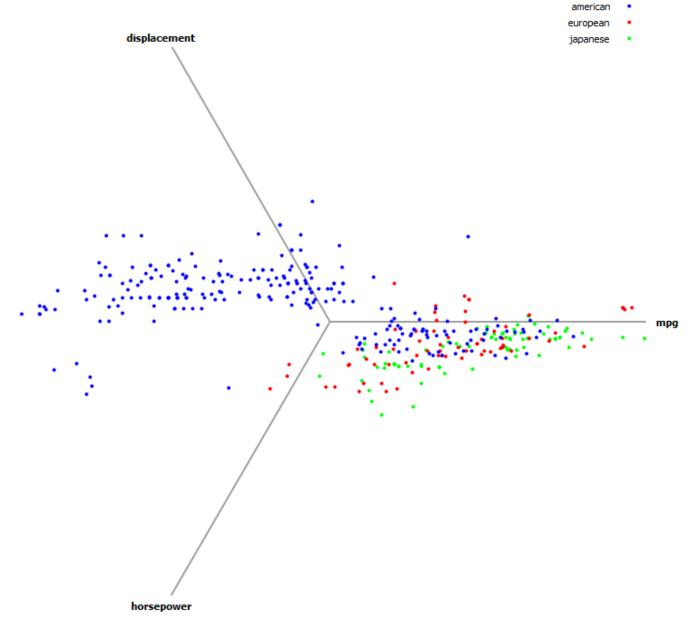
	mpg	displacement	horsepower	weight	acceleration	origin
L	35	72	69	1613	18	japanese
2	31	76	52	1649	17	japanese
3	39	79	58	1755	17	japanese
4	35	<u>8</u> 1	60	<u>1</u> 760	16	japanese
5	31	71	<u>65</u>	<u>1</u> 773	19	japanese
6	33	<u>9</u> 1	53	<u>1</u> 795	18	japanese
7	33	<u>9</u> 1	<u>5</u> 3	<u>1</u> 795	17	japanese
8	36	<u>98</u>	<u>66</u>	1800	14	american
9	36	<u>9</u> 1	<u>60</u>	1800	16	japanese
10	30	<u>9</u> 7	71	1825	12	european
11	36	<b>7</b> 9	<u>58</u>	1825	19	european
12	27	<u>9</u> 7	60	1834	19	european
13	26	<u>9</u> 7	46	1835	21	european
14	32	71	<u>65</u>	1836	21	japanese
15	30	89	<u>62</u>	1845	15	european
16	45	91	<u>67</u>	1850	14	japanese
17	29	68	49	1867	20	european
18	39	<u>8</u> 6	<u>64</u>	1875	16	american
19	36	<u>98</u>	80	1915	14	american
20	32	<u>8</u> 9	71	1925	14	european
21	29	<u>9</u> 0	70	1937	14	european
22	29	<u>9</u> 0	70	1937	14	european
23	29	<u>9</u> 7	78	1940	15	european





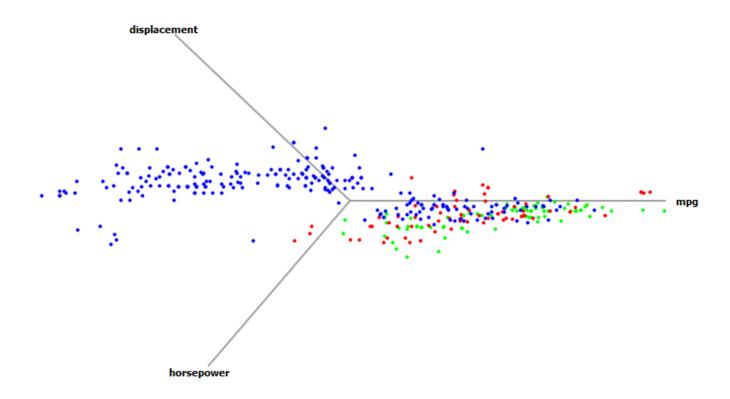


origin



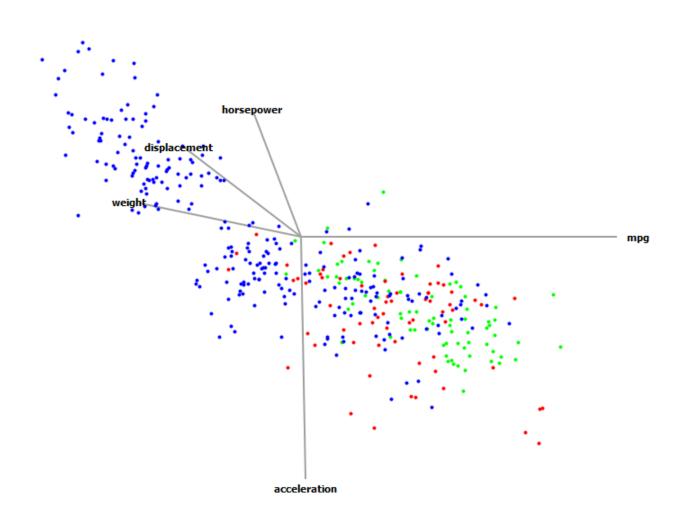


origin american european japanese







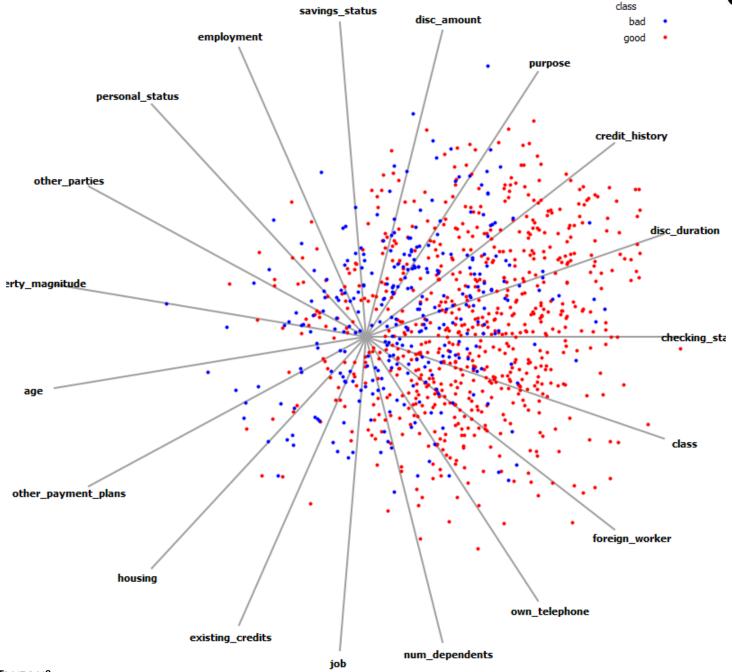




#### ENTREPÔT DE DONNÉES – PLUSIEURS PROJECTIONS POSSIBLES

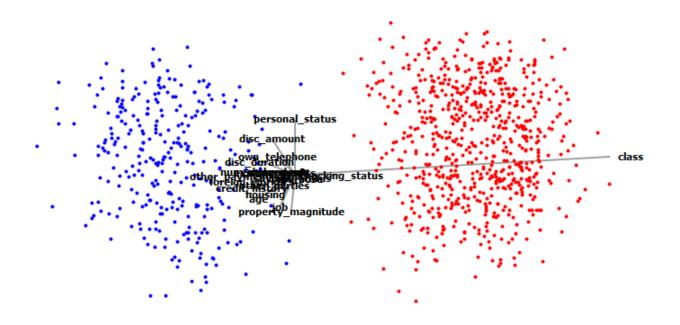
	hecking_statu	lisc_duration	credit_history	purpose	lisc_amoun	savings_status	employment	personal_st
1	<0	lo_1_year	critical/other existing	radio/tv	1000_2000	no known savings	>=7	male single
2	0<=X<200	up_2_years	existing paid	radio/tv	up_2000	<100	1<=X<4	female div/de
3	no checking	lo_1_year	critical/other existing	education	up_2000	<100	4<=X<7	male single
4	<0	up_2_years	existing paid	furniture/equipment	up_2000	<100	4<=X<7	male single
5	<0	1_2_years	delayed previously	new car	up_2000	<100	1<=X<4	male single
6	no checking	up_2_years	existing paid	education	up_2000	no known savings	1<=X<4	male single
7	no checking	1_2_years	existing paid	furniture/equipment	up_2000	500<=X<1000	>=7	male single
8	0<=X<200	up_2_years	existing paid	used car	up_2000	<100	1<=X<4	male single
9	no checking	lo_1_year	existing paid	radio/tv	up_2000	>=1000	4<=X<7	male div/sep
10	0<=X<200	up_2_years	critical/other existing	new car	up_2000	<100	unemployed	male mar/wid
11	0<=X<200	lo_1_year	existing paid	new car	1000_2000	<100	<1	female div/de
12	<0	up_2_years	existing paid	business	up_2000	<100	<1	female div/de
13	0<=X<200	lo_1_year	existing paid	radio/tv	1000_2000	<100	1<=X<4	female div/de
14	<0	1_2_years	critical/other existing	new car	1000_2000	<100	>=7	male single
15	<0	1_2_years	existing paid	new car	1000_2000	<100	1<=X<4	female div/de
16	<0	1_2_years	existing paid	radio/tv	1000_2000	100<=X<500	1<=X<4	female div/de
17	no checking	1_2_years	critical/other existing	radio/tv	up_2000	no known savings	>=7	male single
18	<0	up_2_years	no credits/all paid	business	up_2000	no known savings	<1	male single
19	0<=X<200	1_2_years	existing paid	used car	up_2000	<100	>=7	female div/de
20	no checking	1_2_years	existing paid	radio/tv	up_2000	500<=X<1000	>=7	male single
21	no checking	lo_1_year	critical/other existing	new car	up_2000	<100	1<=X<4	male single
22	<0	lo_1_year	existing paid	radio/tv	up_2000	500<=X<1000	1<=X<4	male single
23	<0	lo_1_year	critical/other existing	new car	up_2000	<100	<1	male single





ESPI Ecole Supérieure Privée d'Impérieure et de Technologies





#### AJUSTEMENT DU NUAGE DE POINTS



### ACP fournit une image simplifiée de N'

- La plus fidèle possible
- À travers le sous-espace qui résume le mieux les données

### La qualité de l'image de N'

- Restitue fidèlement la forme générale du nuage
- Meilleur représentation de la diversité et de la variabilité

La quantification de la qualité de l'image :

Inertie totale = variance généralisée à plusieurs dimensions

## Diagonalisation de données sous forme matricielle

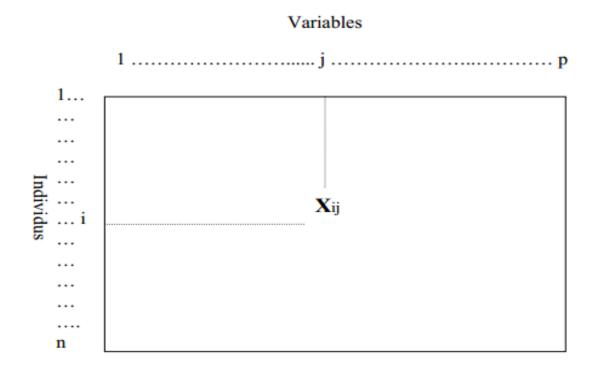


### **MATRICE DE DONNÉES**



On possède un tableau rectangulaire dont :

- ✓ les colonnes sont des variables quantitatives (mensurations, taux,...)
- ✓ les lignes représentent des individus statistiques (unités élémentaires telles que des êtres humains, des pays, des années...)



X: Tableau de données
Xij: Valeur de la ième observation pour la jième variable
Xi.: ième observation du tableau
X.j: jième variable du tableau
n: effectif des individus
p: nombre de variables

# **EXEMPLES D'APPLICATIONS**















#### **ETAPES DE L'ACP**



#### Choix du tableau X

Analyse directe: Construction de l'espace factoriel du nuage de <u>points individus</u> associé au tableau. On garde pour l'instant les p axes factoriels

Analyse duale : Construction de l'espace factoriel du nuage de <u>points variables</u> : elle est déduite de la première

Interprétation de ces analyses : choix du nombre d'axes q à retenir, construction des nuages de points projetés sur ces axes, interprétation des axes principaux et étude des proximités entre points.

Synthèse des résultats, construction éventuelle du tableau réduit (tableau des composantes principales) et visualisation des nuages de points associés.

#### LES COMPOSANTES PRINCIPALES



Les composantes principales : permettent d'exprimer les variables initiales selon de nouveaux axes: les axes principaux, qui sont les vecteurs propres de la matrice :

- des covariances si on a des données hétérogènes, avec des ordres de grandeur différents
- des corrélations lorsque les unités de mesure ne sont pas les mêmes pour toutes les variables



#### ETUDE D'UN ENTREPÔT DE MODÈLE DE VOITURES

- On propose d'analyser un échantillon de 37 véhicules caractérisés par 11 variables quantitatives via une Analyse en Composantes Principales.
- Liste des variables actives :

PUIS puissance,
CYLI cylindrée,
VITE vitesse,
LONG longueur,
LARG largeur,
HAUT hauteur,
POID poids,
COFF coffre,
RESE réservoir,
CONS consommation,
CO2 emission CO2

Variable illustrative : PRIX prix



## **TABLEAU DES VALEURS PROPRES**

NUMERO	 	VALEUR PROPRE	POURCI	ENTAGE	RCENTAGE CUMULE	 
1		7.4633	6'	7.85	67.85	***************************************
2		1.4790	1	3.45	81.29	********
3		0.9396		3.54	89.84	*******
4		0.3534	į :	3,21	93.05	****
5		0.2764		2.51	95.56	***
6		0.1995	:	1.81	97.38	***
7		0.1255	:	1.14	98.52	**
8		0.0892	(	0.81	99.33	*
9		0.0554	(	0.50	99.83	*
10		0.0148	(	).13	99.97	*
11		0.0038	(	0.03	100.00	*



