

L'analyse des données à l'usage des non mathématiciens

2^{ème} Partie: L'analyse en composantes principales

AGRO.M - INRA - Formation Permanente
Janvier 2006

André Bouchier

Copyright © André Bouchier.

© 2006, André Bouchier (20 Janvier 2006)

Permission est accordée de copier et distribuer ce document, en partie ou en totalité, dans n'importe quelle langue, sur n'importe quel support, à condition que la notice © ci-dessus soit incluse dans toutes les copies. Permission est accordée de traduire ce document, en partie ou en totalité, dans n'importe quelle langue, à condition que la notice © ci-dessus soit incluse.

1. Quantifier la variabilité contenue dans un tableau de données :

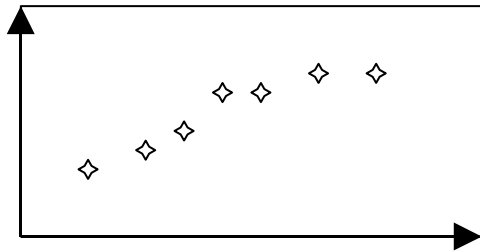
- On appelle *inertie* la quantité d'information contenue dans un tableau de données.
- Une *inertie* nulle signifie que tous les individus sont presque identiques.
- L'*inertie* du nuage sera égale à la somme des variances des j caractères.
- Si les j caractères sont centrés-réduits, l'*inertie* sera égale à j .

2. Projeter sur un plan un tableau de données à j dimensions

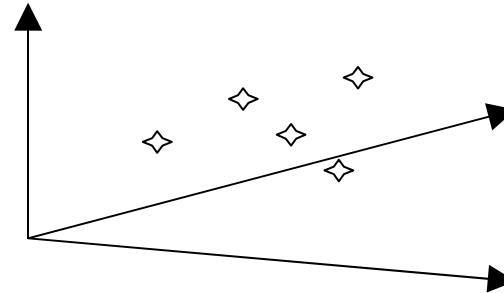
- L'ACP est une méthode descriptive.
- Son objectif est de représenter sous forme graphique l'essentiel de l'information contenue dans un tableau de données quantitatif.
- Dans un tableau de données à j variables, les individus se trouvent dans un espace à j dimensions.

3. La représentation graphique

Lorsqu'il n'y a que deux dimensions (largeur et longueur par exemple), il est facile de représenter les données sur un plan :



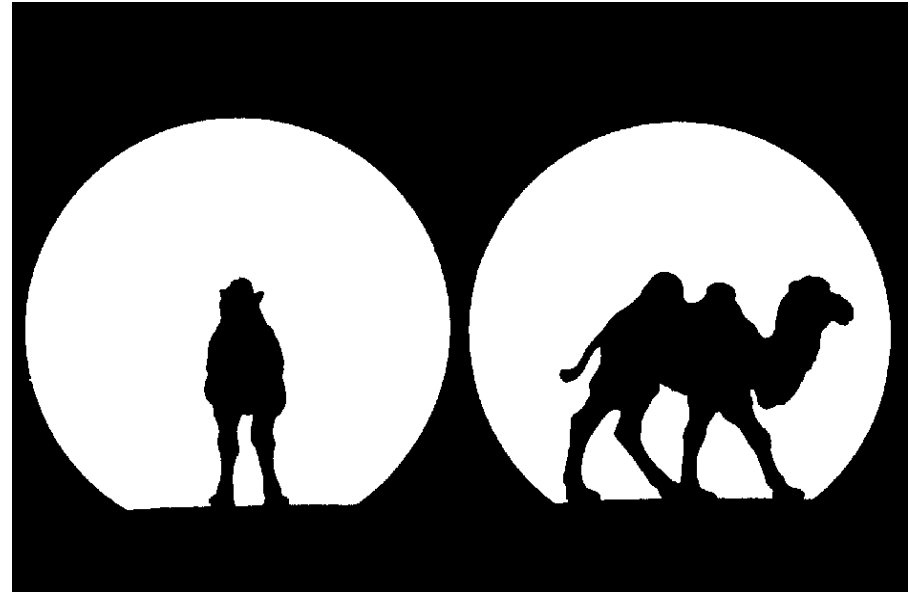
Avec trois dimensions (largeur, hauteur et profondeur par ex.), c'est déjà plus difficile :



- Mais au delà de 3 dimensions, il est impossible de représenter les données sur un plan ou même de les visualiser mentalement.

4. Projeter la réalité sur un plan

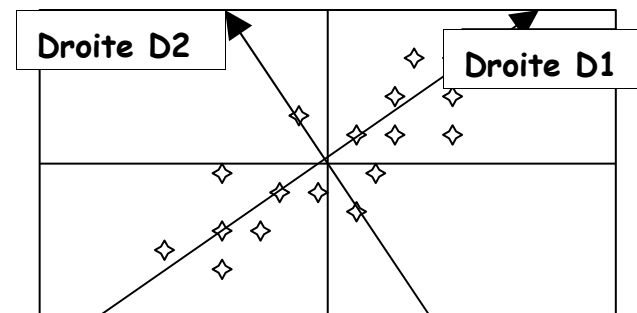
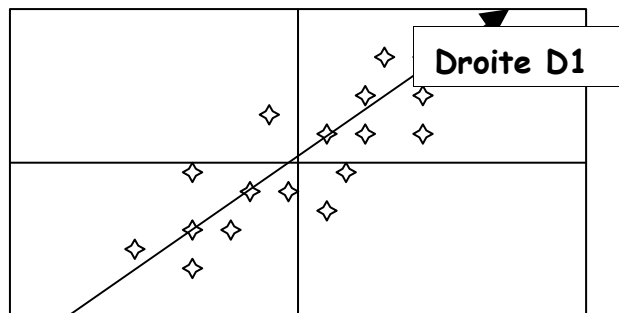
Figure de J.P. Fenelon



- Nous avons l'habitude de dessiner ou photographier la réalité.
- Nous naturellement passons d'un espace à 3 dimensions à un espace à 2 dimensions.
- Selon le point de vue, l'information retenue ne sera pas la même.
- L'ACP nous propose un point de vue permettant de voir au mieux les individus d'un tableau.

5. Résumer les données

- Lorsqu'on projette les données sur un plan, on obtient un graphique déformé de la réalité.
- Le rôle de l'ACP est de trouver des espaces de dimensions plus petites minimisant ces déformations.
- On utilise un espace à 2 dimensions (un plan). Ce plan est appelé le plan principal. Il est constitué de deux droites perpendiculaires.
- La méthode consiste à calculer la première droite D1 de façon à **maximiser** les carrés des distances de projection des points sur la droite.



- Puis une 2ème droite D2 perpendiculaire à la première.

6. Les composantes principales

- Les droites D1 et D2 sont des caractères synthétiques obtenus par des combinaisons linéaires avec les variables d'origines.
- Ces droites sont appelées **composantes principales**, ou **axes principaux**.
- La première composante principale doit "**capturer**" le maximum d'inertie du tableau des données. La variance des individus doit être maximale.
- Il reste un résidu non expliqué par cette première composante. C'est sur ce résidu qu'est calculée la deuxième composante principale.

7. Caractères des composantes principales

- La première composante principale "**capture**" le maximum d'inertie du tableau des données.
- La deuxième composante principale est un complément, une correction de la première.
- La deuxième composante principale doit avoir une corrélation linéaire nulle avec la première (orthogonalité).
- Il n'y a pas de redondance d'information entre deux composantes principales.
- On calcule les autres composantes de la même manière.

8.L' ACP : combien de dimensions ?

- Un tableau de données à j dimensions donnera j composantes principales.
- Nous sommes donc passés d'un tableau de données à j dimensions (impossible à projeter sur un plan) à un tableau de j composantes principales.

On pourrait penser que nous voici bien avancé !

9. Un exemple d'utilisation de l'ACP (les données)

- Le jeu de données est fourni avec le logiciel WinStat (CIRAD)

Données techniques sur 62 véhicules - année modèle 1994

Variables quantitatives : Puiss_admi, Cylindree, Longueur, Largeur,
Surface, Poids_Tota, Vit_Maxi, Dep_arret,
Conso_Moye

NOMBRE D'INDIVIDUS SELECTIONNES : 62

Individus manquants : 10

Effectif pris en compte : 52

10. Un exemple d'utilisation (les valeurs propres)

ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES 06/02/2002 13:34:14

Données centrées réduites

Variables actives	:	9	supplémentaires	:	0
Individus actifs	:	52	supplémentaires	:	0
Individus manquants	:	10	Hors norme	:	0

	VALEUR PROPRE	%	% CUMULE	HISTOGRAMME
001	6.447	71.635	71.635	=====
002	1.140	12.663	84.298	=====
003	0.660	7.337	91.635	====
004	0.332	3.684	95.319	=
005	0.236	2.627	97.946	=
006	0.101	1.117	99.063	
007	0.044	0.483	99.547	
008	0.040	0.449	99.996	
009	0.000	0.004	100.000	
TOTAL	9.000			

11. Un exemple d'utilisation (inertie du plan principal)

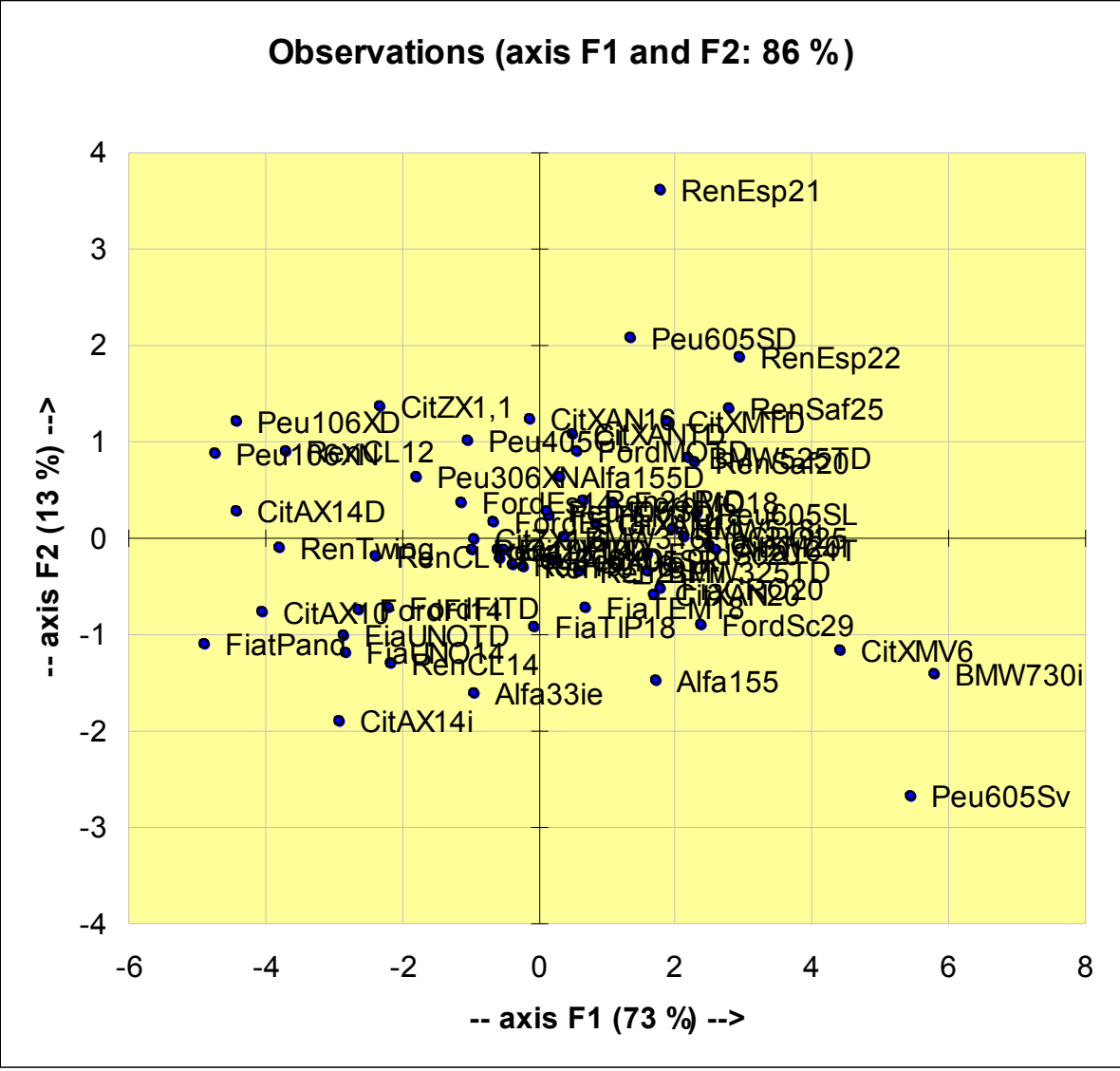
- Le plan principal représente 84% de l'inertie du tableau de données

Remarque : Il y a 9 variables centrées réduites (de variance=1) dans le tableau de données :

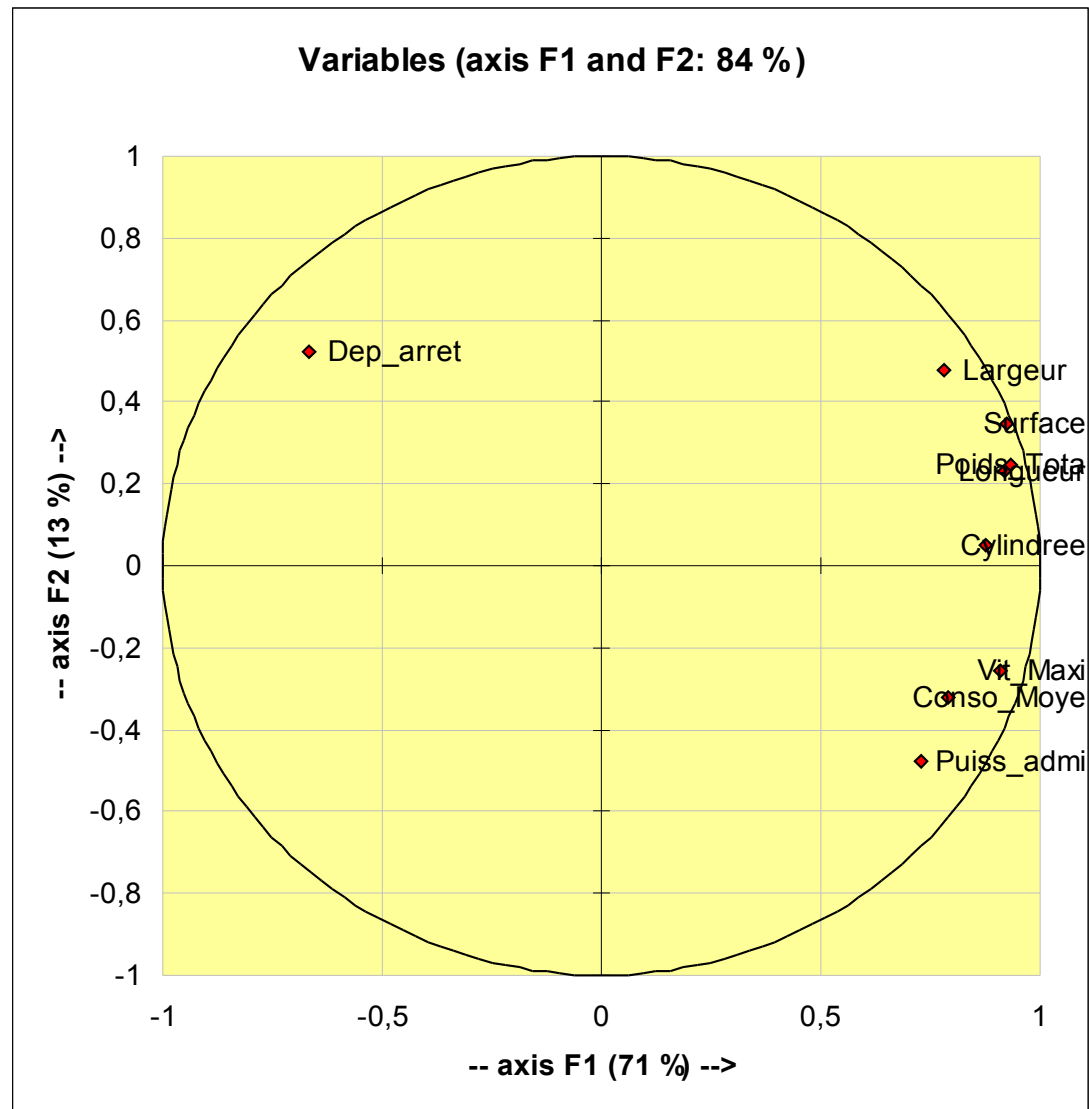


La somme des valeurs propre = 9

12. Un exemple d'utilisation (l'espace des individus)



13. Un exemple d'utilisation (l'espace des variables)



14. Un exemple d'utilisation (les contributions des variables)

COORD : COORDONNEES DES VARIABLES SUR LES AXES
 COS2 : COORD*COORD (COSINUS CARRES)
 CTR : PART (en %) DE LA VARIABLE DANS LA CONSTRUCTION DU FACTEUR
 QLT : QUALITE DE LA REPRESENTATION D'UNE VARIABLE SUR LES AXES SELECTIONNES

VARIABLES ACTIVES	QLT	FACTEUR 01			FACTEUR 02		
		COORD	COS2	CTR	COORD	COS2	CTR
Puiss_admi	74.1	0.730	53.22	8.26	0.457	20.86	18.30
Cylindree	77.5	0.880	77.49	12.02	-0.012	0.02	0.01
Longueur	89.2	0.913	83.27	12.92	-0.244	5.97	5.24
Largeur	85.4	0.774	59.89	9.29	-0.505	25.55	22.42
Surface	98.0	0.919	84.37	13.09	-0.369	13.63	11.96
Poids_Tota	92.2	0.932	86.79	13.46	-0.233	5.45	4.78
Vit_Maxi	91.4	0.903	81.52	12.64	0.315	9.91	8.70
Dep_arret	78.6	-0.717	51.42	7.98	-0.521	27.14	23.81
Conso_Moye	72.2	0.817	66.73	10.35	0.233	5.45	4.78
TOTAL				100.00			100.00

15. Un exemple d'utilisation (interpréter les contributions des variables)

- **COORD** est la corrélation entre les variables d'origine et les nouvelles variables synthétiques (axes principaux). On interprète ce coefficient comme n'importe quelle corrélation linéaire.
- **COS2** représente la répartition de la variables sur les différents facteurs. La somme horizontale sera égale à 100%

Exemple : la variable *cylindree* est représentée à 77.49% sur le premier facteur.

- **CTR** représente la contribution de chaque variable à la construction du facteur. La somme verticale est de 100%

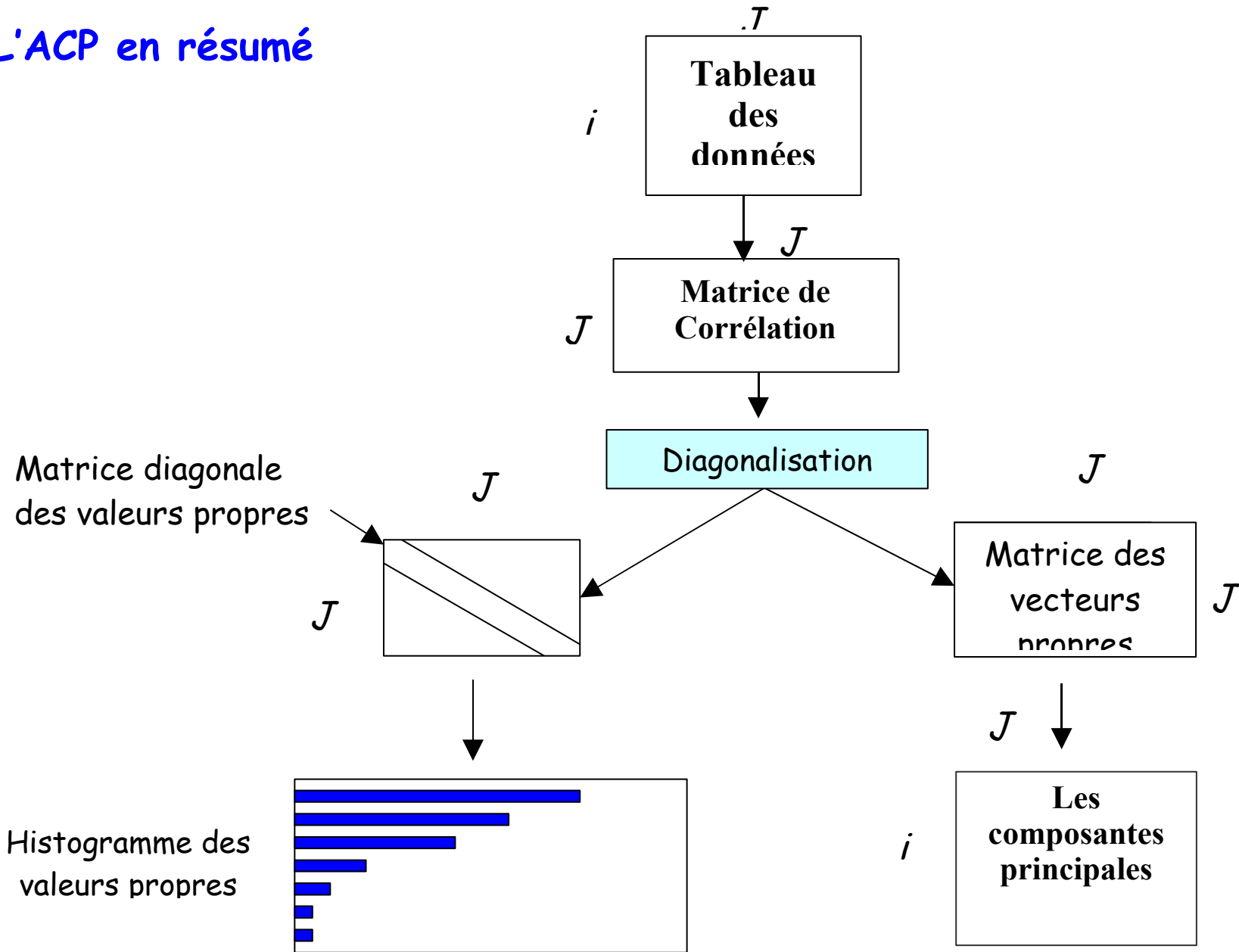
Exemple : La variable *cylindree* ne contribue pas à la construction de l'axe 2 (0.01%)

16. Un exemple d'utilisation (les contributions des individus)

COORD : COORDONNEES DES INDIVIDUS SUR LES AXES
 COS2 : COSINUS CARRE
 CTR : PART (en %) DE L'INDIVIDU DANS LA CONSTRUCTION DU FACTEUR
 QLT : QUALITE DE LA REPRESENTATION DE L'INDIVIDU SUR LES AXES AFFICHES
 INR : INERTIE RELATIVE DE L'INDIVIDU
 P : POIDS DE L'INDIVIDU

INDIVIDUS ACTIFS	FACTEUR 01						FACTEUR 02		
	Poids	INR	QLT	COORD	COS2	CTR	COORD	COS2	CTR
Alfa-Romé	1.00	0.99	70.00	-0.806	13.96	0.19	1.615	56.04	4.40
Alfa-Romé	1.00	1.13	93.12	1.759	58.67	0.92	1.348	34.45	3.07
Alfa-Romé	1.00	0.30	41.67	0.033	0.08	0.00	-0.765	41.60	0.99
Alfa-Romé	1.00	1.77	63.41	2.266	62.15	1.53	-0.322	1.26	0.18
BMW 316i	1.00	0.25	17.39	0.442	16.69	0.06	0.090	0.70	0.01
BMW 325 T	1.00	0.93	42.05	1.351	41.97	0.54	0.057	0.07	0.01
BMW 518i	1.00	1.23	78.92	2.130	78.92	1.35	-0.012	0.00	0.00
BMW 730i	1.00	9.22	97.15	6.312	92.33	11.88	1.443	4.83	3.51
Citroën A	1.00	5.48	71.83	-4.104	65.65	5.02	1.260	6.19	2.68
Citroën A	1.00	2.72	97.82	-2.870	64.79	2.46	2.049	33.04	7.08
.../...									
Renault 2	1.00	0.47	23.22	0.250	2.86	0.02	-0.668	20.35	0.75
Renault S	1.00	1.52	91.36	2.450	84.42	1.79	-0.702	6.94	0.83
Renault S	1.00	2.30	85.81	2.607	63.06	2.03	-1.566	22.75	4.14
Renault E	1.00	4.43	63.27	3.261	51.23	3.17	-1.581	12.04	4.22
Renault E	1.00	3.89	81.78	1.632	14.62	0.79	-3.499	67.16	20.65
TOTAL						100.00		100.00	

17.L'ACP en résumé



18. Transformer les données - Centrage et réduction

- L'importance que prendront les variables dans le calcul des composantes principales est fonction de leur ordre de grandeur.
- Une variable ayant un écart-type important aura plus de poids qu'une variable de faible écart-type.
- Des variables de fort écart-type "construiront" les premières composantes.
- Les calculs ne sont pas faux, mais la lecture des résultats d'une ACP peut devenir compliquée.
- C'est pour remédier à ça qu'il convient de centrer et réduire les variables.

19. Transformer les données - transformation en rang

Exemple :

Identif	Prix HT	Quantité	Prix HT (rang)	Quantité (rang)
001	41.5	27	4	3
002	28.6	42	3	4
003	19.3	51	1	5
004	52.9	12	5	1
005	28.2	14	2	2

- La transformation en rang peut permettre de détecter des relations non linéaires et rapproche les valeurs extrêmes.
- En cas de transformation en rang des données, il n'est pas utile de les réduire.

20. Les vecteurs propres (1)

- Coefficient des variables centrées réduites dans l'équation linéaire des axes

	FACTEUR 1	FACTEUR 2	FACTEUR 3
Puiss_admi	0.287	0.428	-0.482
Cylindree	0.347	-0.012	0.371
Longueur	0.359	-0.229	0.101
Largeur	0.305	-0.473	-0.284
Surface	0.362	-0.346	-0.065
Poids_Tota	0.367	-0.219	0.203
Vit_Maxi	0.356	0.295	0.259
Dep_arret	-0.282	-0.488	-0.341
Conso_Moye	0.322	0.219	-0.557

- Les vecteurs propres sont les coefficients à affecter aux variables initiales pour obtenir les composantes principales.
- Par exemple la première composante s'obtient (*pour chaque individu*):
 $0.287 * \text{Puiss_admin} + 0.347 * \text{Cylindree} + \dots + 0.322 * \text{Conso_Moye}$

21. Les vecteurs propres (2)

- L'utilisation et l'étude des vecteurs propres n'est pas d'un grand intérêt pratique
- A moins que vous ne vouliez calculer les composantes principales à la main...

(armez vous de patience...)

22. Les variables supplémentaires

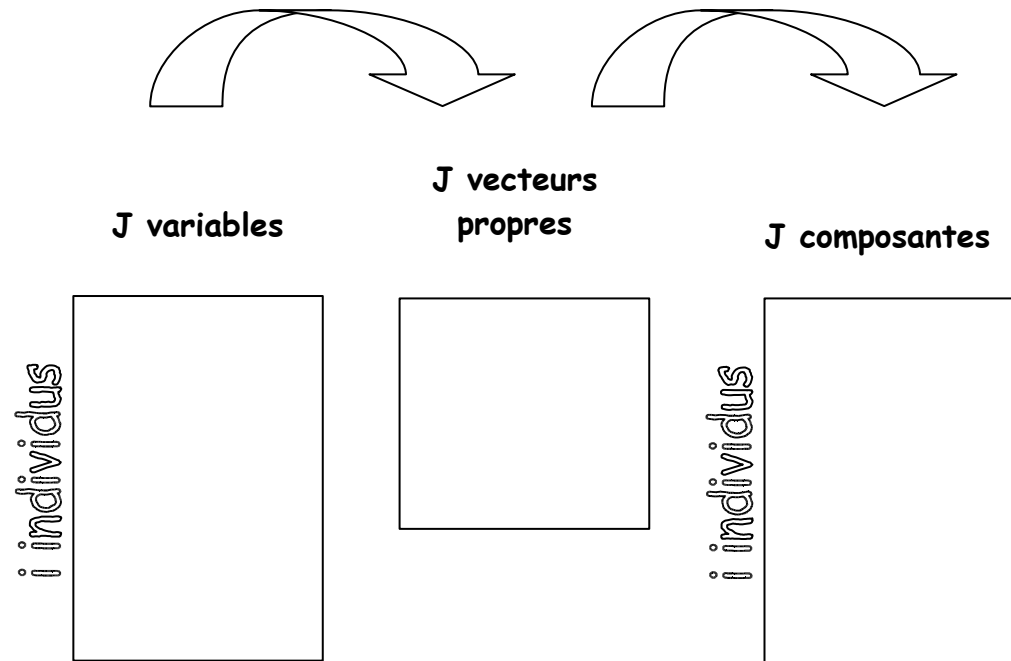
- Nous pouvons considérer que certaines variables sont des variables explicatives et d'autres des variables à expliquer. Nous mettrons en supplémentaire les variables à expliquer
- Par exemple, le rendement d'une culture dépend de la fertilisation, du climat, etc. Dans ce cas, le calcul des composantes principales se fera sans la variable rendement.
- Celle-ci sera introduite à la fin de l'analyse afin de la positionner sur le plan principal.
- D'autres variables peuvent manquer de fiabilité. On peut légitimement hésiter à les introduire dans l'analyse. Elles peuvent être utilisées comme variables supplémentaires

23. Les individus supplémentaires

- De nouveaux individus pourront être mis en supplémentaires dans l'analyse (nouvelles variétés, nouveaux traitement, etc...)
- On peut aussi mettre en supplémentaires des données (variables ou individus) dont on doute de la fiabilité. Ces données seront positionnées sur le plan principal sans participer activement aux calculs.
- Par opposition, les individus ou les variables qui ne sont pas supplémentaires sont dits **actifs** (sous entendu: actifs dans le calcul des composantes principales)

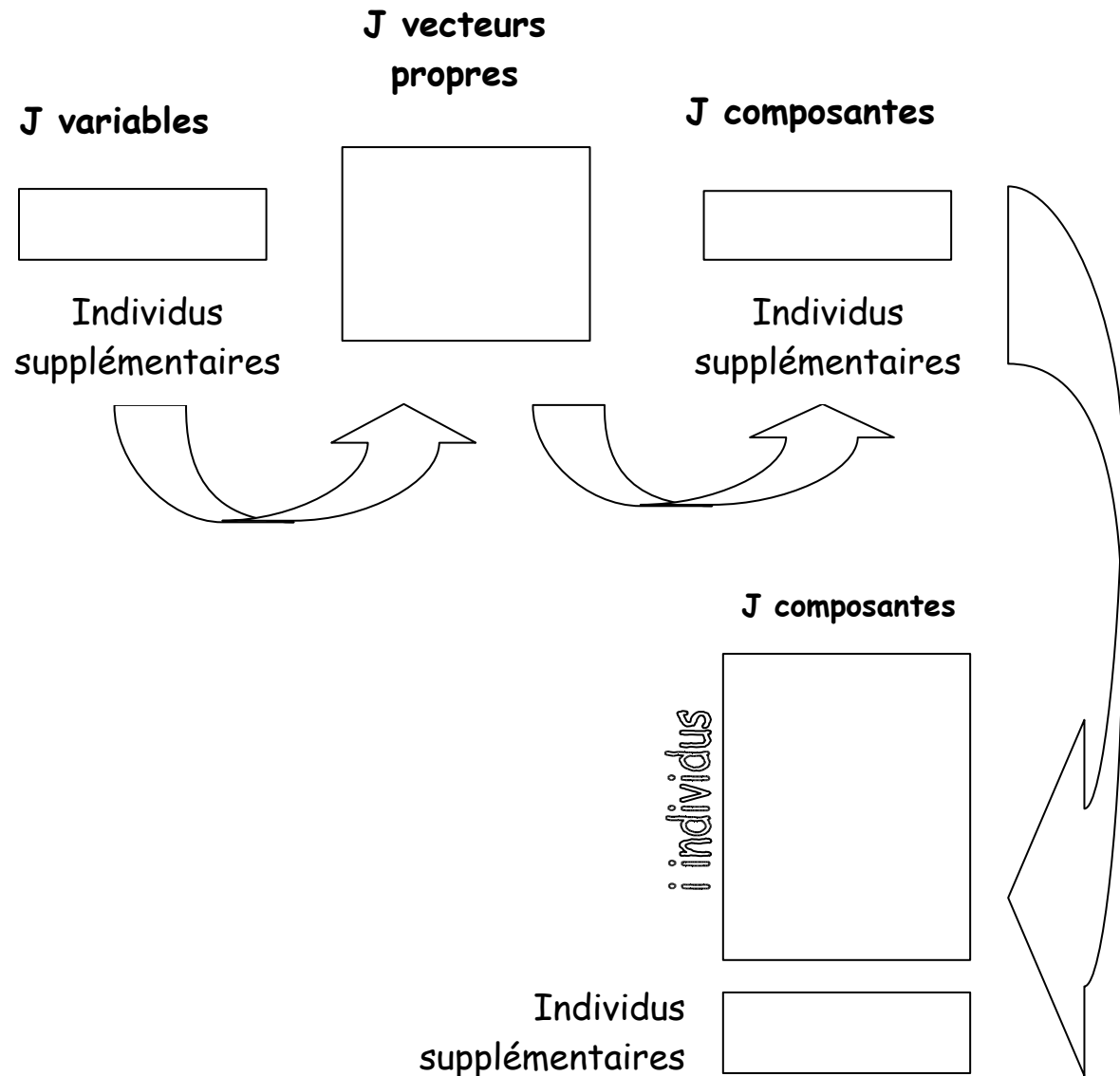
24. Calcul des individus supplémentaires (1)

- Les composantes principales sont de nouvelles variables synthétiques. Elles sont calculées grâce aux individus actifs.



- Les individus supplémentaires sont introduits en fin d'analyse, après le calcul des vecteurs propres.

25. Calcul des individus supplémentaires (2)



On utilise la matrice des vecteurs propres (déjà connue) pour calculer la valeur des composantes principales pour les individus supplémentaires

26. Calcul des variables supplémentaires (1)

- Les composantes principales sont calculées à partir de toutes les variables actives.
- L'espace des variables représente les corrélations entre les variables du tableau de données et les composantes principales.

Tableau des corrélations

	Comp 1	Comp 2
Puiss_admi	0.730	0.457
Cylindree	0.880	-0.012
Longueur	0.913	-0.244
Largeur	0.774	-0.505
Surface	0.919	-0.369
Poids_Tota	0.932	-0.233
Vit_Maxi	0.903	0.315
Dep_arret	-0.717	-0.521
Conso_Moye	0.817	0.233

27. Calcul des variables supplémentaires (2)

- Les variables supplémentaires ne participent pas à l'élaboration des vecteurs propres (donc des composantes principales).

VARIABLES ACTIVES + SUPPLEMENTAIRES

	Comp 1	Comp 2
Puiss_admi	0.730	0.457
Cylindree	0.880	-0.012
Longueur	0.913	-0.244
Largeur	0.774	-0.505
Surface	0.919	-0.369
Poids_Tota	0.932	-0.233
Vit_Maxi	0.903	0.315
Dep_arret	-0.717	-0.521
Conso_Moye	0.817	0.233
<i>Assurance</i>	<i>0.852</i>	<i>-0.325</i>

28.L' ACP : Rotation des axes

Pour faciliter l'interprétation de la nature des facteurs, on peut faire subir des rotations aux axes obtenus par l'analyse factorielle. Il existe plusieurs procédés pour effectuer ces rotations.

La méthode **quartimax**, consiste à maximiser la variance des carrés; cette méthode exige la maximisation de la somme des saturations à la quatrième puissance.

Une autre méthode repose sur la maximisation de la somme des variances des carrés des saturations dans chaque colonne. Cette méthode dite **varimax** est la plus largement employée.

Il existe d'autres rotations qui rendent les axes obliques et par conséquent les facteurs deviennent corrélés : les méthodes **oblimin**, **promax**, etc.

Ces méthodes de rotation ne seront pas abordées dans cette formation

29.L' ACP : exercices pratiques

- Le but de ces exercices est de vous familiariser avec un logiciel statistique.
- Mettez en œuvre une ACP
- Les données sont présentées dans le fascicule "Les données et exercices"
- Nous interpréterons ensemble les sorties logiciel.

