



STATISTIQUE

Analyse en composantes principales

Exemple 1

Les joueurs de tennis

- 34 joueurs de tennis
- 15 caractéristiques des joueurs :
Coup droit, Reverse, Service, Volée, Retour, Smash, Jambe, Lob, Amorti, Passing, Régularité, Touche, Psychisme, Physique, Double
- Nombre des titres remportés

G. Gaudio

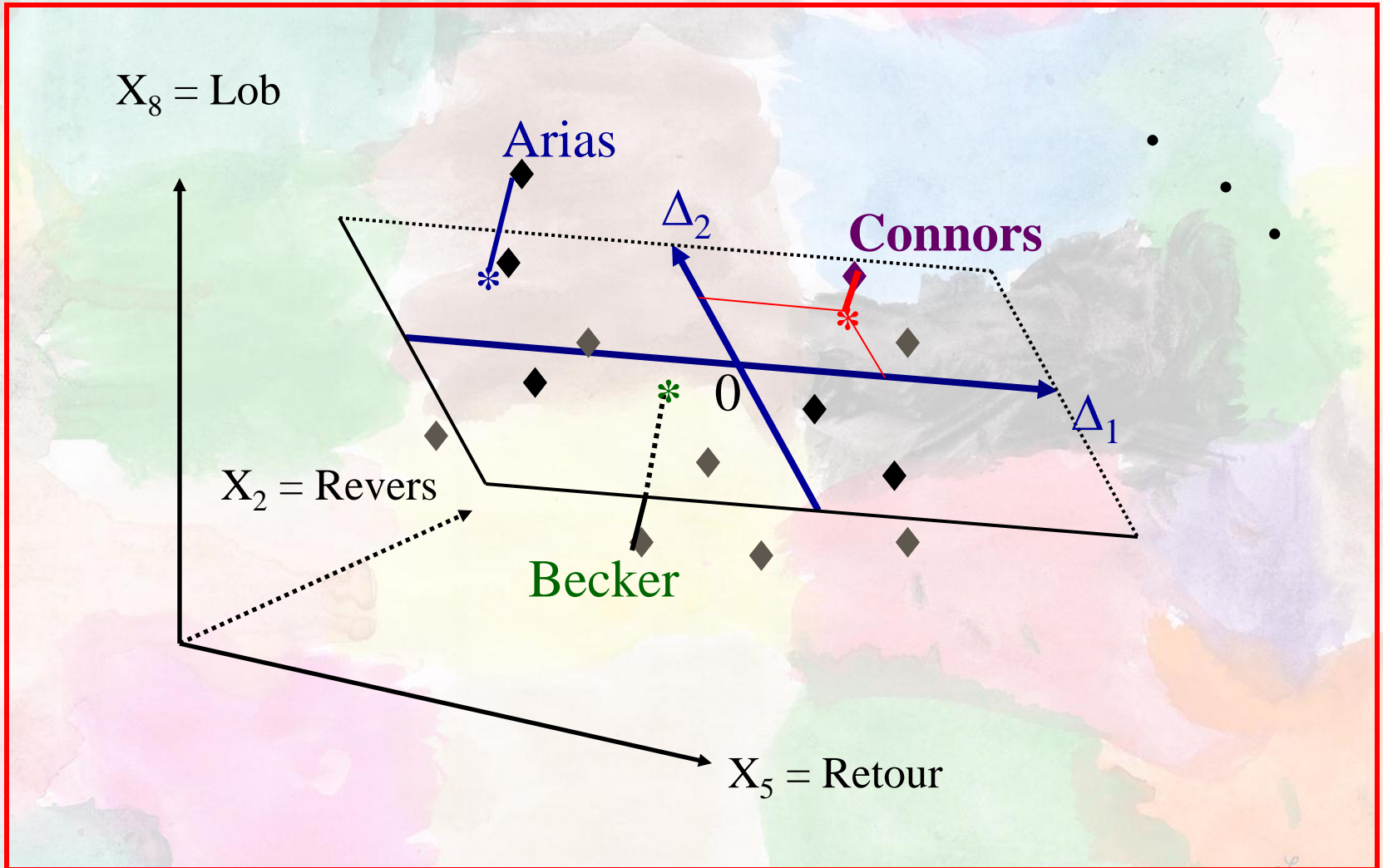
Vainqueur du Roland Garros 2004



Extrait des données

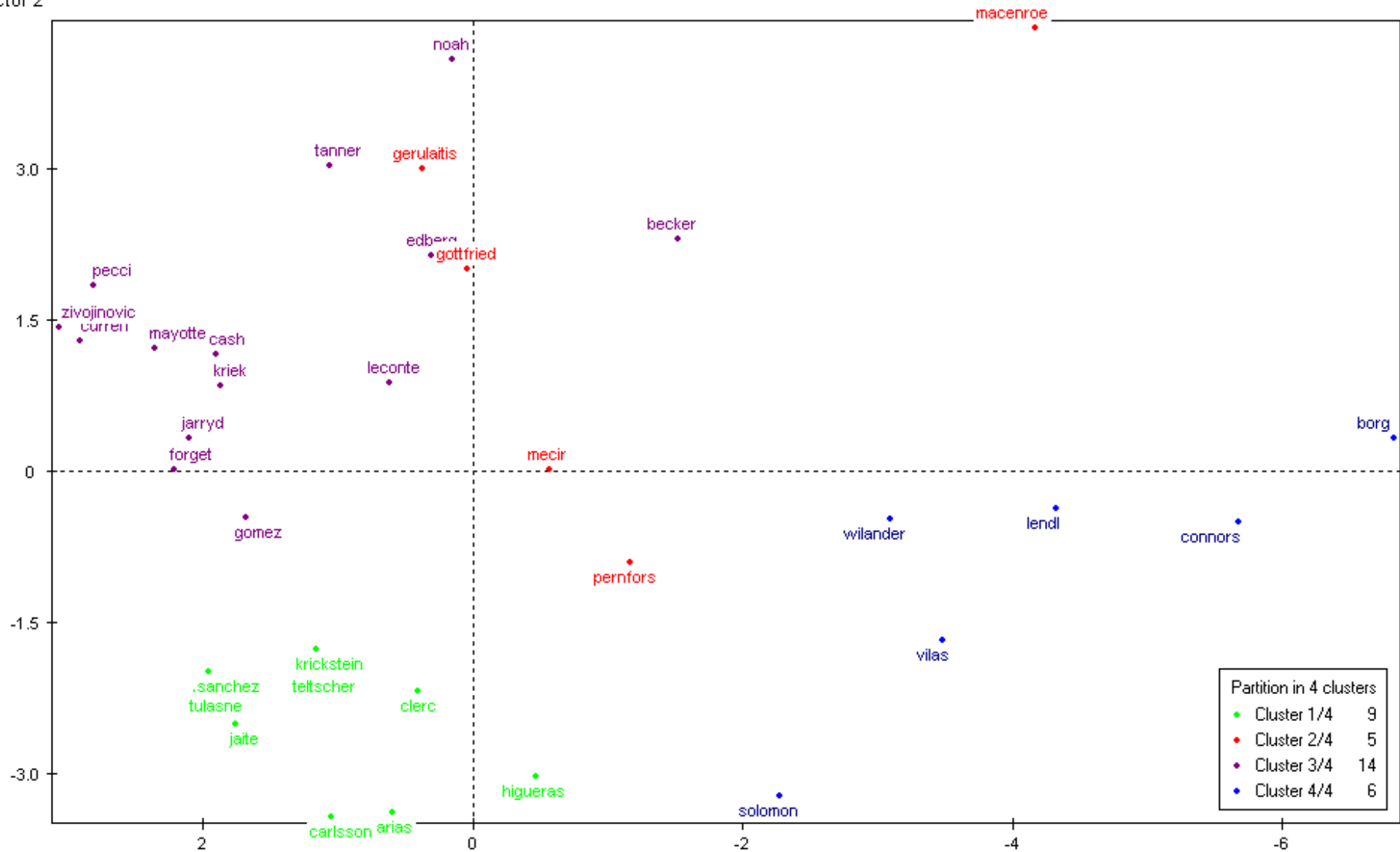
Joueur	Titres remportés	X1 = Coup Droit	X2 = Revers	X3 = Service	X4 = Volée	X5 = Retour	X6 = Smash	X7 = Jambe	X8 = Lob	X9 = Amorti	X10 = Passing	X11 = Régularité	X12 = Touche	X13 = Psychisme	X14 = Physique	X15 = Double
ARIAS	5	8	5	3	2	6	3	6	5	2	6	6	3	4	3	0
BECKER	64	8	7	10	8	7	9	6	5	3	6	5	5	7	8	6
BORG	56	10	9	7	5	9	9	10	7	4	10	10	5	10	10	2
CARLSSON	9	6	5	2	2	6	2	6	4	2	6	6	2	4	5	0
CASH	19	5	4	6	7	5	7	6	3	3	4	4	3	7	6	7
CLERC	26	7	6	4	3	5	3	6	6	3	6	6	4	3	5	2
CONNORS	86	7	10	4	6	10	7	9	9	6	9	8	6	9	6	0
CURREN	31	4	4	8	7	6	6	5	3	4	4	4	4	2	5	7
EDBERG	60	5	8	8	8	6	6	6	3	4	5	5	6	5	7	8

Analyse factorielle des joueurs



Analyse Factorielle des joueurs

Factor 2

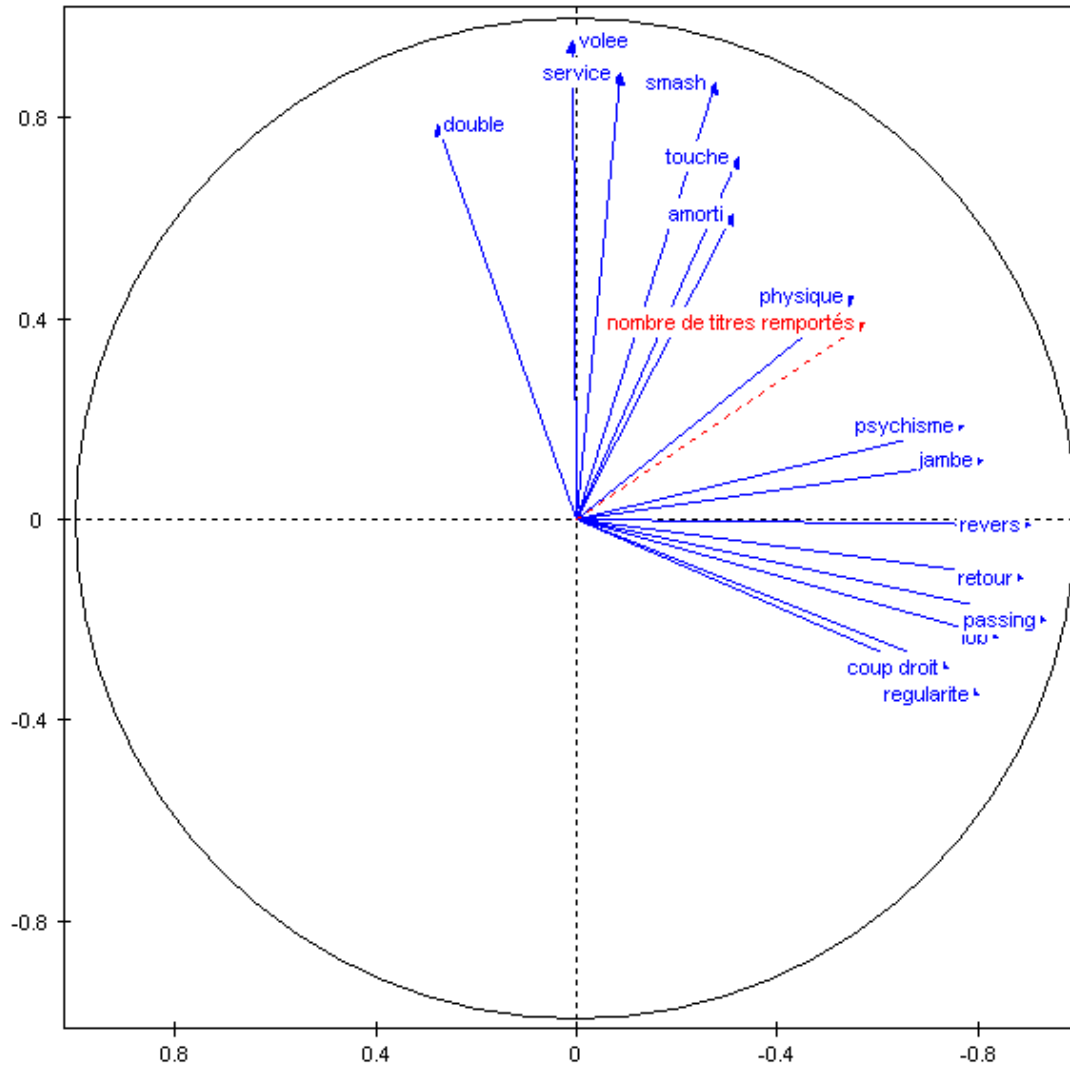


Factor 1

Analyse Factorielle des joueurs

Carte des caractéristiques utilisées pour l'analyse

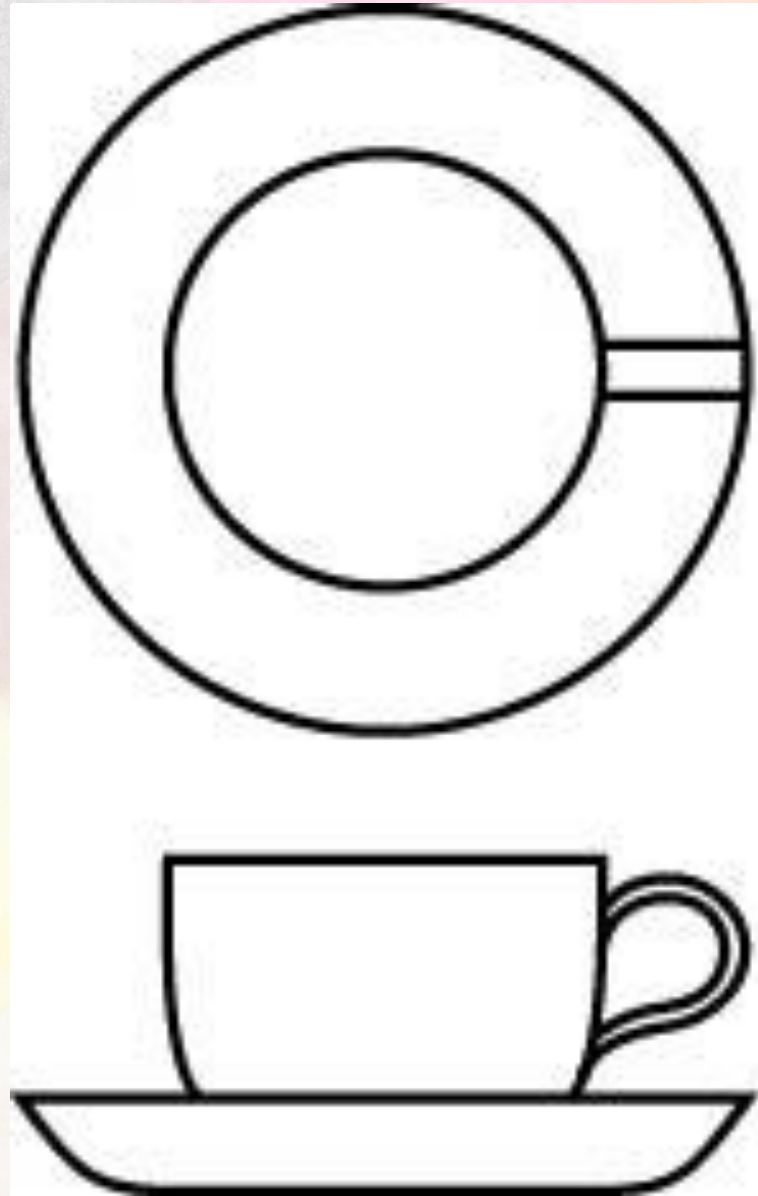
Factor 2



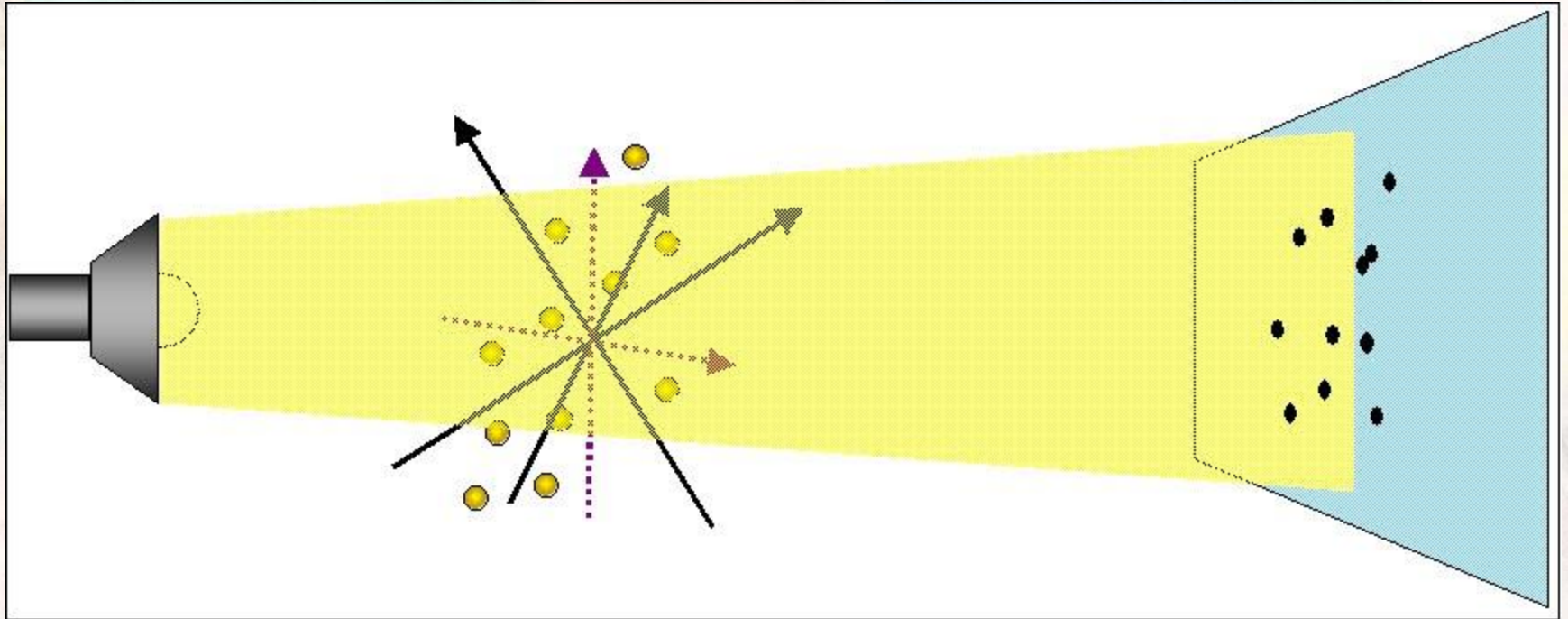
Factor 1

Les variables fléchées en pointillés sont illustratives.

Qu'est ce que
c'est ???



Construction du premier plan principal (Source : S. Wold, 2004)



2. Les objectifs de l'analyse factorielle (option composantes principales)

Décrire un tableau individus×variables :

- Visualiser le positionnement des individus les uns par rapport aux autres
- Visualiser les corrélations entre les variables
- Interpréter les axes factoriels

Visualisation des données

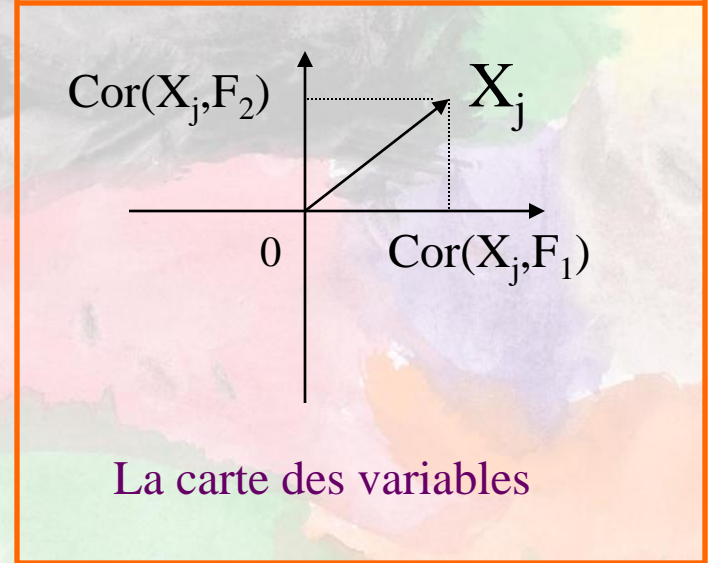
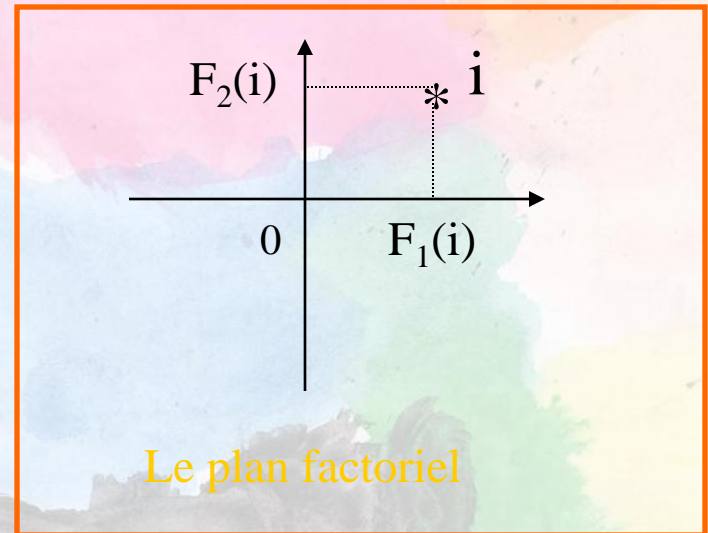
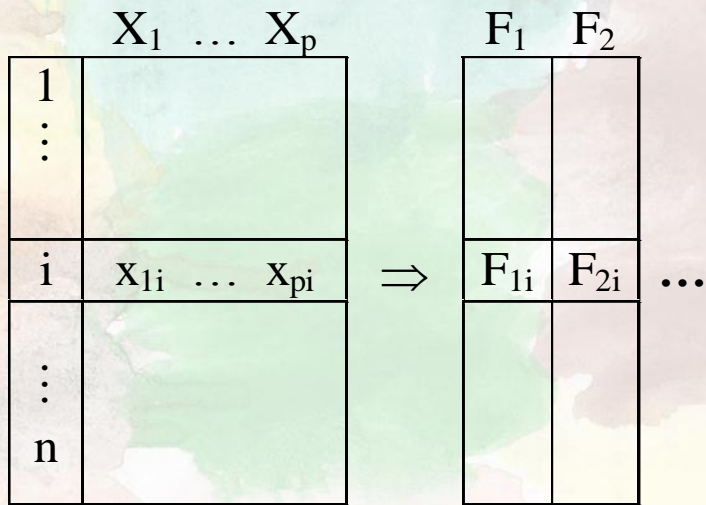


Tableau des données

Facteurs centrés-réduits résumant les données

$$F_h = \sum_{j=1}^p u_{hj} X_j$$

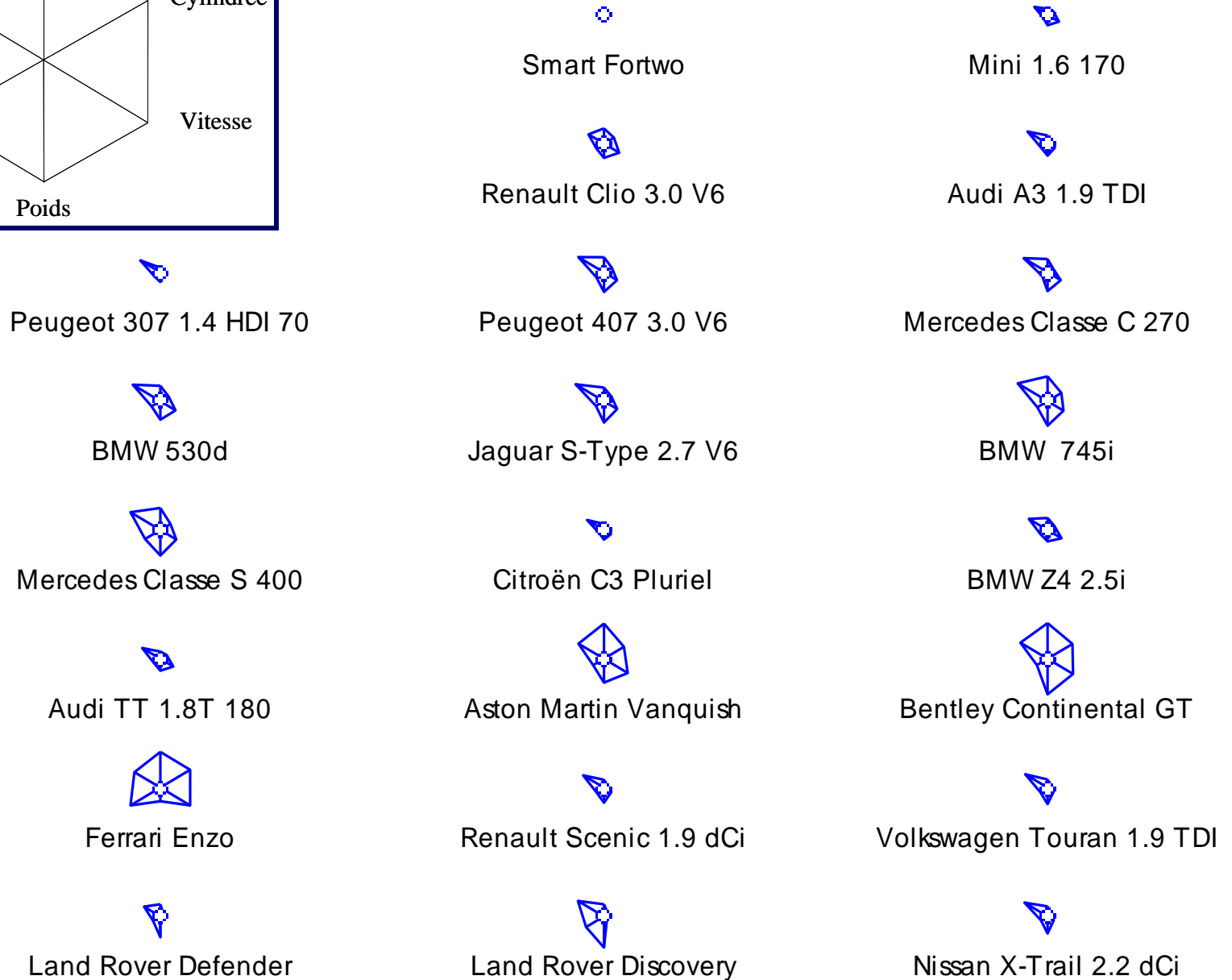
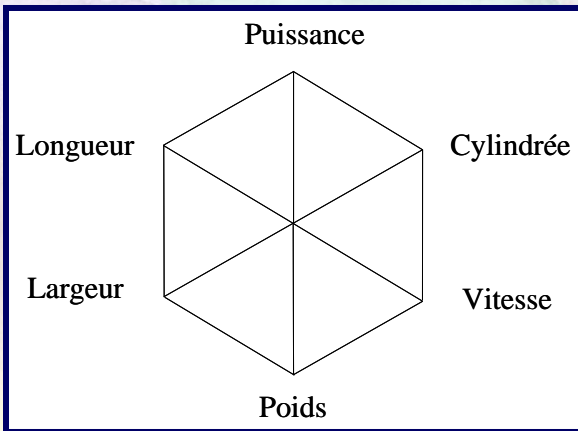
(non corrélés entre eux)

3. Un exemple de positionnement de produits

Caractéristiques de 24 modèles de voiture (Source : L'argus de l'automobile, 2004)

Modèle	Cylindrée (cm ³)	Puissance (ch)	Vitesse (km/h)	Poids (kg)	Largeur (mm)	Longueur (mm)
Citroën C2 1.1 Base	1124	61	158	932	1659	3666
Smart Fortwo Coupé	698	52	135	730	1515	2500
Mini 1.6 170	1598	170	218	1215	1690	3625
Nissan Micra 1.2 65	1240	65	154	965	1660	3715
Renault Clio 3.0 V6	2946	255	245	1400	1810	3812
Audi A3 1.9 TDI	1896	105	187	1295	1765	4203
Peugeot 307 1.4 HDI 70	1398	70	160	1179	1746	4202
Peugeot 407 3.0 V6 BVA	2946	211	229	1640	1811	4676
Mercedes Classe C 270 CDI	2685	170	230	1600	1728	4528
BMW 530d	2993	218	245	1595	1846	4841
Jaguar S-Type 2.7 V6 Bi-Turbo	2720	207	230	1722	1818	4905
BMW 745i	4398	333	250	1870	1902	5029
Mercedes Classe S 400 CDI	3966	260	250	1915	2092	5038
Citroën C3 Pluriel 1.6i	1587	110	185	1177	1700	3934
BMW Z4 2.5i	2494	192	235	1260	1781	4091
Audi TT 1.8T 180	1781	180	228	1280	1764	4041
Aston Martin Vanquish	5935	460	306	1835	1923	4665
Bentley Continental GT	5998	560	318	2385	1918	4804
Ferrari Enzo	5998	660	350	1365	2650	4700
Renault Scenic 1.9 dCi 120	1870	120	188	1430	1805	4259
Volkswagen Touran 1.9 TDI 105	1896	105	180	1498	1794	4391
Land Rover Defender Td5	2495	122	135	1695	1790	3883
Land Rover Discovery Td5	2495	138	157	2175	2190	4705
Nissan X-Trail 2.2 dCi	2184	136	180	1520	1765	4455

Graphiques en étoile des voitures



4. Résumé des données

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Cylindrée	24	698	5998	2722.54	1516.445
Puissance	24	52	660	206.67	155.721
Vitesse	24	135	350	214.71	56.572
Poids	24	730	2385	1486.58	387.507
Largeur	24	1515	2650	1838.42	220.842
Longueur	24	2500	5038	4277.83	581.497

Formule utilisée pour l'écart-type :

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Tableau des corrélations

	Cylindrée	Puissance	Vitesse	Poids	Largeur	Longueur
Cylindrée	1.000	0.954	0.885	0.692	0.706	0.664
Puissance	0.954	1.000	0.934	0.529	0.730	0.527
Vitesse	0.885	0.934	1.000	0.466	0.619	0.578
Poids	0.692	0.529	0.466	1.000	0.477	0.795
Largeur	0.706	0.730	0.619	0.477	1.000	0.591
Longueur	0.664	0.527	0.578	0.795	0.591	1.000

5. Réduction des données

Pour neutraliser le problème des unités on remplace les données d'origine par les données centrées-réduites :

$$\begin{aligned} X_1^* &= \frac{X_1 - \bar{X}_1}{S_1} \\ &\vdots \\ X_p^* &= \frac{X_p - \bar{X}_p}{S_p} \end{aligned}$$

de moyenne 0 et d'écart-type 1.

Les données centrées-réduites

Case Summaries

	MODÈLE	Zscore: Cylindrée	Zscore: Puissance	Zscore: Vitesse	Zscore: Poids	Zscore: Largeur	Zscore: Longueur
1	Citroën C2 1.1 Base	-1.054	-.935	-1.002	-1.431	-.812	-1.052
2	Smart Fortwo Coupé	-1.335	-.993	-1.409	-1.952	-1.464	-3.057
3	Mini 1.6 170	-.742	-.235	.058	-.701	-.672	-1.123
4	Nissan Micra 1.2 65	-.978	-.910	-1.073	-1.346	-.808	-.968
5	Renault Clio 3.0 V6	.147	.310	.535	-.223	-.129	-.801
6	Audi A3 1.9 TDI	-.545	-.653	-.490	-.494	-.332	-.129
7	Peugeot 307 1.4 HDI 70	-.873	-.878	-.967	-.794	-.418	-.130
8	Peugeot 407 3.0 V6 BVA	.147	.028	.253	.396	-.124	.685
9	Mercedes Classe C 270 CDI	-.025	-.235	.270	.293	-.500	.430
10	BMW 530d	.178	.073	.535	.280	.034	.968
11	Jaguar S-Type 2.7 V6 Bi-Turbo	-.002	.002	.270	.608	-.092	1.079
12	BMW 745i	1.105	.811	.624	.989	.288	1.292
13	Mercedes Classe S 400 CDI	.820	.342	.624	1.106	1.148	1.307
14	Citroën C3 Pluriel 1.6i	-.749	-.621	-.525	-.799	-.627	-.591
15	BMW Z4 2.5i	-.151	-.094	.359	-.585	-.260	-.321
16	Audi TT 1.8T 180	-.621	-.171	.235	-.533	-.337	-.407
17	Aston Martin Vanquish	2.118	1.627	1.614	.899	.383	.666
18	Bentley Continental GT	2.160	2.269	1.826	2.318	.360	.905
19	Ferrari Enzo	2.160	2.911	2.391	-.314	3.675	.726
20	Renault Scenic 1.9 dCi 120	-.562	-.557	-.472	-.146	-.151	-.032
21	Volkswagen Touran 1.9 TDI 105	-.545	-.653	-.614	.029	-.201	.195
22	Land Rover Defender Td5	-.150	-.544	-1.409	.538	-.219	-.679
23	Land Rover Discovery Td5	-.150	-.441	-1.020	1.777	1.592	.735
24	Nissan X-Trail 2.2 dCi	-.355	-.454	-.614	.086	-.332	.305
Total	Mean	.000	.000	.000	.000	.000	.000
	Std. Deviation	1.000	1.000	1.000	1.000	1.000	1.000

6. Recherche du premier facteur

On recherche le facteur centré-réduit (moyenne = 0, écart-type = 1)

$$F_1 = \sum_{j=1}^p u_{1j} X_j^*$$

maximisant le critère « Part de la variance totale expliquée par F_1 »

$$\sum_{j=1}^p \text{cor}^2(X_j, F_1)$$

Le facteur F_1 résume aussi bien que possible le tableau de données X .

Résultats

- Le vecteur u_1 est vecteur propre (*eigenvector*) de la matrice des corrélations R associé à la plus grande valeur propre (*eigenvalue*) λ_1 .

- Le critère

$$\sum_{j=1}^p \text{cor}^2(X_j, F_1)$$

est égal à λ_1 .

Résultat SPSS : Valeurs propres

Total Variance Explained

Component	Eigenv alues		
1	4.411	73.521	73.521
2	.853	14.223	87.745
3	.436	7.261	95.006
4	.236	3.931	98.937
5	.051	.857	99.794
6	.012	.206	100.000

Extraction Method: Principal Component Analysis.

Somme des valeurs propres = Nombre de X = p

Résultat SPSS : Les vecteurs propres u_h

Component Score Coefficient Matrix

	Component					
	1	2	3	4	5	6
Cylindrée	.218	-.149	-.325	-.478	-2.877	-4.459
Puissance	.209	-.413	-.207	-.356	-.416	6.990
Vitesse	.201	-.397	-.474	.844	2.507	-2.823
Poids	.172	.675	-.338	-1.090	1.716	-.068
Largeur	.182	-.130	1.338	-.288	.675	-1.187
Longueur	.180	.591	.136	1.379	-1.142	1.685

Extraction Method: Principal Component Analysis.
Component Scores.

$$F_1 = \sum_{j=1}^p u_{1j} X_j^*$$

$$F_1 = .218 \text{ Cyl}^* + .209 \text{ Puis}^* + .201 \text{ Vit}^* + .172 \text{ Poids}^* + .182 \text{ Larg}^* + .180 \text{ Long}^*$$

Quelles seront les voitures à F_1 négatif ? À F_1 positif ?

Résultats SPSS : Les facteurs

	MODÈLE	Facteur 1
1	Citroën C2 1.1 Base	-1.210
2	Smart Fortwo Coupé	-1.934
3	Mini 1.6 170	-.644
4	Nissan Micra 1.2 65	-1.171
5	Renault Clio 3.0 V6	-.001
6	Audi A3 1.9 TDI	-.522
7	Peugeot 307 1.4 HDI 70	-.804
8	Peugeot 407 3.0 V6 BVA	.258
9	Mercedes Classe C 270 CDI	.037
10	BMW 530d	.391
11	Jaguar S-Type 2.7 V6 Bi-Turbo	.336
12	BMW 745i	.991
13	Mercedes Classe S 400 CDI	1.010
14	Citroën C3 Pluriel 1.6i	-.756
15	BMW Z4 2.5i	-.186
16	Audi TT 1.8T 180	-.350
17	Aston Martin Vanquish	1.471
18	Bentley Continental GT	1.939
19	Ferrari Enzo	2.306
20	Renault Scenic 1.9 dCi 120	-.392
21	Volkswagen Touran 1.9 TDI 105	-.375
22	Land Rover Defender Td5	-.500
23	Land Rover Discovery Td5	.396
24	Nissan X-Trail 2.2 dCi	-.286
Total	Mean	.000
	Std. Deviation	1.000

Corrélations entre les variables et les facteurs

	Component					
	1	2	3	4	5	6
Cylindrée	.962	-.127	-.142	-.113	-.148	-.055
Puissance	.923	-.353	-.090	-.084	-.021	.086
Vitesse	.886	-.339	-.206	.199	.129	-.035
Poids	.757	.576	-.147	-.257	.088	-.001
Largeur	.801	-.111	.583	-.068	.035	-.015
Longueur	.795	.504	.059	.325	-.059	.021

Extraction Method: Principal Component Analysis.

Propriétés du premier facteur F_1

- $F_1 = u_{11}X_1^* + u_{12}X_2^* + \dots + u_{1p}X_p^*$

- Moyenne de $F_1 = 0$

- Variance de $F_1 = 1$

- $\text{Cor}(X_j, F_1) = \lambda_1 u_{1j}$

- $\sum_{j=1}^p \text{cor}^2(X_j, F_1) = \lambda_1$ est maximum

Mesure de la qualité du premier facteur F_1

- La variance totale du tableau des données centrées-réduites est définie par :

$$\text{Variance totale} = \sum_{j=1}^p \text{Var}(X_j^*) = p$$

- La part de la variance de X_j^* expliquée par F_1 est égale à $\text{Cor}^2(X_j, F_1)$.
- La part de la variance totale expliquée par F_1 est égale à :

$$\sum_{j=1}^p \text{Cor}^2(X_j^*, F_1) = \lambda_1$$

Qualité du premier facteur

- Variance totale = $p = 6$
- Variance expliquée par le premier facteur

$$\lambda_1 = 4.411$$

- Proportion de variance expliquée par le premier facteur :

$$\frac{\text{Variance expliquée}}{\text{Variance totale}} = \frac{\lambda_1}{p} = \frac{4.411}{6} = 0.73521$$

- Le premier facteur explique 73,521% de la variance totale.

7. Deuxième facteur F_2

- On recherche le deuxième facteur centré-réduit

$$F_2 = \sum_{j=1}^p u_{2j} X_j^*$$

non corrélé à F_1 et résumant au mieux le tableau X.

- Le facteur F_2 maximise

$$\sum_{j=1}^p \text{cor}^2(X_j, F_2)$$

sous la contrainte $\text{cor}(F_1, F_2) = 0$.

Résultats

- Le vecteur u_2 est vecteur propre de la matrice des corrélations R associé à la deuxième plus grande valeur propre λ_2 .
- $F_2 = u_{21}X_1^* + u_{22}X_2^* + \dots + u_{2p}X_p^*$
- F_2 est centré-réduit
- $\text{Cor}(X_j, F_2) = \lambda_2 u_{2j}$
- $\sum_{j=1}^p \text{cor}^2(X_j, F_2) = \lambda_2$ est maximum
sous la contrainte $\text{cor}(F_1, F_2) = 0$.

Résultat SPSS : Valeurs propres

Total Variance Explained

Component	Eigenvalues		
1	4.411	73.521	73.521
2	.853	14.223	87.745
3	.436	7.261	95.006
4	.236	3.931	98.937
5	.051	.857	99.794
6	.012	.206	100.000

Extraction Method: Principal Component Analysis.

Somme des valeurs propres = Nombre de X = p

Résultat SPSS : Les vecteurs propres u_h

Component Score Coefficient Matrix

	Component					
	1	2	3	4	5	6
Cylindrée	.218	-.149	-.325	-.478	-2.877	-4.459
Puissance	.209	-.413	-.207	-.356	-.416	6.990
Vitesse	.201	-.397	-.474	.844	2.507	-2.823
Poids	.172	.675	-.338	-1.090	1.716	-.068
Largeur	.182	-.130	1.338	-.288	.675	-1.187
Longueur	.180	.591	.136	1.379	-1.142	1.685

Extraction Method: Principal Component Analysis.
Component Scores.

$$F_2 = \sum_{j=1}^p u_{2j} X_j^*$$

$$F_2 = -.149 \text{ Cyl}^* - .413 \text{ Puis}^* - .397 \text{ Vit}^* + .675 \text{ Poids}^* - .130 \text{ Larg}^* + .591 \text{ Long}^*$$

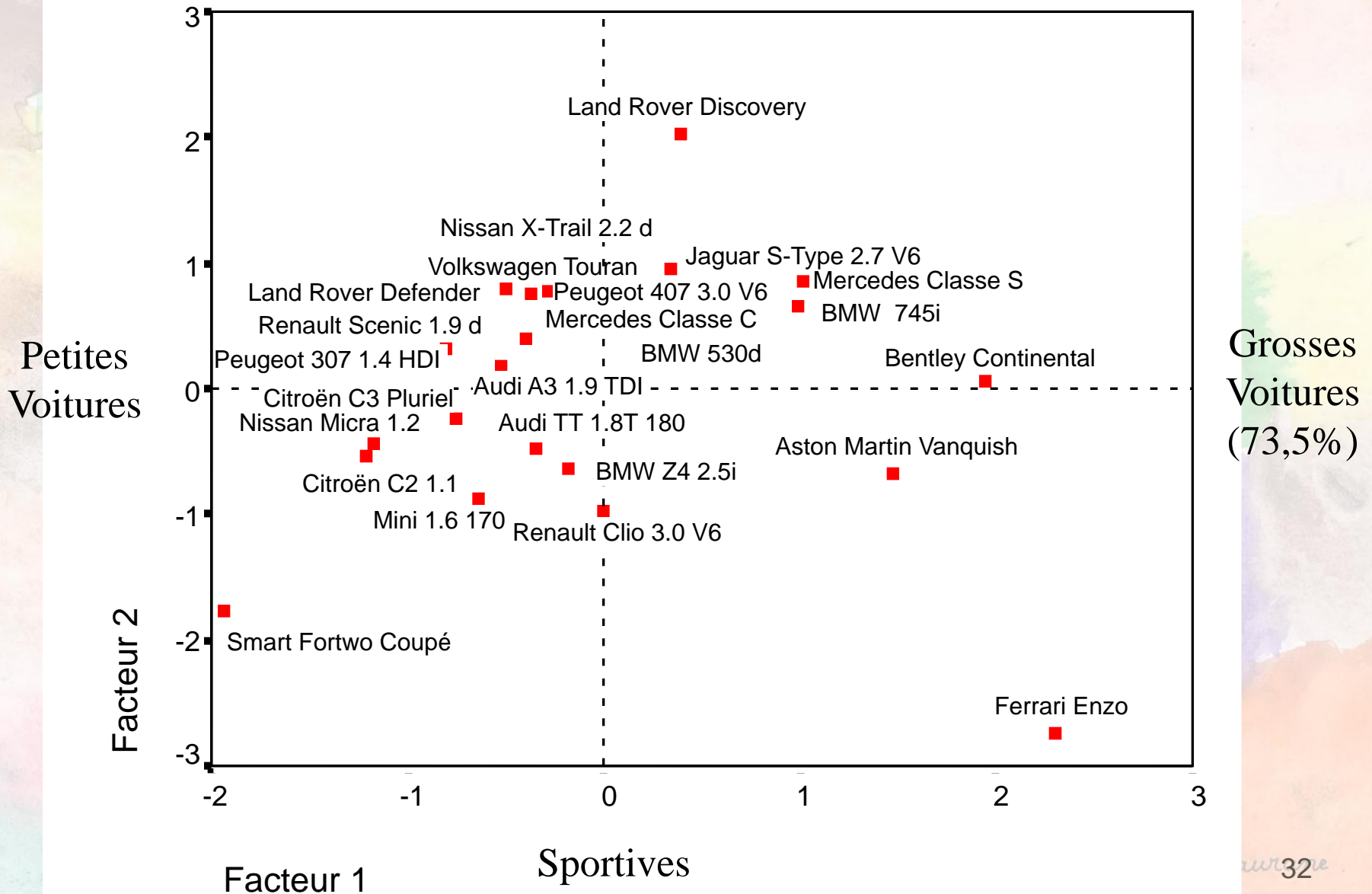
Quelles seront les voitures à F2 négatif ? À F2 positif ?

Le deuxième facteur F_2

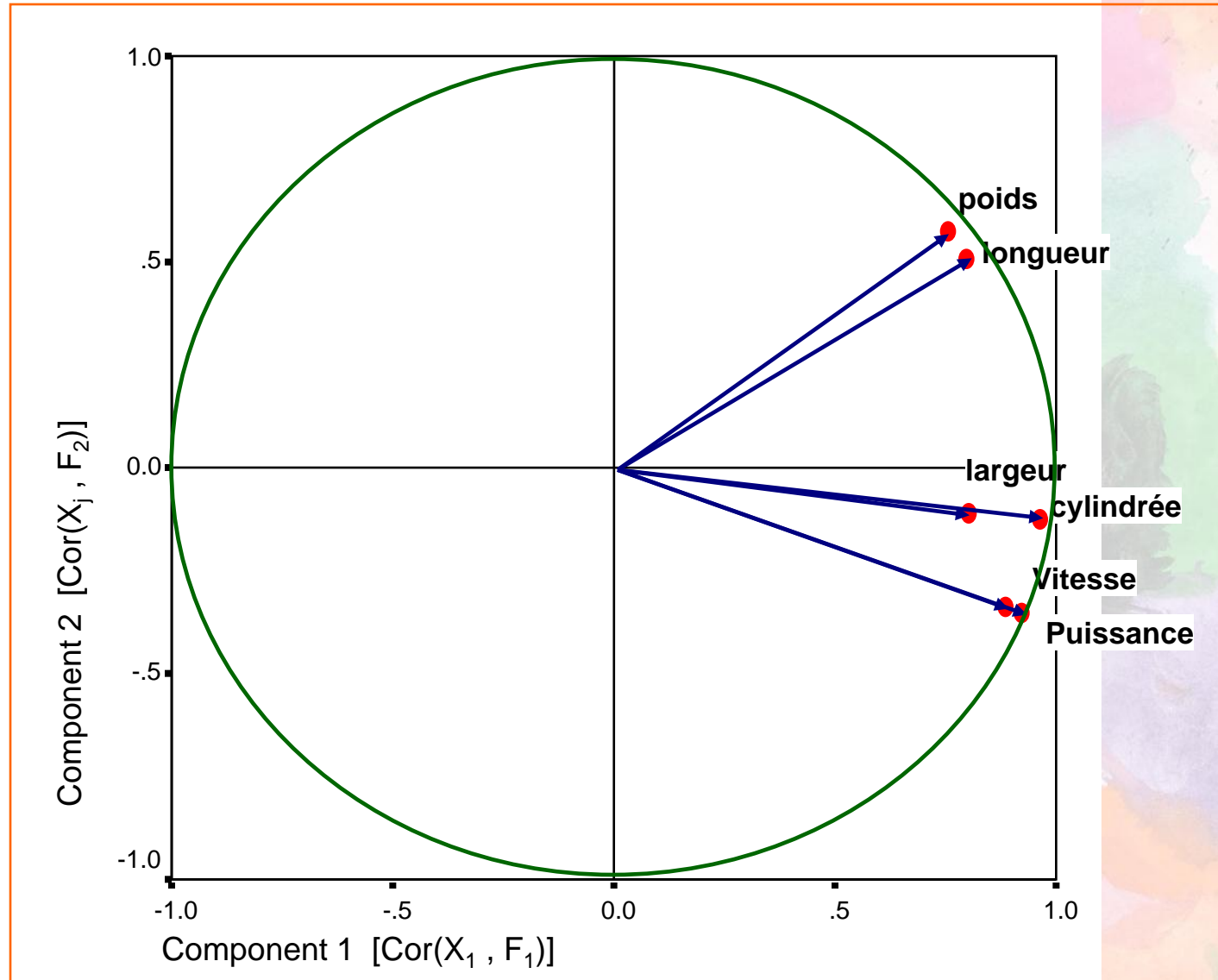
	MODÈLE	Facteur 2
1	Land Rover Discovery Td5	2.035
2	Jaguar S-Type 2.7 V6 Bi-Turbo	.951
3	Mercedes Classe S 400 CDI	.858
4	Land Rover Defender Td5	.796
5	Nissan X-Trail 2.2 dCi	.765
6	Volkswagen Touran 1.9 TDI 105	.755
7	BMW 745i	.646
8	Peugeot 407 3.0 V6 BVA	.554
9	Mercedes Classe C 270 CDI	.510
10	BMW 530d	.488
11	Renault Scenic 1.9 dCi 120	.403
12	Peugeot 307 1.4 HDI 70	.318
13	Audi A3 1.9 TDI	.179
14	Bentley Continental GT	.068
15	Citroën C3 Pluriel 1.6i	-.231
16	Nissan Micra 1.2 65	-.428
17	Audi TT 1.8T 180	-.487
18	Citroën C2 1.1 Base	-.540
19	BMW Z4 2.5i	-.632
20	Aston Martin Vanquish	-.678
21	Mini 1.6 170	-.864
22	Renault Clio 3.0 V6	-.970
23	Smart Fortwo Coupé	-1.765
24	Ferrari Enzo	-2.734

Exemple Auto 2004 : Le premier plan factoriel

Familiales (14,2%)



La carte des variables



Longueur d'une flèche = $R(X_j; F_1, F_2)$

Mesure de la qualité des deux premiers facteurs F_1 et F_2

- La variance totale du tableau des données centrées-réduites est définie par :

$$\text{Variance totale} = \sum_{j=1}^p \text{Var}(X_j^*) = p$$

- La part de la variance de X_j^* expliquée par F_1 et F_2 est égale à

$$R^2(X_j; F_1, F_2) = \text{Cor}^2(X_j, F_1) + \text{Cor}^2(X_j, F_2), \text{ car } \text{Cor}(F_1, F_2) = 0.$$

- La part de la variance totale expliquée par F_1 et F_2 est égale à :

$$\sum_{j=1}^p \left[\text{Cor}^2(X_j^*, F_1) + \text{Cor}^2(X_j^*, F_2) \right] = \lambda_1 + \lambda_2$$

Qualité globale de l'analyse

- Variance totale = p
- Proportion de variance expliquée par le facteur 1 = $\frac{\lambda_1}{p}$
- Proportion de variance expliquée par le facteur 2 = $\frac{\lambda_2}{p}$
- Proportion de variance expliquée par les facteurs 1 et 2
= $\frac{\lambda_1 + \lambda_2}{p}$

Et ainsi de suite pour les autres dimensions...

9. Construction d'une typologie des individus

- Rechercher des groupes d'individus homogènes dans la population :
 - Deux individus appartenant au même groupe sont proches
 - Deux individus appartenant à des groupes différents sont éloignés
- Construire une partition de la population en groupes homogènes et différents les uns des autres.

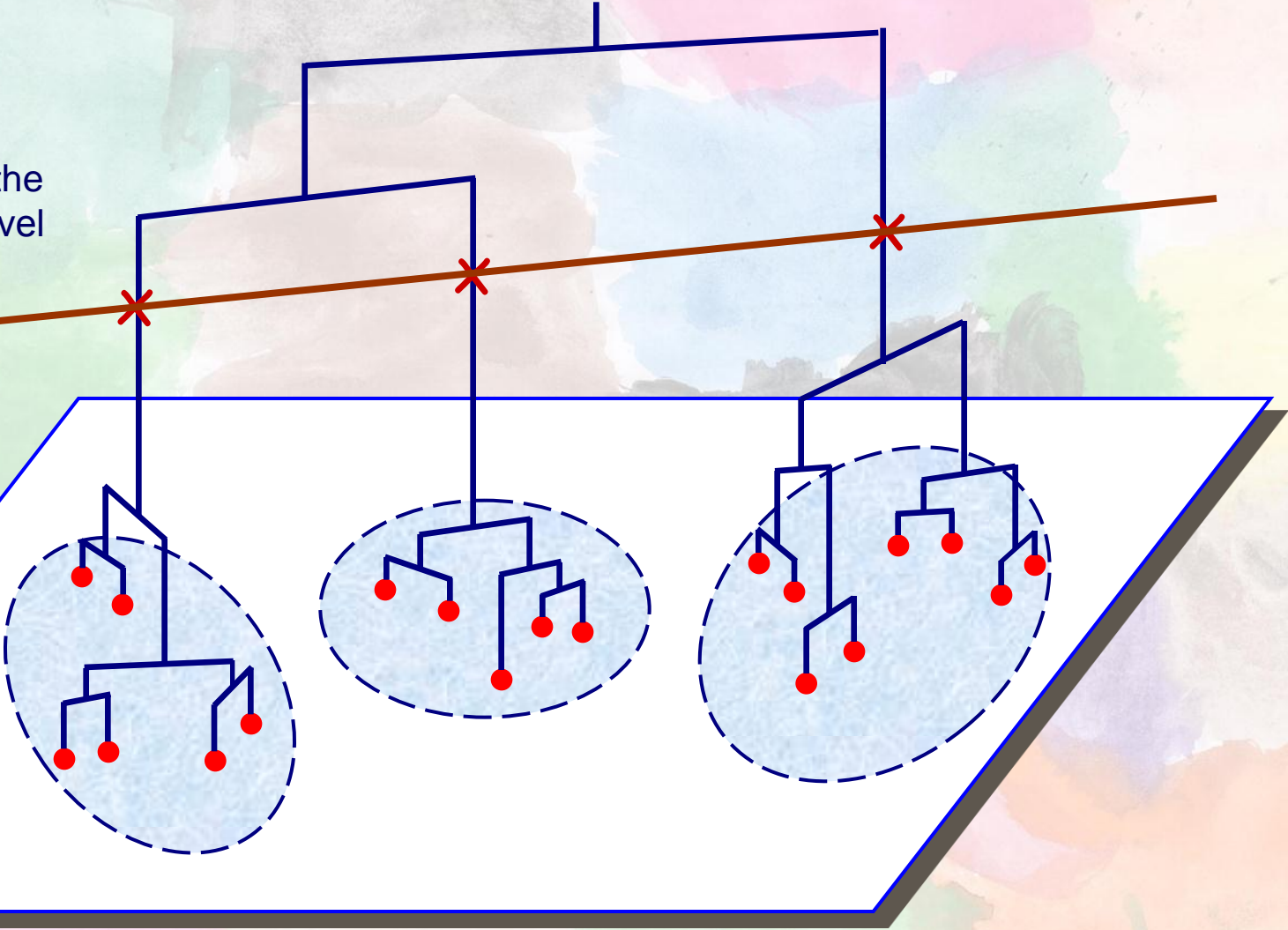
Dendrogramme

1 groupe

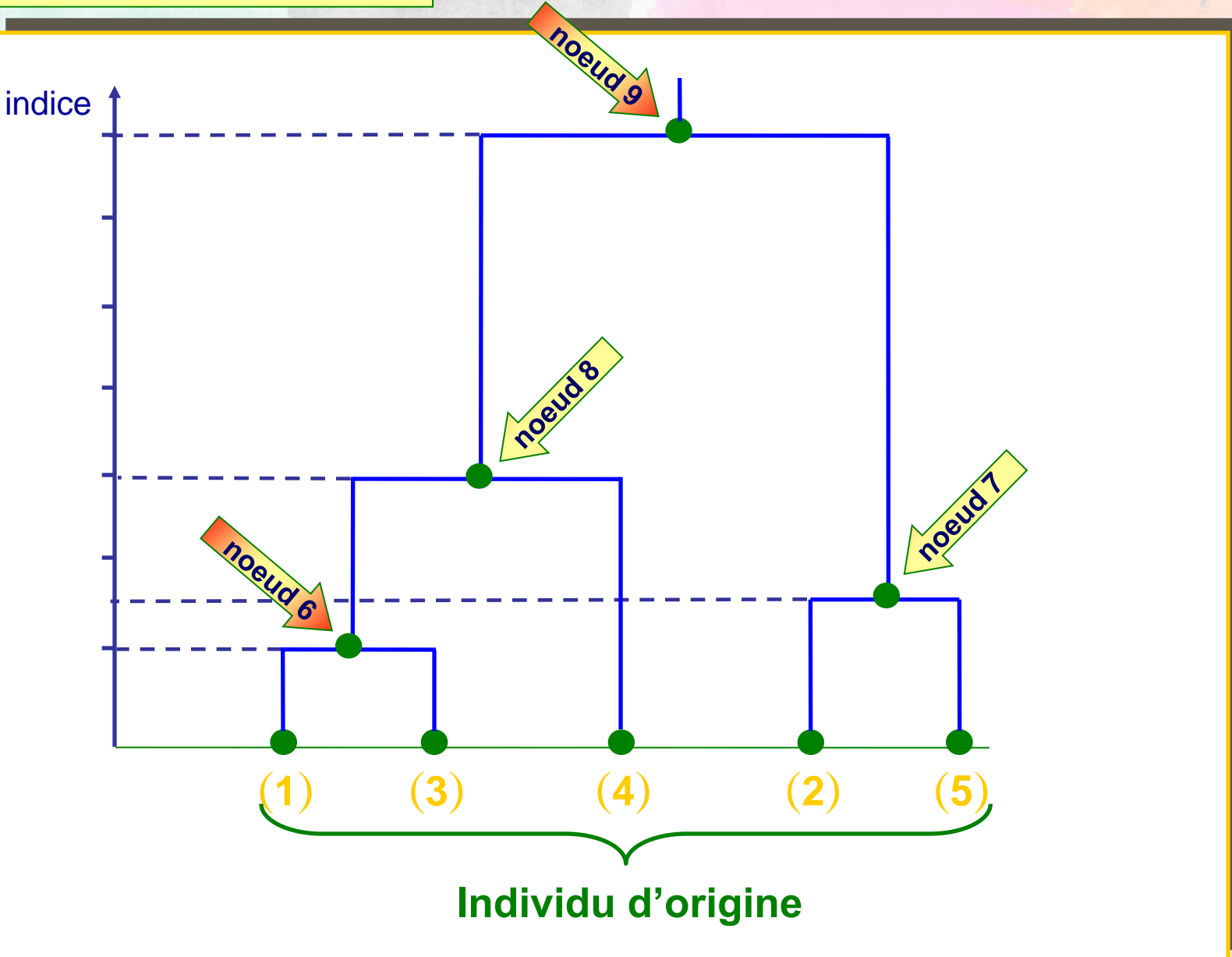


Choosing the "cutting" level

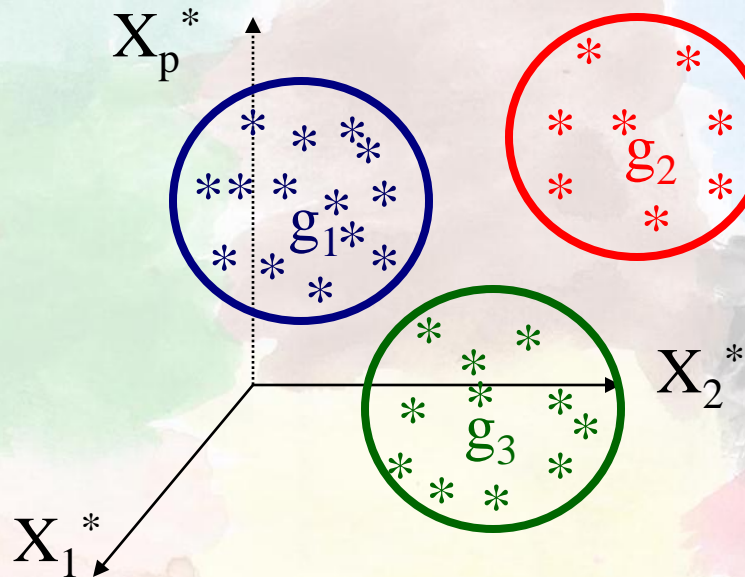
Definition of the clusters



Dendrogramme



Classification ascendante hiérarchique (Méthode de Ward)



$$\text{Distance de Ward : } D(G_i, G_j) = \frac{n_i n_j}{(n_i + n_j)} d^2(g_i, g_j)$$

n_i = effectif de la classe G_i

Tableau des distances entre les voitures

Proximity Matrix

Case	Squared Euclidean Distance						
	1:Citroën C2 1.1 Base	2:Smart Fortwo Coupé	3:Mini 1.6 170	4:Nissan Micra 1.2 65	...	23:Land Rover Discovery	24:Nissan X-Trail 2.2 d
1:Citroën C2 1.1 Base	.000	4.965	2.271	.026	...	20.325	5.246
2:Smart Fortwo Coupé	4.965	.000	9.016	5.412	...	39.487	18.625
3:Mini 1.6 170	2.271	9.016	.000	2.249	...	16.268	3.420
4:Nissan Micra 1.2 65	.026	5.412	2.249	.000	...	19.316	4.703
.
.
.
23:Land Rover Discovery	20.325	39.487	16.268	19.316000	6.953
24:Nissan X-Trail 2.2 d	5.246	18.625	3.420	4.703	...	6.953	.000

This is a dissimilarity matrix

$$d^2(x_k^*, x_l^*) = \sum_{j=1}^p (x_{jk}^* - x_{jl}^*)^2$$

$$D_{\text{Ward}}(\text{Citroën C2}, \text{Nissan Micra}) = \frac{1 \times 1}{(1 + 1)} \times .026 = .013$$

Classification Ascendante Hiérarchique

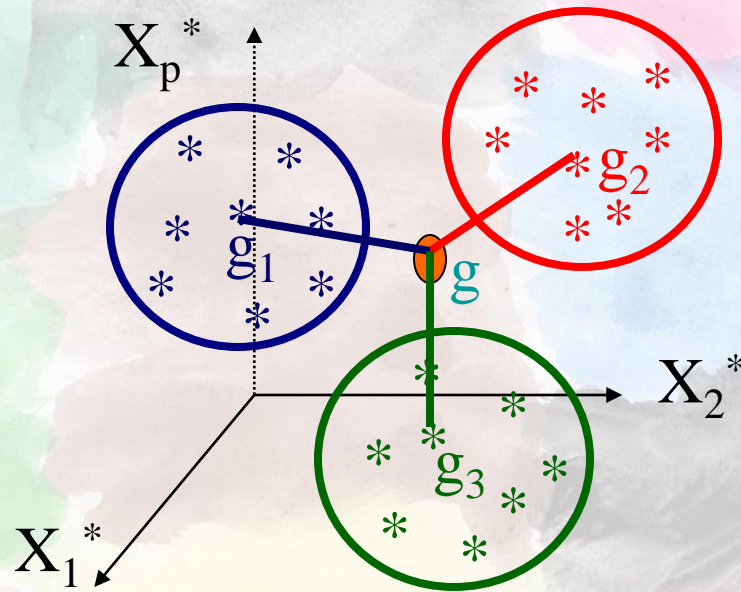
Étape initiale

Chaque individu forme une classe. On regroupe les deux individus les plus proches.

Étape courante

A chaque étape, on regroupe les deux classes G_i et G_j minimisant le critère de Ward $D(G_i, G_j)$.

Décomposition de la somme des carrés totale



$$\sum_{i=1}^n d^2(x_i^*, g) = \sum_{k=1}^K n_k d^2(g_k, g) + \sum_{k=1}^K \sum_{i \in G_k} d^2(x_i^*, g_k)$$

Somme des carrés
totale = $(n-1)*p$

Somme des carrés
inter-classes

Somme des carrés
intra-classes

Qualité de la typologie en K classes

- La somme des carrés expliquée par la typologie en K classes est égale à la somme des carrés inter-classes de la typologie en K classes.
- La qualité de la typologie est mesurée par la proportion de la somme des carrés totale expliquée par la typologie.

Coefficient : Somme des carrés
intra-classes de la typologie en K classes

Résultats SPSS :
Somme des carrés intra-
classes

Distance de Ward(1,4)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	.013	0	0	13
2	21	24	.067	0	0	8
3	6	20	.154	0	0	8
4	8	10	.255	0	0	5
5	8	11	.377	4	0	9
6	15	16	.506	0	0	11
7	7	14	.772	0	0	13
8	6	21	1.056	3	2	15
9	8	9	1.460	5	0	16
10	12	13	1.988	0	0	16
11	5	15	2.567	0	6	12
12	3	5	3.373	0	11	17
13	1	7	4.384	1	7	18
14	17	18	5.650	0	0	20
15	6	22	7.170	8	0	17
16	8	12	10.798	9	10	19
17	3	6	15.117	12	15	18
18	1	3	20.448	13	17	21
19	8	23	25.850	16	0	22
20	17	19	36.511	14	0	22
21	1	2	47.523	18	0	23
22	8	17	73.816	19	20	23
23	1	8	138.000	21	22	0

Part de somme des carrés
totale expliquée par la
typologie en K classes :
 $(138 - \text{Coeff}[n-K])/138$

Part de somme des carrés
totale expliquée par la
typologie en 2 classes :
 $(138 - 73.816)/138 = 0.465$

Somme des carrés
intra-classes pour
la typologie en K=2 classes

Somme des carrés
totale = $p*(n-1)$

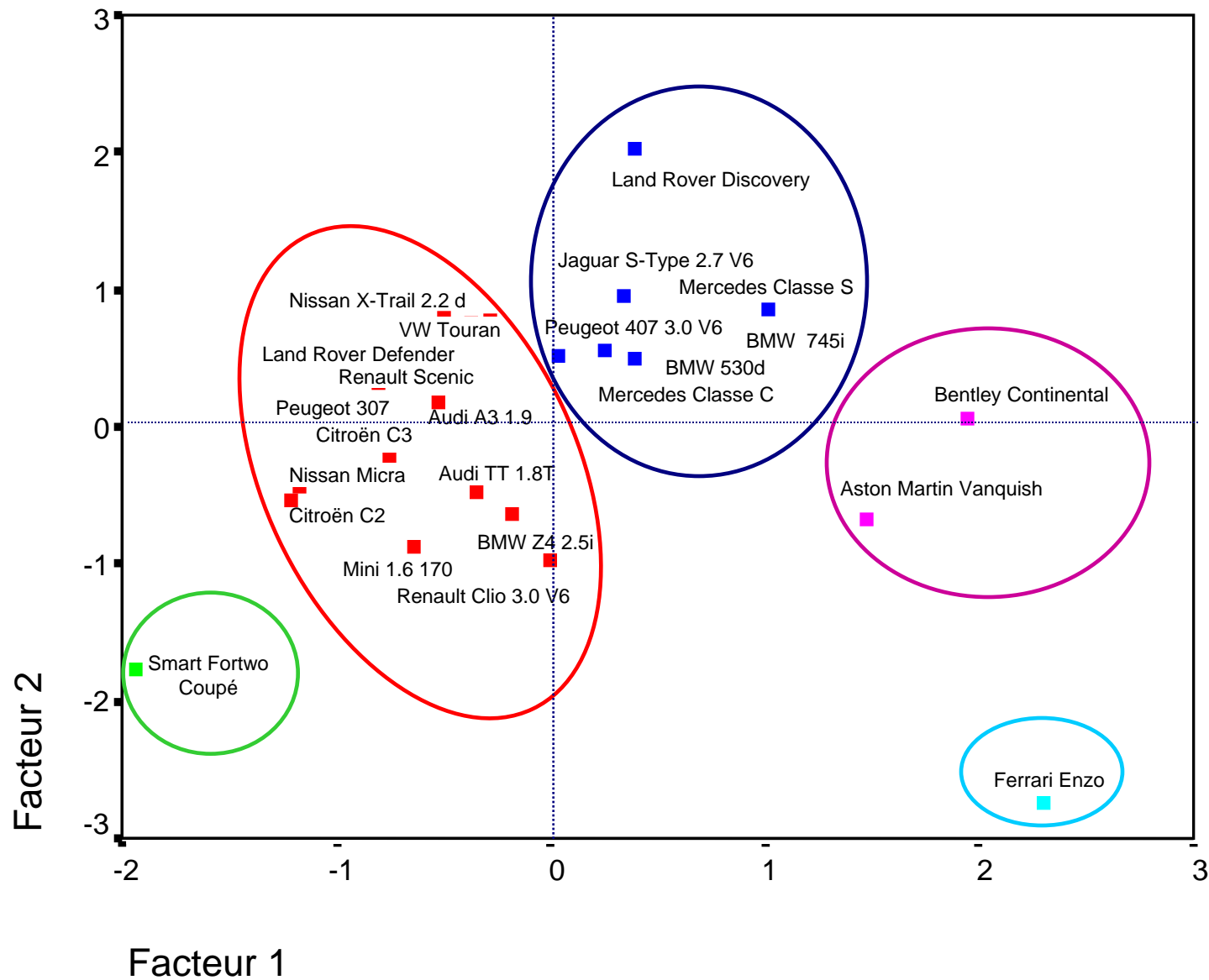
Groupe contenant 1

Choix du nombre de groupes

La typologie en 5 groupes explique 81,27 % de la S.C. totale



Premier plan factoriel et typologie



Interprétation des classes

Report

Ward Method		Cylindrée	Puissance	Vitesse	Poids	Largeur	Longueur
1	Mean	1885.31	130.08	188.69	1295.85	1748.38	4021.31
	N	13	13	13	13	13	13
2	Mean	698.00	52.00	135.00	730.00	1515.00	2500.00
	N	1	1	1	1	1	1
3	Mean	3171.86	219.57	227.29	1788.14	1912.43	4817.43
	N	7	7	7	7	7	7
4	Mean	5966.50	510.00	312.00	2110.00	1920.50	4734.50
	N	2	2	2	2	2	2
5	Mean	5998.00	660.00	350.00	1365.00	2650.00	4700.00
	N	1	1	1	1	1	1
Total	Mean	2722.54	206.67	214.71	1486.58	1838.42	4277.83
	N	24	24	24	24	24	24