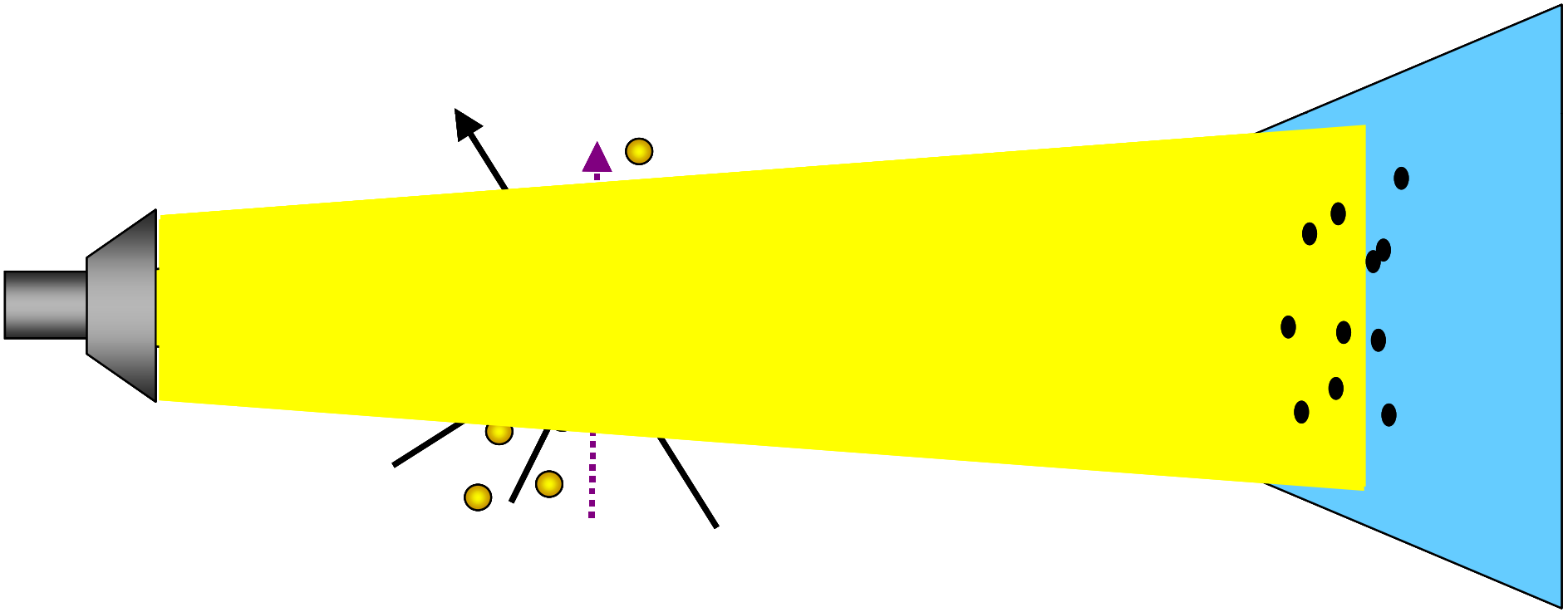
Analyse en Composantes Principales (avec SPAD) et Classification Ascendante Hiérarchique

Michel Tenenhaus

Peinture représentant un étang
(Tombeau de Thèbes, 1400 av. J.-C.)
extrait de l'Histoire de l'Art de Ernst Gombrich

Visualiser



2. Les objectifs de l'analyse en composantes principales

Décrire un tableau individus×variables :

- Résumer le tableau à l'aide d'un petit nombre de facteurs
- Visualiser le positionnement des individus les uns par rapport aux autres
- Visualiser les corrélations entre les variables
- Interpréter les facteurs

Visualisation des données

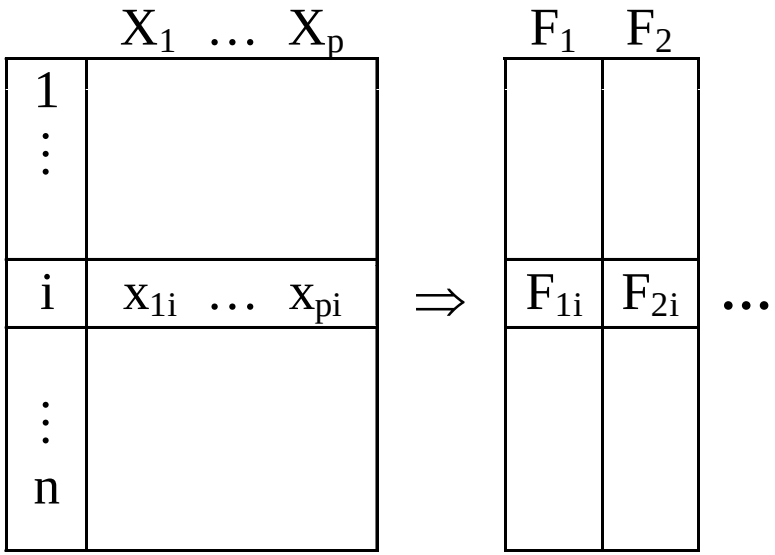
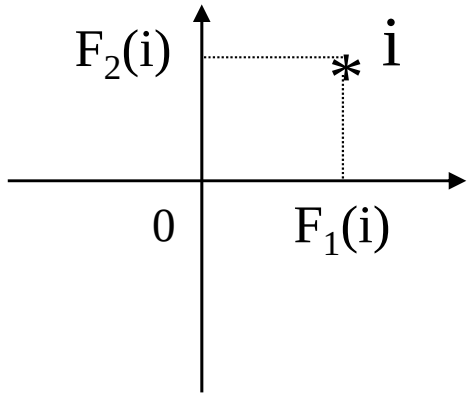


Tableau des données

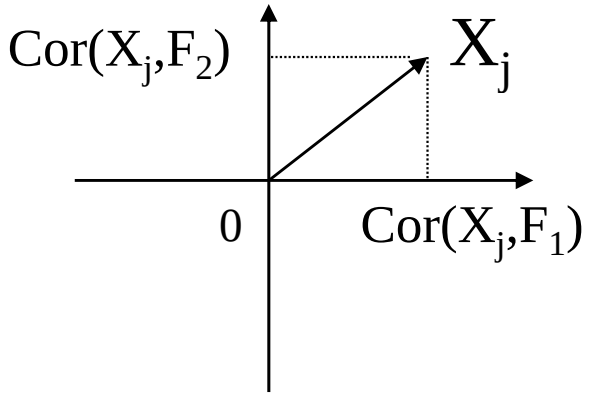
Facteurs centrés-réduits résumant les données

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

(non corrélés entre eux)



Le plan factoriel

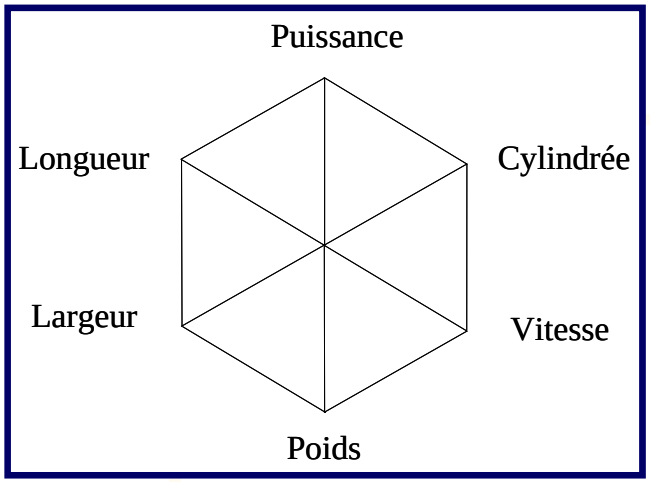


La carte des variables

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

Smart Fortwo

Mini 1.6 170

Renault Clio 3.0 V6

Audi A3 1.9 TDI

Peugeot 307 1.4 HDI 70

Peugeot 407 3.0 V6

Mercedes Classe C 270

BMW 530d

Jaguar S-Type 2.7 V6

BMW 745i

Mercedes Classe S 400

Citroën C3 Pluriel

BMW Z4 2.5i

Audi TT 1.8T 180

Aston Martin Vanquish

Bentley Continental GT

Ferrari Enzo

Renault Scenic 1.9 dCi

Volkswagen Touran 1.9 TDI

Land Rover Defender

Land Rover Discovery

Nissan X-Trail 2.2 dCi

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 :

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

Tableau des corrélations

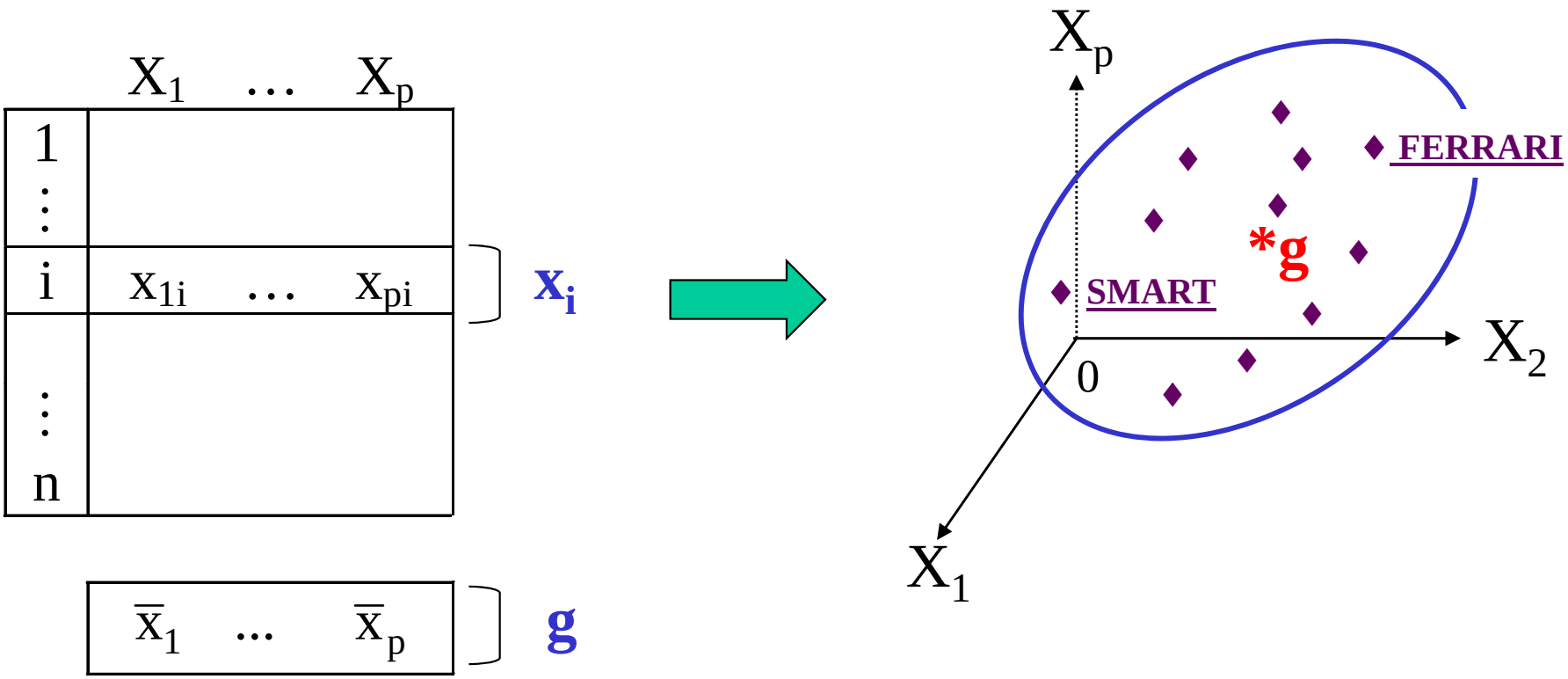
	Cylindrée	Puissance	Vitesse	Poids	Largeur	Longueur
Cylindrée	1.000	0.954	0.885	0.692	0.706	0.664
Puissance	0.954	1.000	0.934	0.529	0.730	0.527
Vitesse	0.885	0.934	1.000	0.466	0.619	0.578
Poids	0.692	0.529	0.466	1.000	0.477	0.795
Largeur	0.706	0.730	0.619	0.477	1.000	0.591
Longueur	0.664	0.527	0.578	0.795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

Toutes les corrélations sont positives.

Toutes les corrélations sont significatives au risque 5%

$$(|R| > 2/\sqrt{n})$$

5. Le nuage de points associé aux données

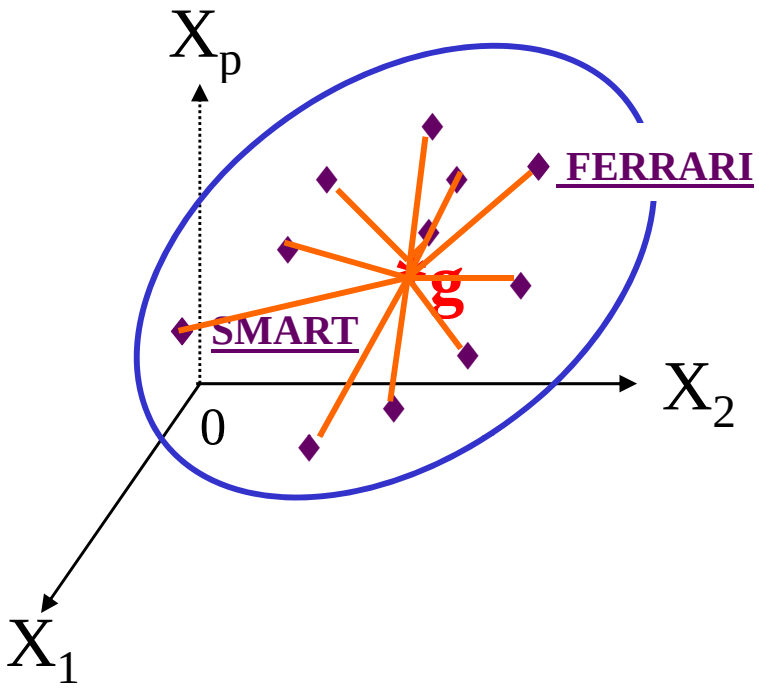


$N = \{x_1, \ldots, x_i, \ldots, x_n\}$ = Nuage de points associé aux données

Centre de gravité du nuage N : $g = \frac{1}{n}\sum_{i=1}^{n} x_i$

6. Inertie totale du nuage de points

	X_1	…	X_p	
1				
⋮				
i	x_{1i}	…	x_{pi}	$\mathbf{x_i}$
⋮				
n				
	$\bar{x}_1$	…	$\bar{x}_p$	**g**



$$\text{Inertie totale} = I(N, g) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ji} - \bar{x}_j)^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 = \sum_{j=1}^{p}\sigma_j^2$$

7. Réduction des données

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

$$X_1^* = \frac{X_1 - \overline{x}_1}{\sigma_1}$$

$$\vdots$$

$$X_p^* = \frac{X_p - \overline{x}_p}{\sigma_p}$$

de moyenne 0 et d'écart-type 1.

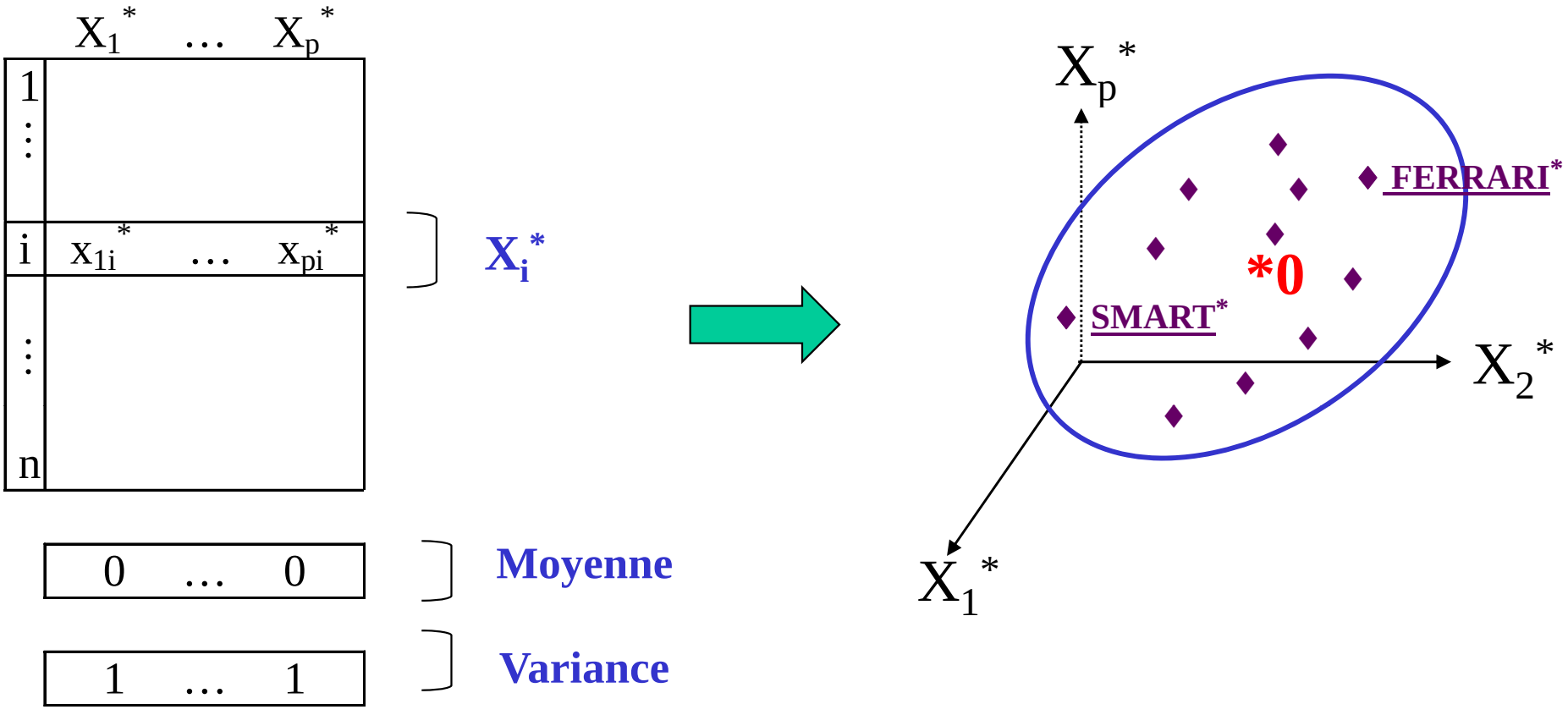
Les données centrées-réduites (SPAD)

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
	Std. Deviation					1.000	1.000	1.000

Outlier si |valeur| > 2

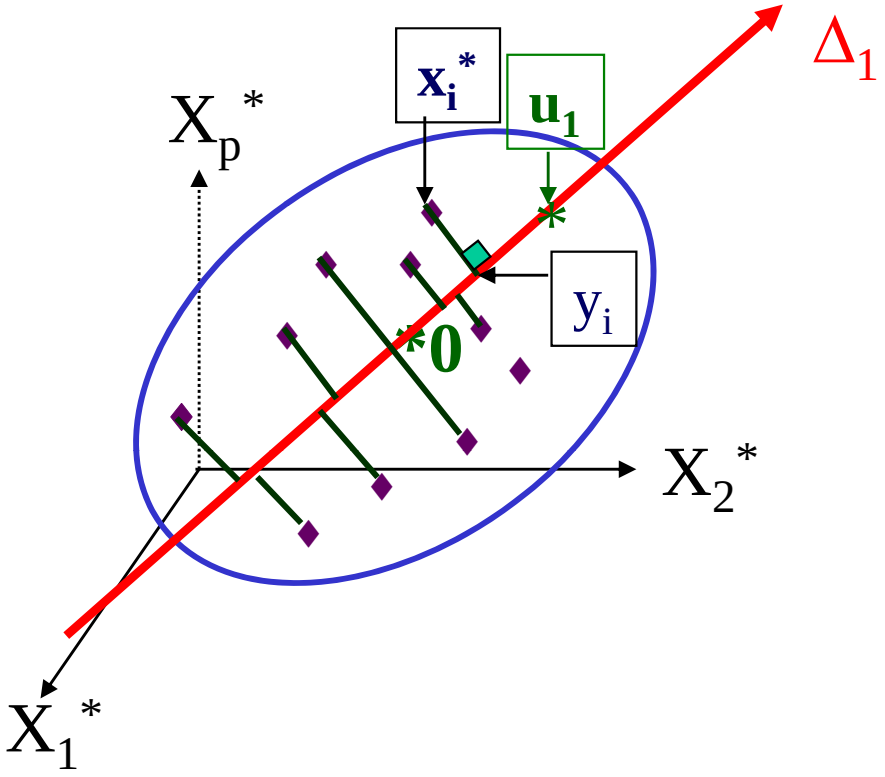
8. Le nuage de points associé aux données réduites



N* = $\{x_1^*, \ldots, x_i^*, \ldots, x_n^*\}$

Centre de gravité : $g^* = 0$, **Inertie totale :** $I(N^*, 0) = p$

9. Premier axe principal Δ_1

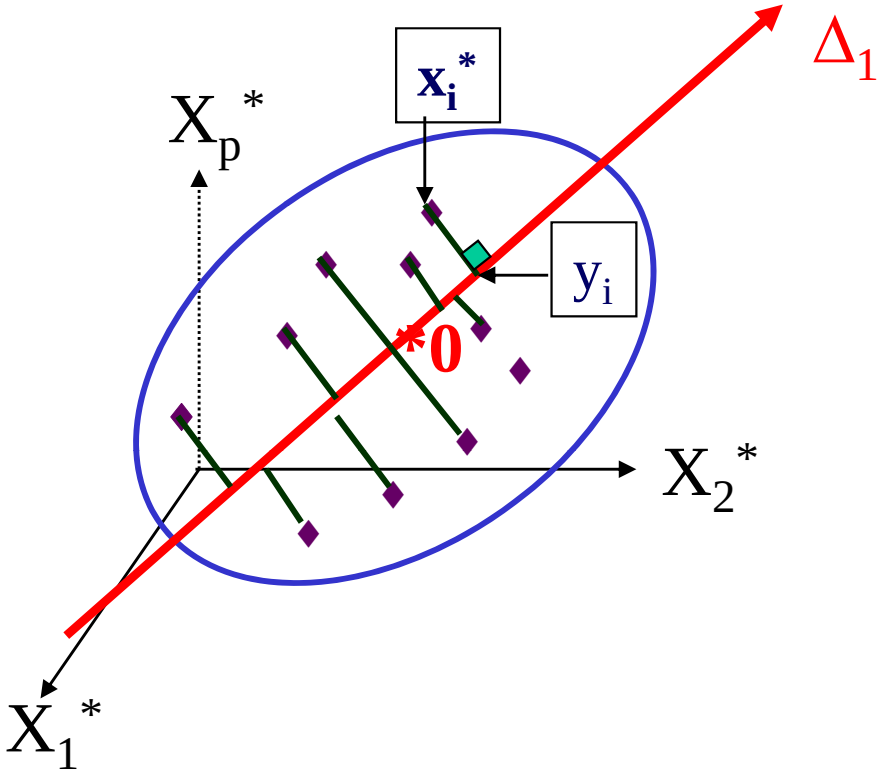


Objectif 1 : On cherche l'axe Δ_1 passant le mieux possible au milieu du nuage N^*.

On cherche à minimiser l'inertie du nuage N^* par rapport à l'axe Δ_1 :

$$I(N^*, \Delta_1) = \frac{1}{n} \sum_{i=1}^{n} d^2(x_i^*, y_i)$$

Premier axe principal Δ_1

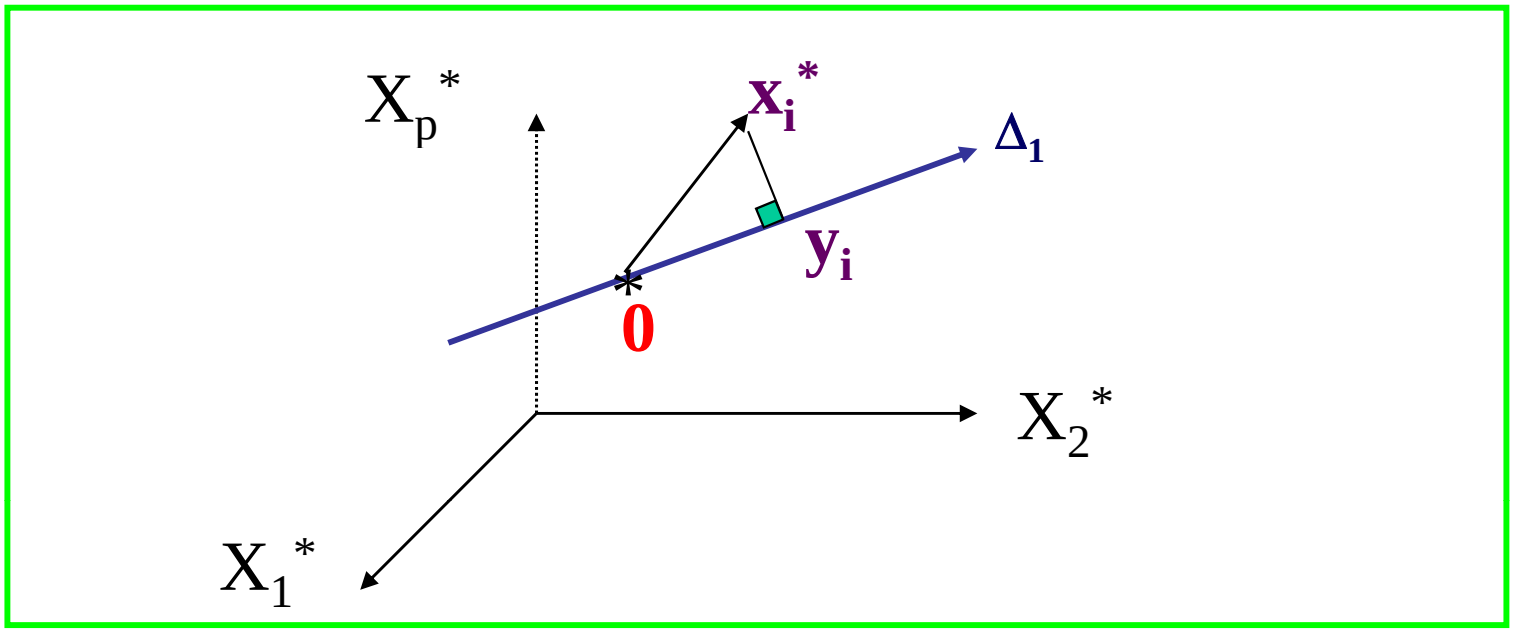


Objectif 2 : On cherche l'axe d'allongement Δ_1 du nuage N^*.

On cherche à maximiser l'inertie du nuage N^* projeté sur l'axe Δ_1 :

$$I(\{y_1,\ldots,y_n\},0)=\frac{1}{n}\sum_{i=1}^{n}d^2(y_i,0)$$

Les objectifs 1 et 2 sont atteints simultanément



De :

$$d^2(x_i^*,0) = d^2(y_i,0) + d^2(x_i^*, y_i)$$

on déduit :

$$\frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*,0) = \frac{1}{n}\sum_{i=1}^{n} d^2(y_i,0) + \frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*, y_i)$$

Inertie totale = p = Inertie expliquée par Δ_1 + Inertie résiduelle

Maximiser Minimiser

Résultats

- L'axe Δ_1 passe par le centre de gravité 0 du nuage de points N^*.
- L'axe Δ_1 est engendré par le vecteur normé u_1, vecteur propre de la matrice des corrélations R associé à la plus grande valeur propre λ_1.
- L'inertie expliquée par l'axe Δ_1 est égal à λ_1.
- La part d'inertie expliquée par le premier axe principal Δ_1 est égal à λ_1/p.

Résultat SPAD

Tableau des valeurs propres

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	4.4113	73.52	73.52
2	0.8534	14.22	87.74
3	0.4357	7.26	95.01
4	0.2359	3.93	98.94
5	0.0514	0.86	99.79
6	0.0124	0.21	100.00

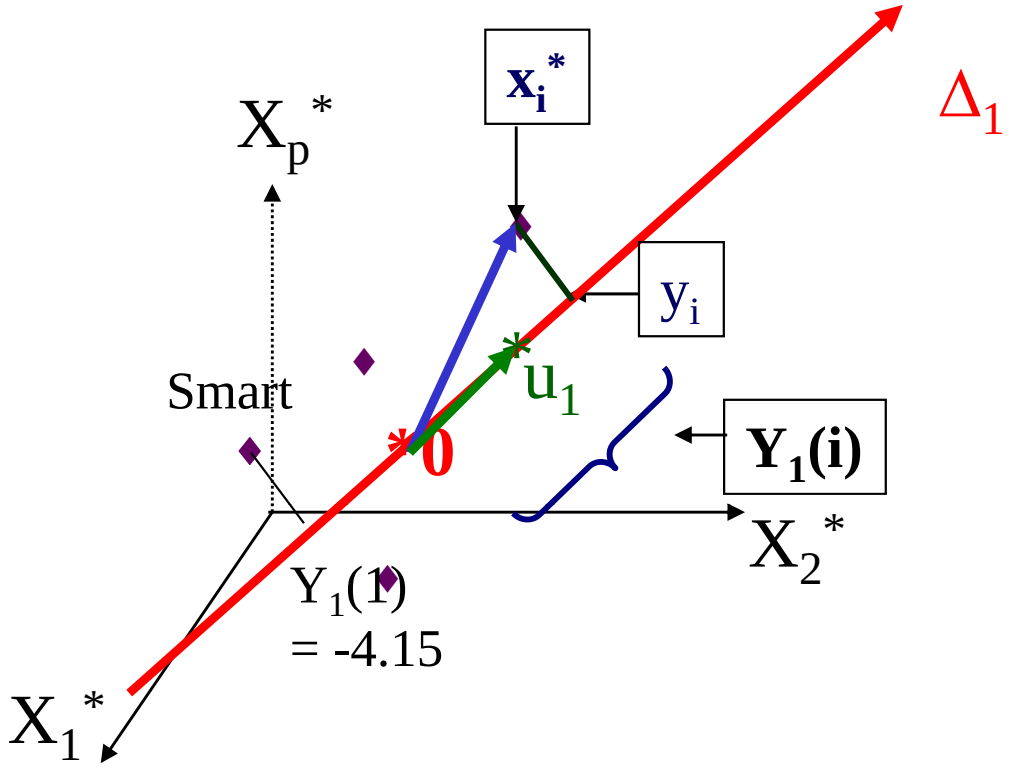
Résultat SPAD

Les vecteurs propres

Libellé de la variable	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Cylindrée	0.46	-0.14	0.21	-0.23	-0.65	0.50
Puissance	0.44	-0.38	0.14	-0.17	-0.09	-0.78
Vitesse	0.42	-0.37	0.31	0.41	0.57	0.31
Poids	0.36	0.62	0.22	-0.53	0.39	0.01
Largeur	0.38	-0.12	-0.88	-0.14	0.15	0.13
Longueur	0.38	0.55	-0.09	0.67	-0.26	-0.19

Normalisation : $.46^2 + .44^2 + \ldots + .38^2 = 1$

10. Première composante principale Y_1



Y_1 est une nouvelle variable définie pour chaque individu i par :

$Y_1(i)$ = longueur algébrique du segment $0y_i$

= coordonnée de y_i sur l'axe Δ_1

= produit scalaire entre les vecteurs x_i^* et u_1

= $\sum_{j=1}^{p} u_{1j} x^*_{ji}$ $\Rightarrow$ $Y_1 = \sum_{j=1}^{p} u_{1j} X_j^*$

Résultats SPAD

Identificateur	Carré de la Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Citroën C2 1.1	7.10	-2.60	-0.51	-0.18	0.17	-0.21	-0.03
Smart Fortwo	20.93	-4.15	-1.67	0.27	-0.92	-0.03	0.03
Mini 1.6 170	2.93	-1.38	-0.82	0.37	-0.05	0.46	-0.05
Nissan Micra 1.2	6.61	-2.51	-0.40	-0.17	0.12	-0.29	-0.05
Renault Clio 3.0 V6	1.16	0.00	-0.92	0.39	-0.27	0.29	0.13
Audi A3 1.9 TDI	1.39	-1.12	0.17	-0.17	0.27	-0.07	0.06
Peugeot 307 1.4 HDI	3.43	-1.73	0.30	-0.41	0.36	-0.24	-0.09
Peugeot 407 3.0 V6	0.76	0.55	0.52	0.26	0.34	0.00	-0.01
Mercedes Classe C 270	0.68	0.08	0.48	0.53	0.37	0.12	0.11
BMW 530d	1.40	0.84	0.46	0.16	0.68	0.05	0.03
Jaguar S-Type 2.7 V6	1.68	0.72	0.90	0.21	0.54	0.10	-0.13
BMW 745i	5.22	2.13	0.61	0.40	0.16	-0.36	-0.09
Mercedes Classe S 400	5.66	2.17	0.81	-0.48	0.14	0.06	0.26
Citroën C3 Pluriel 1.6i	2.72	-1.62	-0.22	0.02	0.18	-0.01	-0.03
BMW Z4 2.5i	0.70	-0.40	-0.60	0.20	0.34	0.13	0.14
Audi TT 1.8T 180	1.08	-0.75	-0.46	0.13	0.33	0.41	-0.07
Aston Martin Vanquish	11.62	3.16	-0.64	1.01	-0.20	-0.39	0.23
Bentley Continental GT	20.32	4.16	0.06	1.49	-0.83	0.14	-0.23
Ferrari Enzo	34.42	4.95	-2.58	-1.81	0.12	-0.07	-0.10
Renault Scenic 1.9 dCi	0.93	-0.84	0.38	-0.25	0.11	0.08	-0.01
Volkswagen Touran 1.9 TDI	1.23	-0.80	0.71	-0.24	0.13	0.00	-0.02
Land Rover Defender	3.24	-1.07	0.75	-0.18	-1.18	-0.31	0.01
Land Rover Discovery	7.81	0.85	1.92	-1.52	-1.00	0.31	0.03
Nissan X-Trail 2.2 dCi	0.96	-0.61	0.72	-0.05	0.12	-0.17	-0.12

DISTO = $d^2(x_i^*, 0)$

Corrélations entre les variables et les composantes principales

Corrélations des variables actives avec les facteurs

Libellé de la variable	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Cylindrée	0.96	-0.13	0.14	-0.11	-0.15	0.06
Puissance	0.92	-0.35	0.09	-0.08	-0.02	-0.09
Vitesse	0.89	-0.34	0.21	0.20	0.13	0.03
Poids	0.76	0.58	0.15	-0.26	0.09	0.00
Largeur	0.80	-0.11	-0.58	-0.07	0.03	0.01
Longueur	0.80	0.50	-0.06	0.33	-0.06	-0.02

Dans SPSS : Component Matrix

Propriétés de la première composante principale Y_1

- $Y_1 = u_{11}X_1^* + u_{12}X_2^* + \ldots + u_{1p}X_p^*$
- Moyenne de $Y_1 = 0$
- Variance de Y_1 = Inertie expliquée par $\Delta_1 = \lambda_1$
- $Cor(X_j, Y_1) = \sqrt{\lambda_1}u_{1j}$
- $\frac{1}{p}\sum_{j=1}^{p} cor^2(X_j, Y_1) = \frac{\lambda_1}{p}$ est maximum

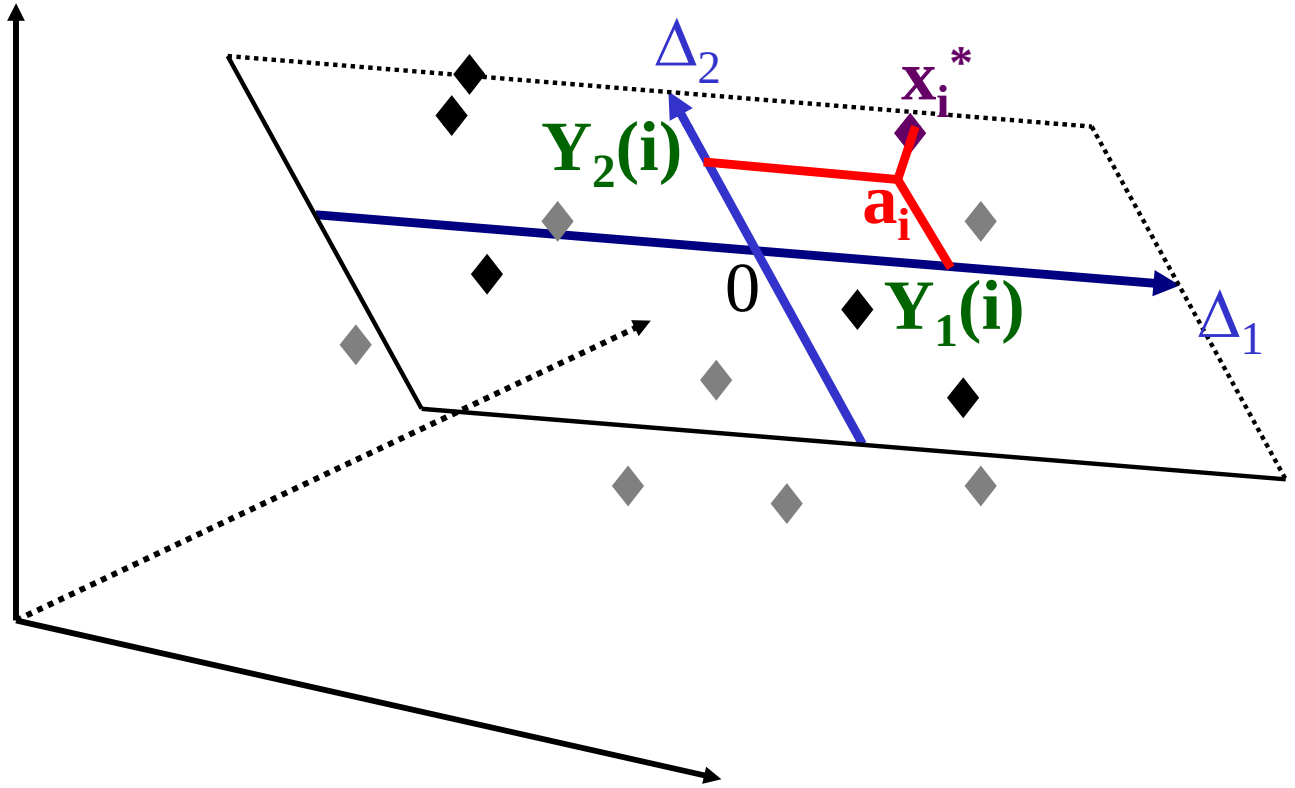
Qualité de la première composante principale

- Inertie totale = 6
- Inertie expliquée par le premier axe principal = $\lambda_1 = 4.4113$
- Part d'inertie expliquée par le premier axe principal :

$$\frac{\lambda_1}{p} = \frac{4.4113}{6} = 0.7352$$

- La première composante principale explique 73,5% de la variance totale.

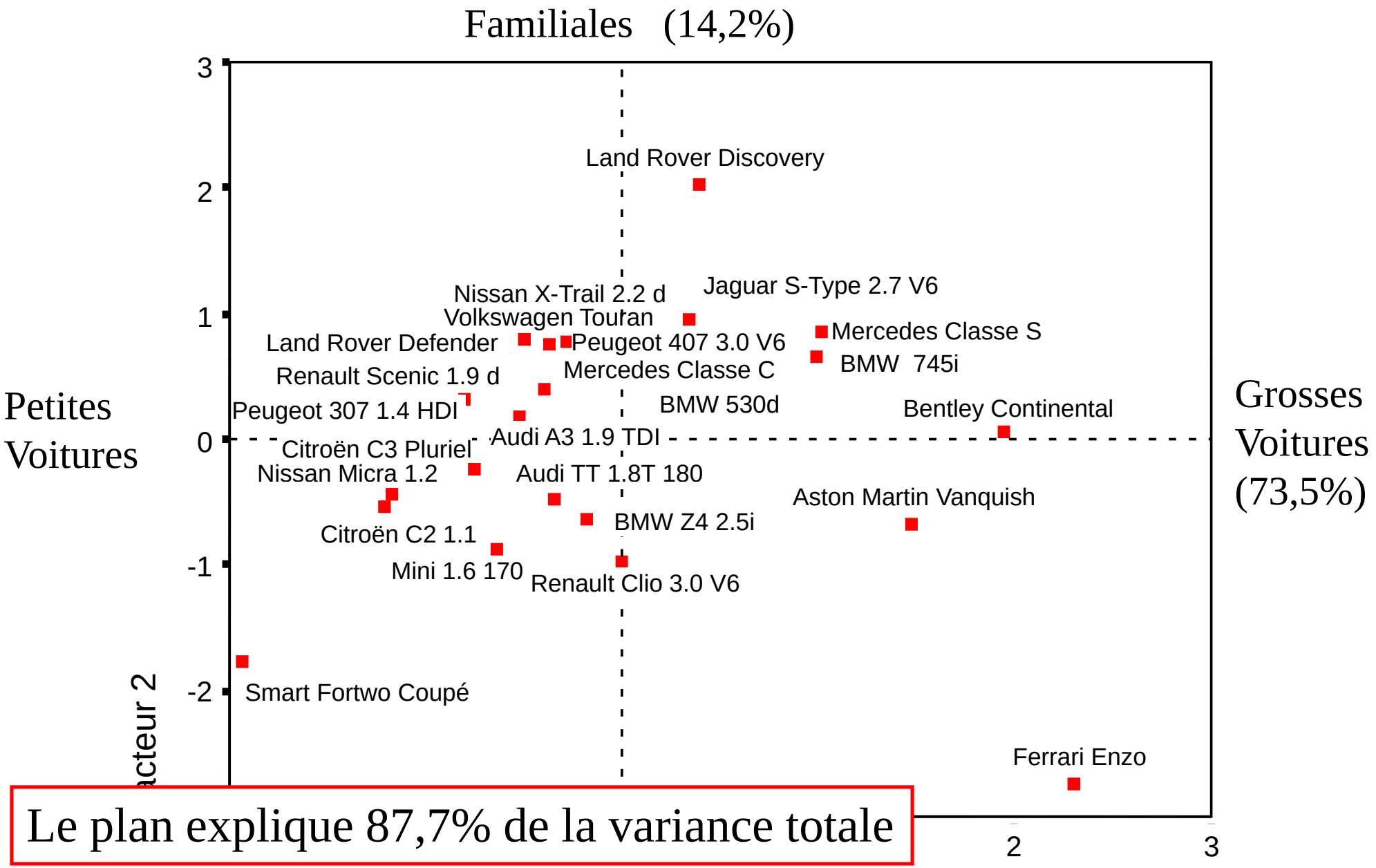
11. Deuxième axe principal Δ_2



Résultats

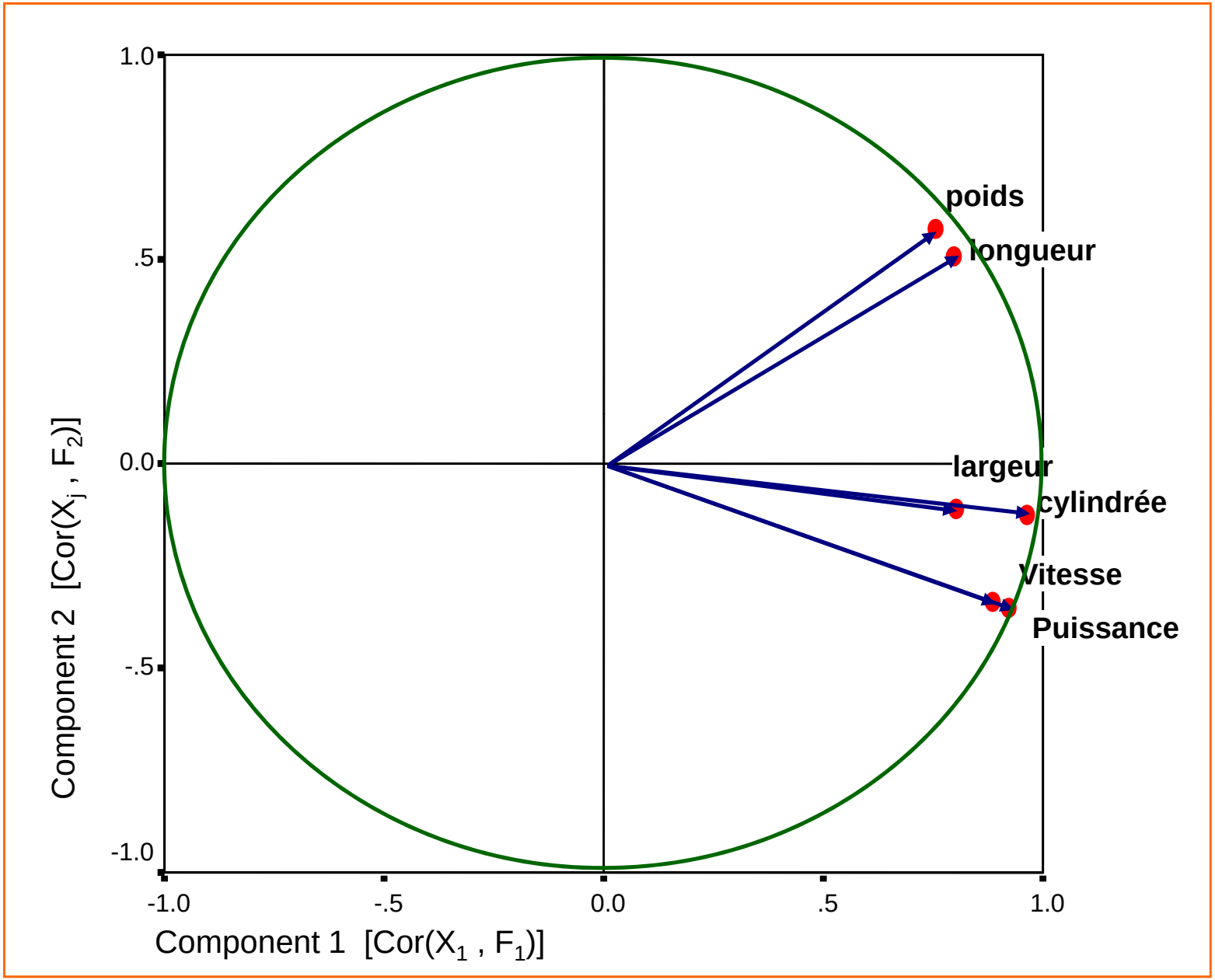
- On recherche le deuxième axe principal Δ_2 orthogonal à Δ_1 et passant le mieux possible au milieu du nuage.
- Il passe par le centre de gravité 0 du nuage de points et est engendré par le vecteur normé u_2, vecteur propre de la matrice des corrélations R associé à la deuxième plus grande valeur propre λ_2.
- La deuxième composante principale Y_2 est définie par projection des points sur le deuxième axe principal.
- La deuxième composante principale Y_2 est centrée, de variance λ_2, et non corrélée à la première composante principale Y_1.

Exemple Auto 2004 : Le premier plan factoriel



Le plan explique 87,7% de la variance totale

La carte des variables



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

Qualité globale de l'analyse

Inertie totale = variance totale = p

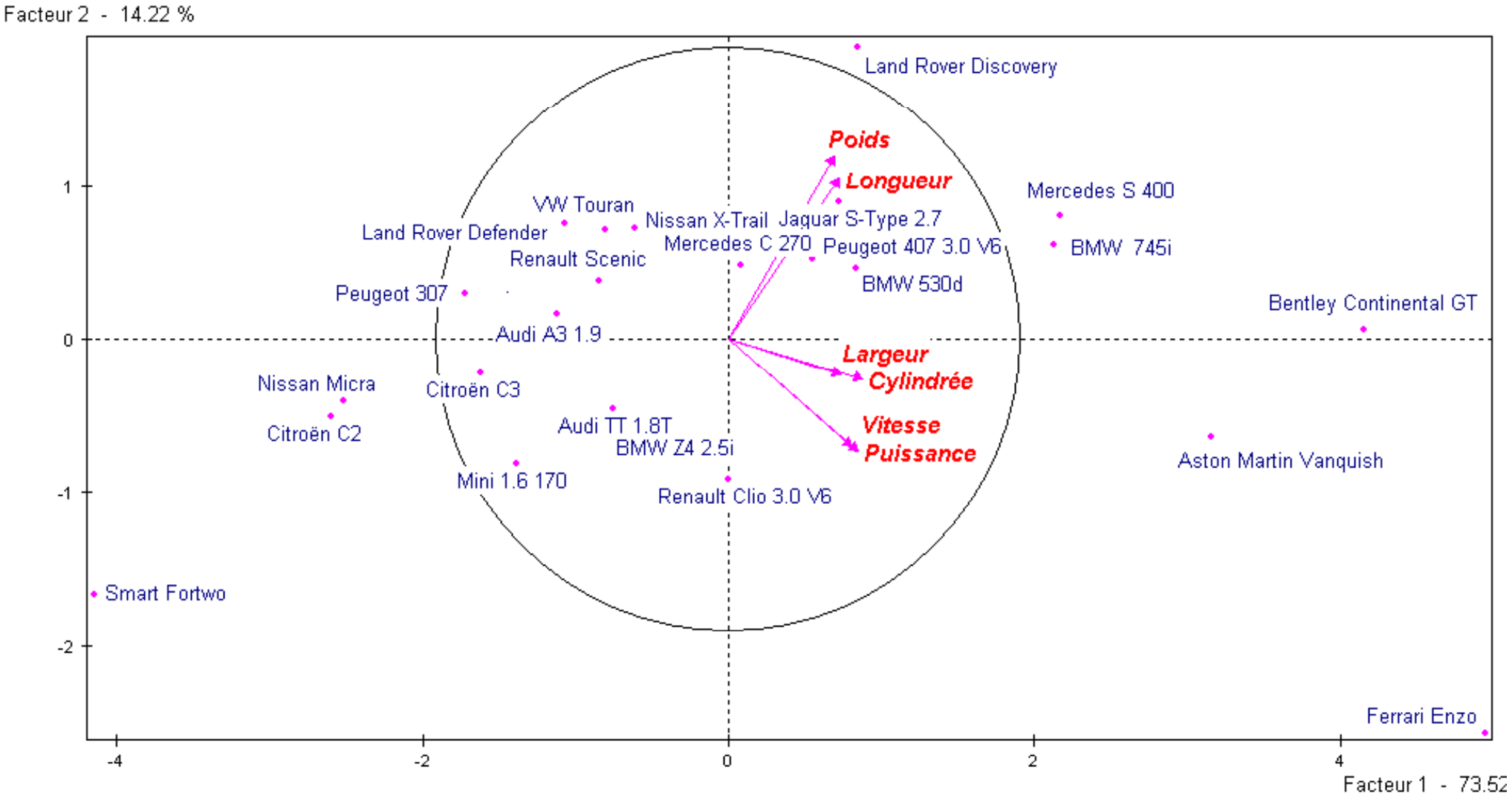
Part de variance expliquée par
la première composante principale = $\frac{\lambda_1}{p}$

Part de variance expliquée par
la deuxième composante principale = $\frac{\lambda_2}{p}$

Part de variance expliquée par
les deux premières composantes principales = $\frac{\lambda_1 + \lambda_2}{p}$

Et ainsi de suite pour les autres dimensions...

12. Le biplot



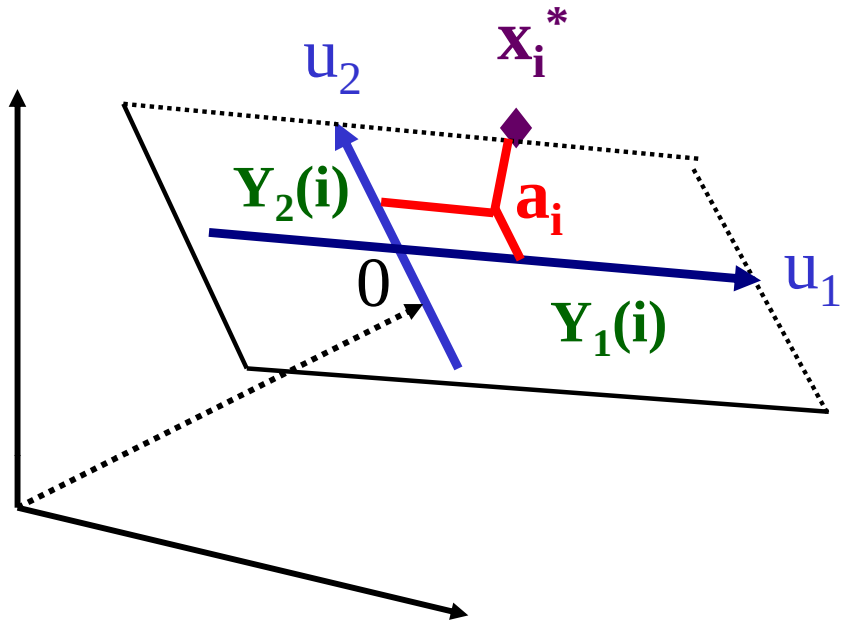
Les échelles doivent être identiques sur les deux axes.
Le cercle des variables doit être un *cercle*.

Interprétation du biplot

- La répartition des projections des individus i sur l'axe variable X_j reflète les valeurs x_{ij}
- Les coordonnées des individus i sont les valeurs des composantes principales :

  $[Y_1(i), Y_2(i)]$.
- Les coordonnées des variables X_j sont les vecteurs propres multipliés par une certaine constante, par exemple 2 : $(2u_{1j}, 2u_{2j})$.

Justification : la formule de reconstitution



De

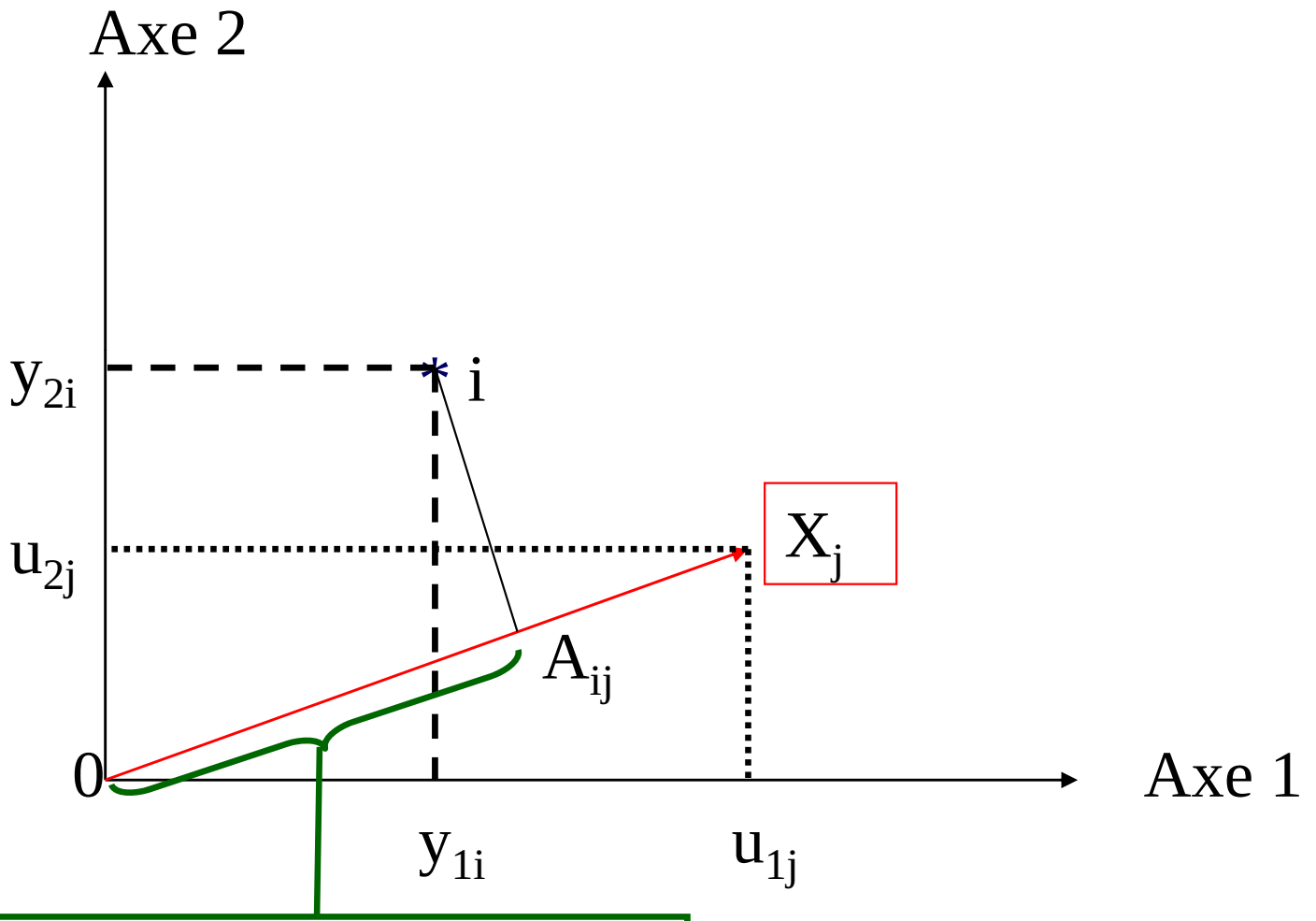
$$x_i^* \approx a_i = Y_1(i)u_1 + Y_2(i)u_2$$

on déduit

$$x_{ij}^* \approx Y_1(i)u_{1j} + Y_2(i)u_{2j} = \left\langle \begin{pmatrix} Y_1(i) \\ Y_2(i) \end{pmatrix}, \begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \right\rangle = < i, X_j >$$

$= \sqrt{u_{1j}^2 + u_{2j}^2} \times$ Coordonnée de la projection de l'individu i sur l'axe variable X_j

Justification de la lecture du bi-plot



$$\overline{0A_{ij}} = \; < i, X_j > / \sqrt{u_{1j}^2 + u_{2j}^2}$$

$$= (y_{1i}u_{1j} + y_{2i}u_{2j}) / \sqrt{u_{1j}^2 + u_{2j}^2} \approx x_{ij}^* / \sqrt{u_{1j}^2 + u_{2j}^2}$$

13. Exemple des races canines

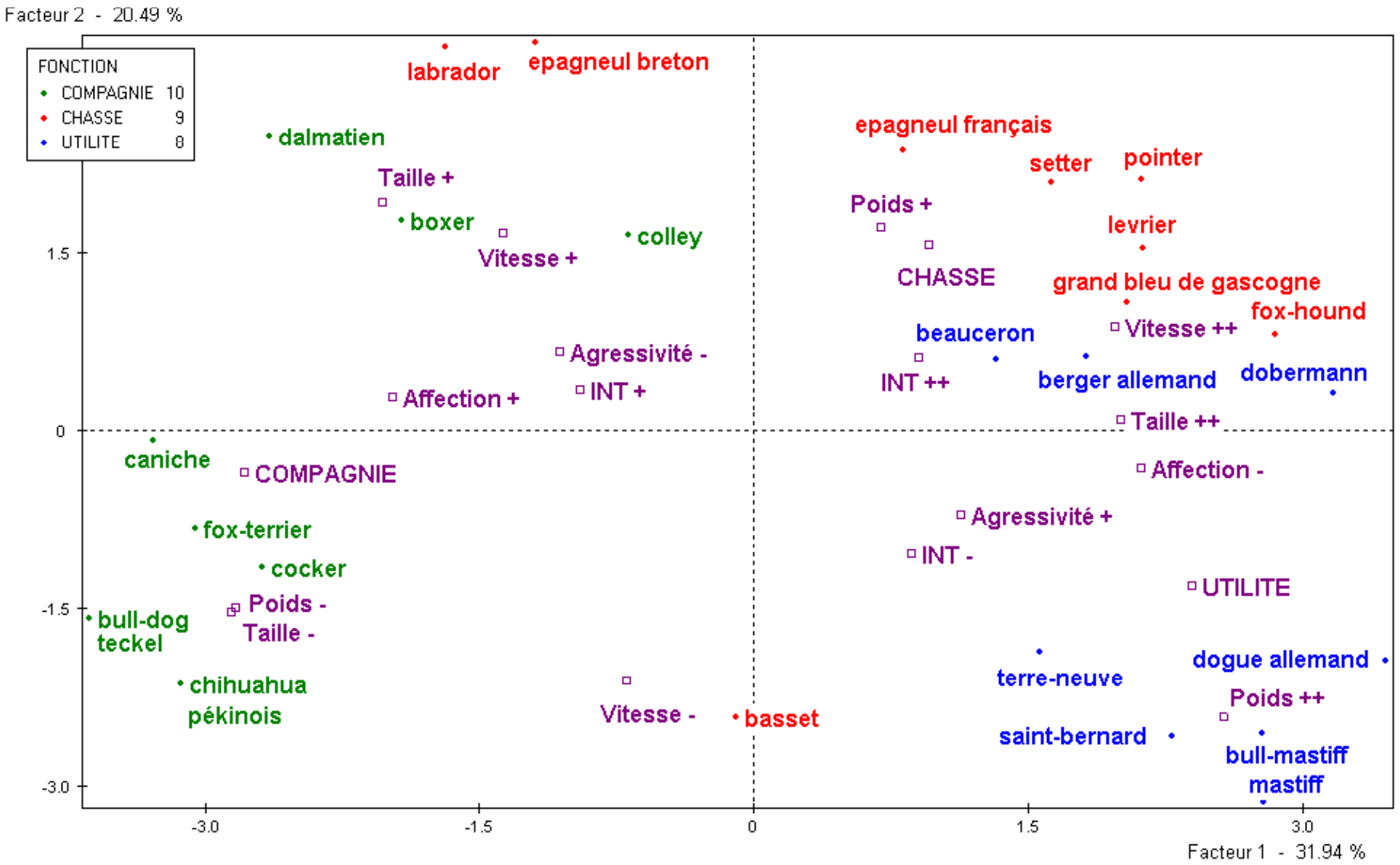
	Race	Taille	Poids	Vitesse	Intell.	Affect.	Agress.	Fonction
1	Beauceron	TA++	PO+	V++	INT+	AF+	AG+	Utilité
2	Basset	TA-	PO-	V-	INT-	AF-	AG+	Chasse
3	Berger-Allemand	TA++	PO+	V++	INT++	AF+	AG+	Utilité
4	Boxer	TA+	PO+	V+	INT+	AF+	AG+	Compagnie
5	Bull-Dog	TA-	PO-	V-	INT+	AF+	AG-	Compagnie
6	Bull-Mastiff	TA++	PO++	V-	INT++	AF-	AG+	Utilité
7	Caniche	TA-	PO-	V+	INT++	AF+	AG-	Compagnie
8	Chihuahua	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
9	Cocker	TA+	PO-	V-	INT+	AF+	AG+	Compagnie
10	Colley	TA++	PO+	V++	INT+	AF+	AG-	Compagnie
11	Dalmatien	TA+	PO+	V+	INT+	AF+	AG-	Compagnie
12	Doberman	TA++	PO+	V++	INT++	AF-	AG+	Utilité
13	Dogue Allemand	TA++	PO++	V++	INT-	AF-	AG+	Utilité
14	Epagneul Breton	TA+	PO+	V+	INT++	AF+	AG-	Chasse
15	Epagneul Français	TA++	PO+	V+	INT+	AF-	AG-	Chasse
16	Fox-Hound	TA++	PO+	V++	INT-	AF-	AG+	Chasse
17	Fox-Terrier	TA-	PO-	V+	INT+	AF+	AG+	Compagnie
18	Grd Bleu de Gascogne	TA++	PO+	V+	INT-	AF-	AG+	Chasse
19	Labrador	TA+	PO+	V+	INT+	AF+	AG-	Chasse
20	Lévrier	TA++	PO+	V++	INT-	AF-	AG-	Chasse
21	Mastiff	TA++	PO++	V-	INT-	AF-	AG+	Utilité
22	Pékinois	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
23	Pointer	TA++	PO+	V++	INT++	AF-	AG-	Chasse
24	Saint-Bernard	TA++	PO++	V-	INT+	AF-	AG+	Utilité
25	Setter	TA++	PO+	V++	INT+	AF-	AG-	Chasse
26	Teckel	TA-	PO-	V-	INT+	AF+	AG-	Compagnie
27	Terre-Neuve	TA++	PO++	V-	INT+	AF-	AG-	Utilité

Le tableau disjonctif complet

Race	T-	T+	T++	P-	P+	P++	V-	V+	V++	I-	I+	I++	Af-	Af+	Ag-	Ag+	Compagnie	Chasse	Utilité
Beauceron	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0	1
Basset	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0
Berger all	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1
Boxer	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	0
Bull-dog	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1	0	0
Bull Mastiff	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	1
Caniche	1	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0
Chihuahua	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
Cocker	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
Colley	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	0
Dalmatien	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	0
Doberman	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1
Dogue all	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1
Epagneul br	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0
Epagneul fr	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0
Fox-Hound	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	0
Fox-Terrier	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0
Grd Bl de G	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0
Labrador	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0
Lévrier	0	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0	0	1	0
Mastiff	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	1
Pékinois	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
Pointer	0	0	1	0	1	0	0	0	1	0	0	1	1	0	1	0	0	1	0
St-Bernard	0	0	1	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	1
Setter	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0
Teckel	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1	0	0
Terre neuve	0	0	1	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	1

x_{ijl} = 1 si l'individu i possède la modalité l de la variable j
= 0 sinon

ACP du tableau disjonctif complet



14. Utilisation de SPSS

Les données centrées-réduites (SPSS)

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
	Std. Deviation					1.000	1.000	1.000

Outlier si |valeur| > 2

Résultats SPSS : Les facteurs

		MODÈLE	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5	Facteur 6
1		Citroën C2 1.1 Base	-1.210	-.540	.266	.334	-.894	.278
2		Smart Fortwo Coupé	-1.934	-1.765	-.407	-1.863	-.126	-.296
3		Mini 1.6 170	-.644	-.864	-.552	-.103	2.003	.449
4		Nissan Micra 1.2 65	-1.171	-.428	.258	.251	-1.249	.447
5		Renault Clio 3.0 V6	-.001	-.970	-.571	-.553	1.234	-1.181
6		Audi A3 1.9 TDI	-.522	.179	.250	.537	-.314	-.540
7		Peugeot 307 1.4 HDI 70	-.804	.318	.615	.719	-1.042	.820
8		Peugeot 407 3.0 V6 BVA	.258	.554	-.380	.681	.012	.099
9		Mercedes Classe C 270 CDI	.037	.510	-.781	.742	.521	-1.000
10		BMW 530d	.391	.488	-.244	1.361	.197	-.225
11		Jaguar S-Type 2.7 V6 Bi-Turbo	.336	.951	-.311	1.080	.431	1.146
12		BMW 745i	.991	.646	-.597	.329	-1.535	.752
13		Mercedes Classe S 400 CDI	1.010	.858	.707	.279	.242	-2.257
14		Citroën C3 Pluriel 1.6i	-.756	-.231	-.028	.372	-.023	.283
15		BMW Z4 2.5i	-.186	-.632	-.295	.678	.560	-1.192
16		Audi TT 1.8T 180	-.350	-.487	-.200	.673	1.769	.658
17		Aston Martin Vanquish	1.471	-.678	-1.491	-.401	-1.685	-2.022
18		Bentley Continental GT	1.939	.068	-2.216	-1.682	.608	2.016
19		Ferrari Enzo	2.306	-2.734	2.683	.235	-.318	.852
20		Renault Scenic 1.9 dCi 120	-.392	.403	.364	.226	.350	.084
21		Volkswagen Touran 1.9 TDI 105	-.375	.755	.350	.269	-.006	.163
22		Land Rover Defender Td5	-.500	.796	.261	-2.383	-1.324	-.075
23		Land Rover Discovery Td5	.396	2.035	2.252	-2.015	1.342	-.305
24		Nissan X-Trail 2.2 dCi	-.286	.765	.068	.235	-.752	1.045
Total	Mean		.000	.000	.000	.000	.000	.000
	Std. Deviation		1.000	1.000	1.000	1.000	1.000	1.000

Propriétés des facteurs de SPSS

Lien entre les composantes principales et les facteurs de SPSS

Les facteurs de SPSS sont les composantes principales réduites.

$$F_h = \frac{1}{\sqrt{\lambda_h}} \sqrt{\frac{n}{n-1}} Y_h$$

Calcul des facteurs de SPSS en fonction des variables $(X_j^*)_{SPSS}$

$$F_h = \sum_{j=1}^{p} w_{hj} (X_j^*)_{SPSS}$$

$$w_h = \frac{1}{\sqrt{\lambda_h}} u_h$$

Tableau des w_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.

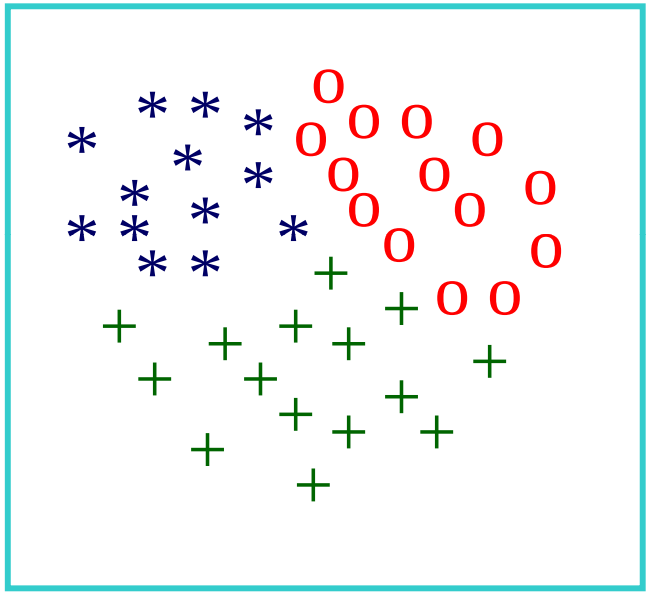
15. Construction d'une typologie des individus

- 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.

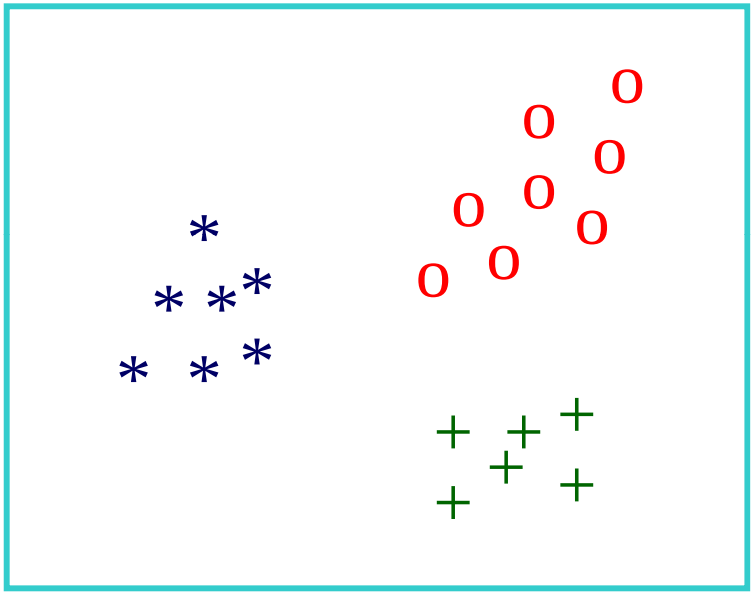
- Construire une partition de la population en groupes homogènes et différents les uns des autres.

- On réalise la typologie au choix
  (1) sur les données centrées-réduites,
  (2) sur les premières composantes principales (SPAD),
  (3) sur les premières composantes principales réduites
  (les facteurs de SPSS).

Construction d'une typologie des individus

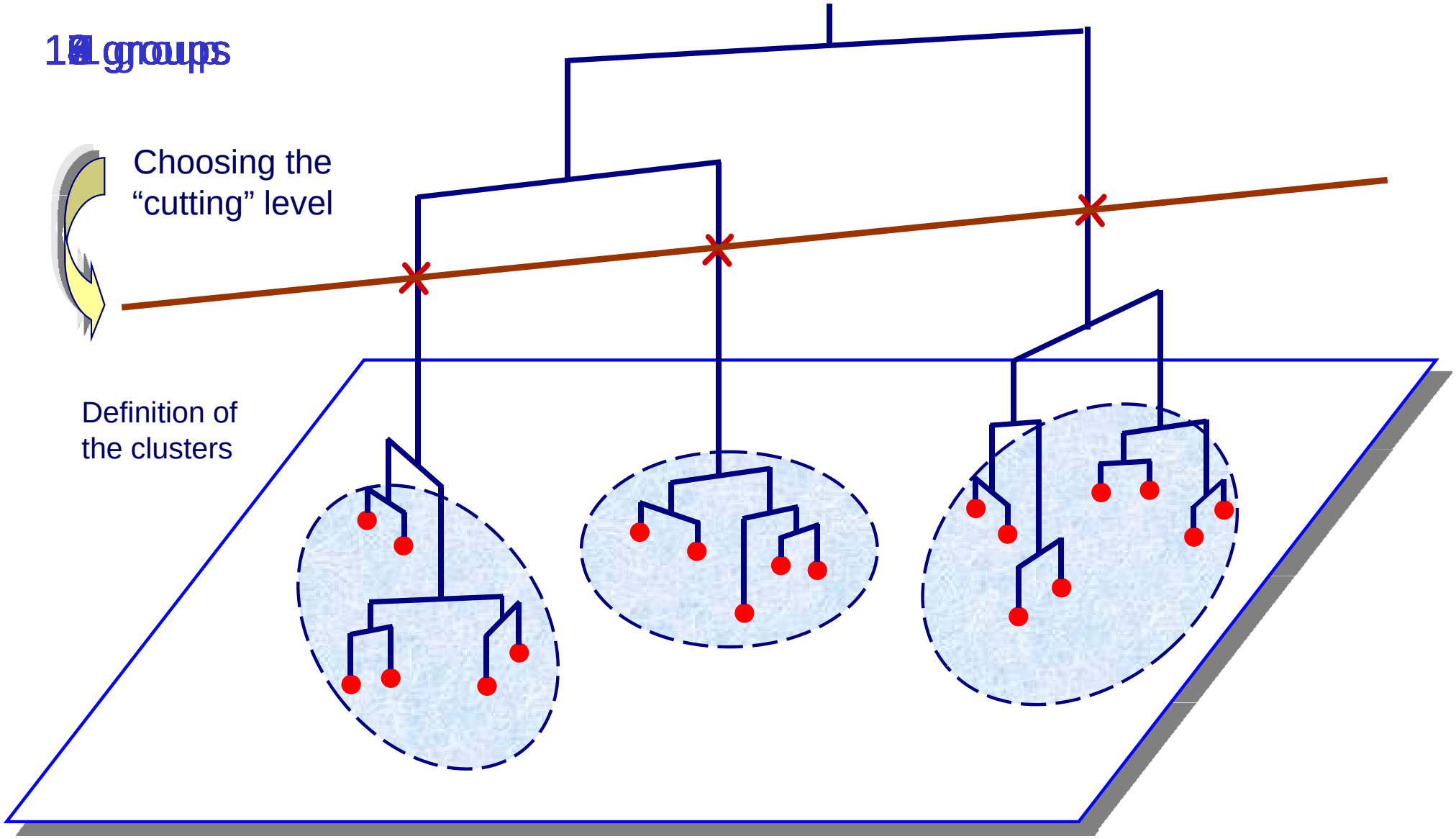


Fabrication de groupes à partir de données uniformément réparties

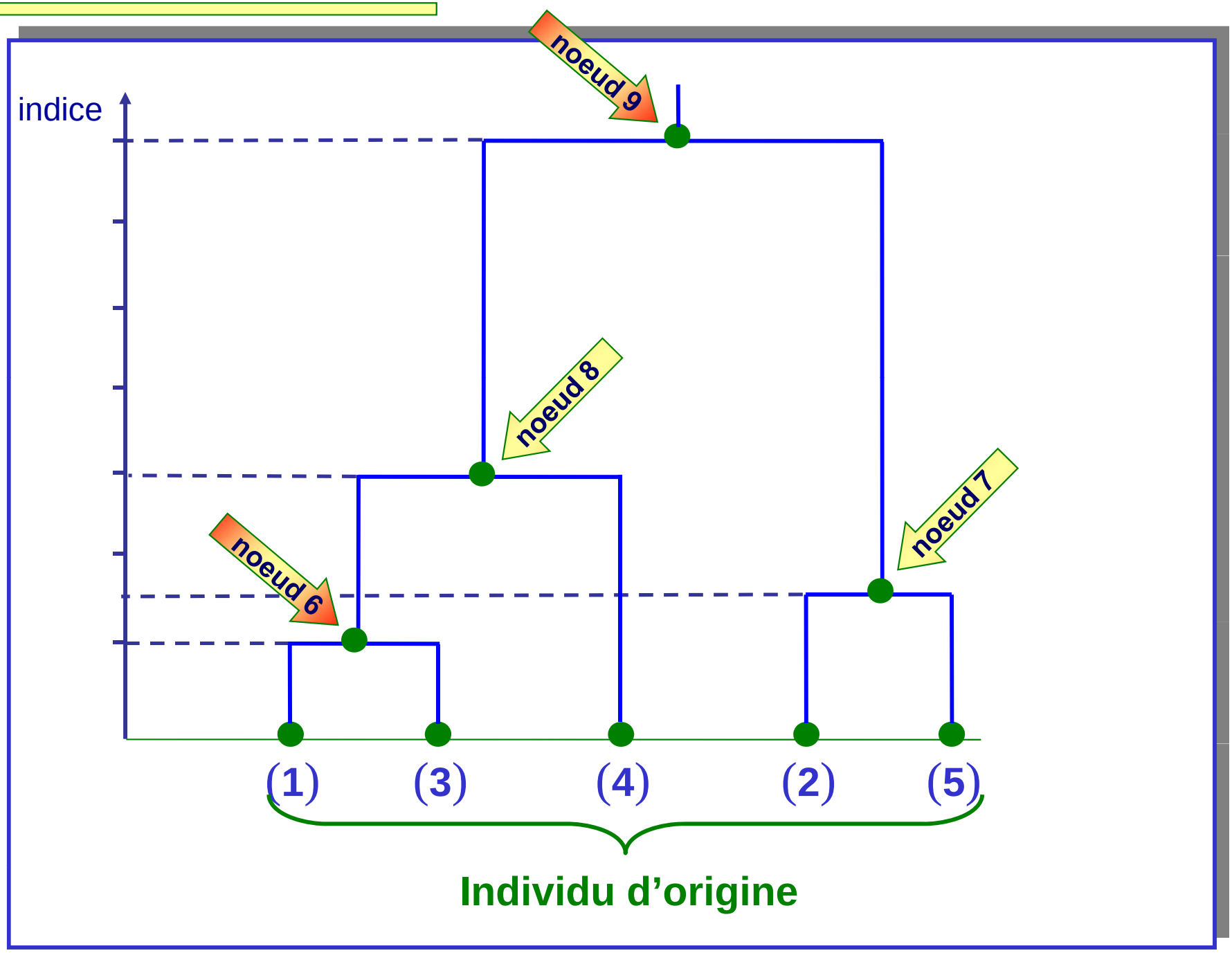


Données structurées en trois groupes

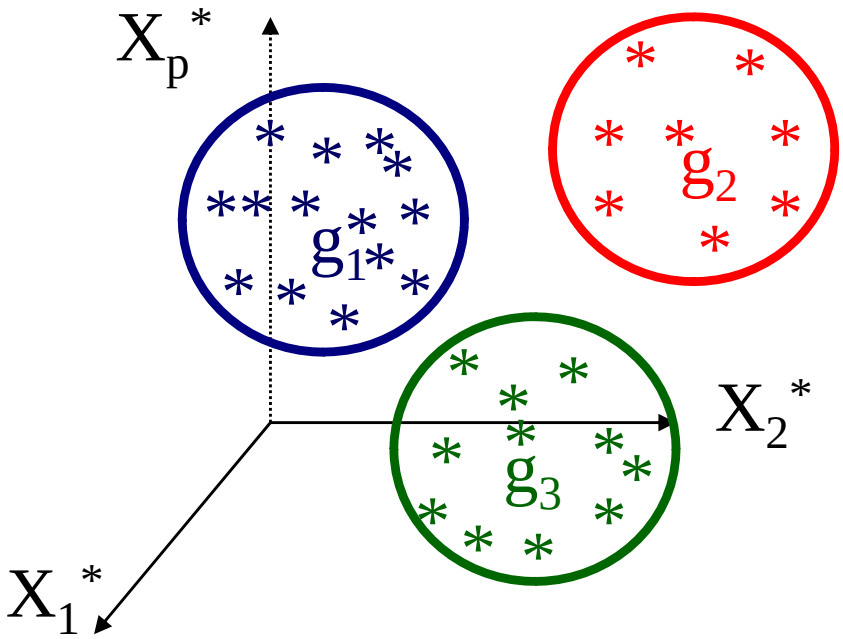
Dendrogramme



Dendrogramme



Classification ascendante hiérarchique (Méthode de Ward)



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	19.316	...	.000	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

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

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

Classification Ascendante Hiérarchique

Étape initiale

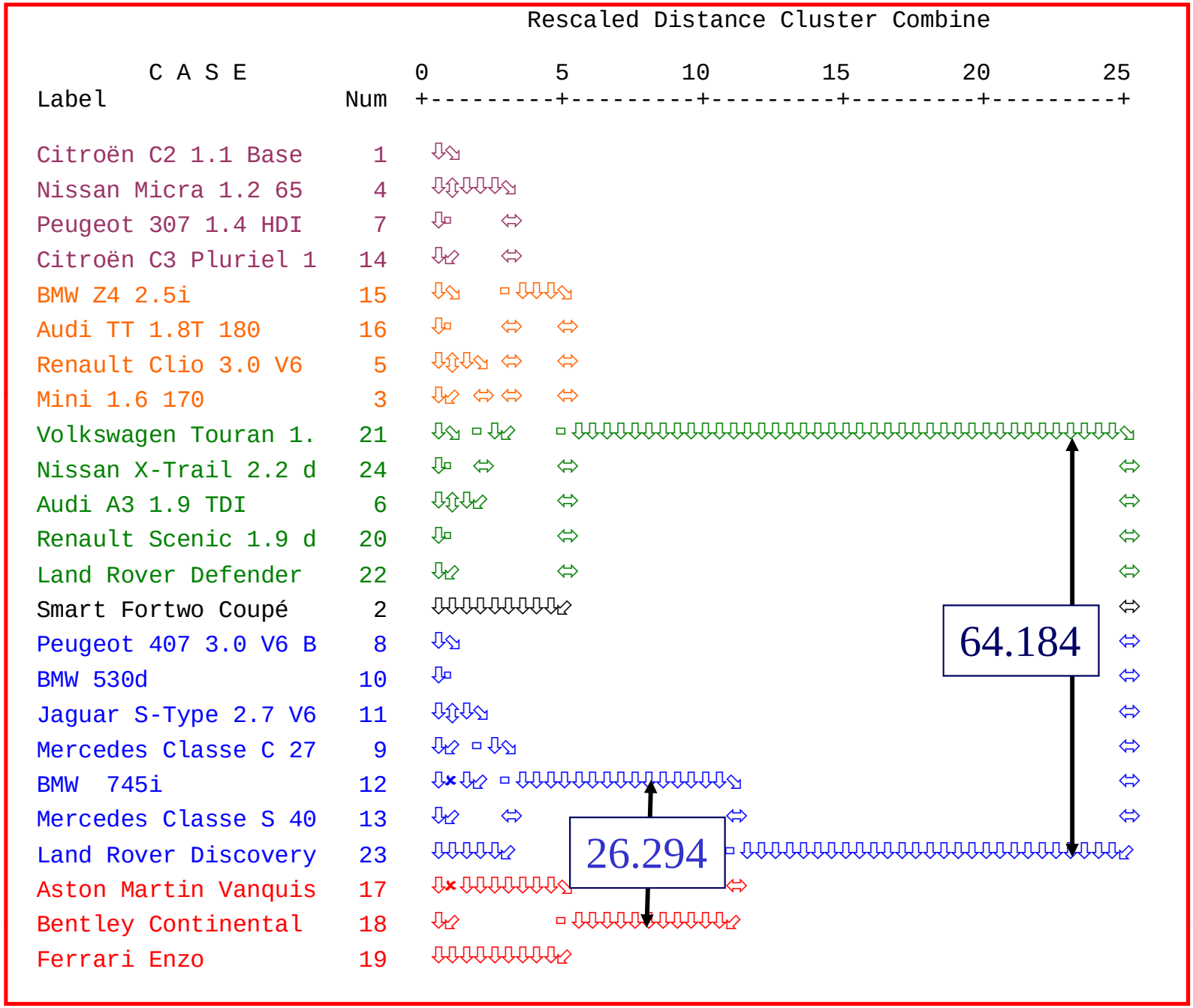
Chaque individu forme une classe. On regroupe les deux individus les plus proches.

Étape courante

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

* * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * *

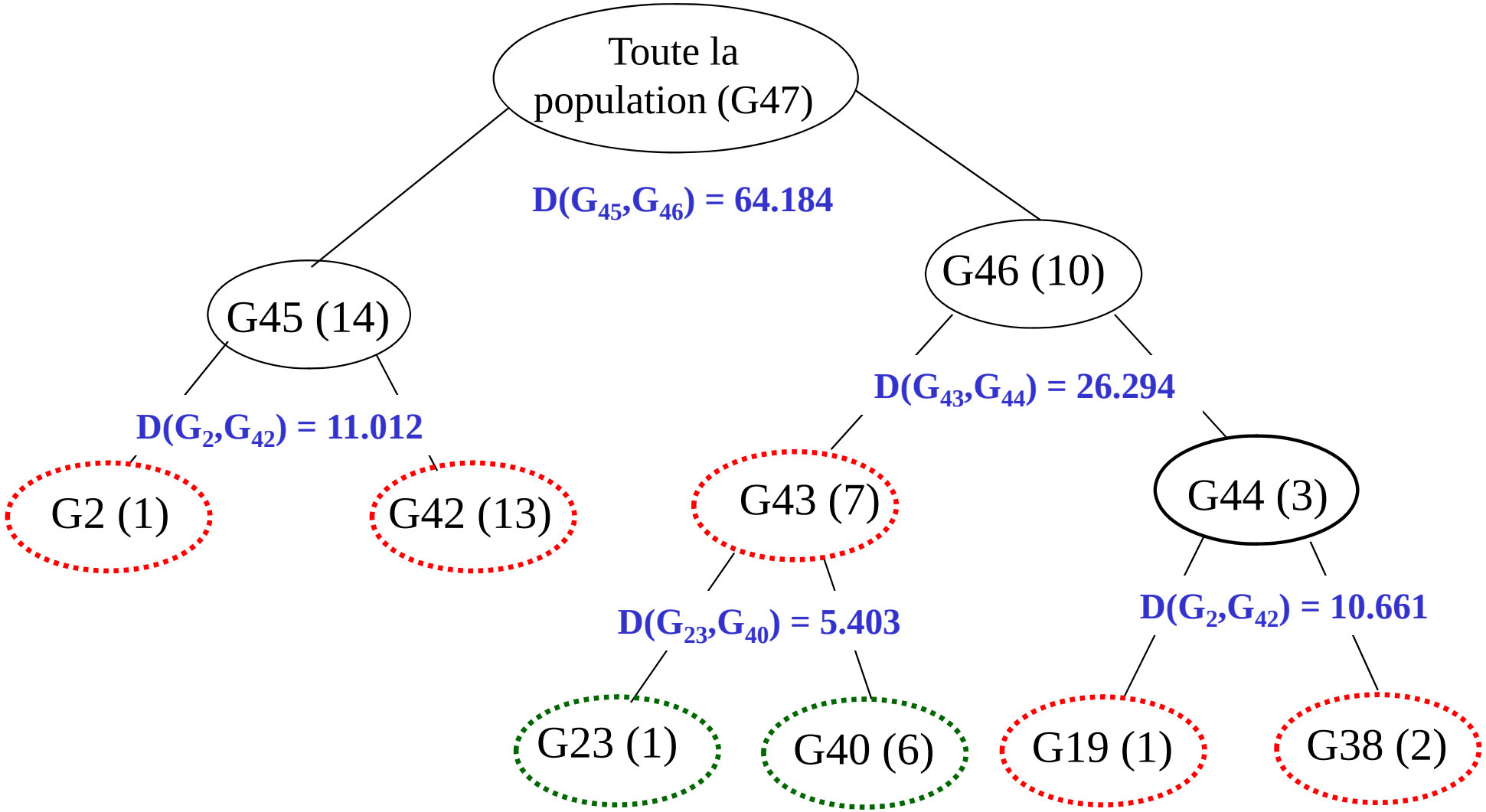
Dendrogram using Ward Method



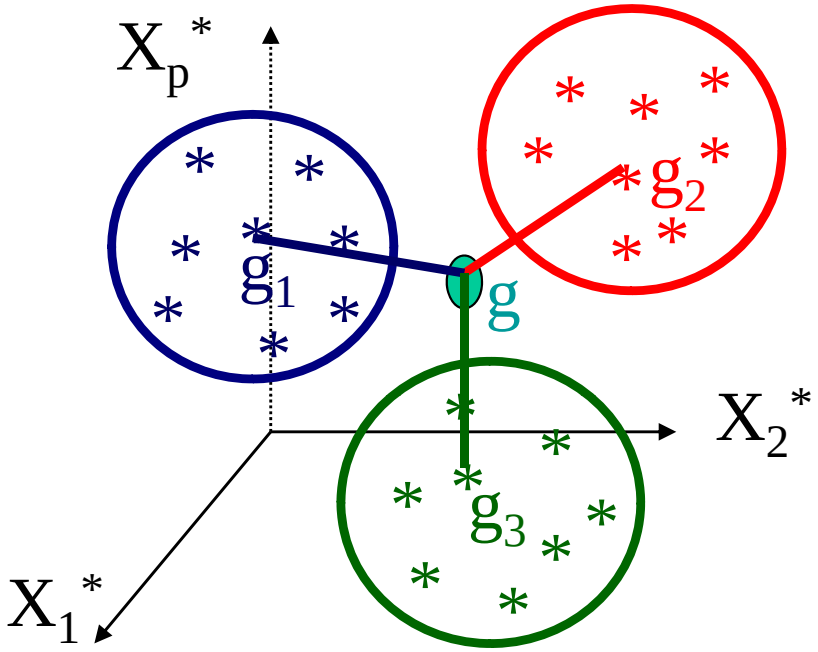
Construction de la classification hiérarchique sur les données centrées-réduites par SPSS

Numéro	Ainé	Benjamin	Nb d'éléments terminaux du noeud	Distance de Ward
25	4	1	2	0.013
26	24	21	2	0.054
27	20	6	2	0.087
28	10	8	2	0.101
29	11	28	3	0.122
30	15	16	2	0.129
31	7	14	2	0.266
32	26	27	4	0.284
33	29	9	4	0.404
34	12	13	2	0.527
35	5	30	3	0.580
36	35	3	4	0.805
37	31	25	4	1.012
38	18	17	2	1.266
39	32	22	5	1.520
40	33	34	6	3.628
41	36	39	9	4.320
42	41	37	13	5.330
43	40	23	7	5.403
44	19	38	3	10.661
45	2	42	14	11.012
46	44	43	10	26.294
47	46	45	24	64.184
Somme des indices de niveau				138.000

Interprétation de la typologie



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 interclasses + Somme des carrés intraclasses

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

Qualité de la typologie en K classes :
(138 - Coeff[n-K])/138

Qualité de 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

Qualité des typologies

Nombre de classes	Somme des carrés intra-classes	Somme des carrés inter-classes	% de Somme des carrés expliquée	Distance de Ward *
24	0	138.00	100.00	
23	0.01	137.99	99.99	0.01
22	0.07	137.93	99.95	0.05
21	0.15	137.85	99.89	0.09
20	0.25	137.75	99.82	0.10
19	0.38	137.62	99.73	0.12
18	0.51	137.49	99.63	0.13
17	0.77	137.23	99.44	0.27
16	1.06	136.94	99.23	0.28
15	1.46	136.54	98.94	0.40
14	1.99	136.01	98.56	0.53
13	2.57	135.43	98.14	0.58
12	3.37	134.63	97.56	0.81
11	4.38	133.62	96.82	1.01
10	5.65	132.35	95.91	1.27
9	7.17	130.83	94.80	1.52
8	10.80	127.20	92.18	3.63
7	15.12	122.88	89.05	4.32
6	20.45	117.55	85.18	5.33
5	25.85	112.15	81.27	5.40
4	36.51	101.49	73.54	10.66
3	47.52	90.48	65.56	11.01
2	73.82	64.18	46.51	26.29
1	138.00	0.00	0.00	64.18

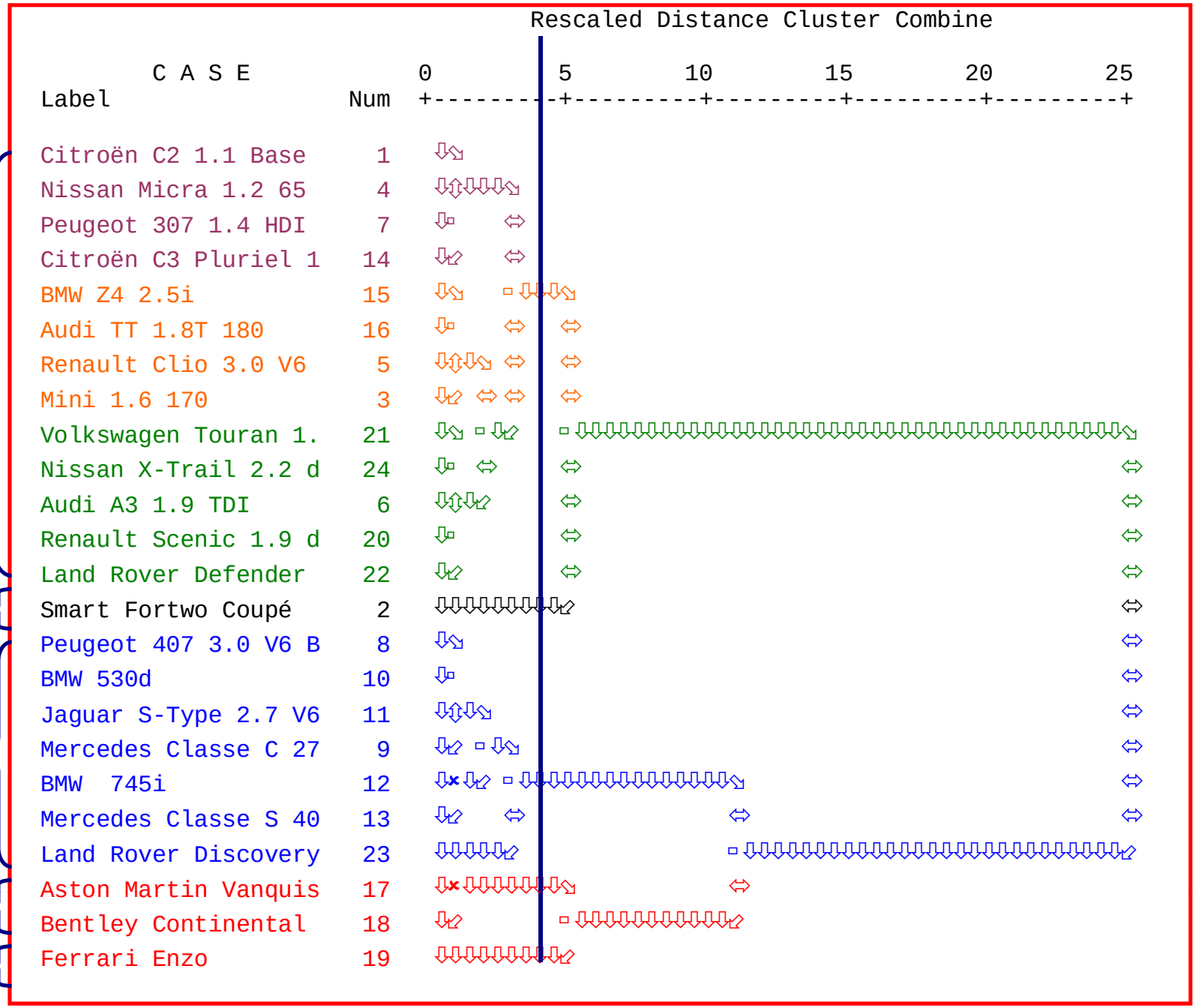
* *distance de Ward entre les groupes fusionnés = Δ (S.C. Intra) = Δ (S.C. Inter)*

Qualité de la typologie en K classes

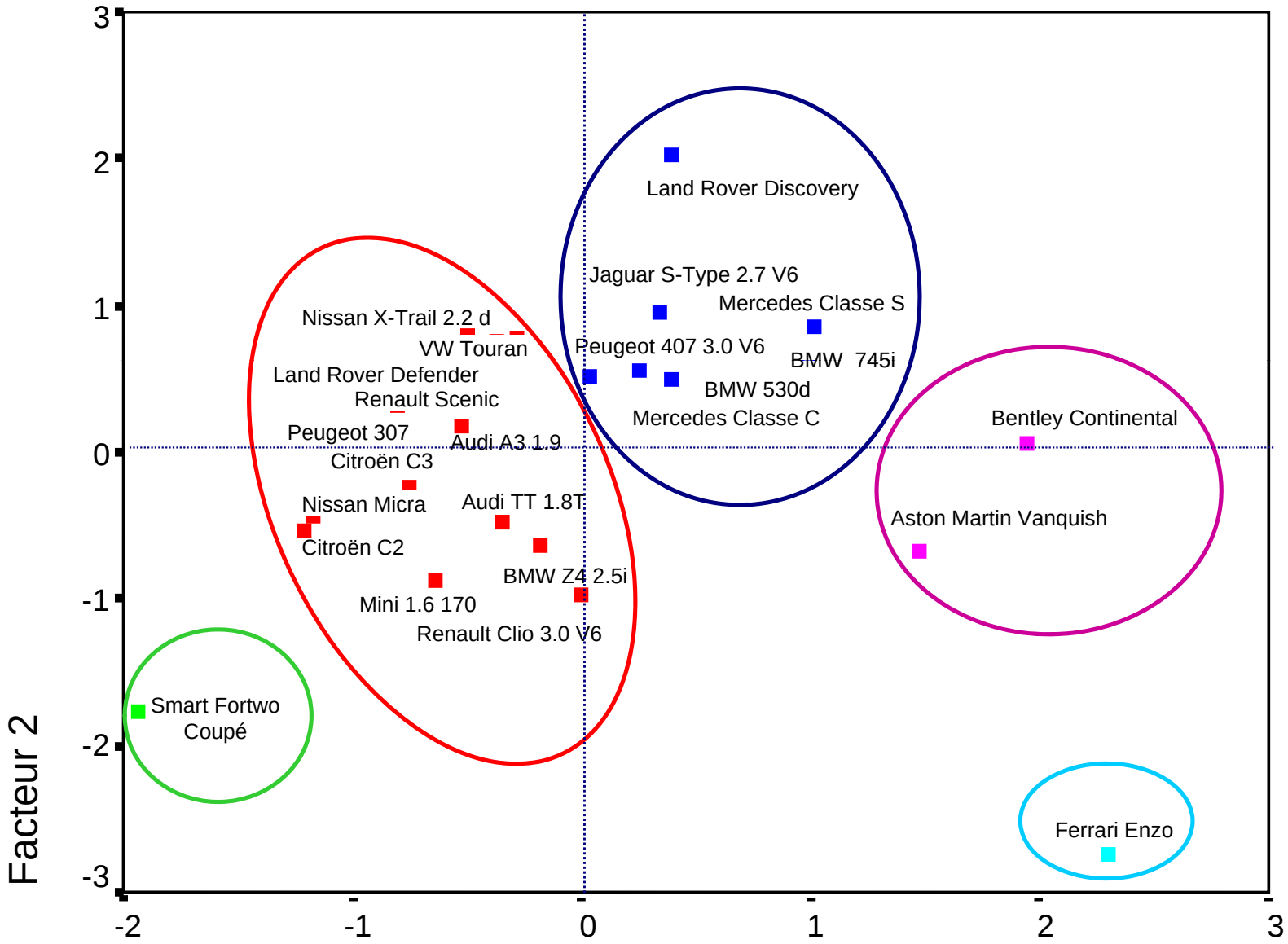
- La somme des carrés expliquée par la typologie en K classes est égale à la somme des carrés inter-classes de la typologie en K classes.
- La qualité de la typologie est mesurée par la proportion de la somme des carrés totale expliquée par la typologie.

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

16. C.A.H. des variables

Les données de Kendall

48 candidats à un certain poste sont évalués sur 15 variables :

(1) Form of letter of application	(9) Experience
(2) Appearance	(10) Drive
(3) Academic ability	(11) Ambition
(4) Likeability	(12) Grasp
(5) Self-confidence	(13) Potential
(6) Lucidity	(14) Keeness to join
(7) Honesty	(15) Suitability
(8) Salesmanship	

Case Summaries

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5
5	6	8	8	8	4	4	9	2	8	5	5	8	8	7	7
6	7	7	7	6	8	7	10	5	9	6	5	8	6	6	6
7	9	9	8	8	8	8	8	8	10	8	10	8	9	8	10
8	9	9	9	8	9	9	8	8	10	9	10	9	9	9	10
9	9	9	7	8	8	8	8	5	9	8	9	8	8	8	10
10	4	7	10	2	10	10	7	10	3	10	10	10	9	3	10
11	4	7	10	0	10	8	3	9	5	9	10	8	10	2	5
12	4	7	10	4	10	10	7	8	2	8	8	10	10	3	7
13	6	9	8	10	5	4	9	4	4	4	5	4	7	6	8
14	8	9	8	9	6	3	8	2	5	2	6	6	7	5	6
15	4	8	8	7	5	4	10	2	7	5	3	6	6	4	6
16	6	9	6	7	8	9	8	9	8	8	7	6	8	6	10
17	8	7	7	7	9	5	8	6	6	7	8	6	6	7	8
18	6	8	8	4	8	8	6	4	3	3	6	7	2	6	4
19	6	7	8	4	7	8	5	4	4	2	6	8	3	5	4
20	4	8	7	8	8	9	10	5	2	6	7	9	8	8	9
21	3	8	6	8	8	8	10	5	3	6	7	8	8	5	8
22	9	8	7	8	9	10	10	10	3	10	8	10	8	10	8
23	7	10	7	9	9	9	10	10	3	9	9	10	9	10	8
24	9	8	7	10	8	10	10	10	2	9	7	9	9	10	8
25	6	9	7	7	4	5	9	3	2	4	4	4	4	5	4
26	7	8	7	8	5	4	8	2	3	4	5	6	5	5	6
27	2	10	7	9	8	9	10	5	3	5	6	7	6	4	5
28	6	3	5	3	5	3	5	0	0	3	3	0	0	5	0
29	4	3	4	3	3	0	0	0	0	4	4	0	0	5	0
30	4	6	5	6	9	4	10	3	1	3	3	2	2	7	3
31	5	5	4	7	8	4	10	3	2	5	5	3	4	8	3
32	3	3	5	7	7	9	10	3	2	5	3	7	5	5	2
33	2	3	5	7	7	9	10	3	2	2	3	6	4	5	2
34	3	4	6	4	3	3	8	1	1	3	3	3	2	5	2
35	6	7	4	3	3	0	9	0	1	0	2	3	1	5	3
36	9	8	5	5	6	6	8	2	2	2	4	5	6	6	3
37	4	9	6	4	10	8	8	9	1	3	9	7	5	3	2
38	4	9	6	6	9	9	7	9	1	2	10	8	5	5	2
39	10	6	9	10	9	10	10	10	10	10	8	10	10	10	10
40	10	6	9	10	9	10	10	10	10	10	10	10	10	10	10
41	10	7	8	0	2	1	2	0	10	2	0	3	0	0	10
42	10	3	8	0	1	1	0	0	10	0	0	0	0	0	10
43	3	4	9	8	2	4	5	3	6	2	1	3	3	3	8
44	7	7	7	6	9	8	8	6	8	8	10	8	8	6	5
45	9	6	10	9	7	7	10	2	1	5	5	7	8	4	5
46	9	8	10	10	7	9	10	3	1	5	7	9	9	4	4
47	0	7	10	3	5	0	10	0	0	2	2	0	0	0	0
48	0	6	10	1	5	0	10	0	0	2	2	0	0	0	0

Tableau des corrélations

Correlation Matrix

	Correlation														
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1.000	.239	.044	.306	.092	.228	-.107	.269	.548	.346	.285	.338	.367	.467	.586
X2	.239	1.000	.123	.380	.431	.371	.354	.477	.141	.341	.550	.506	.507	.284	.384
X3	.044	.123	1.000	.002	.001	.077	-.030	.046	.266	.094	.044	.198	.290	-.323	.140
X4	.306	.380	.002	1.000	.302	.483	.645	.347	.141	.393	.347	.503	.606	.685	.327
X5	.092	.431	.001	.302	1.000	**.808**	.410	**.816**	.015	.704	**.842**	.721	.672	.482	.250
X6	.228	.371	.077	.483	.808	1.000	.356	**.826**	.147	.698	.758	**.883**	.777	.527	.416
X7	-.107	.354	-.030	.645	.410	.356	1.000	.231	-.156	.280	.215	.386	.416	.448	.003
X8	.269	.477	.046	.347	.816	.826	.231	1.000	.233	**.811**	**.860**	.766	.735	.549	.548
X9	.548	.141	.266	.141	.015	.147	-.156	.233	1.000	.337	.195	.299	.348	.215	.693
X10	.346	.341	.094	.393	.704	.698	.280	.811	.337	1.000	.780	.714	.788	.613	.623
X11	.285	.550	.044	.347	.842	.758	.215	.860	.195	.780	1.000	.784	.769	.547	.435
X12	.338	.506	.198	.503	.721	.883	.386	.766	.299	.714	.784	1.000	**.876**	.549	.528
X13	.367	.507	.290	.606	.672	.777	.416	.735	.348	.788	.769	.876	1.000	.539	.574
X14	.467	.284	-.323	.685	.482	.527	.448	.549	.215	.613	.547	.549	.539	1.000	.396
X15	.586	.384	.140	.327	.250	.416	.003	.548	.693	.623	.435	.528	.574	.396	1.000

One of the questions of interest here is how the variables cluster, in the sense that some of the qualities may be correlated or confused in the judge's mind. (There was no purpose in clustering the candidates - only one was to be chosen).

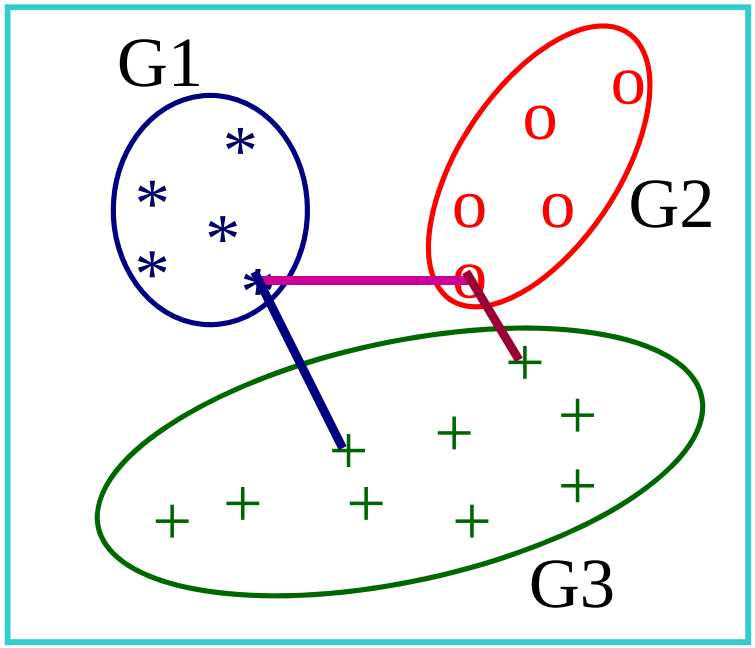
Classification Ascendante Hiérarchique des variables

Méthode des plus proches voisins

A chaque étape, on fusionne les deux groupes G_i et G_j maximisant :

$$\underset{X_a \in G_i, X_b \in G_j}{Max} \left| Cor(X_a, X_b) \right|$$

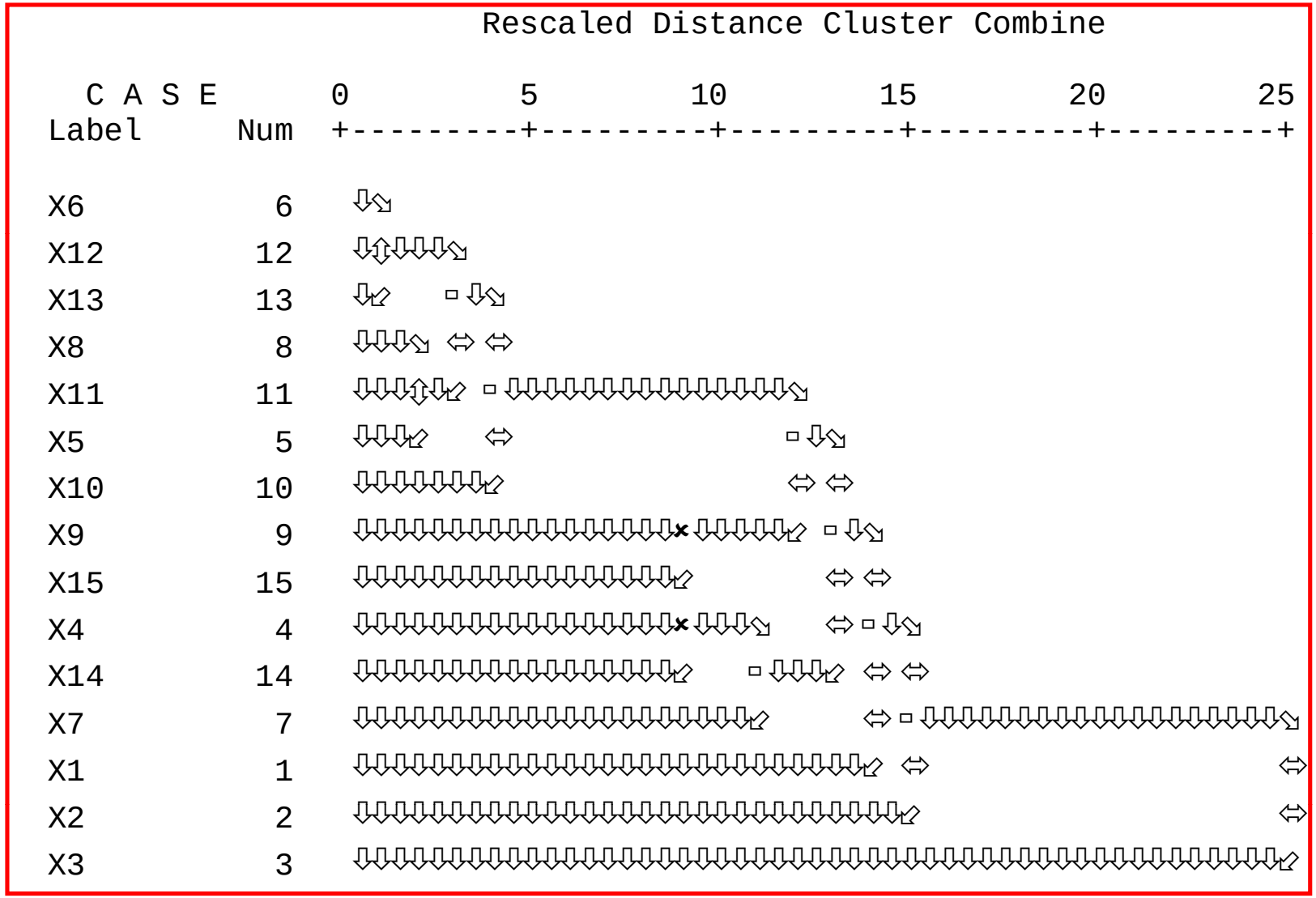
On fusionne G2 et G3.



Classification Ascendante Hiérarchique des variables

*** * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * ***

Dendrogram using Single Linkage (*VOISINS LES PLUS PROCHES*)



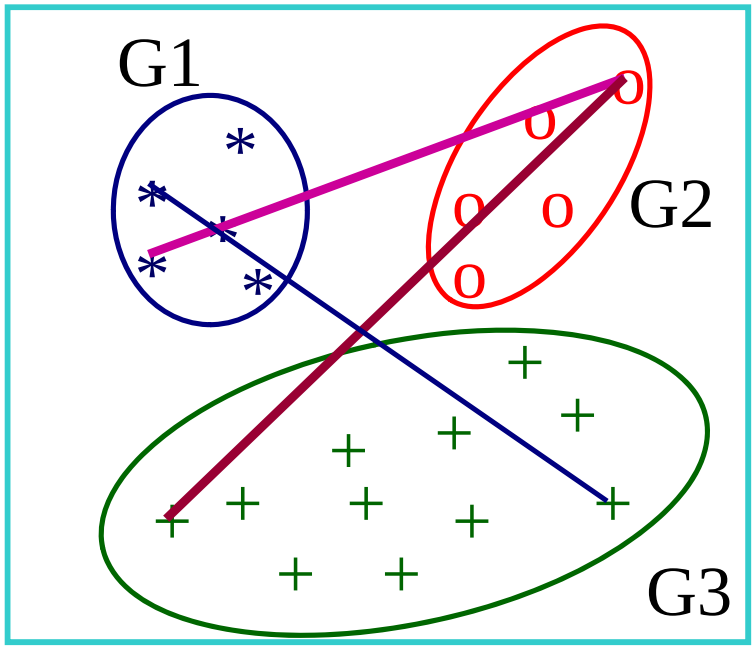
Classification Ascendante Hiérarchique des variables

Méthode des voisins les plus éloignés

A chaque étape, on fusionne les deux groupes G_i et G_j maximisant :

$$\underset{X_a \in G_i, X_b \in G_j}{Min} \left| Cor(X_a, X_b) \right|$$

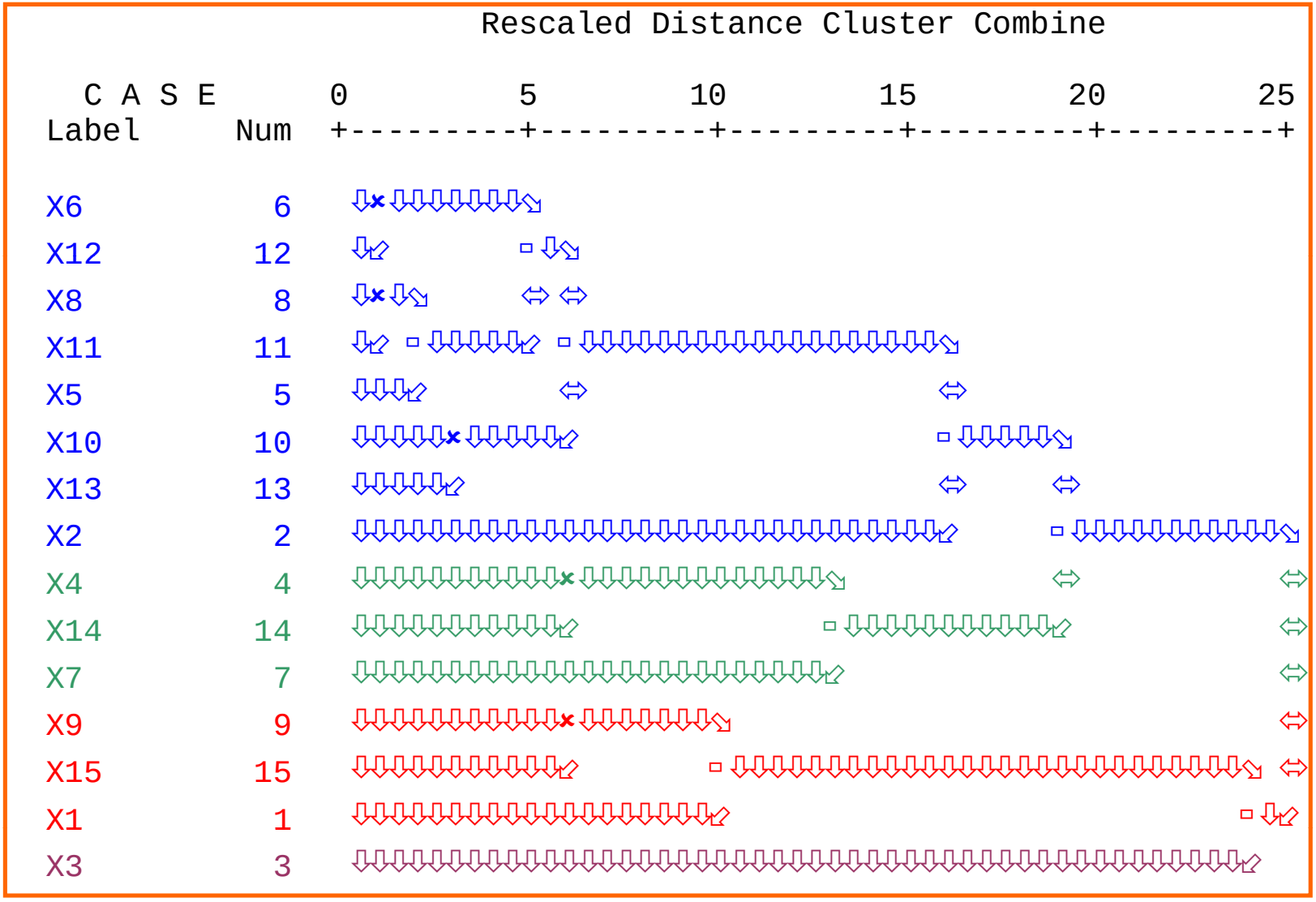
On fusionne G1 et G2.



Classification Ascendante Hiérarchique des variables

*** * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * ***

Dendrogram using Complete Linkage (*VOISINS LES PLUS ELOIGNES*)



Bloc 1

Correlation Matrix

	Correlation							
	X2	X5	X6	X8	X10	X11	X12	X13
X2	1.000	.431	.371	.477	.341	.550	.506	.507
X5	.431	1.000	.808	.816	.704	.842	.721	.672
X6	.371	.808	1.000	.826	.698	.758	.883	.777
X8	.477	.816	.826	1.000	.811	.860	.766	.735
X10	.341	.704	.698	.811	1.000	.780	.714	.788
X11	.550	.842	.758	.860	.780	1.000	.784	.769
X12	.506	.721	.883	.766	.714	.784	1.000	.876
X13	.507	.672	.777	.735	.788	.769	.876	1.000

Les corrélations sont toutes positives.

Bloc 2

Correlation Matrix

	Correlation		
	X4	X7	X14
X4	1.000	.645	.685
X7	.645	1.000	.448
X14	.685	.448	1.000

Bloc 3

Correlation Matrix

	Correlation		
	X1	X9	X15
X1	1.000	.548	.586
X9	.548	1.000	.693
X15	.586	.693	1.000

Interprétation des blocs

Bloc 1 : Qualités humaines favorables au poste

Appearance, Self-confidence, Lucidity, Salesmanship, Drive, Ambition, Grasp, Potential

Bloc 2 : Qualités de franchise et de communication

Likeability, Honesty, Keenness to join

Bloc 3 : Expérience

Form of letter of application, Experience, Suitability

Bloc 4 : Diplôme

Academic ability