

# Cours 7

## ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Master 2 – 2006/2007

1

# 1. Les données

NOMS	PUISS	CYLI	CoupleMaxi	LONG	LARG	HAUT	COFFRE	RESE	POIDS	VITE	CONS
ALF 147 1.9 JTD Distinctive	115	1910	28	4.17	1.73	1.44	280	60	1270	191	5.8
ALF 166 2.5 V6 24V Progression	188	2492	22.5	4.72	1.81	1.42	490	72	1490	225	11.9
ASTMAR D67 Volante	420	5935	65	4.66	1.83	1.26	150	89	1875	285	14
AUD A4 3.0 Quattro Pack	220	2976	30.6	4.55	1.77	1.43	445	66	1515	243	10.5
AUD A8 S8 Pack Avus	360	4172	43.9	5.03	1.88	1.44	525	90	1750	250	14
AUD TT Roadster 1.8 T225 Quattro	225	1781	28.5	4.04	1.76	1.34	270	55	1395	243	9.2
AUDI A4 Cabriolet 2.4	170	2393	23.4	4.57	1.78	1.39	315	70	1600	224	9.7
BEN Continental T	426	6750	88.2	5.22	1.95	1.45	350	100	2450	245	18.7
BMW 316i	115	1796	17.8	4.47	1.74	1.41	440	63	1310	206	7.1
BMW X5 3.0d Pack Luxe	183	2926	41.8	4.67	1.87	1.71	465	93	2095	200	9.7
BMW Z8	400	4941	51	4.4	1.83	1.32	203	73	1585	250	14.5
CAD Seville STS	305	4565	40.8	4.99	1.9	1.43	445	70	1857	241	14.1
CHR Grand Voyager 2.5 CRD Limited	140	2500	31.8	5.09	2	1.75	580	75	2000	165	7.7
CHR PT Cruiser 2.0 Classic	140	1995	19	4.29	1.7	1.6	520	57	1412	170	7.8
CIT Berlingo 1.6i 16V SX	110	1587	15.3	4.11	1.72	1.8	664	55	1252	172	7.4
CIT C3 1.4 HDI 70 ch SXPack Clim	70	1398	15.3	3.85	1.67	1.52	305	45	1022	165	4.2
CIT Picasso 1.6i SX	90	1587	14	4.28	1.75	1.64	515	55	1240	168	7.8
CIT Saxo 1.1i Bic 2	60	1124	9.1	3.72	1.59	1.37	280	45	805	162	6.7
CIT Xsara 2.0 Hdi 110 ch Exclusive	110	1997	26	4.19	1.7	1.4	408	54	1210	191	5.2
DAE Leganza 2.0 CDX	133	1998	19.8	4.67	1.78	1.44	560	65	1403	206	9.2
DAI Sirion 1.3x	102	1298	12.2	3.67	1.59	1.45	235	40	850	180	5.7
FIA Multipla JTD 115 ELX	115	1910	20.7	3.99	1.87	1.67	430	63	1370	176	6.4
FIA Seicento S	55	1108	8.9	3.32	1.51	1.44	170	35	750	150	5.8
FIA Sileo 1.9 JTD 80 ch Active	80	1910	20	4.25	1.76	1.52	335	58	1305	170	5.5
FOR Fiesta 1.4 TDCI Ghia	68	1398	16.3	3.92	1.68	1.42	284	45	1065	164	4.3
FOR Focus ST 170	172	1989	20	4.17	1.7	1.43	390	55	1283	216	9.1
FOR Ka 1.3 Original	60	1299	10.7	3.62	1.63	1.37	185	42	890	155	6.3
FOR Mondeo 2.0 DTCl 130 ch Ghia	130	1998	33.7	4.73	1.81	1.43	500	56	1501	200	5.9
FOR Puma 1.6	103	1596	14.8	3.98	1.67	1.34	240	42	1035	190	7.3
HON Accord 2.3iES	152	2254	21	4.59	1.75	1.43	427	65	1423	212	8.7

NOMS  
ALF147 1.9 JTD Distinctive  
ALF166 2.5 V6 24V Progression  
ASTMAR D67 Volante  
AUD A4 3.0 Quattro Pack

+ encore 63 modèles de voitures

2

# 2. Résumé des données

Sortie SPAD

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES

EFFECTIF TOTAL : 91 POIDS TOTAL : 91.00

NUM .	IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
1 .	C2 - PUISS	91	91.00	147.81	90.78	45.00	476.00
2 .	C3 - CYLI	91	91.00	2253.71	1103.69	599.00	6750.00
3 .	C4 - CoupleMaxi	91	91.00	28.49	31.81	7.10	299.00
4 .	C5 - LONG	91	91.00	4.33	0.47	2.50	5.39
5 .	C6 - LARG	91	91.00	1.75	0.09	1.51	2.00
6 .	C7 - HAUT	91	91.00	1.48	0.13	1.14	1.86
7 .	C8 - COFFRE	91	91.00	379.26	144.16	110.00	900.00
8 .	C9 - RESE	91	91.00	60.81	15.33	22.00	100.00
9 .	C10 - POIDS	91	91.00	1382.49	380.01	680.00	2450.00
10 .	C11 - VITE	91	91.00	195.69	30.66	135.00	305.00
11 .	C12 - CONS	91	91.00	7.95	2.94	4.20	18.70
12 .	C13 - PRIX	91	91.00	36117.64	48234.54	7290.00	342798.00

3

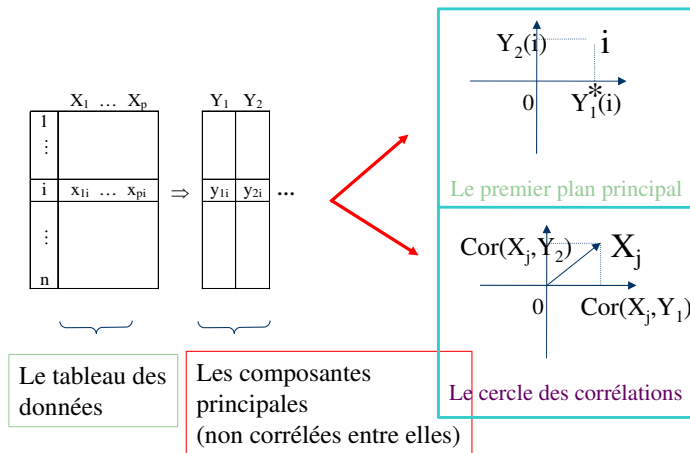
# 3. Tableau des corrélations

	PUISS	CYLI	CoupleMaxi	LONG	LARG	HAUT	COFFRE	RESE	POIDS	VITE	CONS	PRIX
PUISS	1	.925*	.491*	.515*	.550*	-.086	.022	.714*	.882*	.869*	.795*	
CYLI		1	.467*	.675*	.732*	.024	.180	.796*	.824*	.724*	.867*	.873*
CoupleMaxi			1	.411*	.404*	-.018	.104	.406*	.434*	.448*	.393*	.375*
LONG				1	.857*	.195	.635*	.863*	.846*	.621*	.526*	.507*
LARG					1	.333*	.620*	.898*	.895*	.592*	.633*	.570*
HAUT						1	.522*	.342*	.427*	-.294*	.058	.055
COFFRE							1	.544*	.549*	.137	.104	.026
RESE								1	.934*	.606*	.697*	.638*
POIDS									1	.583*	.706*	.655*
VITE										1	.639*	.554*
CONS											1	.788*
PRIX												1

\*\* Correlation is significant at the 0.01 level (2-tailed).

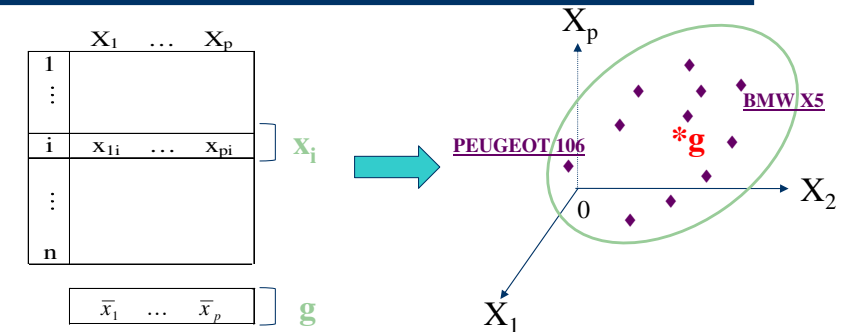
4

## 4. Visualisation des données



5

## 5. Le nuage de points associé aux données

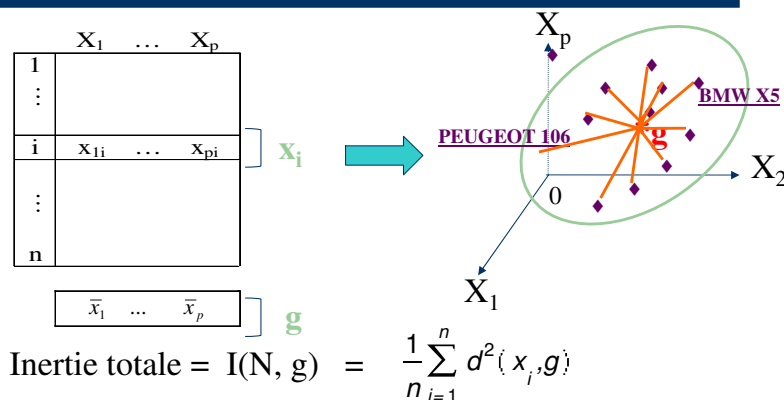


$N = \{x_1, \dots, x_i, \dots, x_n\} =$  Nuage de points associé aux données

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

6

## 6. Inertie totale du nuage de points



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

7

## 7. Réduction des données

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

$$X_1^* = \frac{X_1 - \bar{x}_1}{s_1}$$

$$\vdots$$

$$X_p^* = \frac{X_p - \bar{x}_p}{s_p}$$

8

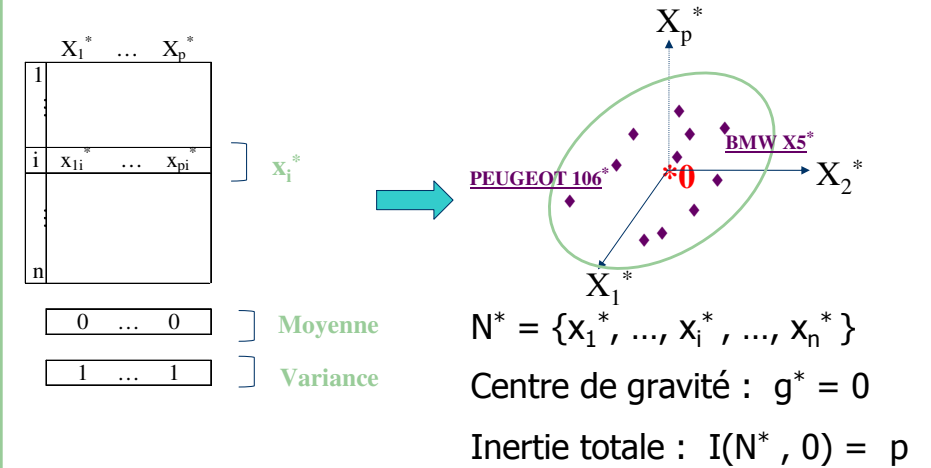
de moyenne 0 et d'écart-type 1.

## Les données centrées-réduites

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
	PUI	CYL	coup	LONG	LA	HA	OF	RES	POI	VIT	COI	PRI	CAF	FR	EQ	
ALF 1	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	-0,	
ALF 1	0,4	0,2	-0,1	0,8	0,7	-0,	0,7	0,7	0,2	0,9	1,3	-0,	0,	2	2	
ASTM	2,9	3,3	0,8	0,7	0,9	-1,	-1,	1,8	1,2	2,2	2,0	2,2	1	2	4	
AUD	0,7	0,6	0,0	0,4	0,2	-0,	0,4	0,3	0,3	1,5	0,8	0,0	1	2	3	
AUD	2,3	1,7	0,4	1,4	1,4	-0,	1,0	1,8	0,9	1,7	2,0	1,2	1	2	3	
AUD	0,8	-0,4	0,0	-0,6	0,1	-1,	-0,	-0,	0,0	1,5	0,4	0,0	1	2	4	
AUDI	0,2	0,1	-0,1	0,5	0,3	-0,	-0,	0,6	0,5	0,9	0,5	0,0	1	2	2	
BENC	3,0	4,0	1,8	1,8	2,2	-0,	-0,	2,5	2,7	1,6	3,6	6,3	1	2	3	
BMW	-0,	-0,4	-0,3	0,3	-0,	-0,	0,4	0,1	-0,	0,3	-0,	-0,1	1	2	1	
BMW	0,3	0,6	0,4	0,7	1,3	1,8	0,5	2,0	1,8	0,1	0,5	0,2	2	2	3	
BMW	2,7	2,4	0,7	0,1	0,9	-1,	-1,	0,7	0,5	1,7	2,2	1,9	1	2	4	
CADS	1,7	2,0	0,3	1,4	1,6	-0,	0,4	0,6	1,2	1,4	2,0	0,6	1	3	3	
CHRC	-0,	0,2	0,1	1,6	2,7	2,1	1,3	0,9	1,6	-0,	-0,	0,1	2	3	3	
CHRF	-0,	-0,2	-0,3	-0,6	-0,	0,	0,9	-0,	0,0	-0,	-0,	-0,3	1	3	2	
CITBe	-0,	-0,6	-0,4	-0,4	-0,	2,	5	1,9	-0,	-0,	-0,	-0,	-0,4	1	1	2
CITC3	-0,	-0,7	-0,4	-1,6	-0,	0,	3	-0,	-1,	-0,	-1,	-1,	-0,4	2	1	2
CITPi	-0,	-0,6	-0,4	-0,	0,	1,	2	0,9	-0,	-0,	-0,	-0,	-0,4	1	1	1
CITSa	-0,	-1,0	-0,6	-1,4	-1,	-0,	-0,	-0,	-1,	-1,	-1,	-0,	-0,5	1	1	1
CITX3	-0,	-0,2	-0,	-0,	-0,	0,	2	-0,	-0,	-0,	-0,	-0,	2	1	3	

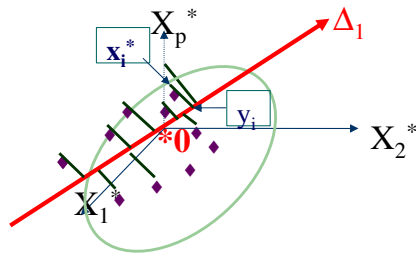
9

## 8. Le nuage de points associé aux données réduites



10

## 9. Premier axe principal $\Delta_1$



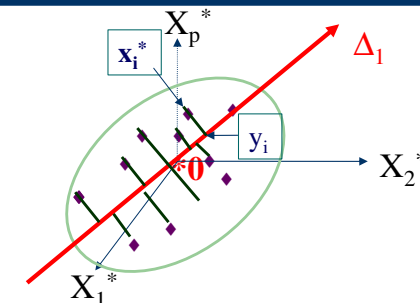
**Objectif 1 :** On cherche l'axe  $\Delta_1$  passant le mieux possible au milieu du nuage  $N^*$ .

On cherche à minimiser l'inertie du nuage  $N^*$  par rapport à l'axe  $\Delta_1$  :

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

11

## Premier axe principal $\Delta_1$ (suite)



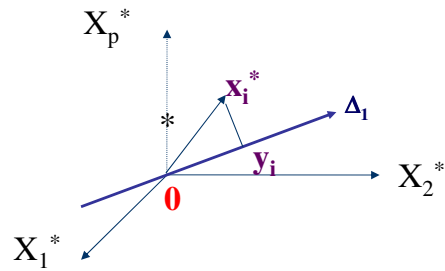
**Objectif 2 :** On cherche l'axe d'allongement  $\Delta_1$  du nuage  $N^*$ .

On cherche à maximiser l'inertie du nuage  $N^*$  projeté sur l'axe  $\Delta_1$  :

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

12

## Les objectifs 1 et 2 sont atteints simultanément



De :  
on déduit :

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

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

Inertie totale = p = **Inertie expliquée par  $\Delta_1$**  + Inertie résiduelle

Maximiser                      Minimiser

13

## 1<sup>er</sup> axe principal $\Delta_1$ : Résultats

- L'axe  $\Delta_1$  passe par le centre de gravité 0 du nuage de points  $N^*$ .
- L'axe  $\Delta_1$  est engendré par le vecteur normé  $u_1$ , vecteur propre de la matrice des corrélations R associé à la plus grande valeur propre  $\lambda_1$ .
- L'inertie du nuage projeté est égal à  $\lambda_1$ .
- La part d'inertie expliquée par le premier axe principal  $\Delta_1$  est égale à  $\lambda_1/p$ .

14

## Résultats SPAD

```

VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 11.0000
                                         SOMME DES VALEURS PROPRES .... 11.0000
HISTOGRAMME DES 11 PREMIERES VALEURS PROPRES
+-----+-----+-----+-----+
| NUMERO | VALEUR | POURCENT. | POURCENT. |
|   |   PROPRE   |   |   CUMULE   |
+-----+-----+-----+-----+
| 1 | 6.6969 | 60.88 | 60.88 | *****
| 2 | 2.0236 | 18.40 | 79.28 | *****
| 3 | 0.7451 | 6.77 | 86.05 | *****
| 4 | 0.6926 | 6.30 | 92.35 | *****
| 5 | 0.2839 | 2.58 | 94.93 | ****
| 6 | 0.2013 | 1.83 | 96.76 | ***
| 7 | 0.1300 | 1.18 | 97.94 | **
| 8 | 0.0893 | 0.81 | 98.75 | **
| 9 | 0.0757 | 0.69 | 99.44 | *
| 10 | 0.0385 | 0.35 | 99.79 | *
| 11 | 0.0230 | 0.21 | 100.00 | *
+-----+-----+-----+-----+
    
```

15

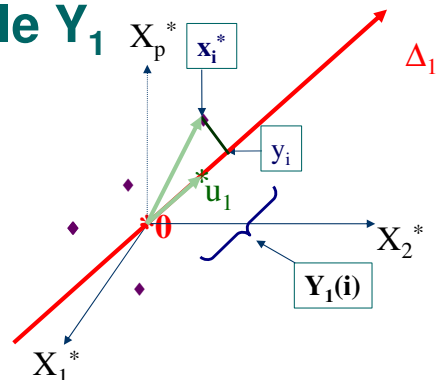
## Résultats SPAD

Direction du vecteur propre associé à la plus grande valeur propre :

-0.86  
-0.91  
-0.54  
-0.87  
-0.90  
-0.21  
-0.46  
-0.94  
-0.94  
-0.76  
-0.81

16

## 10. Première composante principale $Y_1$



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

$Y_1(i)$  = coordonnée de  $y_i$  sur l'axe  $\Delta_1$   
 = produit scalaire entre les vecteurs  $x_i^*$  et  $u_1$

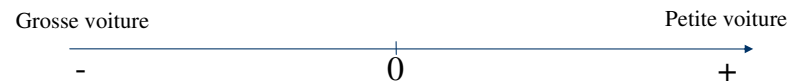
$$= \sum_{j=1}^p u_{1j} x_{ji} \quad \longrightarrow \quad Y_1 = \sum_{j=1}^p u_{1j} X_j$$

17

## Interprétation de la première composante principale $Y_1$

$Y_1 =$

- 0.86 PUISS
- 0.91 CYLI
- 0.54 Couplemaxi
- 0.87 LONG
- 0.90 LARG
- 0.21 HAUT
- 0.46 COFFRE
- 0.94 RESE
- 0.94 POIDS
- 0.76 VITE
- 0.81 CONS



18

## Résultats SPAD

```

COORDONNEES DES INDIVIDUS
AXE 1
-----+-----
|          INDIVIDUS          | COORDONNEES |
|-----+-----|
| IDENTIFICATEUR | P.REL | DISTO |
|-----+-----|
| ALF 147 1,9 JTD Distinct | 1.10 | 1.59 | 0.95 |
| ALF 166 2,5 V6 24V Progr | 1.10 | 5.61 | -1.88 |
| ASTMAR DB7 Volante       | 1.10 | 42.11 | -4.92 |
| AUD A4 3,0 Quattro Pack  | 1.10 | 5.09 | -1.79 |
| AUD A8 S8 Pack Avus      | 1.10 | 26.11 | -4.86 |
| AUD TT Roadster 1,8 T225 | 1.10 | 5.83 | -0.22 |
| AUDIA4 Cabriolet 2,4     | 1.10 | 3.11 | -1.14 |
| BEN Continental T        | 1.10 | 68.44 | -7.76 |
| BMW 316i                 | 1.10 | 1.25 | 0.23 |
| BMW X5 3,0d Pack Luxe    | 1.10 | 14.90 | -3.06 |
| BMW Z8                   | 1.10 | 27.12 | -3.68 |
| CAD Seville STS         | 1.10 | 21.14 | -4.26 |
| CHR Grand Voyager 2,5 CR | 1.10 | 20.40 | -2.80 |
    
```

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

19

## Propriétés de la première composante principale $Y_1$

- Moyenne de  $Y_1 = 0$
- Variance de  $Y_1 = \frac{1}{n} \sum_{i=1}^n Y_1(i)^2 = \frac{1}{n} \sum_{i=1}^n d^2(y_i, 0) = \lambda_1$
- $\text{Cor}(X_j, Y_1) = \sqrt{\lambda_1} u_{1j}$
- $\frac{1}{p} \sum_{j=1}^p \text{cor}^2(X_j, Y_1) = \frac{\lambda_1}{p}$  est maximum

20

## Qualité de la première composante principale $Y_1$

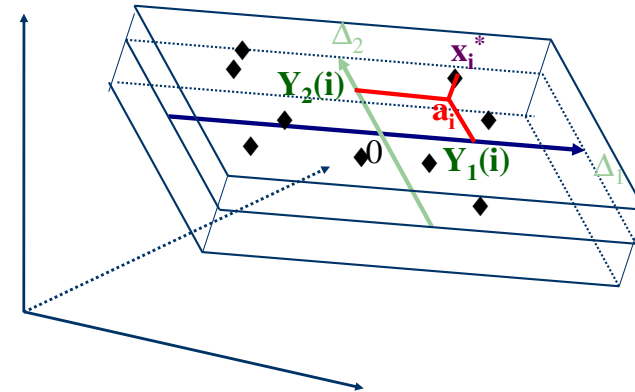
- Inertie totale = 11
- Inertie expliquée par le premier axe principal =  $\lambda_1 = 6.69$
- Part d'inertie expliquée par le premier axe principal :

$$\frac{\lambda_1}{p} = \frac{6.69}{11} = 60.88$$

- La première composante principale explique 60.88 % de la variance totale.

21

## 11. Deuxième axe principal $\Delta_2$



22

## 2<sup>ème</sup> axe principal $\Delta_2$ : Résultats

- On recherche le deuxième axe principal  $\Delta_2$  orthogonal à  $\Delta_1$  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  $\lambda_2$ .
- 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  $\lambda_2$ , et non corrélée à la première composante principale  $Y_1$ .

23

## Résultats SPAD

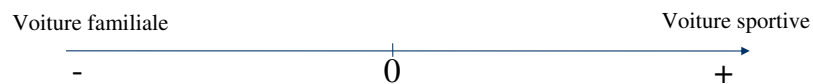
COORDONNEES DES VARIABLES SUR LES AXES 1 A 5  
VARIABLES ACTIVES

IDEN - LIBELLE COURT	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR				
	1	2	3	4	5	1	2	3	4	5
C2 - PUISS	-0.86	0.43	-0.15	-0.01	0.15	-0.86	0.43	-0.15	-0.01	0.15
C3 - CYLI	-0.91	0.26	-0.20	-0.04	-0.01	-0.91	0.26	-0.20	-0.04	-0.01
C4 - CoupleMaxi	-0.54	0.22	0.53	-0.61	0.02	-0.54	0.22	0.53	-0.61	0.02
C5 - LONG	-0.87	-0.23	0.26	0.23	-0.15	-0.87	-0.23	0.26	0.23	-0.15
C6 - LARG	-0.90	-0.22	0.02	0.06	-0.25	-0.90	-0.22	0.02	0.06	-0.25
C7 - HAUT	-0.21	-0.83	-0.30	-0.35	0.08	-0.21	-0.83	-0.30	-0.35	0.08
C8 - COFFRE	-0.46	-0.73	0.26	0.21	0.33	-0.46	-0.73	0.26	0.21	0.33
C9 - RESE	-0.94	-0.20	-0.04	0.05	-0.12	-0.94	-0.20	-0.04	0.05	-0.12
C10 - POIDS	-0.94	-0.23	-0.09	-0.05	-0.07	-0.94	-0.23	-0.09	-0.05	-0.07
C11 - VITE	-0.76	0.45	0.18	0.26	0.16	-0.76	0.45	0.18	0.26	0.16
C12 - CONS	-0.81	0.29	-0.37	-0.12	0.13	-0.81	0.29	-0.37	-0.12	0.13

24

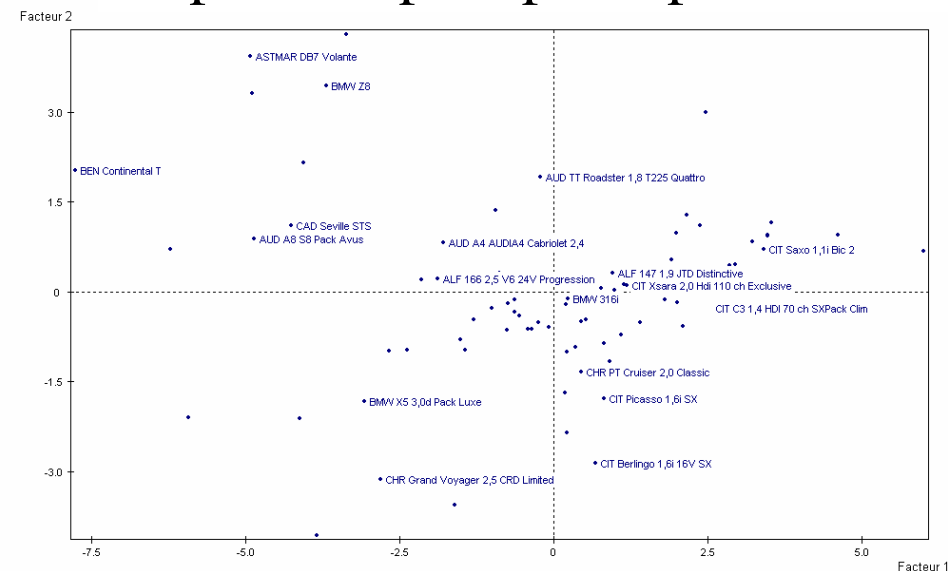
## Interprétation de la deuxième composante principale $Y_2$

$$Y_2 = \begin{array}{l} 0.43 \text{ PUISS} \\ +0.26 \text{ CYLI} \\ +0.22 \text{ Couplemaxi} \\ -0.23 \text{ LONG} \\ -0.22 \text{ LARG} \\ -0.83 \text{ HAUT} \\ -0.73 \text{ COFFRE} \\ -0.20 \text{ RESE} \\ -0.23 \text{ POIDS} \\ +0.45 \text{ VITE} \\ +0.29 \text{ CONS} \end{array}$$



25

## 12. Exemple Auto 2002 Le premier plan principal



## 13. 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 =  $\frac{\lambda_1 + \lambda_2}{p}$

28

Et ainsi de suite pour les autres dimensions...

## Le cercle des corrélations

