Analyse de Données

Par: Houda Benbrahim

Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- Chapitre II: Régression Linéaire Simple et Multiple
- Chapitre III: Analyse Factorielle: ACP
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques

Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- Chapitre II: Régression Linéaire Simple et Multiple
- **Chapitre III: Analyse Factorielle: ACP**
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques

Chapitre III: Analyse Factorielle: Analyse en Composantes Principales

- Section I: Introduction et Définition
- Section II: Rappel: Algèbre Linéaire
- Section III: ACP
- Section IV: ACP sous SPSS
- Section V: Récapitulatif et Conclusion

Section I: Introduction et Définition

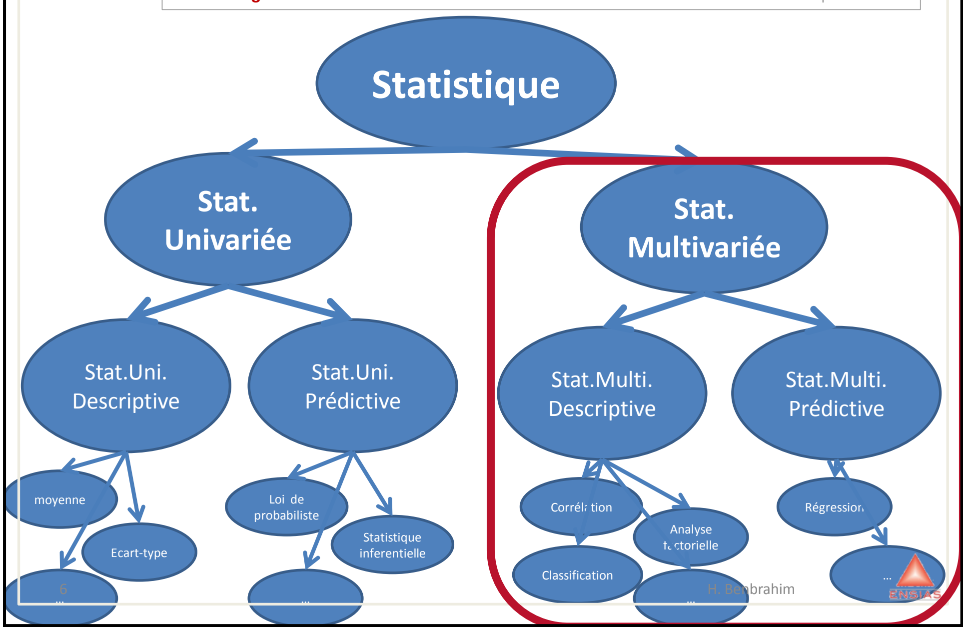
Section II: Rappel: Algèbre Linéaire

**Chapitre III:
Analyse Factorielle: Analyse en Composantes Principales**

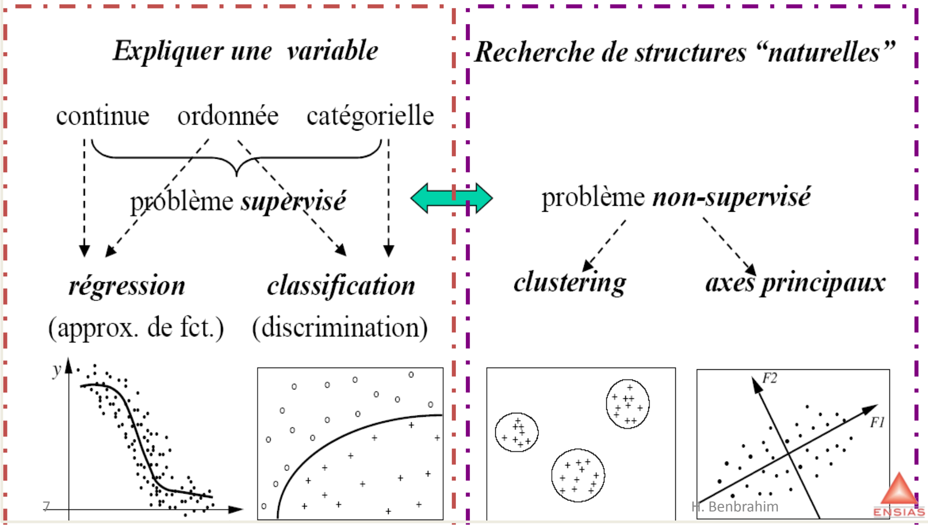
Section III: ACP

Section IV: ACP sous SPSS

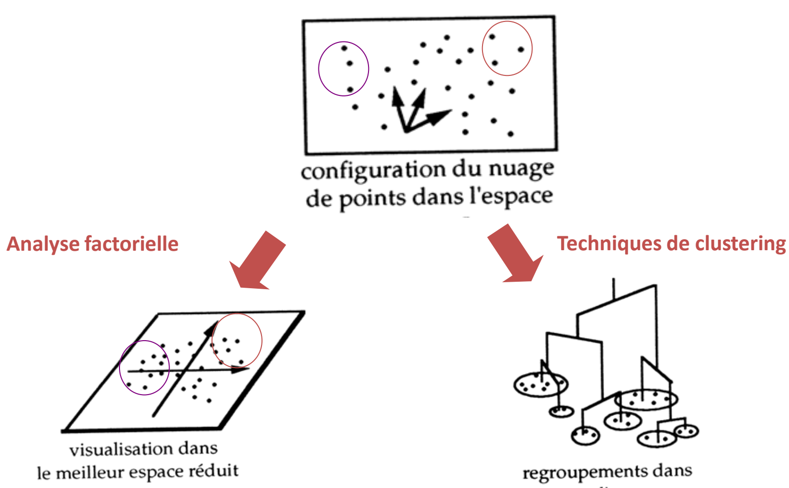
Section V: Récapitulatif et Conclusion



Contexte générale de l'analyse multidimensionnelle des données



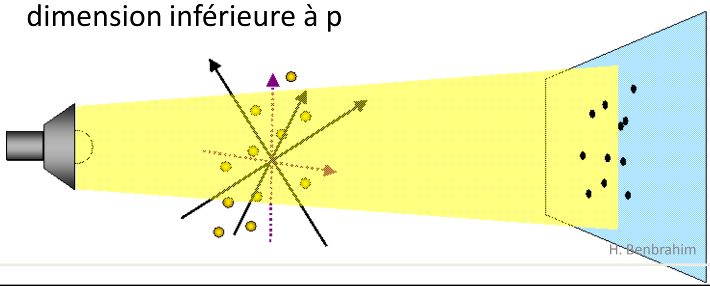
Recherche de structures naturelles



Analyse factorielle

- BUT?
  ➔ Synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables).

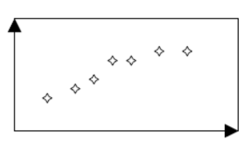
- COMMENT RESUMER?
  ➔ Projeter ces points d'individus dans un espace de dimension inférieure à p



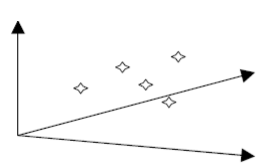
Analyse factorielle: Représentation Graphique

- POURQUOI PROJETER?

Lorsqu'il n'y a que deux dimensions (largeur et longueur par ex.), 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.

Ex. Projeter la réalité sur un plan

- 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.

➔ Le photographe cherche le meilleur angle de vue pour transcrire en dimension 2 (le plan de sa photo) une scène située en dimension 3 (notre espace ambiant).

L' Analyse Factorielle

- Lorsqu'on projette les données (n individus, p variables) sur un plan, on obtient un graphique déformé de la réalité.

➔Le rôle de l' analyse factorielle est de:

- ➢ trouver des espaces de dimensions plus petites minimisant ces déformations.
- ➢ obtenir le résumé le plus pertinent des données initiales.

L' Analyse Factorielle

- Données:
  - Quantitatives → Analyse en Composantes Principales (ACP)
  - 2 Qualitative → Analyse Factorielle des Correspondances (AFC)
  - n Qualitatives → Analyse des correspondances multiples (ACM)

Analyse en Composantes Principales (ACP)

- 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 quantitatives.

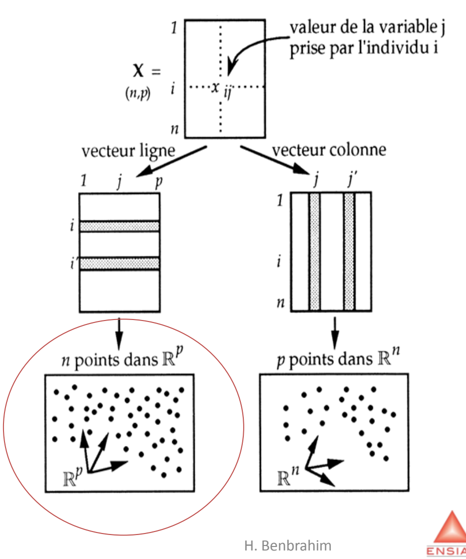
Analyse en Composantes Principales

- Exemple:

Considérons l'ensemble des notes des élèves de l'ENSIAS durant une année. Le nombre d'élèves est environ de 240, et nous pouvons considérer qu'ils obtiennent environ 32 notes chacun. Ainsi le tableau représentant l'ensemble des notes est constitué de 7680 valeurs. Une ACP permet de réduire ce nombre à 270 valeurs sans perte d'information.

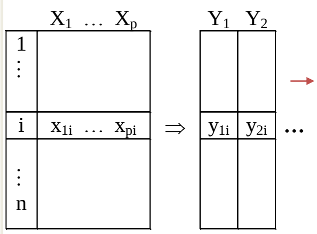
Visualisation des données

- Une analyse factorielle sert à résumer et à hiérarchiser l'information contenue dans un tableau de n lignes (les individus) et p colonne (les variables).
- Les n individus sont décrits par un nuage de p variables.
- L'information représentée par ce nuage correspond à la dispersion des n points.

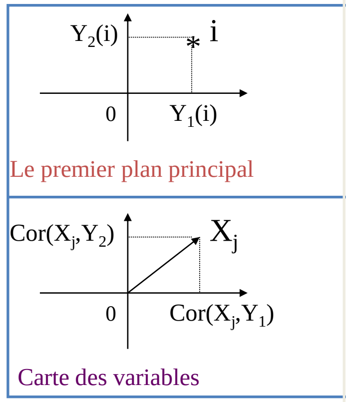


Visualisation des données

Résumer l'information ?
projeter ces points d'individus dans un espace de dimension inférieure à p. Les axes de ce sous-espace sont dits "axes factoriels" ou "facteurs".



Composantes principales
$$Y_h = \sum_{j=1}^{p} u_{hj} X_j$$
non corrélées entre elles



Le résumé est possible dans le cas où les variables ne sont pas totalement indépendantes

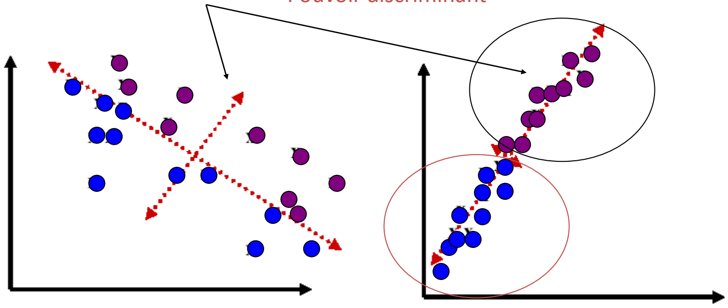
Chaque variable « **X**" porte en elle :
- une part d'information originale ;
- une part d'information redondante avec les autres.

C'est cette **part d'information redondante** que l'on va **regrouper** dans le **résumé factoriel**.

- Chaque **facteur** est la **combinaison linéaire** des "**p**" variables.

Visualisation des données

But:
- Eliminer la redondance entre les caractéristiques
- Eliminer les caractéristiques non-significatives.

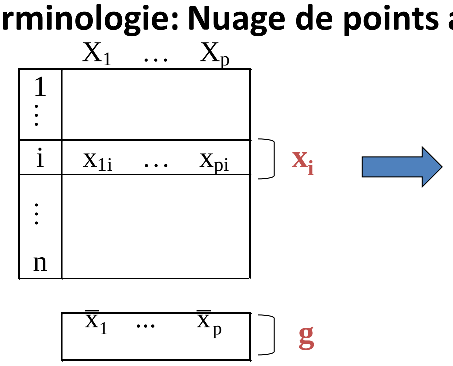


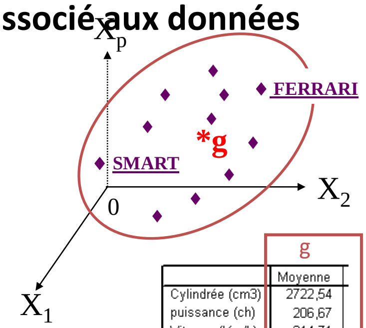
Un exemple de positionnement de produits

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

Terminologie: Nuage de points associé aux données





g

	Moyenne
Cylindrée (cm3)	2722,54
puissance (ch)	206,67
Vitesse (Km/h)	214,71
Poids (Kg)	1486,58
Largeur (mm)	1838,42
Longeur (mm)	4277,83

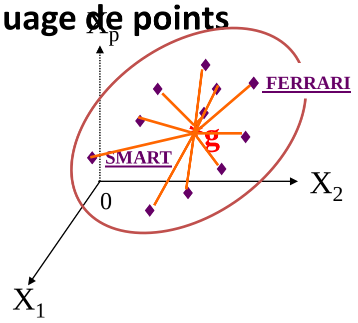
N = $\{x_1, \ldots, x_i, \ldots, x_n\}$ = Nuage de points associé aux données

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

Terminologie: Inertie totale du nuage de points

$$\begin{array}{c|ccc|l} & X_1 & \dots & X_p & \\ \hline 1 & & & & \\ \vdots & & & & \\ i & x_{1i} & \dots & x_{pi} & \} \ x_i \\ \vdots & & & & \\ n & & & & \end{array}$$

$$\begin{array}{|ccc|} \hline \bar{x}_1 & \dots & \bar{x}_p \\ \hline \end{array} \ \} \ g$$



Ecart-type	
Cylindrée (cm3)	1516,445
puissance (ch)	155,721
Vitesse (Km/h)	56,572
Poids (Kg)	387,507
Largeur (mm)	220,842
Longeur (mm)	581,497

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

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

Terminologie: Réduction des données

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

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

$$\vdots$$

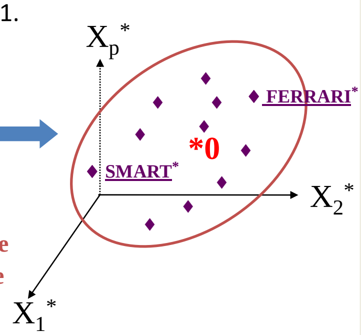
$$X_p^* = \frac{X_p - \bar{x}_p}{\sigma_p}$$

de moyenne 0 et d'écart-type 1.

$$\begin{array}{c|ccc|l} & X_1^* & \dots & X_p^* & \\ \hline 1 & & & & \\ \vdots & & & & \\ i & x_{1i}^* & \dots & x_{pi}^* & \} \ x_i^* \\ \vdots & & & & \\ n & & & & \end{array}$$

$$\begin{array}{|ccc|} \hline 0 & \dots & 0 \\ \hline \end{array} \ \} \ \text{Moyenne}$$

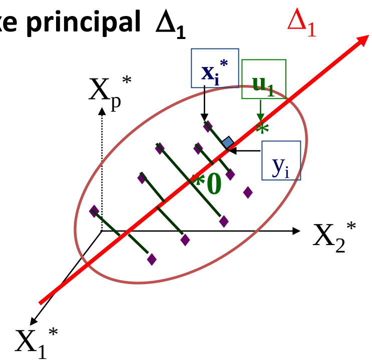
$$\begin{array}{|ccc|} \hline 1 & \dots & 1 \\ \hline \end{array} \ \} \ \text{Variance}$$



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

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

Terminologies: Premier axe principal Δ_1



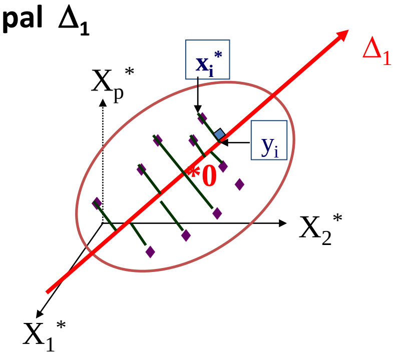
Objectif 1 : On cherche l'axe Δ_1 passant le mieux possible au milieu du nuage N^*.

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

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

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion
Contexte général | Définition | ACP- Géométrique

Terminologie: Premier axe principal Δ_1

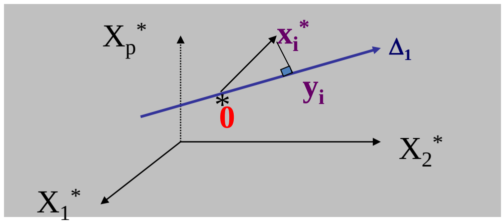


Objectif 2 : On cherche l'axe d'allongement Δ_1 du nuage N^*.

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

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

Les objectifs 1 et 2 sont atteints simultanément



De : $\boxed{d^2(x_i^*,0) = d^2(y_i,0) + d^2(x_i^*, y_i)}$

on déduit :

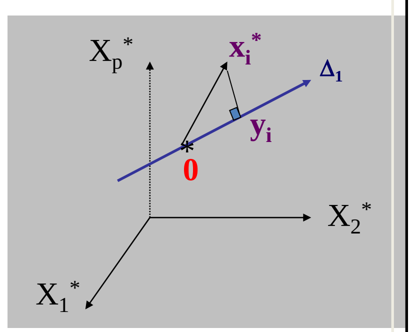
$$\boxed{\frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*,0) = \frac{1}{n}\sum_{i=1}^{n} d^2(y_i,0) + \frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*, y_i)}$$

$\boxed{\text{Inertie totale} = p}$ = $\boxed{\text{Inertie expliquée par } \Delta_1}$ + $\boxed{\text{Inertie résiduelle}}$

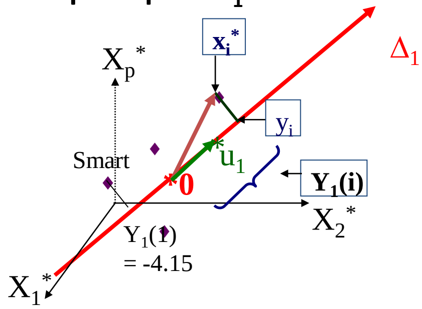
Maximiser | Minimiser

Propriétés de l'axe Δ_1

- 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.



Terminologies: 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

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

Résultats de ACP

- 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.

Propriétés de la première composante principale Y_1

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

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion
Contexte général | Définition | **ACP- Géométrique**

Qualité de la première composante principale

- Inertie totale = 6
- Inertie expliquée par le premier axe principal = λ_1 = 4.4113
- Part d'inertie expliquée par le premier axe principal :

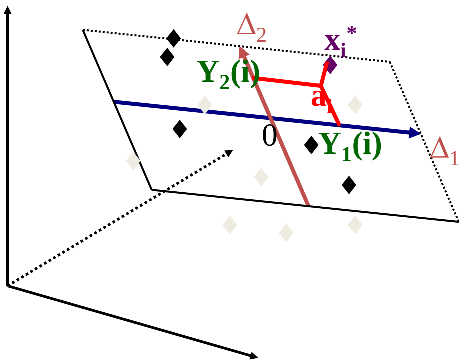
$$\frac{\lambda_1}{p} = \frac{4.4113}{6} = 0.7352$$

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

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

Contexte général | Définition | **ACP- Géométrique**

Terminologies: Deuxième axe principal Δ_2



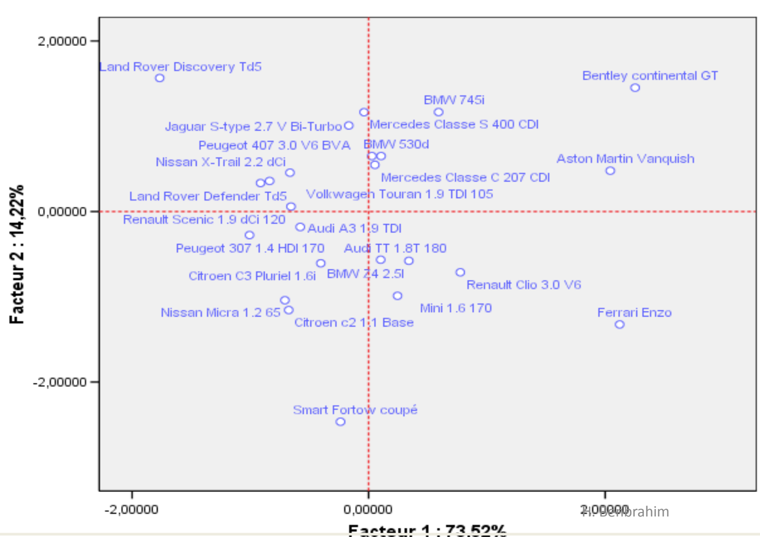
- On recherche le deuxième axe principal Δ_2 **orthogonal à Δ_1** et passant le mieux possible au milieu du nuage.
- Il **passe par le centre de gravité 0** du nuage de points et est engendré par le vecteur normé u_2, **vecteur propre de la matrice des corrélations R associé à la deuxième plus grande 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.

31 H. Benbrahim ENSIAS

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

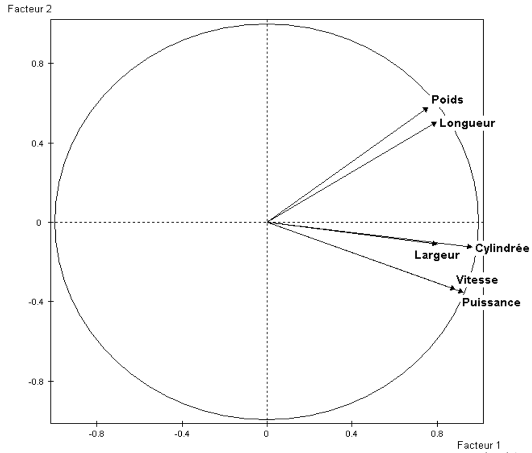
Contexte général | Définition | **ACP- Géométrique**

Terminologies: Le premier plan principal



32 32 H. Benbrahim ENSIAS

Terminologies: Le cercle des corrélations



En résumé : Qualité globale de l'analyse

- Inertie totale = variance totale = p
- Part de variance expliquée par la première composante principale $= \dfrac{\lambda_1}{p}$
- Part de variance expliquée par la deuxième composante principale $= \dfrac{\lambda_2}{p}$
- Part de variance expliquée par les deux premières composantes principales =

$$\frac{\lambda_1 + \lambda_2}{p}$$

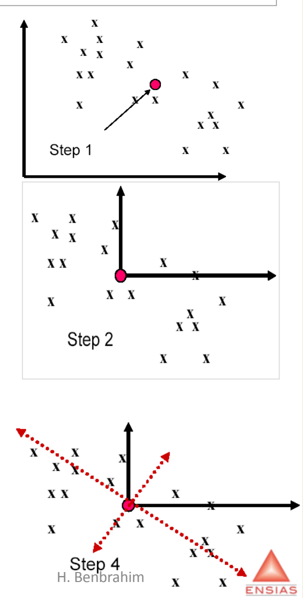
- Et ainsi de suite pour les autres dimensions…

Les techniques d'analyses factorielles

- Réduire les dimensions du tableau initial nécessite le calcul de distances entre les éléments de ce tableau.
- Selon le type des variables que l'on cherche à synthétiser, la distance à utiliser ne sera pas la même.
- Si les variables sont de type échelle, une *distance euclidienne* standard peut s'appliquer.
- Si les variables sont nominales, on utiliser une distance de type Khi-deux.
- A chaque grand type de variables va correspondre une analyse statistique particulière.

Mise en oeuvre d'une ACP

- **Etape 1**: Calculer la moyenne de chaque vecteur de caractéristiques.
- **Etape 2** : Centrer et réduire chaque vecteur de caractéristiques.
- **Etape 3** : Calculer la matrice des corrélations
- **Etape 4** : Calculer les valeurs et vecteurs propres de la matrice des corrélations.
- **Etape 5 :** Ne conserver que les valeurs propres (+ vecteurs) les plus grandes.
- **Etape 6** : Projeter les données dans ce nouvel espace propre.



Comment appliquer une ACP ?

- Il faut répondre aux questions suivantes :

  ✓ Combien de facteurs sont nécessaires pour donner une représentation juste et parcimonieuse des données ?

  ✓ Quelle est la nature de ces facteurs, comment peut-on les interpréter ?

  ✓ Quelle proportion de la variance des données peut être expliquée par un certain nombre de dimensions majeures ?

  ✓ La structure factorielle est-elle la même pour divers groupes?

**Chapitre III:
Analyse Factorielle: Analyse en Composantes Principales**

Algèbre Linéaire

- les données sont vues de manière abstraites comme un nuage de points dans un espace vectoriel. On utilise:
  - Des matrices qui permettent de manipuler un ensemble de variables comme un objet mathématique unique ;
  - Des valeurs et vecteurs propres qui permettent de décrire la structure d'une matrice.
  - Des métriques : permettent de définir la distance entre deux points de l'espace vectoriel ; on utilise aussi des produits scalaires.

Variance et écart-type

- **Définition**
  la variance de x est définie par
  $$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad ou \qquad s_x^2 = \sum_{i=1}^{n} p_i (x_i - \overline{x})^2$$
  L'écart type s_x est la racine carrée de la variance.
- **Propriétés**
  La variance satisfait la formule suivante
  $$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} p_i x_i^2 - (\overline{x})^2$$
  La variance est « la moyenne des carres moins le carre de la moyenne ».
  L'ecart-type, qui a la même unité que x, est une mesure de dispersion.

Mesure de liaison entre deux variables

- Définitions la covariance observée entre deux variables x et y est

$$s_{xy} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \overline{xy}$$

et le cœfficient de r de Bravais-Pearson ou coefficient de corrélation est donnée par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} p_i (y_i - \bar{y})^2}}$$

Corrélation et liaison significative

- **Problème**
  A partir de quelle valeur de r_{xy} peut-on considérer que les variables x et y sont liées?
- **Domaine d'application**
  on se place dans le cas ou le nombre d'individus est n > 30.
- **Méthode**
  si x et y sont deux variables gaussiennes indépendantes, alors on peut montrer que

$$\frac{(n-2)r_{xy}^2}{1-r_{xy}^2}$$

suit une loi de Fischer-Snedecor F(1; n-2). Le résultat est valable dans le cas non gaussien pour n > 30.

Le test

- on se fixe un risque d'erreur (0,01 ou 0,05 en général) et on calcule la probabilité

$$P(F(1,n-2) > \frac{(n-2)r_{xy}^2}{1-r_{xy}^2}) = \pi$$

- Si $\pi < \alpha$ on considère que l'événement est trop improbable et que donc que l'hypothèse originale d'indépendance doit être rejetée au seuil . On trouvera en général ces valeurs dans une table pré-calculée de la loi F.

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

Notation matricielle

- **Matrice**
  tableau de données carre ou rectangulaire.
- **Vecteur**
  matrice a une seule colonne.
- **Cas particuliers**

$$I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \qquad 1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- **Transposition de matrice**
  échange des lignes et des colonnes d'une matrice ; on note M’ la transposée de M.

Tableau de données

- Pour n individus et p variables, on a le tableau
  X est une matrice rectangulaire a n lignes et p colonnes

$$X = (x^1, \dots, x^p) = \begin{bmatrix} x_1^1 & x_1^2 & & \cdots & & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & \ddots & & & \\ \vdots & & & x_i^j & & \vdots \\ & & & & \ddots & \\ x_n^1 & & & \cdots & & x_n^p \end{bmatrix}$$

Vecteurs variable et individu

- **Variable**
  Une colonne du tableau

$$x^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ x_n^j \end{bmatrix}$$

- **Individu**
  Une ligne du tableau

$$e_i' = (x_i^1 \quad x_i^2 \qquad x_i^p)$$

La matrice des poids

- **Pourquoi**
  utile quand les individus n'ont pas la même importance

- **Comment**
  on associe aux individus un poids pi tel que
  $$p_1 + p_2 + \ldots + p_n = 1$$
  et on représente ces poids dans la matrice diagonale de taille n
  $$D = \begin{bmatrix} p_1 & & \ldots & 0 \\ & p_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \ldots & p_n \end{bmatrix}$$

- **Cas uniforme**
  tous les individus ont le même poids $p_i = 1 / n$ et D = I / n

Point moyen et tableau centre

- **Point moyen**
  c'est le vecteur g des moyennes arithmétiques de chaque variable :
  $$g' = (\bar{x}^1 \quad \ldots \quad \bar{x}^p)$$
  ou $$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j$$
  On peut aussi écrire $$g = X'D1$$

  Tableau centré
  il est obtenu en centrant les variables autour de leur moyenne
  $$y_i^j = x_i^j - \bar{x}^j$$
  ou, en notation matricielle,
  $$Y = X - 1g' = (I - 11'D)X$$

Matrice de variance covariance

- **Définition**
  c'est une matrice carrée de dimension p

$$V = \begin{bmatrix} s_1^1 & s_1^2 & \dots & s_1^p \\ s_2^1 & s_2^2 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & s_p^p \end{bmatrix}$$

ou s_{kl} est la covariance des variables x^k et x^l et s^2_j est la variance de la variable x^j

- **Formule matricielle**

$$V = X'DX - gg' = Y'DY$$

Matrice de corrélation

- **Définition**
  Si l'on note

$$r_{kl} = \frac{s_{kl}}{s_k s_l} \qquad R = \begin{bmatrix} 1 & r_1^2 & \dots & r_1^p \\ r_2^1 & 1 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & 1 \end{bmatrix}$$

- **Formule matricielle**

$$R = D_{1/s} V D_{1/s} \qquad D_{1/s} = \begin{bmatrix} \frac{1}{s_1} & & & 0 \\ \vdots & \frac{1}{s_2} & & \vdots \\ & & \ddots & \\ 0 & & & \frac{1}{s_p} \end{bmatrix}$$

Section I: Introduction et Définition

Section II: Rappel: Algèbre Linéaire

**Chapitre III:
Analyse Factorielle: Analyse en Composantes Principales**

Section III: ACP

Section IV: ACP sous SPSS

Section V: Récapitulatif et Conclusion

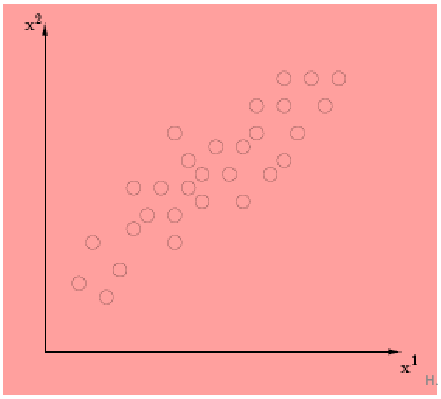
L'analyse de composantes principales

- **Contexte**
chaque individu est considéré comme un point d'un espace vectoriel F de dimension p . L'ensemble des individus est un nuage de points dans F et g est son centre de gravite.

- **Principe**
on cherche a réduire le nombre p de variables tout en préservant au maximum la structure du problème. Pour cela on projette le nuage de points sur un sous-espace de dimension inférieure

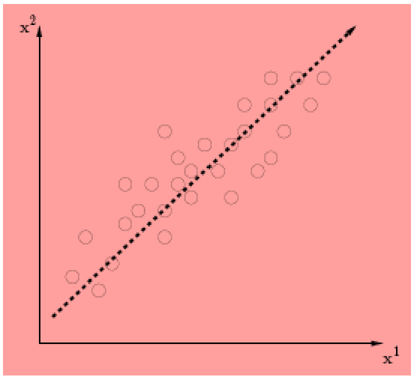
Exemple en dimension 2

- On veut passer de 2 variables a 1 seule.



Exemple en dimension 2

- On cherche la direction qui différencie le plus les points entre eux.



Distance entre individus

- **Motivation**
  afin de pouvoir considérer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.

- **Distance euclidienne classique**
  la distance la plus simple entre deux points de Rp est définie par

$$d^2(u,v) = \sum_{j=0}^{p} (u_j - v_j)^2 = \|u - v\|$$

- **Généralisation simple**
  on multiplie la variable j par $\sqrt{a_j}$

$$d^2(u,v) = \sum_{j=0}^{p} a_j (u_j - v_j)^2$$

Métrique

- **Matrice définie positive**
  c'est une matrice symétrique telle que, pour tout u non nul, u'Mu > 0.

- **Définition**
  soit M = (m_{jk}) définie positive de dimension p. On pose

$$\|u\|_M^2 = u'Mu = \sum_{j=0}^{p} \sum_{k=1}^{p} m_{jk} u_j u_k \quad et \quad d_M^2(u,v) = \|u - v\|_M^2$$

- **Espace métrique**
  il est défini par le produit scalaire

$$\langle u,v \rangle_M = u'Mu = \sum_{j=0}^{p} \sum_{k=1}^{p} m_{jk} u_j u_k$$

On dit que u et v sont orthogonaux si $\langle u,v \rangle_M = 0$

Comparaison avec le cas usuel

- Norme

$$\|u\|^2 = u'u = \sum_{j=0}^{p} u_j^2 = u'Iu$$

$$\|u\|_M^2 = u'Mu = \sum_{j=0}^{p}\sum_{k=1}^{p} m_{jk}u_j u_k$$

- Produit scalaire

$$\langle u,v\rangle = u'u = \sum_{j=0}^{p} u_j v_k = u'Iu$$

$$\langle u,v\rangle_M = u'Mu = \sum_{j=0}^{p}\sum_{k=1}^{p} m_{jk}u_j u_k$$

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

Inertie

- **Définition**
  l'inertie en un point a du nuage de points est

$$I_a = \sum_{i=1}^{n} p_i \|e_i - a\|_M^2 = \sum_{i=1}^{n} p_i (e_i - a)' M (e_i - a)$$

- **Autres relations**
  l'inertie totale Ig est la moitie de la moyenne des carres des distances entre les individus

$$2I_g = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \|e_i - e_j\|_M^2$$

- L'inertie totale est aussi donnée par la trace de la matrice MV
  (la trace d'une matrice étant la somme de ses éléments diagonaux).

$$I_g = Tr(MV)$$

Métriques particulières

- **Métrique usuelle**
  M = I correspond au produit scalaire usuel et
  $$I_g = Tr(V) = \sum_{j=1}^{p} s_i^2$$
- **Problèmes**
  - la distance entre individus dépend de l'unité de mesure.
  - la distance privilégie les variables les plus dispersées.
- **Métrique réduite**
  c'est la plus courante ; on prend la matrice diagonale des inverses des variances
  $$M = D_{1/s^2} = \begin{bmatrix} \frac{1}{s_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p^2} \end{bmatrix}$$

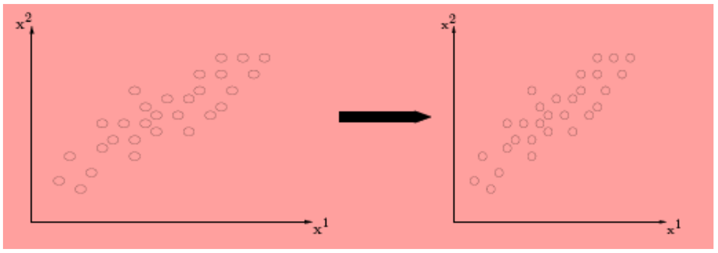
$$I_g = Tr(D_{1/s^2} V) = Tr(D_{1/s} V D_{1/s}) = Tr(R) = p$$

Métriques et tableaux transformés

- Utiliser la métrique M = T'T sur le tableau X est équivalent a travailler avec la métrique classique I sur le tableau transforme XT'.
- **Tableau transformé**
  Si on travaille sur le tableau transforme XT' (changement de variables) au lieu de X, alors les nouveaux individus seront de la forme Te_i et
  $$\langle Te_{i_1}, Te_{i_2} \rangle = (Te_{i_1})'(Te_{i_2}) = e_{i_1}'T'Te_{i_2} = e_{i_1}'Me_{i_2} = \langle e_{i_1}, e_{i_2} \rangle_M$$
- **Réciproque**
  pour toute matrice symétrique positive M, il existe une matrice T (racine carrée de M) telle que
  $$M = T'T$$
  et donc on peut ramener l'utilisation de la métrique a un changement de variables.

Métriques et tableaux transformés (suite)

- Utiliser une métrique est donc équivalent à tordre les données pour les rendre comparables



- Exemple utiliser la métrique réduite est équivalent a travailler sur les données centrées réduites $Z = YD_{1/s}$.

L'analyse de composantes principales (2)

- **Principe**
  on cherche a projeter le nuage de points sur un espace F_k de dimension $k < p$.

- **Critère**
  on veut que la moyenne des carres des distances entre les points projetés soit maximale (elle est toujours plus petite que pour le nuage original).

  Pour cela on cherche F_k, sous espace de dimension k de R^k, tel que l'inertie du nuage projeté sur F_k soit maximale.

Interlude : valeurs et vecteurs propres

- **Définition**
  un vecteur v de taille p est un vecteur propre d'une matrice A de taille p x p s'il existe λ Є C telle que
  $$Av = \lambda v$$
  est une valeur propre de A associée à v.
- **Domaine**
  En général, les vecteurs propres et valeurs propres sont complexes; dans tous les cas qui nous intéressent, ils seront réels.
- **Interprétation des vecteurs propres**
  ce sont les directions dans lesquelles la matrice agit.
- **Interprétation des valeurs propres**
  c'est le facteur multiplicatif associe a une direction donnée.

Exemple: valeurs et vecteurs propres

La matrice

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

a pour vecteurs propres

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2 \qquad \lambda_2 = 4 \qquad \lambda_3 = 6$$

Cas particuliers: Valeurs et vecteurs propres

- **Matrice nulle**
  sa seule valeur propre est 0, et tout vecteur est vecteur propre.
- **Matrice identité**
  tout vecteur est vecteur propre de I avec valeur propre 1, puisque Iv = v.
- **Matrice diagonale**
  si D_λ est une matrice diagonale avec les coefficients $\lambda_1, \lambda_2, \ldots \lambda_p$, alors le i-eme vecteur coordonnée est vecteur propre de D_λ associe a la valeur propre λ_i.
  L'action d'une matrice diagonale est de multiplier chacune des coordonnées d'un vecteur par la valeur propre correspondante.
- **Matrice diagonalisable**
  c'est une matrice dont les vecteurs propres forment une base de l'espace vectoriel : tout vecteur peut être représenté de manière unique comme combinaison linéaire des vecteurs propres. Une matrice de taille p x p qui a p valeurs propres réelles distinctes est diagonalisable dans R.

Quelques matrices diagonalisables

- **Matrice symétrique**
  une matrice symétrique réelle (A' = A) possède une base de vecteurs propres orthogonaux et ses valeurs propres sont réelles

$$\langle v_i, v_j \rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice M-symetrique**
  une matrice M-symetrique réelle (A'M = MA) possède une base de vecteurs propres M-orthogonaux et ses valeurs propres sont positives ou nulles

$$\langle v_i, v_j \rangle_M = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice définie positive**
  c'est une matrice symétrique dont les valeurs propres sont strictement positives et donc

$$\langle v_i, v_j \rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i > 0$$

Analyse de la matrice de variance: VM

- **Valeurs propres**
  la matrice VM est M-symetrique: elle est donc diagonalisable et ses valeurs propres λ_1, λ_2, λ_p sont réelles.
- **Vecteurs propres**
  il existe donc p vecteurs $a_{1,\ \dots,}\ a_p$ tels que

$$VMa_i = \lambda a_i \qquad avec \quad \langle a_i, a_j \rangle_M = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{sinon} \end{cases}$$

  Les a_i sont les axes principaux d'inertie de VM. Ils sont M-orthonormaux.
- **Signe des valeurs propres**
  les valeurs propres de VM sont positives et on peut les classer par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

- **Idée du lien avec l'inertie**
  on sait que .

$$Tr(VM) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

  Si on ne garde que les données relatives a $a_{1,\ \dots,}\ a_p$ on gardera l'inertie $\lambda_1 + \lambda_2 + \ldots + \lambda_p$, et c'est le mieux qu'on puisse faire.

Résultat principal

- **Théorème principal (Admis)**

1. Si F_k est le sous-espace de dimension k portant l'inertie principale, alors

$$F_{k+1} = F_k \oplus f_{k+1}$$

   ou $f_{k+}1$ est le sous espace de dimension 1 M-orthogonal a F_k portant l'inertie maximale : les solutions sont emboîtées;

2. F_k est engendre par les k vecteurs propres de VM associes aux k plus grandes valeurs propres.

- **Interprétation du théorème**
  l'ACP sur k + 1 variables est obtenue par ajout d'une variable d'inertie maximale a l'ACP sur k variables. Il n'est pas nécessaire de refaire tout le calcul.

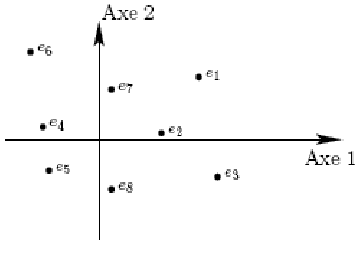
Les composantes principales

- **Coordonnées des individus**
  supposons que
  $$e_i - g = \sum_{k=1}^{p} c_{ij} a_k$$
  alors
  $$\langle e_i - g, a_j \rangle_M = \sum_{k=1}^{p} c_{ij} \langle a_k, a_j \rangle_M = c_{ij}$$
  La coordonnée de l'individu centre e_i - g sur l'axe principal a_j est donc donné par la projection M-orthogonale
  $$c_{ij} = \langle e_i - g, a_j \rangle_M = (e_i - g)' M a_j$$

- **Composantes principales**
  ce sont les variables c_j de taille n définies par
  $$c_j = YMa_j$$
  Chaque c_j contient les coordonnées des projections M-orthogonales des individus centres sur l'axe défini par les a_j.

Représentation des individus dans un plan principal

- **Qu'est-ce que c'est**?
  C'est une représentation ou, pour deux composantes principales c_1 et c_2, on représente chaque individu i par un point d'abscisse c_{i1} et d'ordonnée c_{i2}.



Propriétés des composantes principales

- **Moyenne arithmétique**
  les composantes principales sont centrées :
  $$\bar{c}_j = c'_j D1 = a'_j MY'D1 = 0 \quad car \quad Y'D1 = 0$$
- **Variance**
  la variance de c_j est j car
  $$V(c_j) = c'_j Dc_j = a'_j MY'DYMa_j$$
  $$= a'_j MVMa_j = \lambda_j a'_j Ma_j = \lambda_j$$
- **Covariance**
  de même, pour $i \neq j$
  $$\mathrm{cov}(c_i, c_j) = c'_i Dc_j = \ldots = \lambda_j a'_j Ma_j = 0$$
  Les composantes principales ne sont pas corrélées entre elles.

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

Facteurs principaux

- **Définition**
  on associe a un axe principal a_j le facteur principal
  $$u_j = Ma_j$$
  de taille p.
  C'est un vecteur propre de MV car
  $$MVu_j = MVMa_j = \lambda_j Ma_j = \lambda_j u_j$$
- **Calcul en pratique**,
  on calcule les u_j par diagonalisation de MV, puis on obtient les
  $$c_j = Yu_j$$
  Les a_j ne sont pas intéressants. La valeur d'une variable c_j pour l'individu e_i est donc
  $$c_{ij} = (e_j - g)'u_j = \sum_{k=1}^{p} y_i^k u_{jk} \quad où \quad u'_j = (u_{i1}, u_{i2}, \ldots, u_{ip})$$

Formules de reconstruction

- Il est possible de reconstruire le tableau centre Y a partir des composantes principales et des facteurs principaux

$$Y = \sum_{j=1}^{p} c_j a'_j = \sum_{j=1}^{p} c_j u'_j M^{-1}$$

- **Preuve**
il suffit de calculer

$$\left(\sum_{j=1}^{p} c_j a'_j\right) Ma_j = \sum_{j=1}^{p} c_j a'_j Ma_j = c_i = YMa_i$$

et, comme M est inversible et que ai est une base, on obtient Y.

- **Approximation**
si on prend les k premiers termes seulement, on obtient la meilleure approximation de Y par une matrice de rang k au sens des moindres carres (théorème de Eckart-Young).

Le cas de la métrique D_{1/s^2}

- **Pourquoi cette métrique ?**
pour que les distances soient indépendantes des unités de mesure et qu'elles ne privilégient pas les variables dispersées.

- **Équivalence avec les données réduites**
on a

$$D_{1/s^2} = D_{1/s} D_{1/s}$$

et donc

$$\langle e_i, e_j \rangle_{D_{1/s^2}} = \langle D_{1/s} e_i, D_{1/s} e_j \rangle$$

Travailler avec la métrique D_{1/s^2} est équivalent a diviser chaque variable par son ecart-type et a utiliser la métrique I.

- **Données centrées réduites**
c'est le tableau Z contenant les données $z_i^j = \dfrac{x_i^j - \bar{x}^j}{s_j}$

qui se calcule matriciellement comme $Z = YD_{1/s}$

L'ACP sur les données centrées réduites

- **Matrice de variance covariance**
  c'est la matrice de corrélation car

$$Z'DZ = D_{1/s} Y'DYD_{1/s} = D_{1/s} VD_{1/s} = R$$

- **Métrique**
  on prend la métrique M = I.

- **Facteurs principaux**
  ce sont les p vecteurs propres orthonormés de R,

$$Ru_i = \lambda_i u_i \qquad avec \quad \langle u_i, u_j \rangle = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{sinon} \end{cases}$$

  dont les valeurs propres sont classes par valeur propre croissante

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$$

- **Composantes principales**
  elles sont données par

$$c_j = Zu_j$$

Définition Algèbre Linéaire ACP ACP sous SPSS Conclusion

Nombre d'axes a retenir

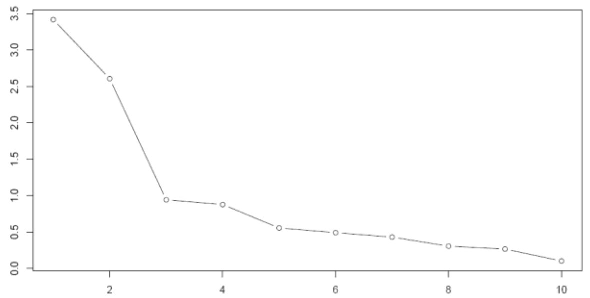
- **Dimension de l'espace** des individus L'ACP visant a réduire la dimension de l'espace des individus, on veut conserver aussi peu d'axes que possible. Il faut pour cela que les variables d'origine soient raisonnablement corrélées entre elles. Les seuls critères utilisables sont empiriques.

- **Interprétation des axes**
  on s'efforce de ne retenir que des axes a propos desquels une forme d'interprétation est possible (soit directement, soit en terme des variables avec lesquels ils sont très corrélés).

- **Critère de Kaiser (variables centrées réduites)**
  on ne retient que les axes associes a des valeurs propres supérieures a 1, c'est-à-dire dont la variance est supérieure a celle des variables d'origine. Une autre interprétation est que la moyenne des valeurs propres étant 1, on ne garde que celles qui sont supérieures a cette moyenne.

Nombre d'axes a retenir (2)

- **Éboulis des valeurs propres**
  on cherche un coude dans le graphe des valeurs propres



L'espace des variables

- **Métrique D**
  il faut munir l'espace des variables d'une métrique raisonnable. On choisit toujours la métrique D des poids :
  $$\langle x, y\rangle_D = x'Dy \qquad \|x\|_D^2 = x'Dx$$
- **Interprétation**
  pour deux variables centrées x et y, on a:
  $$\operatorname{cov}(x,y) = \langle x, y\rangle_D \qquad V(x) = \|x\|_D^2 \qquad cor(x,y) = \frac{\langle x, y\rangle_D}{\|x\|_D \|y\|_D}$$
- **Exemple**
  les vecteurs $\frac{c_k}{\sqrt{\lambda_k}}$ forment une base D-orthonormale
  $$\left\langle \frac{c_k}{\sqrt{\lambda_k}}, \frac{c_l}{\sqrt{\lambda_l}} \right\rangle_D = cor(c_k, c_l) = \begin{cases} 1 & \text{si} \quad k = l \\ 0 & \text{sinon} \end{cases}$$

Corrélation entre composantes et variables initiales

- Quand on travaille sur les variables centrées-réduites, la corrélation entre une composante principale c_k et une variable z_j est

$$r(z^j, c_k) = \frac{\mathrm{cov}(z^j, c_k)}{V(c_k)} = \frac{(z^j)' D c_k}{\sqrt{\lambda_k}}$$

et donc le vecteur des corrélations de c_k avec Z est

$$r(Z, c_k) = (r(z^1, c_k), r(z^2, c_k), \ldots, r(z^p, c_k))' = \frac{Z' D c_k}{\sqrt{\lambda_k}}$$

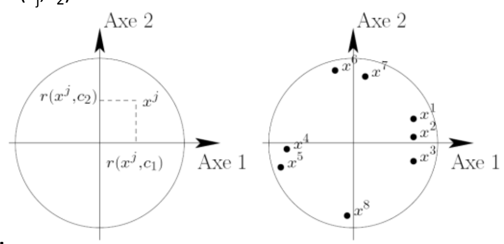
- Comme $Z' D c_k = Z' D Z u_k = R u_k = \lambda_k u_k$

on a finalement

$$r(Z, c_k) = \sqrt{\lambda_k}\, u_k$$

Le cercle des corrélations

- **Qu'est-ce que c'est?**
c'est une représentation ou, pour deux composantes principales, par exemple c_1 et c_2, on représente chaque variable z_j par un point d'abscisse $r(z_j; c_1)$ et d'ordonnée $r(z_j;c_2)$.



- **Effet « taille »**
cela arrive quand toutes les variables sont corrélées positivement avec la première composante principale. Cette composante est alors appelée facteur de taille, la seconde facteur de forme.

Le cercle des corrélations (2)

- **Pourquoi un cercle?**
  comme les ck=pk forment une base D-orthonormale,

$$z^j = \sum_{k=1}^{p} \left\langle \frac{c_k}{\lambda_k}, z^j \right\rangle_D \frac{c_k}{\lambda_k} = \sum_{i=1}^{p} r(c_k, z^j) \frac{c_k}{\lambda_k}$$

$$\|z_j\|_D^2 = 1 = \sum_{k=1}^{p} r^2(c_k, z^j)$$

  Les points sont bien a l'intérieur d'un cercle de rayon 1.

- **Interprétation**
  - les points sont la projection orthogonale dans D des variables dans le plan défini par les composantes principales c_1 et c_2.
  - Il ne faut interpréter la proximité des points que s'ils sont proches de la circonférence.

Contribution d'un individu a une composante

- Définition
  On sait que

$$V(c_k) = \lambda_k = \sum_{i=1}^{n} p_i c_{ik}^2$$

  La contribution de l'individu i a la composante k est donc

$$\frac{p_i c_{ik}^2}{\lambda_k}$$

- **Interprétation**
  la contribution d'un individu est importante si elle excède le poids pi de l'individu concerne, c'est-à-dire

$$\frac{p_i c_{ik}^2}{\lambda_k} \geq p_i \Rightarrow |c_{ik}| \geq \sqrt{\lambda_k}$$

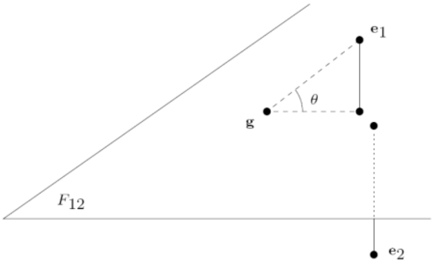
- **Individus sur-representés**
  ce sont les individus qui jouent un rôle trop fort dans la définition d'un axe (par exemple > 0;25). Il « tire a lui » l'axe k et risque de perturber les représentations des autres points sur les axes de rang k. Un tel individu peut ^être le signe de données erronées.

Qualité globale de la représentation

- **Calcul de l'inertie**
  on se souvient que
  $$I_g = Tr(VM)$$
  comme la trace d'une matrice est la somme de ses valeurs propres, on a
  $$I_g = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$
- **Définition**
  la qualité de la représentation obtenue par k valeurs propres est la proportion de l'inertie expliquée
  $$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$
- **Utilisation**
  si par exemple $\lambda_1 + \lambda_2$ est égal 90% de Ig, on en déduit que le nuage de points est aplati autour du premier plan principal.

Qualité locale de la représentation

- **But**
  on cherche a déterminer si le nuage de points est très aplati par la projection sur les sous-espaces principaux. Dans ce cas, deux individus éloignes pourraient artificiellement sembler proches les uns des autres.



Angle entre un individu et un axe principal

- Il est défini par son cosinus carre. Le cosinus de l'angle entre l'individu centre i et l'axe principal j est

$$\cos^2(e_i, a_j) = \frac{\langle e_i - g, a_j \rangle_M}{\|e_i - g\|_M}$$

car les a_j forment une base orthonormale.

Comme $$\langle e_i - g, a_j \rangle_M = c_{ij}$$

$$\cos^2(e_i, a_j) = \frac{c_{ij}}{\sum_{k=1}^{p} c_{ik}^2}$$

Cette grandeur mesure la qualité de la représentation de l'individu i sur l'axe principal a_j.

Angle entre un individu et un sous-espace principal

- C'est l'angle entre l'individu et sa projection orthogonale sur le sous-espace. La projection de $e_i - g$ sur le sous-espace F_q est, $q \le p$

$$\sum_{k=1}^{q} c_{ik} a_k$$

et donc

$$\cos^2(e_i, F_q) = \frac{\sum_{k=1}^{q} c_{ik}^2}{\sum_{k=1}^{p} c_{ik}^2}$$

- La qualité de la représentation de l'individu i sur le plan F_q est donc la somme des qualités de représentation sur les axes formant F_q. Il est significatif quand le point e n'est pas trop près de g.

L'ACP en trois transparents (1)

- **Données**
  les données représentent les valeurs de p variables mesurées sur n individus ; les individus peuvent avoir un poids. En général on travaille sur des données centrées réduites Z (on retranche la moyenne et on divise par l'écart type).

- **Matrice de corrélation**
  c'est la matrice R de variance covariance des variables centrées réduites. Elle possède p valeurs propres:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$$

- **Facteurs principaux u_k**
  ce sont les vecteurs propres orthonormés de R (de dimension p) associes aux valeurs propres k. Leur j-ieme composante u_{kj} est le poids de la variable j dans la composante k.

- **Composantes principales c_k**
  ce sont les vecteurs Zu_k de dimension n. Leur i-ieme coordonnée c_{ki} est la valeur de la composante k pour l'individu i. Les ck sont decorrelees et leur variance est :

$$V(c_k) = j$$

Définition Algèbre Linéaire ACP ACP sous SPSS Conclusion

L'ACP en trois transparents (2)

- **Nombre d'axes**
  on se contente souvent de garder les axes interprétables de valeur propre supérieure a 1. La qualité de la représentation retenue est mesure par la part d'inertie expliquée par ces composantes.

- **Cercle des corrélations**
  il permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. A partir de la, on peut soit trouver une signification physique a chaque composante, soit montrer que les composantes séparent les variables en paquets. Seules les variables bien représentées (situées près du bord du cercle) doivent être interprétées.

L'ACP en trois transparents (3)

- **Représentation des individus pour un plan principal donné**, la représentation des projections des individus permet de conformer l'interprétation des variables. On peut aussi visualiser les individus aberrants (erreur de donnée ou individu atypique).
- **Contribution d'un individu a une composante** c'est la part de la variance d'une composante principale qui provient d'un individu donne. Si cette contribution est très supérieure aux autres, on peut avoir intérêt a mettre l'individu en donnée supplémentaire.
- **Qualité globale de la représentation** c'est la part de l'inertie totale Ig qui est expliquée par les axes principaux qui ont été retenus. Elle permet de mesurer la précision et la pertinence de l'ACP.
- **Qualité de la représentation d'un individu** elle permet de vérifier que tous les individus sont bien représentes par le sous-espace principal choisi; elle s'exprime comme le carre du cosinus de l'angle entre l'individu et sa projection orthogonale.

ENSIAS

Chapitre III: Analyse Factorielle: Analyse en Composantes Principales

Illustration sur les données : voiture.sav (extrait de Saporta)

voiture - Editeur de données SPSS

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Fenêtre Aide

7 :

	num	modele	model2	cylindre	puissanc	vitesse	poids	longueur	largeur
2	22	Peugeot 205 Rallye	P205Ral	1294,00	103,00	189,00	805,00	370,00	157,00
3	23	Seat Ibiza SX I	SIbizSxi	1461,00	100,00	181,00	925,00	363,00	161,00
4	24	Citroën AX Sport	AXSport	1294,00	95,00	184,00	730,00	350,00	160,00
5	2	Renault 19	R19	1721,00	92,00	180,00	965,00	415,00	169,00
6	3	Fiat Tipo	FiatTipo	1580,00	83,00	170,00	970,00	395,00	170,00
7	4	Peugeot 405	P405	1769,00	90,00	180,00	1080,00	440,00	169,00
8	5	Renault 21	R21	2068,00	88,00	180,00	1135,00	446,00	170,00
9	6	Citroën BX	BX	1769,00	90,00	182,00	1060,00	424,00	168,00
10	10	Opel Omega	OOmega	1998,00	122,00	190,00	1255,00	473,00	177,00
11	11	Peugeot 405 Break	405break	1905,00	125,00	194,00	1120,00	439,00	171,00
12	12	Ford Sierra	FSierra	1993,00	115,00	185,00	1190,00	451,00	172,00
13	16	Renault Espace	REspace	1995,00	120,00	177,00	1265,00	436,00	177,00
14	17	Nissan Vanette	NVanette	1952,00	87,00	144,00	1430,00	436,00	169,00
15	18	VW Caravelle	VWCarav	2109,00	112,00	149,00	1320,00	457,00	184,00
16	14	Audi 90 Quattro	Audi90Q	1994,00	160,00	214,00	1220,00	439,00	169,00
17	7	BMW 530i	BMW 530i	2986,00	188,00	226,00	1510,00	472,00	175,00
18	8	Rover 827i	Rov827i	2675,00	177,00	222,00	1365,00	469,00	175,00
19	9	Renault 25	R25	2548,00	182,00	226,00	1350,00	471,00	180,00
20	13	BMW 325iX	BMW 325i	2494,00	171,00	208,00	1600,00	432,00	164,00
21	15	Ford Scorpio	FScorpio	2933,00	150,00	200,00	1345,00	466,00	176,00
22	20	Fiat Uno	Fiat Uno	1116,00	58,00	145,00	780,00	364,00	155,00
23	21	Peugeot 205	P205	1580,00	80,00	159,00	880,00	370,00	156,00
24	19	Ford Fiesta	FFiesta	1117,00	50,00	135,00	810,00	371,00	162,00

ENSIAS

Définition | Algèbre Linéaire | ACP | ACP sous SPSS | Conclusion

VOITURE.SAV - SPSS Editeur de données

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Fenêtre Aide

32 : num

Analyse:
- Rapports
- Statistiques descriptives
- Tableaux
- Comparer les moyennes
- Modèle linéaire général
- Modèles mixtes
- Corrélation
- Régression
- Analyse log-linéaire
- Classification
- Factorisation
  - Analyse factorielle...
  - Analyse des correspondances...
  - Codage optimal...
- Positionnement
- Tests non paramétriques
- Séries chronologiques
- Survie
- Réponses multiples
- Analyse des valeurs manquantes...
- Echantillons complexes

	num	modele				uissanc	vitesse	poids
3	23,00					100,00	181,00	925,00
4	24,00					95,00	184,00	730,00
5	2,00					92,00	180,00	965,00
6	3,00					83,00	170,00	970,00
7	4,00					90,00	180,00	1080,00
8	5,00							
9	6,00							
10	10,00							
11	11,00					125,00	194,00	1120,00
12	12,00					115,00	185,00	1190,00
13	16,00					120,00	177,00	1265,00
14	17,00					87,00	144,00	1430,00
15	18,00		18	VWCarav	2109,00	112,00	149,00	1320,00
16	14,00		14	Audi90Q	1994,00	160,00	214,00	1220,00
17	7,00		7	BMW 530i	2986,00	188,00	226,00	1510,00
18	8,00		8	Rov827i	2675,00	177,00	222,00	1365,00

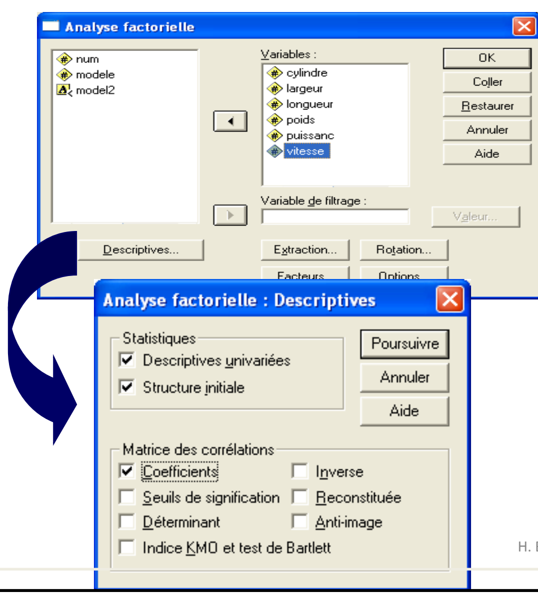
ENSIAS

LES ÉTAPES D'UNE ANALYSE FACTORIELLE sous SPSS

étape 1	***Examen de la matrice des corrélation*** • mettre en évidence les relations entre les variables; • évaluation des propriétés du modèle factoriel; • décider du traitement des valeurs manquantes.
étape 2	***Extraction des facteurs*** • déterminer le nombre defacteurs requis; • choix de la méthode d'extraction des facteurs.
étape 3	***Transformation par rotation des facteurs*** • rendre les facteurs plus interprétables.
étape 4	***Calcul des scores*** • calcul des coefficients associés à chaque facteur pour servir à d'autres analyses.

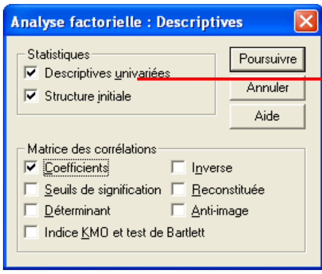
ETAPE 1 : ETUDE DES LIENS ENTRE LES VARIABLES

Étape 1



ETAPE 1 : exploration des variables d'analyse

1.1 Caractéristiques univariées



Statistiques descriptives

	Moyenne	Ecart-type	n analyse
cylindre	1906,1250	527,90870	24
largeur	168,8333	7,65374	24
longueur	421,5833	41,34049	24
poids	1123,3333	248,43277	24
puissanc	113,6667	38,78443	24
vitesse	183,0833	25,21545	24

L'intérêt principal de ce type de statistiques est d'étudier la variabilité des quantités qui vont composer l'ACP.

On peut également obtenir davantage d'informations statistiques en cliquant dans *Analyse>Statistiques descriptives>Caractéristiques.*

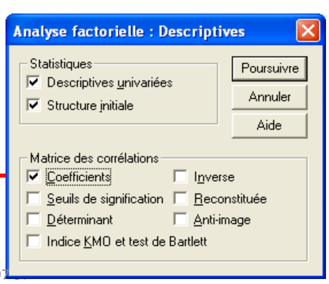
Nécessité de centrage et de réduction

ETAPE 1 : ETUDE DES LIENS ENTRE LES VARIABLES

1.2 Analyse de la Matrice de corrélation

Correlation Matrix:

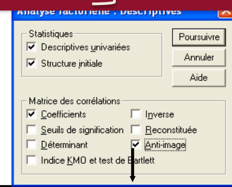
	CYLINDRE	LARGEUR	LONGUEUR	POIDS	PUISSANC	VITESSE
CYLINDRE	1.00000					
LARGEUR	.70906	1.00000				
LONGUEUR	.86420	.86381	1.00000			
POIDS	.89732	.70000	.86337	1.00000		
PUISSANC	.86098	.55228	.68851	.76921	1.00000	
VITESSE	.69332	.36323	.53191	.50742	.89399	1.00000



Les variables sont fortement corrélées. Il paraît pertinent de chercher à synthétiser l'information en réduisant le nombre de variables en un petit nombre de facteurs deux à deux non corrélés.

ETAPE 1 : ETUDE DES LIENS ENTRE LES VARIABLES

1.3 Matrice des corrélations partielles : anti-image

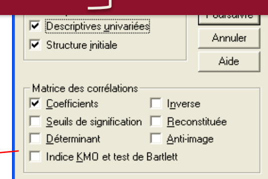


Les corrélations partielles totales donnent une idée **de la force intrinsèque** qui relie deux variables en supprimant les effets linéaires induits par les autres variables.

Des coefficients proches de 0 implique la présence d'inter relations transitant par toutes les variables du modèle.

CRITERES DE PERTINENCE D'UNE ACP

1.4 Test de Kaiser Meyer Olkin : KMO



$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}$$

avec r_{ij}^2, les corrélations totales et a_{ij}^2, les corrélations partielles.

Le critère nous permet de poursuivre l'ACP si KMO est proche de 1, i.e. lorsque les corrélations partielles sont faibles. En effet, en ACP, on souhaite que les corrélations soient expliquées par d'autres variables que celles concernées. Il ne serait pas intéressant d'étudier des variables uniquement corrélées deux à deux.

Echelles de mesure

KMO	évaluation de l'AF	KMO	évaluation de l'AF
0,9	Merveilleux	0,6	médiocre
0,8	méritoire	0,5	misérable
0,7	moyen	moins de 0,5	inacceptable

Kaiser-Meyer-Olkin Measure of Sampling Adequacy = .66938

CRITERES DE PERTINENCE D'UNE ACP

1.5 Measure of Sampling Adequacy : MSA

$$msa_i = \frac{\sum_j r_{ij}^2}{\sum_j r_{ij}^2 + \sum_j a_{ij}^2}$$

calculée variable par variable

Anti-image Correlation Matrix:

	CYLINDRE	LARGEUR	LONGUEUR	POIDS	PUISSANC	VITESSE
CYLINDRE	.93403					
LARGEUR	.00453	.66050				
LONGUEUR	-.26598	-.76061	.65758			
POIDS	-.32958	.39762	-.59316	.65060		
PUISSANC	-.21941	-.41220	.53269	-.65901	.62545	
VITESSE	-.06258	.46028	-.52436	.70642	-.89849	.52070

Measures of Sampling Adequacy (MSA) are printed on the diagonal.

Plus msa_i est élevé et proche de 1, plus la variable correspondante contribue fortement dans la construction des facteurs.

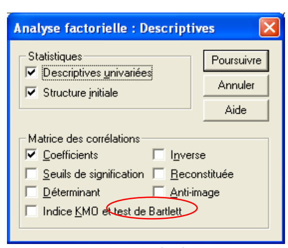
ETAPE 1 : ETUDE DES LIENS ENTRE LES VARIABLES

1.6 Le test de sphéricité de Bartlett

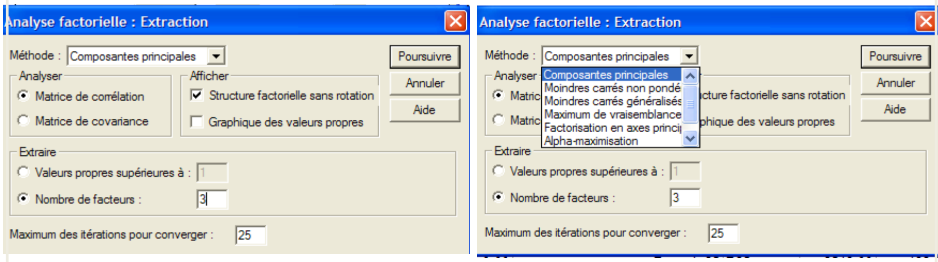
Permet de tester l'hypothèse nulle selon laquelle la matrice des corrélations est égale à la matrice identité. En d'autres termes, on cherche à savoir si les variables sont corrélées entre elles. La sphéricité implique un nuage de points qui se dilate dans tous les sens. Les points sont alors représentés par une sphère.

Bartlett Test of Sphericity = 171.18761, Significance = .00000

On rejette l'hypothèse que la matrice de corrélation est l'identité



ETAPE 2 : DETERMINATION DES COMPOSANTES PRINCIPALES SOUS SPSS

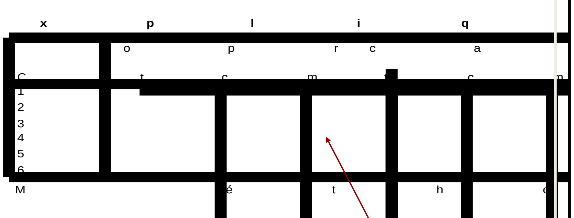


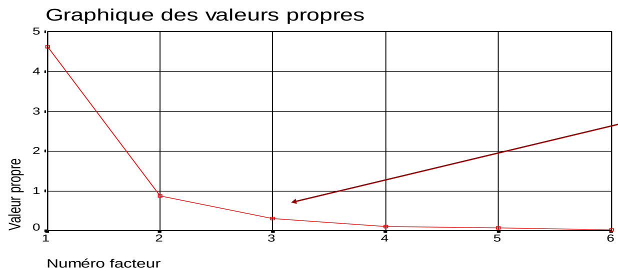
SPSS propose plusieurs techniques d'extractions de facteurs, la plus connue et la plus utilisée étant l'analyse en composantes principales. Cette méthode détermine les facteurs principaux en constituant des combinaisons linéaires non corrélées deux à deux des variables initiales.

Définition Algèbre Linéaire ACP ACP sous SPSS Conclusion

2.1 NOMBRE DE FACTEURS À EXTRAIRE

Formellement, il est possible d'extraire autant de facteurs que de variables. Cependant, la méthode de construction des axes fait que **les derniers ont un pouvoir explicatif très faible**. Graphiquement, vous pouvez représenter la décomposition de l'inertie sur les axes factoriels



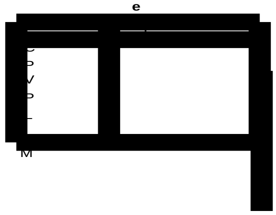


3 facteurs

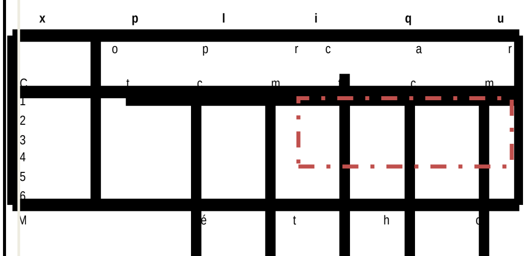
Graphiquement, on peut représenter la décomposition de l'inertie sur les axes factoriels

2.1 NOMBRE DE FACTEURS À EXTRAIRE

Qualité de représentation des variables



Permet de savoir comment les variables sont expliquées par les axes retenus (ici, 3). On peut considérer ces extractions comme étant la somme des contributions des axes à la variable. **Nous constatons qu'elles sont toutes quasiment parfaitement expliquées.**



Algébriquement, la transformation linéaire subie par les données est une décomposition en valeurs singulières.

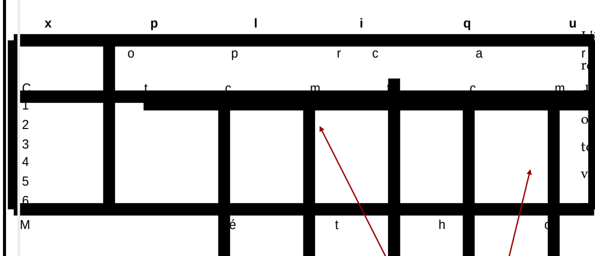
On effectue un changement de base **orthonormée de vecteurs propres**.

La somme des 6 valeurs propres = 6 = nb. de de variables de départ.

C'est également la trace de la matrice des corrélations ie la somme des termes diagonaux de la matrice .

2.1 NOMBRE DE FACTEURS À EXTRAIRE

Qualité de représentation des variables



L'inertie du nuage de points des coordonnées centrées réduites (représentée par la matrice des corrélations) est décomposée en la projetant sur des directions orthogonales. Nous rappelons que nous cherchons toujours à expliquer une variance totale et que chaque valeur propre contribue à expliquer un certain pourcentage.

Par exemple, le premier axe explique 77% de la variance totale . Si on se réfère aux pourcentages cumulés, on remarque que 96.7% de la variance totale est expliquée par les trois premiers axes.

Nous avons donc réussi sur 3 variables la presque totalité de la variabilité de l'échantillon.

2.1 NOMBRE DE FACTEURS À EXTRAIRE

Corrélations reproduites

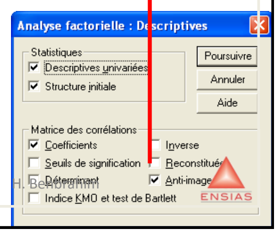
		cylindre	puissanc	vitesse	poids	longueur	largeur
Corrélation reproduite	cylindre	,947[b]	,876	,698	,925	,873	,710
	puissanc	,876	,965[b]	,912	,761	,716	,534
	vitesse	,698	,912	,987[b]	,504	,514	,373
	poids	,925	,761	,504	,980[b]	,872	,692
	longueur	,873	,716	,514	,872	,941[b]	,897
	largeur	,710	,534	,373	,692	,897	,980[b]
Résidu[a]	cylindre		-,015	-,005	-,028	-,009	-,001
	puissanc	-,015		-,018	,008	-,027	,018
	vitesse	-,005	-,018		,003	,017	-,010
	poids	-,028	,008	,003		-,008	,008
	longueur	-,009	-,027	,017	-,008		-,033
	largeur	-,001	,018	-,010	,008	-,033	

Méthode d'extraction : Analyse en composantes principales.

a. Les résidus sont calculés entre la covariance observée et la covariance reproduite. Il y a 0 (,0%) résidu non redondants avec des valeurs absolues supérieures à 0,05.

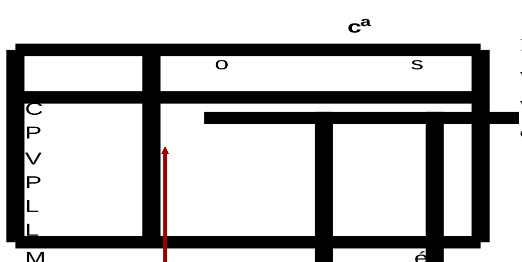
b. Qualité de représentation reproduite

Nombre de facteurs à retenir = 3



ETAPE 3: INTERPRETATION DES FACTEURS

3.1 Matrice des composantes

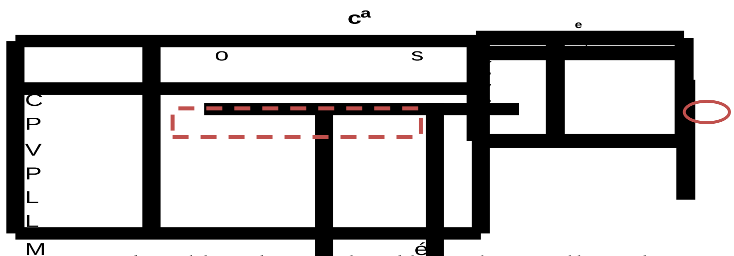


Les coordonnées des variables sur les axes. Nous voyons par quelles variables sont faits les axes. Ces valeurs s'interprètent également comme des corrélations.

On remarque que le premier axe est celui sur lequel les coefficients de corrélation sont les plus élevés. Nous sommes en présence **d'un effet taille**. Nous verrons qu'il existe des **méthodes de rotations** nous permettant de pallier ce problème.

ETAPE 3: INTERPRETATION DES FACTEURS

3.1 Matrice des composantes



Les proportions de variabilité expliquées par le modèle pour chaque variable sont les sommes des carrés des coefficients de corrélation sur les facteurs. Nous rappelons que les variables sont centrées-réduites

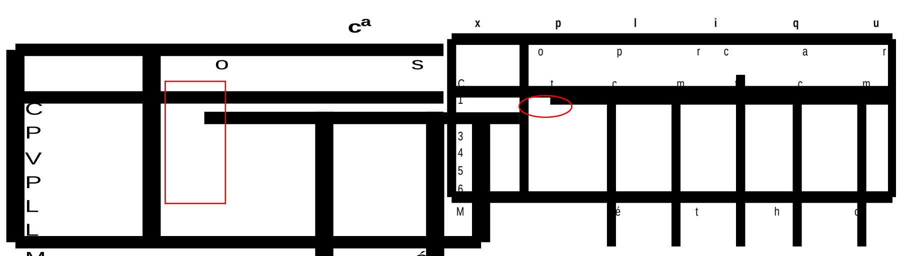
La qualité de représentation de la variable **VITESSE** formée par les trois premiers axes est égale à

$$0.751^2 + 0.618^2 + 0.213^2 = 0{,}987.$$

Définition Algèbre Linéaire ACP ACP sous SPSS Conclusion

ETAPE 3: INTERPRETATION DES FACTEURS

3.1 Matrice des composantes



Nous pouvons retrouver la décomposition des valeurs propres.
Par exemple, la première valeur propre est égale à $\lambda_1 = 4.617$

$$\lambda_1 = 0{,}961^2 + 0{,}905^2 + 0{,}751^2 + 0{,}910^2 + 0{,}920^2 + 0{,}798^2$$

SPSS n'affiche pas les contributions de chaque variable à l'inertie de chaque axe factoriel. Il est possible de les recalculer en divisant le carré des coordonnées par chaque valeur propre. Par exemple, pour $\lambda_1 = 4.617$, nous aurons :

$$CTR(CYLINDRE) = \frac{0.961^2}{4.617} = 0.2$$

ETAPE 3: INTERPRETATION DES FACTEURS

Diagramme de composantes et Rotation des facteurs

Diagramme de composantes

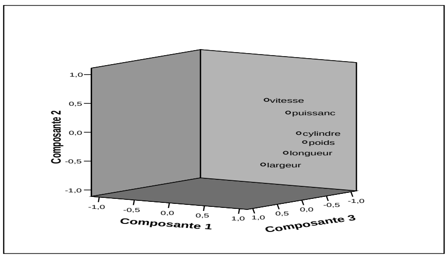
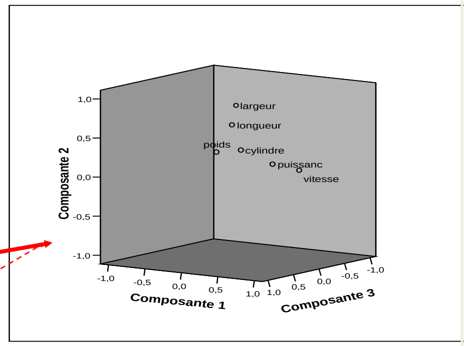
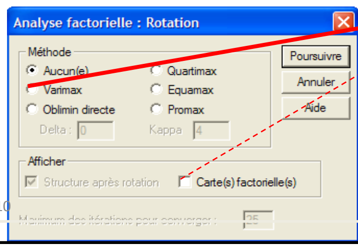


Diagramme de composantes dans l'espace après rotation





ETAPE 3: INTERPRETATION DES FACTEURS

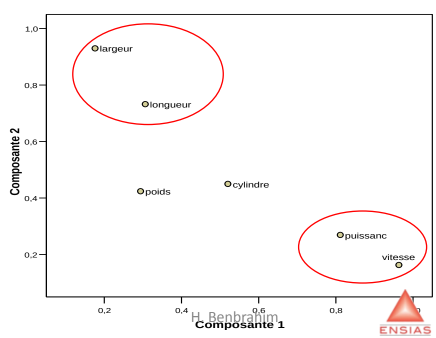
Diagramme de composantes et Rotation des facteurs

```
FACTOR
 /VARIABLES cylindre puissanc vitesse poids
longueur largeur  /MISSING LISTWISE
/ANALYSIS cylindre puissanc vitesse poids
 longueur largeur
 /PRINT INITIAL REPR EXTRACTION ROTATION
 /FORMAT SORT
 /PLOT ROTATION
 /CRITERIA FACTORS(3) ITERATE(25)
 /EXTRACTION PC
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX
 /METHOD=CORRELATION .
```

Pour obtenir des diagrammes à deux dimensions, exécutez la syntaxe

/PLOT ROTATION (1,2) (1,3) (2,3)

Diagramme de composantes dans l'espace après rotation



ETAPE 3: INTERPRETATION DES FACTEURS

Diagramme de composantes et Rotation des facteurs

Diagramme de composantes dans l'espace après rotation

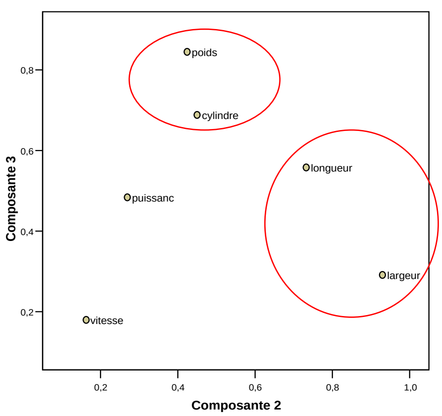
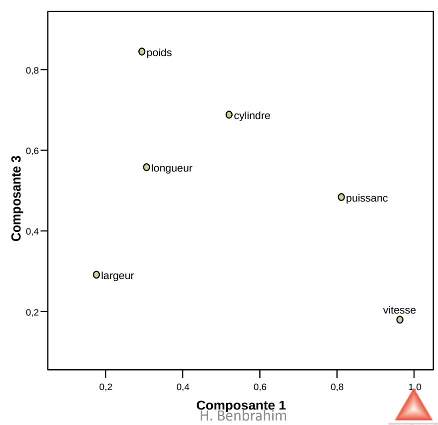


Diagramme de composantes dans l'espace après rotation



ETAPE 3: INTERPRETATION DES FACTEURS

Rotation varimax des facteurs

Matrice des composantes[a]

	Composante		
	1	2	3
cylindre	,961	,011	-,150
longueur	,920	-,303	,054
poids	,910	-,176	-,347
puissanc	,905	,382	-,015
largeur	,798	-,477	,341
vitesse	,751	,618	,203

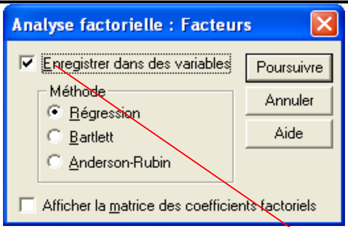
Méthode d'extraction : Analyse en composantes principale
a. 3 composantes extraites.

Matrice des composantes après rotation[a]

	Composante		
	1	2	3
vitesse	,963	,163	,180
puissanc	,811	,269	,484
largeur	,176	,930	,291
longueur	,306	,732	,558
poids	,294	,424	,845
cylindre	,520	,450	,688

Méthode d'extraction : Analyse en composantes principales
Méthode de rotation : Varimax avec normalisation de Kaise
a. La rotation a convergé en 5 itérations.

- La méthode **VARIMAX** s'applique lorsque la plupart des variables sont représentées sur un seul axe. Il s'agit d'une méthode de rotation orthogonale qui minimise le nombre de variables qui ont des corrélations importantes avec un facteur.
- La méthode **QUARTIMAX** s'utilise lorsqu'une variable est fortement corrélée à plusieurs axes à la fois. C'est une méthode de rotation qui minimise le nombre de facteurs requis pour expliquer le nombre de facteurs.
- La méthode **EQUAMAX** est une combinaison des deux méthodes précédentes. Il s'agit d'une méthode de rotation, qui minimise à la fois le nombre de variables qui pèsent fortement sur un facteur et le nombre de facteurs requis pour expliquer une variable.
- La méthode **OBLIMIN** permet d'effectuer des rotations obliques sur les axes factoriels. Les facteurs ne sont plus fortement non corrélés. OBLIMIN permet de mettre en évidence des phénomènes qui déterminent des directions d'allongement non orthogonales des nuages de points.



Facteurs de régression : ils ont une moyenne de 0 et une variance égale au carré de la corrélation multiple entre les facteurs estimés et les valeurs vraies des facteurs.

Facteurs de Bartlett : ils ont une moyenne de 0. La somme des carrés des facteurs uniques sur l'intervalle des variables minimisées.

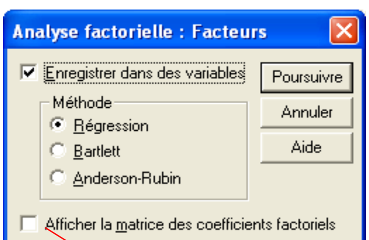
Facteurs d'Anderson-Rubin : ils sont une variante des facteurs de Bartlett permettant d'assurer l'orthogonalité des facteurs estimés. Ils ont une moyenne de 0 et un écart-type de 1.

voiture - Editeur de données SPSS

	puissanc	vitesse	poids	longueur	largeur	FAC1_1	FAC2_1	FAC3_1	var
1	90,00	174,00	850,00	369,00	166,00	-,03615	-,21758	-1,2941	
2	103,00	189,00	805,00	370,00	157,00	,68780	-1,37947	-1,0803	
3	100,00	181,00	925,00	363,00	161,00	,23323	-1,14951	-,60655	
4	95,00	184,00	730,00	350,00	160,00	,58000	-1,01621	-1,5902	
5	92,00	180,00	965,00	415,00	169,00	-,10875	,37307	-,84603	
6	83,00	170,00	970,00	395,00	170,00	-,49179	,36933	-,86907	
7	90,00	180,00	1080,00	440,00	169,00	-,35431	,43229	-,29506	
8	88,00	180,00	1135,00	446,00	170,00	-,47230	,45435	,13860	
9	90,00	182,00	1060,00	424,00	168,00	-,21714	,16714	-,34884	
10	122,00	190,00	1255,00	473,00	177,00	-,02754	1,37882	-,16955	
11	125,00	194,00	1120,00	439,00	171,00	,43169	,44826	-,40725	
12	115,00	185,00	1190,00	451,00	172,00	-,10282	,59683	,07129	
13	120,00	177,00	1265,00	436,00	177,00	-,38813	,97305	,11545	
14	87,00	144,00	1430,00	436,00	169,00	-2,02274	-,29178	2,13637	
15	112,00	149,00	1320,00	457,00	184,00	-1,53868	2,01778	,44846	
16	160,00	214,00	1220,00	439,00	169,00	1,34121	-,15225	-,08511	

Affichage des données / Affichage des variables

SPSS processeur est prêt



Matrice des composantes après rotation[a]

	Composante		
	1	2	3
vitesse	,963	,163	,180
puissanc	,811	,269	,484
largeur	,176	,930	,291
longueur	,306	,732	,558
poids	,294	,424	,845
cylindre	,520	,450	,688

Méthode d'extraction : Analyse en composantes principales
Méthode de rotation : Varimax avec normalisation de Kaise

a. La rotation a convergé en 5 itérations.

Toute les variables varient dans le même sens que les axes factoriels

Matrice des coefficients des coordonnées des composantes

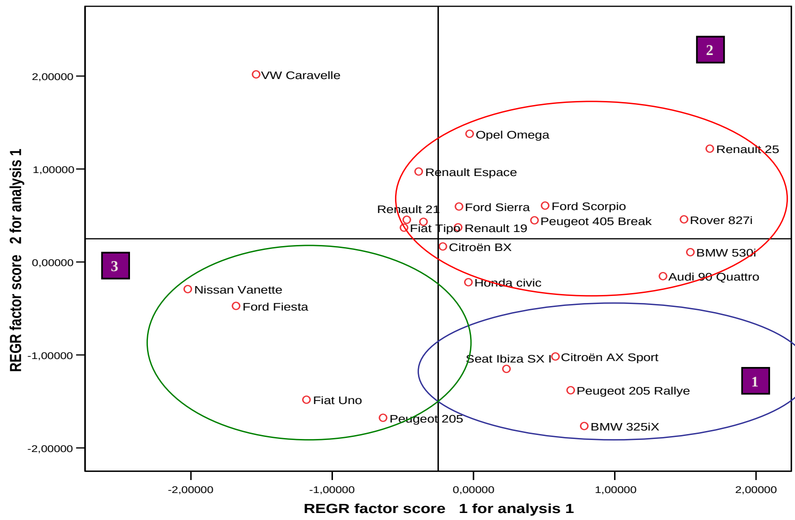
	Composante		
	1	2	3
cylindre	-,001	-,169	,511
puissanc	,437	-,182	,080
vitesse	,814	,042	-,551
poids	-,342	-,411	1,049
longueur	-,107	,422	,038
largeur	-,029	1,057	-,682

Méthode d'extraction : Analyse en composantes principales
Méthode de rotation : Varimax avec normalisation de Kaiser
Scores composante.

Groupe 1 : véhicules de petites tailles, puissants et rapides (type sport)

Groupe 2 : véhicules de tailles élevées, puissants et rapides (type berline)

Groupe 3 : véhicules lents, peu puissants.



ENSIAS

Section I: Introduction et Définition

Section II: Rappel: Algèbre Linéaire

Chapitre III: Analyse Factorielle: Analyse en Composantes Principales

Section III: ACP

Section IV: ACP sous SPSS

Section V: Récapitulatif et Conclusion

RECAPUTILATIF?

CONCLUSION?