

Variance et écart-type

- Définition  
la variance de  $X$  est définie par
$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

L'écart type  $\sigma$ , est la racine

  - Propriétés

**Variance et écart-type**

La variance de  $X$  est définie par

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

ou  $\text{Var}(X) = \sum p_i (x_i - \bar{x})^2$

L'écart type  $\sigma$ , est la racine carrée de l' :

La variance satisfait la formule suivante

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ». L'écart-type, qui a la même unité que  $x$ , est une mesure de dispersion.

## La matrice des poids

- Pourquoi : utile quand les individus n'ont pas la même importance
  - Comment? on associe aux individus un poids  $p_i$  tel que
$$p_1 + p_2 + \dots + p_n = 1$$

et on représente ces poids dans la matrice diagonale de tailles

$$D = \begin{bmatrix} p_1 & & & \\ \vdots & p_2 & & \\ & \ddots & \vdots & \\ 0 & & & p_n \end{bmatrix}$$

- Cas uniforme tous les individus ont le même poids  $\mu_I = I/n$  et  $D = I/n$

## Interlude : notation matricielle

**Matrice**  
tableau de données carre ou rectangulaire

- Vecteur matrice à une seule colonne.
  - Cas particuliers
  - Transposition de matrice  
échange des lignes et des colonnes d'une matrice ; on note  $M'$  la transposée de  $M$ .
$$I = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \quad I = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

BIBLIOTHEQUE  
1990

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

Forme générale de données

- Tableau sous forme d'une matrice formée par p variables statistiques notée par  $X_1, X_2, \dots, X_p$  et un échantillon  $x_1, x_2, \dots, x_n$  de taille n.
  - Pour n individus et p variables, on a le tableau X est une matrice rectangulaire à n lignes et p colonnes

et on représente ces poids dans la matrice diagonale de taille  $n$

$$p_1 + p_2 + \dots + p_n = 1$$

tous les individus ont le même poids  $p_j = 1/n$  et  $D = 1/n$

où  $x_i$  représente la valeur de  $X_i$  prise par l'individu  $i$ .

monotonus

٦٧

## Vecteurs variable et individu

- La Variable  $X_j$  : Une colonne du tableau

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \in R^n$$

où:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

- L'individu  $x_i$  : Une ligne du tableau

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in R^p$$

- Les  $n$  individus sont décrits par un nuage de  $p$  variables.  
L'information représentée par un nuage correspond à la dispersion des  $n$  points.

L'individu  $x_i$  : Une ligne du tableau  
On cherche à représenter le nuage des individus.

Exemple: cas  $p=3$  et  $N=10$ .

Tableau centré: il est obtenu en centrant les variables autour de leur moyenne:

$$\bar{x}_i = x_{ij} - \bar{x}_j$$

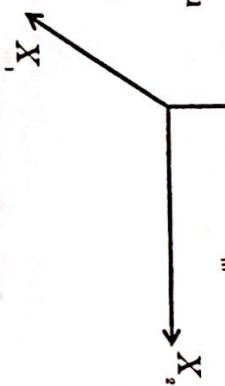
- Point moyen: c'est le vecteur  $g$  des moyennes arithmétiques de chaque variable:

$$g = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T \in R^p$$

On peut aussi écrire:

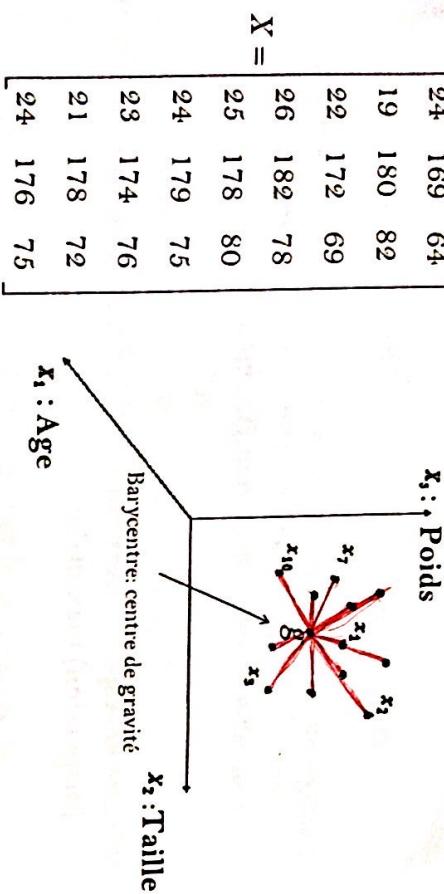
$$g = X^T D_1$$

- A chaque individu  $x_i$ , on peut associer un point dans  $R^p$  = espace des individus.
- Les axes de ce sous-espace de dimension réduite sont dits "axes factoriels".
- La figure suivante présente le nuage de points correspondant aux  $n$  individus.
- A chaque variable  $X_j$  du tableau est associé un axe de  $R^p$ .



Individu	AGE: X <sub>1</sub> (ans)	TAILLE: X <sub>2</sub> (cm)	POIDS: X <sub>3</sub> (kg)
1	25	169	64
2	24	180	82
3	23	172	69
4	24	175	68
5	26	182	78
6	25	178	80
7	24	179	75
8	23	174	76
9	26	178	72
10	24	176	75

## Centre d'inertie et tableau centre



- La dispersion du nuage de points présente l'information de l'échantillon
- Le centre de gravité a comme coordonnées  $\bar{g}(\bar{x}_1, \bar{x}_2, \bar{x}_3)$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\Sigma = \begin{bmatrix} V(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_p, X_1) & \text{cov}(X_p, X_2) & \dots & V(X_p) \end{bmatrix}$$

avec

$$\text{cov}(X_i, X_j) = \frac{1}{n} \sum_{t=1}^n (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)$$

On peut montrer que:

$$\Sigma = \frac{1}{n} X' X$$

## Matrice de données

- Pour une matrice de données on a trois types d'analyses:
- Analyse univariée: On traite (étudie) chaque variable indépendamment des autres (statistique descriptive).
- Analyse bivariée: traitement de deux variables X et Y dépendantes.
- Analyse multivariée: Traitement de l'information de p variables statistiques  $X_1, X_2, \dots, X_p$  présentées dans une matrice de données.



## Mesure de liaison entre deux variables

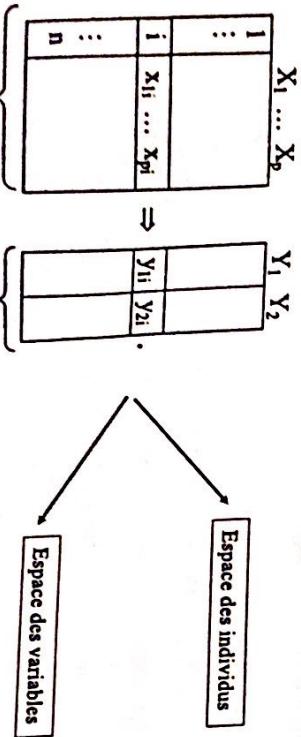
- le coefficient de corrélation entre deux variables X et Y est donnée par

$$\rho_{1,2} = \frac{\text{cov}(X_1, Y)}{\sigma_{X_1} \sigma_Y}$$

- On a toujours (inégalité de Cauchy)  $-1 \leq \rho_{1,2} \leq 1$
- Si  $|\rho_{1,2}| > 0.7$ , les variables X et Y sont fortement corrélées.
- Si  $|\rho_{1,2}| < 0.5$ , les variables X et Y sont faiblement corrélées.
- Si  $|\rho_{1,2}| = 0$ , les variables X et Y sont non corrélées.

## ACP: Analyse en Composantes Principales

- Définition L'analyse en composantes principales (ACP) est une analyse multivariée de traitement de  $p$  variables dépendantes.
  - Principe On cherche une représentation des  $n$  individus, dans un sous-espace  $R'$  de  $R^p$  de dimension  $k$  ( $k$  petit : 2, 3 ... ; par exemple un plan).
- Autrement dit, on cherche à définir  $k$  nouvelles variables combinaisons linéaires des  $p$  variables initiales qui feront perdre le moins d'information possible.
- Ces variables seront appelées «composantes principales».
  - Les axes qu'elles déterminent : « axes principaux ».
  - Les formes linéaires associées : « facteurs principaux ».



- L'objectif de la méthode ACP est de construire ces nouveaux axes ( $Y_1, Y_2, \dots, Y_p$ ) qui sont des combinaisons linéaires des variables origines. Chaque axe factoriel peut être exprimée par la formule suivante :

$$Y_h = \sum_{j=1}^p a_{hj} X_j, \quad \text{pour } h = 1, 2, \dots, k$$

### Choix d'une distance

- La distance utilisée par l'ACP est la distance euclidienne.
- La distance entre deux points est égale à :

$$d^2(x_\ell, x_h) = \sum_{j=1}^p (x_{\ell j} - x_{hj})^2, \quad \ell, h \in \{1, \dots, n\}.$$

- Réduire les dimensions du tableau initial nécessite le calcul de distances entre les éléments de ce tableau.
- Pour qualifier (définir un indicateur) d'information du nuage de points on calcule l'inertie totale.

## Perdre le moins d'information possible

1. Le sous-espace obtenu devra être « ajusté » le mieux possible au nuage des individus: la somme des carrés des distances des individus au sous-espace doit être minimale.



2. Le sous-espace obtenu sur lequel le nuage projeté ait une inertie (dispersion) maximale.

1 et 2 sont basées sur les notions de :

- distance
- projection orthogonale

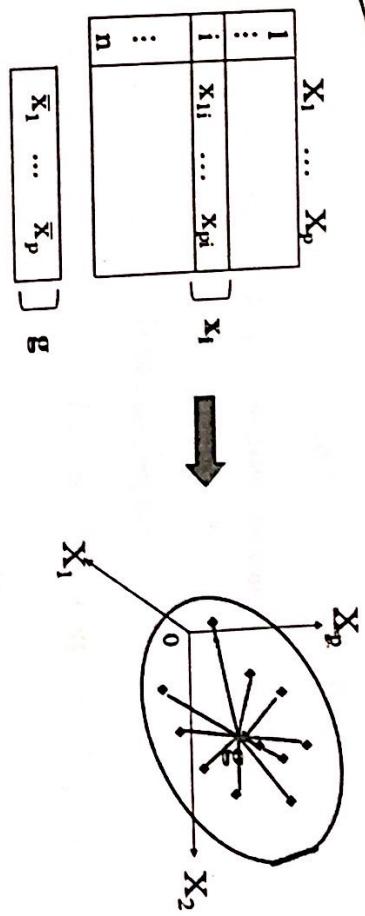
## Inertie totale du nuage des individus

- On note  $I_g$  le moment d'inertie du nuage des individus par rapport au centre de gravité

$$I_g = \frac{1}{n} \sum_{i=1}^n d^2(x_i, g)$$

- Ce moment d'inertie totale est intéressant puisqu'il nous permet de mesurer la dispersion du nuage des individus par rapport à son centre de gravité.

- Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé,
- Si ce moment d'inertie est petit, alors le nuage est très concentré sur son centre de gravité.



## Normalisation des données :

- Pour neutraliser le problème des unités on remplace les données d'origine par les données centrées-réduites. Dans ce cas nous procéderons par une étape de centralisation et de réduction des données.
- La première phase de centralisation se fait par rapport au centre de gravité.
- La deuxième étape consiste à réduire les données en divisant chaque colonne de la matrice par l'écart type correspond à la variable. La matrice de donnée obtenue sera utilisée pour déterminer les axes factoriels.

## Détermination de la première composante principale: $Y_1$

- La forme de  $Y_1$  est:

$$Y_1 = \sum_{j=1}^p \alpha_{1j} X_j$$

- La variance de  $Y_1$  est:

$$\text{Var}(Y_1) = \alpha_1^T \Sigma \alpha_1,$$

où:  $\alpha_1 = (\alpha_{11}, \dots, \alpha_{1p})^T$ .

- On a aussi  $\text{Var}(Y_1) \leq \sum_{j=1}^p \text{Var}(X_j) = I_g$

$\Rightarrow$  Trouver  $Y_1$  revient à chercher  $\alpha_1$  sous les contraintes:

$$\|\alpha_1\| = \alpha_1^T \alpha_1 = 1 \quad \text{et} \quad \text{Var}(Y_1) \quad \text{soit maximale}$$

## Problème de maximisation

- Sans contrainte sur  $\alpha_i$ , la maximisation de  $\alpha_i^T \Sigma \alpha_i$  n'aurait aucun sens.
- Définissons la fonction de Lagrange:

$$F(\alpha_i, \lambda) = \alpha_i^T \Sigma \alpha_i - \lambda(\alpha_i^T \alpha_i - 1)$$

avec  $\lambda \in \mathbb{R}$  est le multiplicateur de Lagrange.

- Le problème revient à maximiser  $F(\alpha_i, \lambda)$ .
- La solution du problème s'obtient en dérivant par rapport à

$$\lambda, \alpha_{i1}, \alpha_{i2}, \dots, \alpha_{ir}$$

si et seulement si:

$$\Sigma \alpha_i = \lambda \alpha_i$$

## Ce qu'il faut retenir sur $Y_1$

Donc:

$$\text{Var}(Y_1) = \alpha_i^T \Sigma \alpha_i = \alpha_i^T \lambda \alpha_i = \lambda.$$

Ainsi pour que  $\text{Var}(Y_1)$  soit maximale il faut prendre  $\lambda$  la plus grande valeur propre de  $\Sigma$ .

**Conclusion:** Pour que  $\text{Var}(Y_1)$  soit maximale, il faut donc prendre

- $\lambda = \lambda_1$ , la plus grande valeur propre de  $\Sigma$ ;
- $\alpha_i$ , le vecteur propre normé correspondant.

## Dérivation

- La contrainte  $\|\alpha_i\| = 1$  est prise en compte par

$$\frac{\partial F}{\partial \lambda}(\alpha_i, \lambda) = 1 - \alpha_i^T \alpha_i = 0.$$

- Notons:  $\frac{\partial F}{\partial \alpha_i}(\alpha_i, \lambda) = \left( \frac{\partial F}{\partial \alpha_{i1}}, \dots, \frac{\partial F}{\partial \alpha_{ir}} \right)^T$ .
- On se convainc sans difficulté que:

$$\frac{\partial F}{\partial \alpha_i}(\alpha_i, \lambda) = 2 \Sigma \alpha_i - 2 \lambda \alpha_i = 0,$$

## Détermination de la deuxième composante principale: $Y_2$

- La forme de  $Y_2$  est:

$$Y_2 = \sum_{j=1}^r \alpha_{j2} X_j$$

- Trouver  $Y_2$  revient à chercher  $\alpha_2$  sous les contraintes :  $\alpha_2^T \alpha_2 = 1$ ,  $\text{cov}(Y_1, Y_2) = 0$  et  $\text{Var}(Y_2)$  soit maximale
- La fonction à maximiser est:

$$F(\alpha_2, \lambda, k) = \alpha_2^T \Sigma \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - k(\alpha_2^T \alpha_2 - 0)$$

Détermination de  $\alpha_2$ : On cherche le vecteur  $\alpha_2$  qui maximise  $F(\alpha_2, \lambda, k)$

- On a:

$$\text{cov}(Y_1, Y_2) = \text{cov}(\alpha_1^T X, \alpha_2^T X) = \alpha_1^T \Sigma \alpha_2 = \alpha_2^T \Sigma \alpha_1 = \lambda_1 \alpha_2^T \alpha_1$$

## Utilité des multiplicateurs:

- D'une part,  $\alpha_i$  est normé puisque:

$$\frac{\partial F}{\partial \lambda}(\alpha_i, \lambda, k) = 1 - \alpha_i \alpha_i^T = 0.$$

- D'autre part  $\alpha_i$  et  $\alpha_j$  sont linéairement indépendants car:

$$\frac{\partial F}{\partial k}(\alpha_i, \lambda, k) = -\alpha_i^T \alpha_i = -\alpha_i^T \alpha_j = 0.$$

- Autres dérivées partielles:

$$\frac{\partial F}{\partial \alpha_i}(\alpha_i, \lambda, k) = \left( \frac{\partial F}{\partial \alpha_{i1}}, \dots, \frac{\partial F}{\partial \alpha_{in}} \right)^T.$$

- Un calcul simple montre que

$$\frac{\partial F}{\partial \alpha_i}(\alpha_i, \lambda, k) = 2\sum \alpha_i - 2\lambda \alpha_i - k \alpha_i = 0$$

## Choix de $\alpha_2$ :

- Puisque:

$$\text{Var}(Y_i) = \alpha_i^T \Sigma \alpha_i = \alpha_i^T \lambda \alpha_i = \lambda$$

- Celle-ci est maximale si  $\lambda = \lambda_1$ , la deuxième plus grande valeur propre de  $\Sigma$ .

Ainsi,  $\alpha_2$  est le vecteur propre normé correspondant.

## Utilité des multiplicateurs:

- De plus,  $k=0$  (voir ci-dessous), de sorte que :

$$\sum \alpha_i = \lambda \alpha_i.$$

- Ainsi,  $\lambda$  est une autre valeur propre de  $\Sigma$  !

- Pourquoi  $k=0$ ? En multipliant l'équation précédente par,  $\alpha_i^T$

on trouve:

$$2\alpha_i^T \Sigma \alpha_i - 2\lambda \alpha_i^T \alpha_i - k \alpha_i^T \alpha_i = 0.$$

Mais

$$\alpha_i^T \Sigma = \lambda_i \alpha_i^T, \quad \alpha_i^T \alpha_i = 0 \quad \text{et} \quad \alpha_i^T \alpha_i = 1$$

d'où:

$$2(\lambda_i - \lambda k) \alpha_i^T \alpha_i - k = -k = 0.$$



## Généralisation

- Procédant par maximisations successives, on conclut que:

$$Y_k = k^{\text{ème}} \text{ composante principale} = \alpha_i^T X.$$

où:  $\alpha_i$  est le vecteur propre normé associé à  $\lambda_i$  avec:

$$\lambda_1 < \lambda_{i-1} < \dots < \lambda_i < \lambda_i.$$

- On note  $\Lambda$  la matrice carrée de taille  $p$ :

$$\Lambda = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_2 & \alpha_3 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_n & \alpha_n & \cdots & \alpha_n \end{pmatrix}$$

- La matrice  $\Lambda$  est orthogonale et a pour colonnes les vecteurs propres de  $\Sigma$  telle que vu précédemment, on a donc

$$\Lambda \Lambda^T = \Lambda^T \Lambda = I_p, \quad \Lambda^T = \Lambda^{-1} \quad \text{et} \quad \Sigma \Lambda = \Lambda \Lambda$$

## Ecriture matricielle

- Soit la sous-matrice de  $A$  de taille  $(p, k)$  notée encore  $A'$  définie par

$$A' = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rn} \end{pmatrix}$$

- Pour définir simultanément et de façon plus compacte les composantes principales, on pose:
- $$Y = (Y_1, Y_2, \dots, Y_k)^T = A'X = A'(X_1, X_2, \dots, X_r)^T$$

$$\forall i \in \{1, \dots, k\}: Y_i = \sum_j \alpha_{ij} X_j \Rightarrow \forall j \in \{1, \dots, p\}: X_j = \sum_i \alpha_{ij} Y_i$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r1} & x_{r2} & \cdots & x_{rr} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rn} \end{pmatrix} = XA$$

- Nous avons aussi

$$\|X_j\| = \sqrt{\alpha_{j1}^2 + \alpha_{j2}^2 + \cdots + \alpha_{jr}^2} \leq 1$$

## Nombre d'axes à retenir

- D'après ce qui précède on a:

$$Tr(\Sigma) = Tr(\Lambda) = \sum_{i=1}^r \lambda_i$$

est une mesure globale de variation.

- La contribution relative de l'axe  $Y_l$  à l'inertie totale du nuage des individus est

$$\frac{Var(Y_l)}{I_g} = \frac{\lambda_l}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

ce qui équivaut en pourcentage

$$C_l = \frac{\lambda_l}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \times 100$$

- Les critères les plus utilisés sont les suivants:

- La nouvelle matrice de données  $Y$  dans la nouvelle base  $(Y_1, Y_2, \dots, Y_k)$  est donnée par:

- Ainsi  $\forall j \in \{1, \dots, p\}$ , la norme de  $X_j$  dans la base  $(Y_1, Y_2, \dots, Y_k)$  vérifie:

## Premier critère :

➤ Ce critère consiste à calculer les contributions des axes à l'inertie totale. Dans ce cas, le nombre de facteurs qu'on doit retenir est égale au nombre de valeurs dont la somme de leurs contributions dépasse 80%. Pourquoi 80% ? C'est purement arbitraire !

- Si  $C_1 \geq 80\%$ , alors  $k=1$ .

- Sinon, si  $C_1 + C_2 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_p} \times 100 \geq 80\%$ , alors  $k=2$ .

- Ainsi l'entier naturel  $k$  est le plus petit entier non nul tel que:

$$C_1 + C_2 + \cdots + C_k = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \times 100 \geq 80\%$$

## Deuxième critère : critère de Kaiser (1960)

- Ce critère consiste à retenir que les valeurs propres qui dépassent la valeur 1. Le nombre de ces valeurs propres détermine le nombre d'axes factoriels.
- Si l'ACP est effectuée sur la matrice des corrélations,

Garder  $y_i \Leftrightarrow \lambda_i \geq 1$

Ainsi  $k$  sera le nombre des valeurs propres qui sont  $\geq 1$ .

Conclusion:

On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les  $k$  premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse à un sous-espace de dimension  $k < p$ .

## Procédure de l'ACP

- On cherche  $X^T$  la transposée de la matrice  $X$ .
- On détermine les valeurs propres de la matrice symétrique  $X^T X$ .
- Soient  $\lambda_1, \lambda_2, \dots, \lambda_p$  ces valeurs propres.
- On classe  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \dots$
- Alors

$$X^T X = A \Lambda A^{-1} \quad \text{où:} \quad \Lambda = \begin{bmatrix} \lambda_1 & & \dots & 0 \\ & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \lambda_p \end{bmatrix}$$

- D'après les propriétés de la trace des matrices: on a:

$$\text{Tr}\left(\frac{1}{n} X^T X\right) = \text{Tr}(A \Lambda A^{-1}) = \text{Tr}(A^{-1} \Lambda A) = \text{Tr}(\Lambda)$$

- En raison des valeurs numériques  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \dots$  la somme des premières valeurs propres représente souvent une proportion importante de trace de  $X^T X$ .

$$d''(x, g) = d''(y, g) + d''(x, y)$$

- On en déduit la formule de décomposition donnée par l'expression suivante :

$$\frac{1}{n} \sum_{i=1}^M d''(x_i, g) = \frac{1}{n} \sum_{i=1}^M d''(y_i, g) + \frac{1}{n} \sum_{i=1}^M d''(x_i, y_i)$$

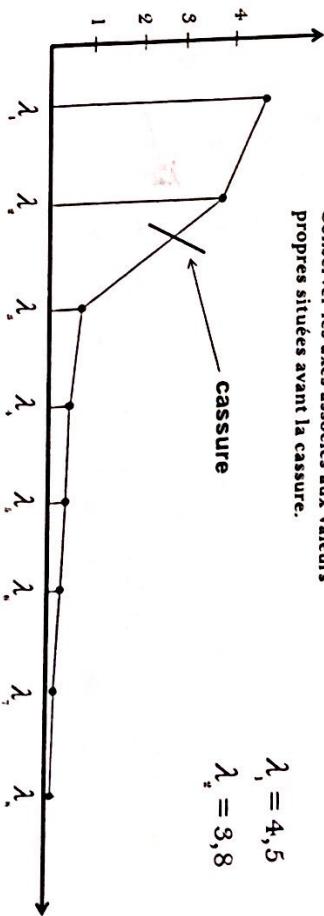
## Troisième critère :

- On traite la distribution des valeurs propres pour détecter une décroissance brutale de l'inertie permettant de considérer les axes résiduels comme négligeables. Pour cela on observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du premier point d'inflexion.

Conserver les axes associés aux valeurs propres situées avant la cassure.

$$\lambda_1 = 4,5$$

$$\lambda_2 = 3,8$$

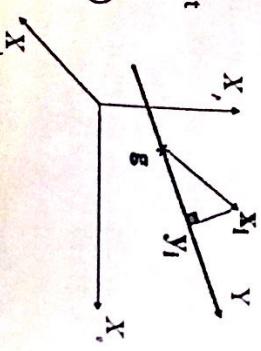


## Décomposition de l'inertie totale:

- En présence de l'axe de projection  $Y$  on peut décomposer l'inertie totale  $I_g = I(n, g)$  en deux parties :

- Le schéma ci-dessous montre la projection du point  $x_i$  sur l'axe  $Y$  où  $y_i$  est la projection orthogonale de  $x_i$  sur l'axe  $Y$ .

- Notons par  $\{x_1, x_2, \dots, x_n\}$  le nuage de points et  $\{y_1, y_2, \dots, y_n\}$  leurs projections sur  $Y$ . En utilisant la formule suivante:



## Décomposition de l'inertie totale :

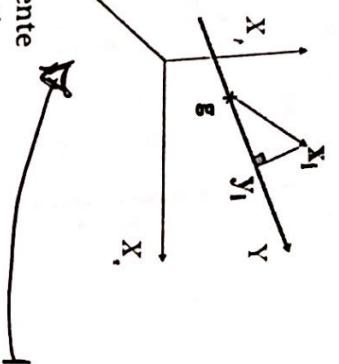
- L'inertie résiduelle du nuage des individus par rapport à un axe Y passant par  $\mathbf{g}$  est:

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

- Donc on aura:

$$I_s(x_1, \dots, x_n) = I_s(y_1, \dots, y_n) + I(n, Y)$$

- La première partie de la décomposition représente l'inertie expliquée par l'axe  $Y$  et la deuxième partie l'inertie résiduelle.



## Mise en œuvre de l'axe $Y_1$ :

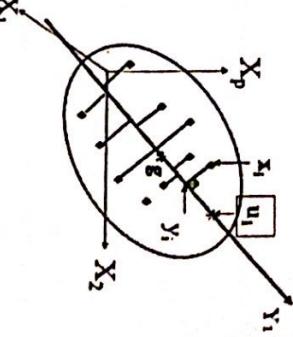
- L'objectif 1 : On cherche l'axe  $Y_1$  passant le mieux possible au milieu du nuage  $n$ . On cherche en premier lieu à minimiser l'inertie du nuage par rapport à l'axe  $Y_1$ , c.à.d. minimiser la quantité  $I(n, Y_1)$ .
- L'objectif 2 : On cherche l'axe d'allongement  $Y_1$  du nuage  $n$ . On cherche en deuxième lieu à maximiser l'inertie du nuage  $n$  projeté sur l'axe  $Y_1$ .
- L'axe factoriel  $Y_1$ , vérifiant ces deux contraintes est déterminé par les résultats suivants :
  - L'axe  $Y_1$  passe par le centre de gravité  $\mathbf{g}$  du nuage de points  $n$ .
  - L'axe  $Y_1$  est engendré par le vecteur normé  $u_1$ , vecteur propre de la matrice des corrélations associé à la plus grande valeur propre  $\lambda_1$ .
  - L'inertie expliquée par l'axe  $Y_1$  est égal à  $\lambda_1$ .
  - La part d'inertie expliquée par le premier axe principal  $Y_1$  est égal à  $\frac{\lambda_1}{I_g}$ .

## Mise en œuvre de l'axe $Y_1$ :

- Géométriquement, on cherche un axe  $Y_1$  passant par  $\mathbf{g}$  d'inertie  $I(n, Y_1)$  minimum, c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du nuage.

Si on utilise la relation entre les inerties données au paragraphe précédent, rechercher  $Y_1$  tel que  $I(n, Y_1)$  est minimum, est équivalent à chercher  $Y_1$  tel que  $I_g$  est maximum. Donc deux objectifs que l'axe  $Y_1$  doit vérifier :

Pour que  $Y_1$  passe par  $\mathbf{g}$



## Recherche des axes suivants:

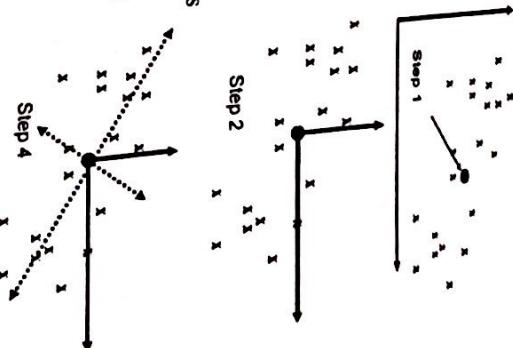
- On cherche ensuite un deuxième axe  $Y_2$  orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe  $Y_2$  passant par  $\mathbf{g}$ . De même L'axe  $Y_2$  est engendré par le vecteur normé  $u_2$ , vecteur propre de la matrice des corrélations associé à la grande valeur propre  $\lambda_2$  qui suit la plus grande. Par conséquent, la détermination des axes factoriels sont hiérarchisés de sorte que:
  - le 1er axe concentre le maximum de l'information;
  - C'est l'axe qui explique la plus grande variabilité du nuage de points (inertie expliquée) dans un espace à une dimension sauf qu'il laisse des résidus (de l'information)

## Recherche des axes suivants:

- ❖ Le 2<sup>ème</sup> axe concentre le maximum de l'information restante :
  - C'est un axe orthogonal au premier (par construction)
  - Il est de la plus grande dimension résiduelle du nuage de points
  - Il est associé au 1er axe, c'est le meilleur résumé dans un espace à deux dimensions.
  - Il laisse aussi des résidus.
- ❖ Le 3<sup>ème</sup> axe prend encore une part d'information moindre ;
  - C'est un axe orthogonal au deux premiers (toujours par construction)
  - Ainsi de suite ..

## Mise en œuvre d'une ACP:

- Etape 1: Calculer la moyenne de chaque vecteur de caractéristiques.
- Etape 2: Soustraire la moyenne de chaque vecteur de caractéristiques.
- Etape 3: Calculer la matrice des covariances  $\Sigma = (\text{Cov}(X_i, X_j))$
- Etape 4: Calculer les valeurs et vecteurs propres de la matrice de covariance.
- Etape 5: Ne conserver que les valeurs propres (+ vecteurs) les plus grandes.
- Etape 6: Projeter les données dans ce nouvel espace propre.



## Représentation des individus dans les nouveaux axes:

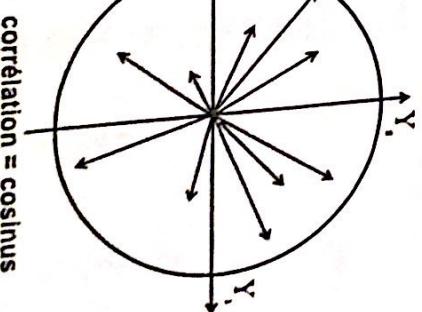
- Pour faire la représentation des individus dans le plan défini par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir les coordonnées  $y_{ij}$  de l'individu  $x_i$  sur l'axe  $Y_j$ , on projette orthogonalement le vecteur  $x_i$  sur cet axe.

## Représentation des variables

- Le cercle des corrélations est la projection du nuage des variables sur le plan des composantes principales.

- Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

- On obtient alors ce que l'on appelle communément le " cercle des corrélations", un dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et +1.



## Mise en pratique de l'ACP pour les

## données infarct

- Nous appliquons la méthode ACP sur des données représentant les informations réelles. Les données sont présentées sous format du fichier SPSS.

- SPSS  « (Statistical Package for the Social Sciences) est un logiciel permettant de réaliser la totalité des analyses statistiques habituellement utilisées en sciences humaines. C'est un logiciel robuste complet.

- Il existe bien d'autres logiciels comme Matlab, S-Plus, R ou SAS qui permettent d'atteindre les mêmes buts, c'est-à-dire faire des analyses statistiques.

Importation des données en SPSS

- Les données sont constituées par un tableau récapitulant l'information d'un certain nombre de patients classées par une variable nommée « pronostic » décrivant l'état du patient par rapport à l'infarctus.

- Le fichier SPSS infarctus.sav comporte les variables suivantes

- Le tableau ci-dessous présente les données infarctus.sav tel que la variable prono indique la classe de chaque observation et la variable prono1 est son codage en format numérique.

## Analyse descriptive des données

- Les données sont importées dans le logiciel SPSS sous format d'un tableau où chaque ligne représente un enregistrement (une passion) et la colonne représente une caractéristique (variable statistique). L'utilisation du logiciel SPSS a abouti aux résultats suivants.

## Analyse descriptive des données

Statistiques descriptives

	N	Méimum	Maximum	Moyenne	Ecart type
Fréquence Cardiaque	101	51	135	92,16	14,28
Index Cardiaque	101	.60	.37	1,0457	.45902
Index Sytologique	101	6,2	6,0	20,818	1,6130
Présion Diastolique	101	8,0	34,0	10,258	5,8093
Présion Artériale Pulmonaire	101	10,0	46,0	26,000	7,3228
Présion Veinale	101	1,0	20,0	9,500	4,3411
Résistance Pulmonaire	101	345,0	5067,0	1324,058	741,3438
N.vale (Intra)	101				

## LICENCE D'EXCELLENCE MAM

### Test de sphéricité de Bartlett.

- L'idée sous-jacente à ces indicateurs est la suivante : est-ce qu'il est possible d'obtenir un bon résumé ?

- En effet, on peut considérer l'ACP comme une compression de l'information. Elle n'est possible que si les données présentent une certaine redondance.
- Si les variables sont parfaitement corrélées, un seul axe factoriel suffit, il restituera 100% de l'information disponible.

- A l'inverse, si elles sont deux à deux orthogonales, a fortiori si elles sont deux à deux indépendantes, le nombre adéquat de facteurs à retenir est égal au nombre de variables.

- Dans ce dernier cas, la matrice de corrélation – impliquée dans le calcul de la solution – est la matrice unité (ou matrice identité).

- Le test de sphéricité de Bartlett propose une mesure globale en s'appuyant sur une démarche statistique. Il est basé sur le déterminant d'une estimation de la matrice de corrélation.

- Il vise à détecter dans quelle mesure la matrice de corrélation  $\Gamma = (\sigma_{ij})$  calculée sur nos données (matrice observée) diverge significativement de la matrice unité (matrice théorique sous hypothèse nulle  $H_0$ ).

## Validité de l'application de la méthode ACP:

- Normalement l'application de la méthode a pour objectif de réduire le nombre de variables qui sont corrélées en un nombre de facteur minimale.
- De plus les nouveaux facteurs sont non corrélés et permet d'éliminer l'information redondante.

- L'analyse de la matrice de corrélations introduite dans le premier chapitre permet de justifier la finalité d'application de la méthode ACP et porte sur le calcul des coefficients de corrélations (analyse bivariée entre différentes variables) donnée par la formule suivante :

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

	N	Méimum	Maximum	Moyenne	Ecart type
Fréquence Cardiaque	101	51	135	92,16	14,28
Index Cardiaque	101	.60	.37	1,0457	.45902
Index Sytologique	101	6,2	6,0	20,818	1,6130
Présion Diastolique	101	8,0	34,0	10,258	5,8093
Présion Artériale Pulmonaire	101	10,0	46,0	26,000	7,3228
Présion Veinale	101	1,0	20,0	9,500	4,3411
Résistance Pulmonaire	101	345,0	5067,0	1324,058	741,3438
N.vale (Intra)	101				

## LICENCE D'EXCELLENCE MAM

### Énoncé:

- Partant d'un échantillon de  $n$  individus d'un ensemble de  $p$  variables aléatoires réelles  $X_1, X_2, \dots, X_p$ , le test concerne la validité de

- $(H_0)$  : (hypothèse nulle) : les variables sont globalement indépendantes.
  - $(H_1)$  : les variables sont globalement dépendantes.
- En se basant sur une estimation de la matrice de corrélation  $\Gamma$ , le test évalue

$$\chi^2 = - \left( n - 1 - \frac{2p + 5}{6} \right) \ln |\det \Gamma|$$

qui, sous  $(H_0)$ , suit «approximativement» une loi de  $\chi^2$  disposant de  $\frac{p(p+1)}{2}$  degrés de liberté, i.e:

$$f_{\chi^2}(x) = \frac{1}{2^{\frac{p(p+1)}{4}}} x^{\frac{p(p+1)}{4}-1} e^{-x/2}, \quad x \in \mathbb{R}^+$$

- On définit la P-value par:

$$P_{\text{value}} = \mathbb{P}(H_0 \text{ est vraie} / H_1 \text{ est fausse})$$

- Si  $P_{\text{value}} \leq 0,05$  ( $= \alpha$  le seuil de test), alors  $(H_1)$  est fausse.
- Si  $P_{\text{value}} > 0,05$ , on peut rien conclure !

## LICENCE EXCELLENCE MATH Indice KMO global

Matrice des corrélations partielles:

- Les corrélations partielles peuvent être déduites de la matrice de corrélation  $\Gamma = (\sigma_{ij}^*)$ . Nous inversons cette dernière, nous obtenons la matrice  $\Gamma' = (\Gamma^{-1})$ .
- La matrice de corrélations partielles  $R = (r_{ij}')$  est formée à l'aide de la formule suivante :

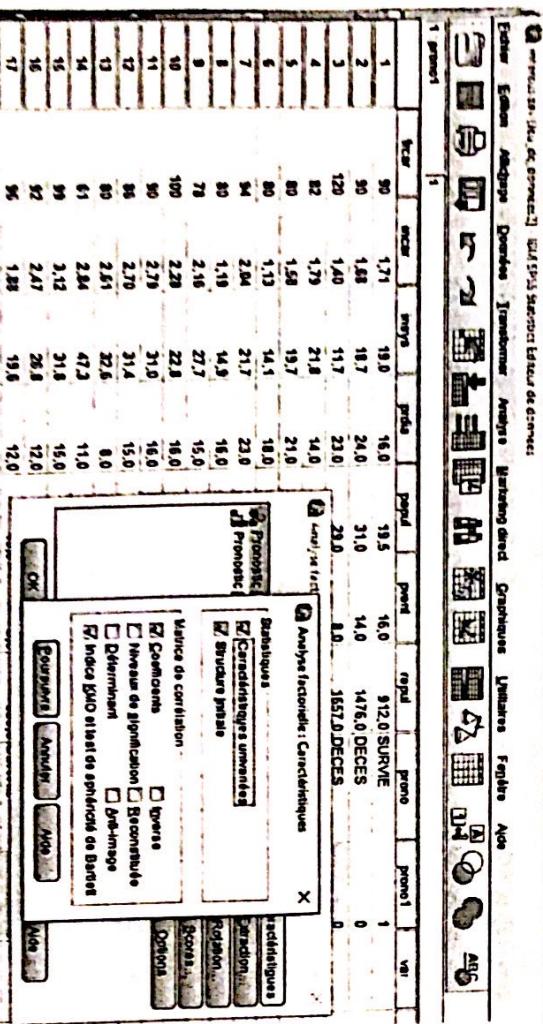
b. Indice KMO global:

$$r_{ij}' = -\frac{\sqrt{v_{ii} \times v_{jj}}}{\sum_{k=1}^n r_{ik}}$$

$$KMO = \frac{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^* + \sum_{i=1}^n \sum_{j=1}^n r_{ij}'}{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^*}$$

- L'indice KMO varie entre 0 et 1. S'il est proche de 0, les corrélations partielles sont identiques aux corrélations brutes. Dans ce cas, une compression efficace n'est pas possible. Les variables sont deux à deux orthogonales. S'il est proche de 1, nous aurons un excellent résumé de l'information sur les premiers axes factoriels.
- On nous donne parfois ici et là des grilles de lecture :

- « inacceptable » en dessous de 0,5;
- « acceptable » entre 0,5 et 0,7;
- « bien » entre 0,7 et 0,8;
- « très bien » entre 0,8 et 0,9;
- « excellent » au dessus de 0,9



## LICENCE EXCELLENCE MATH Application de la méthode ACP:

- La figure ci-contre montre la démarche de calculer les deux indices sous le logiciel.

	Analyse	Marketing direct	Graphes	Unit
Rapports				
statistiques descriptives				
Tables				
Comparer les moyennes				
Modèle linéaire général				
Modèles linéaires généralisés				
Modèles Mixture				
Corrélation				
Régression				
Log Linéaire				
Réseaux neuronaux				
Classification				
Réduction des dimensions				
Echelle				
Tests non paramétriques				
Prévisions				
Suivi				
Réponses multiples				
Analyses des valeurs manquantes				
Imputation multiple				
Echantillons complexes				
Contrôle de qualité				
Courbe ROC...				

## LICENCE EXCELLENCE MATH Caractéristiques univariées :

- Les caractéristiques univariées sont données par le calcul de la moyenne et l'écart-type de chaque variable.
- L'analyse de la moyenne obtenue indique que l'échelle de mesure est différente pour chaque variable.

### Statistiques descriptives

	Moyenne	Ecart type	Analyse N
Fréquence Cardiaque	92,16	16,428	101
Index Cardiaque	1,8457	,65902	101
Index Systolique	20,816	8,8130	101
Pression Diastolique	19,259	5,8093	101
Pression Artérielle	26,000	7,3226	101
Pulmonaire			
Pression Ventriculaire	9,500	4,3411	101
Résistance Pulmonaire	1324,059	741,3438	101

Pour répon  
matrice des

Si

## Les données sont-elles factorisables?

- Pour répondre à cette question, dans un premier temps, il convient d'observer la matrice des corrélations «Correlation Matrix».
- Si plusieurs variables sont corrélées ( $>0,5$ ), la factorisation est possible. Sinon, la factorisation n'a pas de sens et n'est donc pas conseillé.
- Le tableau suivant donne les coefficients de corrélation calculées entre les 7 variables.

Matrice de corrélation						
	Féquence Cardiaque	Index Cardiaque	Index Systolique	Pression Diastolique	Pression Artérielle Pulmonaire	Résistance Pulmonaire
Corrélation						
Index Cardiaque	1,000	-,112	-,003	,399	,370	,005
Index Systolique	-,112	1,000	,887	,361	,269	,167
Pression Diastolique	-,503	,887	1,000	,483	,405	,201
Pression Artérielle Pulmonaire	,370	,361	,483	1,000	,928	,701
Pression Ventilatoire Pulmonaire	-,005	,405	,928	,000	,244	,850
Résistance Pulmonaire	,247	,282	,201	,244	,000	,258
				,735	,701	,000

LICENCE D'EXCELLENCE MAM

08/09

## Les données sont-elles factorisables?

- Le tableau suivant donne la valeur de KMO calculé pour les données utilisées et aussi le résultat de test de sphéricité.
- La valeur de KMO obtenue (0,6) et le test de sphéricité de Bartlett est significatif ( $0,00 < 0,05$ ) ce qui permet de conclure que l'ACP satisfait les conditions de l'application et a un sens justificatif pour notre cas d'étude.

### Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.	,595
Test de sphéricité de Bartlett	
Khi-deux approximé	704,587
ddl	21
Signification de Bartlett	,000

## Analyse descriptive des données\*

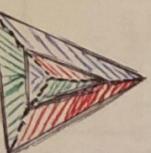
(P)

- L'analyse de ce tableau nous permet de conclure qu'il existe une forte liaison (ici  $\rho > 0,7$ ) entre : Index cardiaque et Index systolique, Index cardiaque et Résistance pulmonaire, Index systolique et Résistance pulmonaire, Pression artérielle pulmonaire et pression diastolique.

- Nous pouvons aussi conclure qu'il existe une liaison (ici  $\rho > 0,5$ ) entre : fréquence cardiaque et Index systolique, Pression artérielle pulmonaire et Résistance pulmonaire.

Résistance pulmonaire.

- Le signe de coefficient de corrélation indique le sens de variation pour les deux variables utilisées.



LICENCE D'EXCELLENCE MAM

08/09

### Extraction des facteurs :

- Le nombre de facteurs à extraire dépend des valeurs propres calculées de la matrice de corrélation obtenus à partir du tableau de données.
- Le schéma suivant indique les étapes d'extraction du nombre de facteurs à retenir pour notre matrice de données.

#### Analyse factorielle

**Méthode :**  Composantes principales

**Analyse :**  Matrice de corrélation  Matrice de covariance

**Variables :**  Amplier  Structure factorielle sans rotation  Diagramme des valeurs propres

**Extraire :**  Basé sur la valeur propre.  Valeurs propres supérieures à : 1  Nombre fixe de facteurs  Facteurs à extraire :

Maximum des itérations pour converger : 25

**Poursuivre** **Annuler** **Aide**

## Détermination analytique des facteurs (axe principal):

- C'est l'étape la plus importante de l'application de la méthode ACP. Il s'agit de représenter chaque axe  $Y_i$ ,  $i=1,2,\dots,k$  en fonction des variables originales  $X_i$ ,  $i=1,2,\dots,p$ . Cette relation est exprimé par des combinaisons linéaires entre chaque  $Y_i$  et les  $X_i$ .
- Le tableau « matrice de composantes » affiché ci-dessous permet d'extraire les coefficients qui relie chaque composante  $Y_i$  par rapport aux 7 variables explicatives introduites dans l'étude de la maladie infarctus.

Matrice des composantes<sup>a</sup>

	Composante		
	1	2	3
Fréquence Cardiaque	.477	.525	-.500
Index Cardiaque	-.759	.592	.219
Index Systolique	-.851	.277	.394
Pression Diastolique	.838	.387	.284
Pression Artérielle Pulmonaire	.782	.469	.320
Pression Ventriculaire	.361	-.351	.676
Résistance Pulmonaire	.903	-.153	.008

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites.

n à deux

## Détermination analytique des facteurs (axe principaux):

- Nous pouvons donner les combinaisons linéaires des trois axes principaux en tenant compte les trois équations suivantes :

$$Y_1 = 0,477X_1 - 0,759X_2 - 0,851X_3 + 0,838X_4 + 0,782X_5 + 0,361X_6 + 0,903X_7$$

$$Y_2 = 0,525X_1 + 0,592X_2 + 0,277X_3 + 0,397X_4 - 0,469X_5 - 0,351X_6 - 0,153X_7$$

$$Y_3 = -0,5X_1 + 0,219X_2 + 0,394X_3 + 0,284X_4 - 0,320X_5 - 0,676X_6 - 0,08X_7$$

- Nous pouvons retrouver la décomposition des valeurs propres. Par exemple, la première valeur propre est égale à

$$\lambda_1 = 0,477^2 + (-0,759)^2 + (-0,851)^2 + 0,838^2 + 0,782^2 + 0,361^2 + 0,903^2 = 3,787$$

exemple, la qualité de représentation de la variable Fréquence Cardiaque est:

$$0,477^2 + 0,525^2 + (-0,5)^2 = 0,753$$

LICENCE EXCELLENCE MAIN

## Analyse des résultats et interprétation :

### 2- Représentation graphique de variables/composantes:

- Les coordonnées factorielles des 7 variables du tableau de données par rapport aux trois composantes sont données par le tableau de la matrice des composantes.

- Les représentations graphiques concernent uniquement l'espace des points-variables sont déterminées en utilisant la matrice de composante. Par défaut, si l'on extrait deux composantes principales, le graphique factoriel représente les points-variables dans le repère orthonormé des deux axes principaux d'inertie.

- En général, on prend les deux premiers axes des composants. On obtient ainsi les projections des variables utilisées de l'ACP sur le graphique-plan composante 1 × composante 2 noté  $Y_1 \times Y_2$ .

## Analyse des résultats et interprétation:

- Le tableau 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. Examinant le tableau ci-dessous, nous constatons qu'elles sont toutes quasiment bien expliquées puisque les coefficients d'extraction dépassent 0,7.

	Initial	Extraction
Fréquence Cardiaque	1,000	.753
Index Cardiaque	1,000	.974
Index Systolique	1,000	.956
Pression Diastolique	1,000	.941
Pression Artérielle	1,000	.933
Pression Pulmonaire	1,000	.710
Pression Ventriculaire	1,000	.838
Résistance Pulmonaire	1,000	

Méthode d'extraction : Analyse en composantes principales.

LICENCE EXCELLENCE MAIN

## Diagramme de composantes : cercle des correlations

- Pour obtenir des diagrammes à deux dimensions, exécutez la syntaxe:

/PLOT ROTATION (1,2)

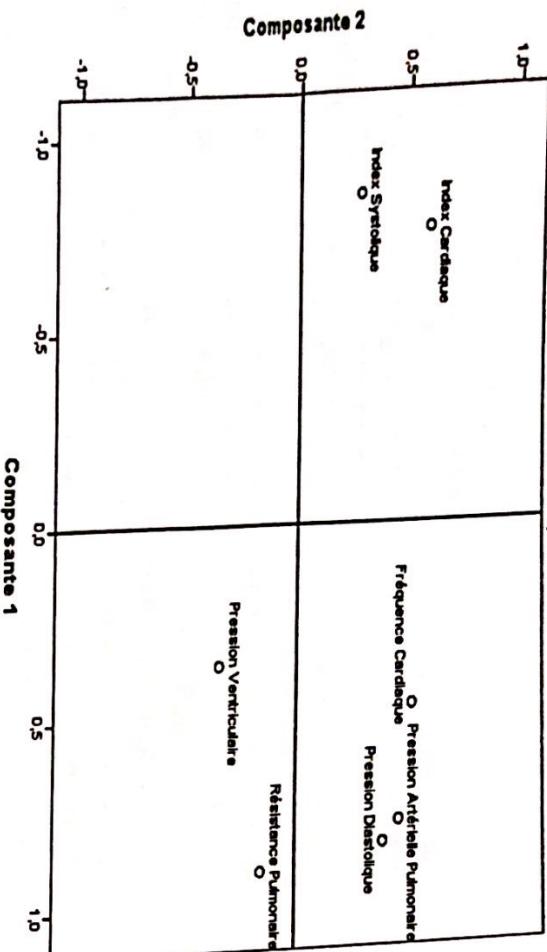
```

1
2 DATASET ACTIVATE Jeu_de_donnees2.
3
4 FACTOR
5 N VARIABLES fcar incar insys prdia papul pvent repul
6 /MISSING LISTWISE
7 /ANALYSIS fcar incar insys prdia papul pvent repul
8 /PRINT UNIVARIATE INITIAL CORRELATION KMO EXTRACTION
9 /CRITERIA FACTORS(2) ITERATE(25)
10 /EXTRACTION PC
11 /PLOT ROTATION (1,2)
12 /ROTATION NOROTATE
13 /SAVE REG(ALL)
14 /METHOD=CORRELATION.

```

## Analyse des résultats et interprétation

Diagramme de composantes



- L'analyse de dispersion des variables sur le premier plan factoriel permet de tirer plusieurs remarques et des hypothèses intéressantes.

- A partir du graphique, nous pouvons conclure qu'on peut regrouper les 7 variables en trois sous groupes.

- Le premier groupe contient les trois variables à savoir: fréquence cardiaque, pression artérielle pulmonaire et pression diastolique.
- Le deuxième groupe regroupe les deux variables, résistance pulmonaire et pression ventriculaire.
- Le troisième groupe englobe les deux variables Index cardiaque et Index systolique.

- Nous constatons que les variables qui forment même groupe ont le même effet sur l'infarctus et que l'influence du 1er groupe et 2eme groupe sont opposés à l'effet de 3eme groupe.

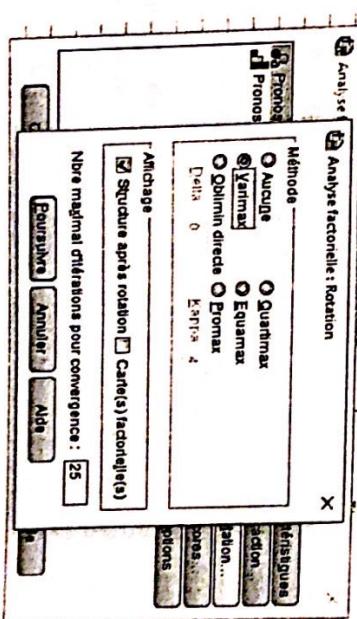
## INTERPRETATION DES FACTEURS

### Rotation des facteurs

- 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 corrélates. OBLIMIN permet de mettre en évidence des phénomènes qui déterminent des directions d'allongement non orthogonales des nuages de points.
- La méthode PROMAX: permet aux facteurs d'être corrélés. Peut être calculée plus rapidement qu'une rotation OBLIMIN directe, aussi elle est utile pour les vastes jeux de données.

## Rotation VARIMAX des facteurs

- 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.

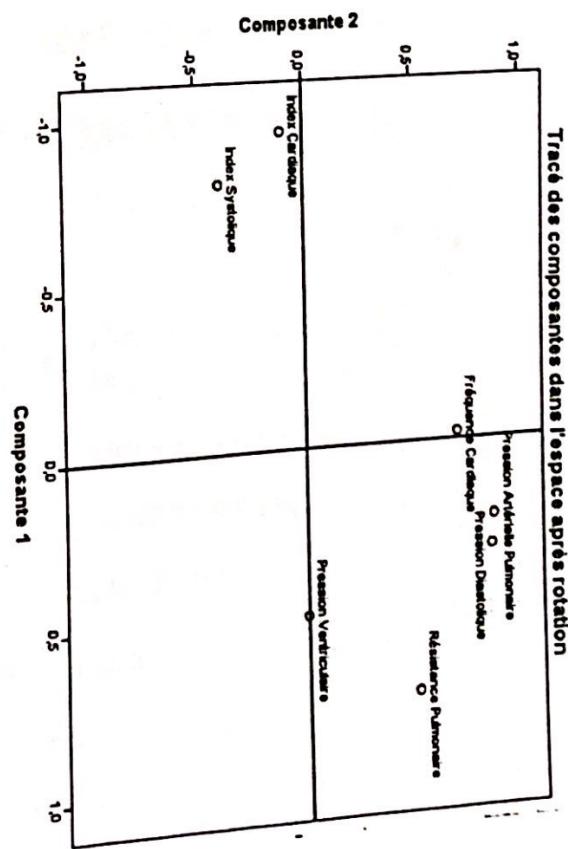
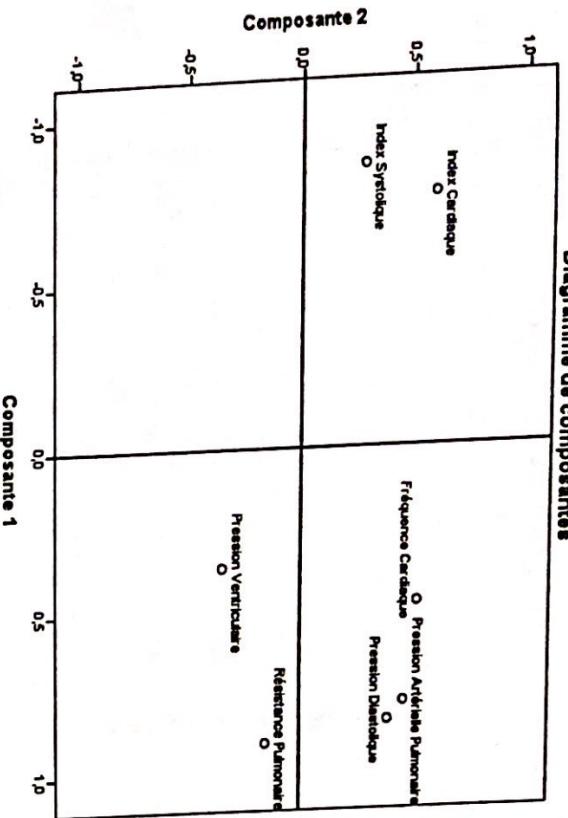


## Interprétation des résultats obtenus

## LICENCE D'EXCELLENCE MAM Rotation VARIMAX des facteurs

LICENCE D'EXCELLENCE MAM

Rotation VARIMAX des facteurs



## LICENCE D'EXCELLENCE MAM Rotation VARIMAX des facteurs

LICENCE D'EXCELLENCE MAM

## LICENCE D'EXCELLENCE MAM Représentation graphique des individus :

- On peut également obtenir la matrice des coordonnées des individus par rapport aux nouveaux axes factoriels. La figure suivante montre comment obtenir ces coordonnées.

**Analyse factorielle**

Enregistrer dans des variables

Méthode

- Régression
- Bartlett
- Anderson-Rubin

Afficher la matrice des coefficients factoriels

Poursuivre

Annuler

Aide

Méthode d'extraction : Analyse en composantes principales.

Méthode de rotation : Varimax avec normalisation Kaiser.

a. Convergence de la rotation dans 4 itérations.

OK

Annuler

Annuler

Annuler

Annuler

Annuler

Annuler

Annuler

## Représentation graphique des individus

- Le tableau suivant présente les nouveaux coordonnées pour chaque observation par rapport aux nouveaux axes construits ( $Y_1, Y_2, Y_3$ ).

	id	age	sexe	pression artérielle	pression ventriculaire	résistance pulmonaire	index cardiaque	survive
1	1	60	0	100	100	100	100	0
2	2	60	0	100	100	100	100	0
3	3	60	0	100	100	100	100	0
4	4	60	0	100	100	100	100	0
5	5	60	0	100	100	100	100	0
6	6	60	0	100	100	100	100	0
7	7	60	0	100	100	100	100	0
8	8	60	0	100	100	100	100	0
9	9	60	0	100	100	100	100	0
10	10	60	0	100	100	100	100	0
11	11	60	0	100	100	100	100	0
12	12	60	0	100	100	100	100	0
13	13	60	0	100	100	100	100	0
14	14	60	0	100	100	100	100	0
15	15	60	0	100	100	100	100	0
16	16	60	0	100	100	100	100	0
17	17	60	0	100	100	100	100	0
18	18	60	0	100	100	100	100	0
19	19	60	0	100	100	100	100	0
20	20	60	0	100	100	100	100	0
21	21	60	0	100	100	100	100	0
22	22	60	0	100	100	100	100	0
23	23	60	0	100	100	100	100	0
24	24	60	0	100	100	100	100	0
25	25	60	0	100	100	100	100	0
26	26	60	0	100	100	100	100	0
27	27	60	0	100	100	100	100	0
28	28	60	0	100	100	100	100	0
29	29	60	0	100	100	100	100	0
30	30	60	0	100	100	100	100	0
31	31	60	0	100	100	100	100	0
32	32	60	0	100	100	100	100	0
33	33	60	0	100	100	100	100	0
34	34	60	0	100	100	100	100	0
35	35	60	0	100	100	100	100	0
36	36	60	0	100	100	100	100	0
37	37	60	0	100	100	100	100	0
38	38	60	0	100	100	100	100	0
39	39	60	0	100	100	100	100	0
40	40	60	0	100	100	100	100	0
41	41	60	0	100	100	100	100	0
42	42	60	0	100	100	100	100	0
43	43	60	0	100	100	100	100	0
44	44	60	0	100	100	100	100	0
45	45	60	0	100	100	100	100	0
46	46	60	0	100	100	100	100	0
47	47	60	0	100	100	100	100	0
48	48	60	0	100	100	100	100	0
49	49	60	0	100	100	100	100	0
50	50	60	0	100	100	100	100	0
51	51	60	0	100	100	100	100	0
52	52	60	0	100	100	100	100	0
53	53	60	0	100	100	100	100	0
54	54	60	0	100	100	100	100	0
55	55	60	0	100	100	100	100	0
56	56	60	0	100	100	100	100	0
57	57	60	0	100	100	100	100	0
58	58	60	0	100	100	100	100	0
59	59	60	0	100	100	100	100	0
60	60	60	0	100	100	100	100	0
61	61	60	0	100	100	100	100	0
62	62	60	0	100	100	100	100	0
63	63	60	0	100	100	100	100	0
64	64	60	0	100	100	100	100	0
65	65	60	0	100	100	100	100	0
66	66	60	0	100	100	100	100	0
67	67	60	0	100	100	100	100	0
68	68	60	0	100	100	100	100	0
69	69	60	0	100	100	100	100	0
70	70	60	0	100	100	100	100	0
71	71	60	0	100	100	100	100	0
72	72	60	0	100	100	100	100	0
73	73	60	0	100	100	100	100	0
74	74	60	0	100	100	100	100	0
75	75	60	0	100	100	100	100	0
76	76	60	0	100	100	100	100	0
77	77	60	0	100	100	100	100	0
78	78	60	0	100	100	100	100	0
79	79	60	0	100	100	100	100	0
80	80	60	0	100	100	100	100	0
81	81	60	0	100	100	100	100	0
82	82	60	0	100	100	100	100	0
83	83	60	0	100	100	100	100	0
84	84	60	0	100	100	100	100	0
85	85	60	0	100	100	100	100	0
86	86	60	0	100	100	100	100	0
87	87	60	0	100	100	100	100	0
88	88	60	0	100	100	100	100	0
89	89	60	0	100	100	100	100	0
90	90	60	0	100	100	100	100	0
91	91	60	0	100	100	100	100	0
92	92	60	0	100	100	100	100	0
93	93	60	0	100	100	100	100	0
94	94	60	0	100	100	100	100	0
95	95	60	0	100	100	100	100	0
96	96	60	0	100	100	100	100	0
97	97	60	0	100	100	100	100	0
98	98	60	0	100	100	100	100	0
99	99	60	0	100	100	100	100	0
100	100	60	0	100	100	100	100	0

	id	age	sexe	pression artérielle	pression ventriculaire	résistance pulmonaire	index cardiaque	survive
1	1	60	0	100	100	100	100	0
2	2	60	0	100	100	100	100	0
3	3	60	0	100	100	100	100	0
4	4	60	0	100	100	100	100	0
5	5	60	0	100	100	100	100	0
6	6	60	0	100	100	100	100	0
7	7	60	0	100	100	100	100	0
8	8	60	0	100	100	100	100	0
9	9	60	0	100	100	100	100	0
10	10	60	0	100	100	100	100	0
11	11	60	0	100	100	100	100	0
12	12	60	0	100	100	100	100	0
13	13	60	0	100	100	100	100	0
14	14	60	0	100	100	100	100	0
15	15	60	0	100	100	100	100	0
16	16	60	0	100	100	100	100	0
17	17	60	0	100	100	100	100	0
18	18	60	0	100	100	100	100	0
19	19	60	0	100	100	100	100	0
20	20	60	0	100	100	100	100	0
21	21	60	0	100	100	100	100	0
22	22	60	0	100	100	100	100	0
23	23	60	0	100	100	100	100	0
24	24	60	0	100	100	100	100	0
25	25	60	0	100	100	100	100	0
26	26	60	0	100	100	100	100	0
27	27	60	0	100	100	100	100	0
28	28	60	0	100	100	100	100	0
29	29	60	0	100	100	100	100	0
30	30	60	0	100	100	100	100	0
31	31	60	0	100	100	100	100	0
32	32	60	0	100	100	100	100	0
33	33	60	0	100	100	100	100	0
34	34	60	0	100	100	100	100	0
35	35	60	0	100	100	100	100	0
36	36	60	0	100	100	100	100	0
37	37	60	0	100	100	100	100	0
38	38	60	0	100	100	100	100	0
39	39	60	0	100	100	100	100	0
40	40	60	0	100	100	100	100	0
41	41	60	0	100	100	100	100	0
42	42	60	0	100	100	100	100	0
43	43	60	0	100	100	100	100	0
44	44	60	0	100	100	100	100	0
45	45	60	0	100	100	100	100	0
46	46	60	0	100	100	100	100	0
47	47	60	0	100	100	100	100	0
48	48	60	0	100	100	100	100	0
49	49	60	0	100	100	100	100	0
50	50	60	0	100	10			