INU Champollion

Licence Info 3ème Année

Maths pour l'I.A.

Thierry Montaut

Couples de séries statistiques.

Définition

Nous parlerons dans ce chapitre d'observation d'individus plutôt que d'issues d'expériences aléatoires, On pourra donc également considérer que l'observation de n individus $i_1, \ldots, i_n$ *nous fournit n observations (de même taille) de deux caractéristiques couplées :* $x = (x_1, \ldots, x_n)$ *et* $y = (y_1, \ldots, y_n)$.

Covariance de deux séries statistiques

Une première façon d'établir un lien entre les observations x et y est d'étudier si elles ont tendance à varier dans le même sens

Définition

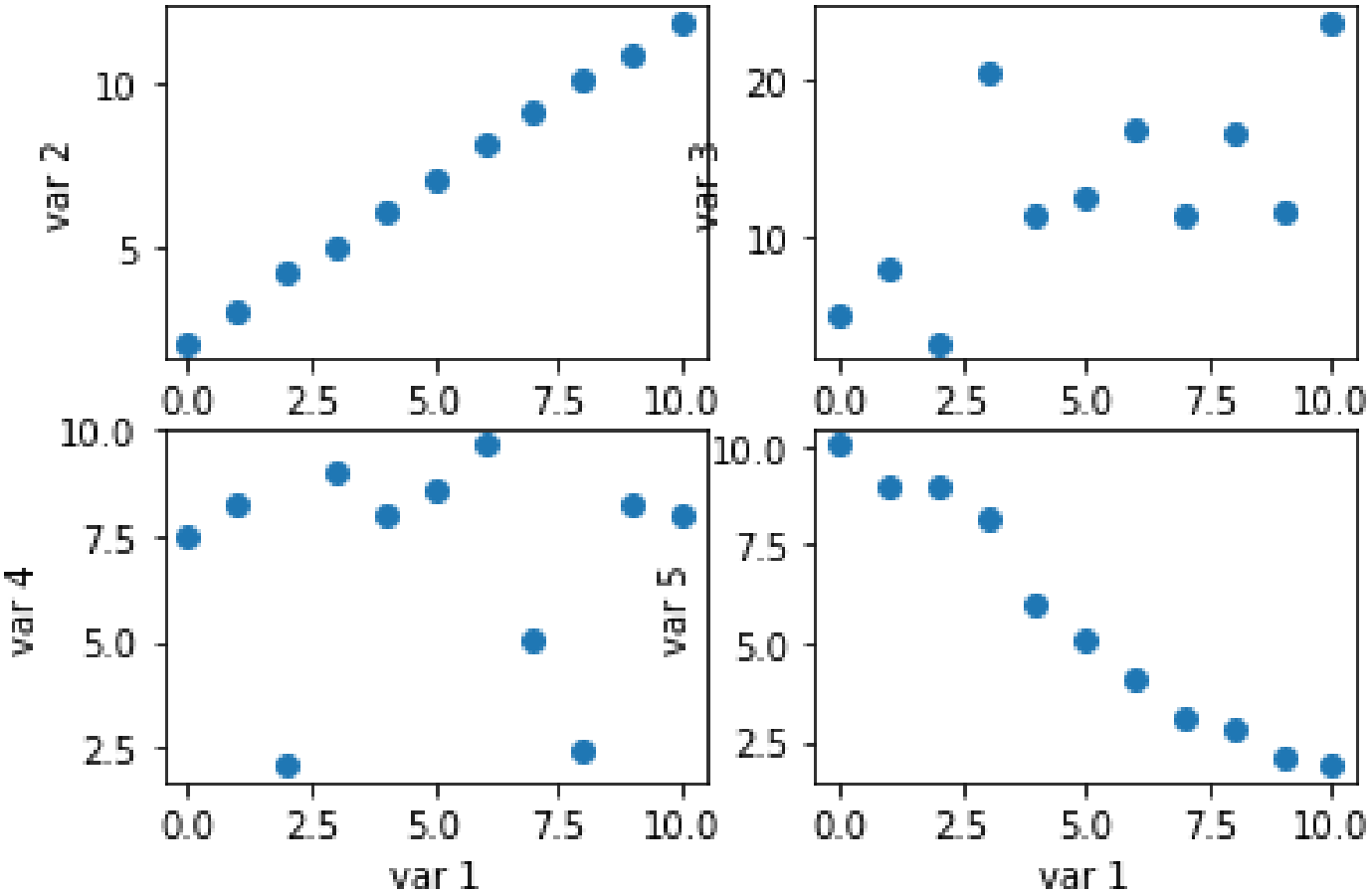
La covariance entre les observations x et y des variables X et Y est définie par :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

où $\overline{x}$ est la moyenne de x, et $\overline{y}$ est la moyenne de y.
Ou de manière équivalente :

$$cov(x,y) = \overline{xy} - \overline{x} \times \overline{y}$$

Exemples



Propriétés

La covariance est bilinéaire, symétrique, définie et positive :

- $cov(ax, y) = a.cov(x, y)$
- $cov(x + y, z) = cov(x, z) + cov(y, z)$
- $cov(x, y) = cov(y, x)$
- $cov(x, x) = var(x) \geq 0.$
- $var(x + y) = var(x) + var(y) + 2cov(x, y),$

C'est donc (presque) un produit scalaire dont la norme associée est l'écart-type.

Corrélation

Si le signe de la covariance nous fournit une information précieuse, ce n'est pas le cas de sa valeur car elle dépend de l'échelle des variables X et Y. Pour améliorer cela nous allons la normaliser :

Définition

La corrélation entre x et y est la covariance normalisée, définie par :

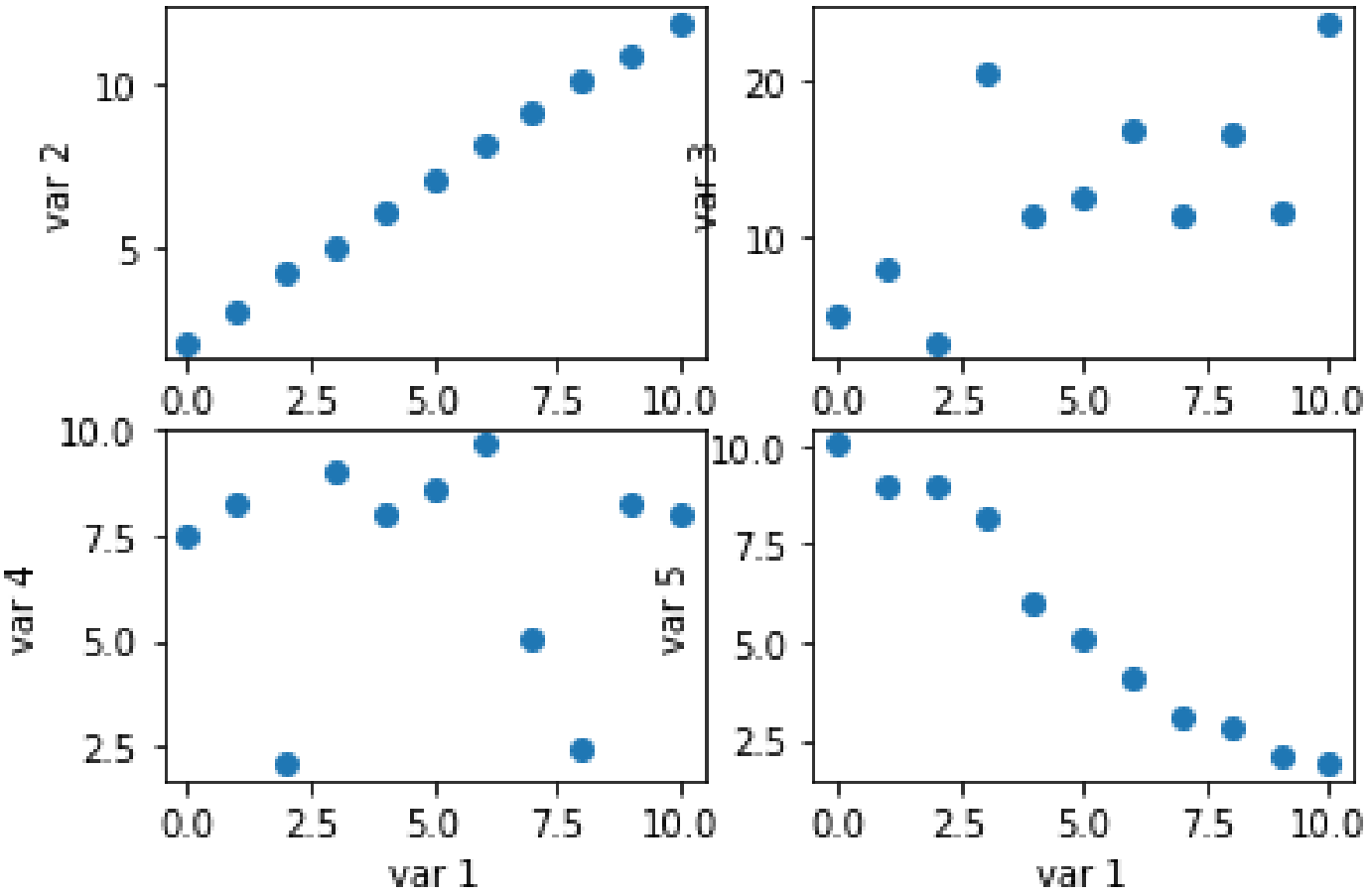
$$\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}.$$

- Elle est maintenant toujours comprise entre -1 et 1.
- $\rho(x, x) = 1$.
- Une corrélation proche de 1 ou -1 indique que les variables sont linéairement liées
- Une corrélation proche de 0 indique une faible corrélation.
- Attention : non corrélées ne signifie par indépendantes. Les variables indépendantes sont bien non corrélées mais la réciproque n'est pas toujours vraie. Il existe d'autre dépendances que la dépendance linéaire !

Représentation graphique

- On a vu que la covariance est un produit scalaire dont la norme associée est l'écart-type. Par analogie avec le produit scalaire de $\mathbb{R}^2$, $(X|Y) = ||X|||Y||cos(X,Y)$, $\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$ peut être vu comme le cosinus d'un "angle" entre les variables aléatoires, exprimant leur plus ou moins grande corrélation.
- Deux variables fortement corrélées auront un coefficient proche de 1 donc un angle proche de 0,
- deux variables corrélées négativement auront un coefficient proche de -1, donc un angle proche de π,
- Deux variable non corrélées auront un coefficient proche de 0 dont un angle proche de $\frac{\pi}{2}$ et seront dites "orthogonales".

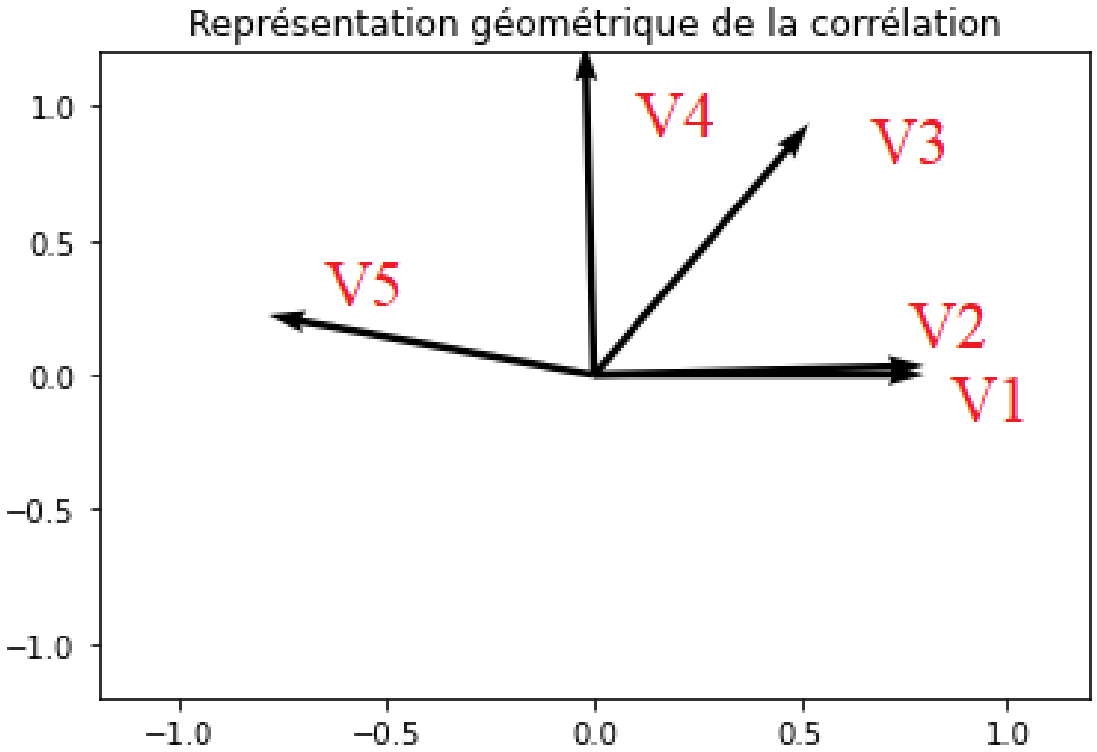
Exemples



$$cor(X_1, X_2) = 0.99, \quad cor(X_1, X_3) = 0.65,$$

$$cor(X_1, X_4) = -0.2, \quad cor(X_1, X_5) = -0.98$$

Représentation graphique des corrélations



Vecteur de variables aléatoires

On suppose maintenant que nous disposons de p séries statistiques concernant p variables X_j, issues de l'observation de n individus. Pour $j \in [\![1,p]\!]$, les n observations de la variable X_j seront notées $x^j = (x_1^j, \ldots, x_n^j)$. Les notions de covariance et de corrélations vont se généraliser au cas de p variables en les considérant 2 à 2. On obtient alors des formes matricielles :

Définition

On appelle matrice de covariance, la matrice carrée à p lignes $Cov = (c_{i,j})_{i,j \in [\![1,p]\!]}$ *où* $\forall i,j \in [\![1,p]\!],\ c_{i,j} = cov(x^i, x^j)$.

$$Cov = \begin{pmatrix} var(x^1) & cov(x^1,x^2) & \ldots & cov(x^1,x^n) \\ cov(x^2,x^1) & var(x^2) & \ldots & cov(x^2,x^n) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x^n,x^1) & cov(x^n,x^2) & \ldots & var(x^n) \end{pmatrix}$$

Matrice de covariance

Propriété

1. *On sait en effet que* $cov(x, x) = var(x)$. *Les coefficients diagonaux sont donc les variances des variables aléatoires (C'est pourquoi on parle parfois de matrice de variance-covariance)*
2. *Par symétrie de la covariance, la matrice est symétrique... donc diagonalisable dans une BON.*

Matrice de corrélation

On définit de même la matrice de corrélation :

Définition

$$Cor = \begin{pmatrix} 1 & \rho(x^1, x^2) & \dots & \rho(x^1, x^n) \\ \rho(x^2, x^1) & 1 & \dots & \rho(x^2, x^n) \\ \vdots & \vdots & \vdots & \vdots \\ \rho(x^n, x^1) & \rho(x^n, x^2) & \dots & 1 \end{pmatrix}$$

Elle n'a que des 1 sur la diagonale et elle est également symétrique.

Distance entre individus

Un individu correspond à une observation, donc une ligne du tableau statistique. C'est donc un vecteur de $\mathbb{R}^p$, $I_k = (x_k^1, \ldots, x_k^p)$. En munissant $\mathbb{R}^k$ de sa structure euclidienne, on définit donc la distance entre deux individus :

Définition

$$d(I_k, I_l) = \sqrt{\sum_{i=1}^{n} (x_k^i - x_l^i)^2}$$

Distance entre individus

Définition

A l'issue de ces observations, on peut définir l'individu moyen :

$$\bar{I} = (\bar{x^1}, \ldots \bar{x^p})$$

et la distance d'un individu à l'individu moyen :

$$d(I_k, \bar{I}) = \sqrt{\sum_{i=1}^{n} (x_k^i - \bar{x^i})^2}$$

Inertie

On sait que la variance des observations d'une variable aléatoire est une mesure de la dispersion de ces observations par rapport à leur moyenne. On cherche à étendre cette mesure au cas de l'observation de p variables aléatoires.

Définition

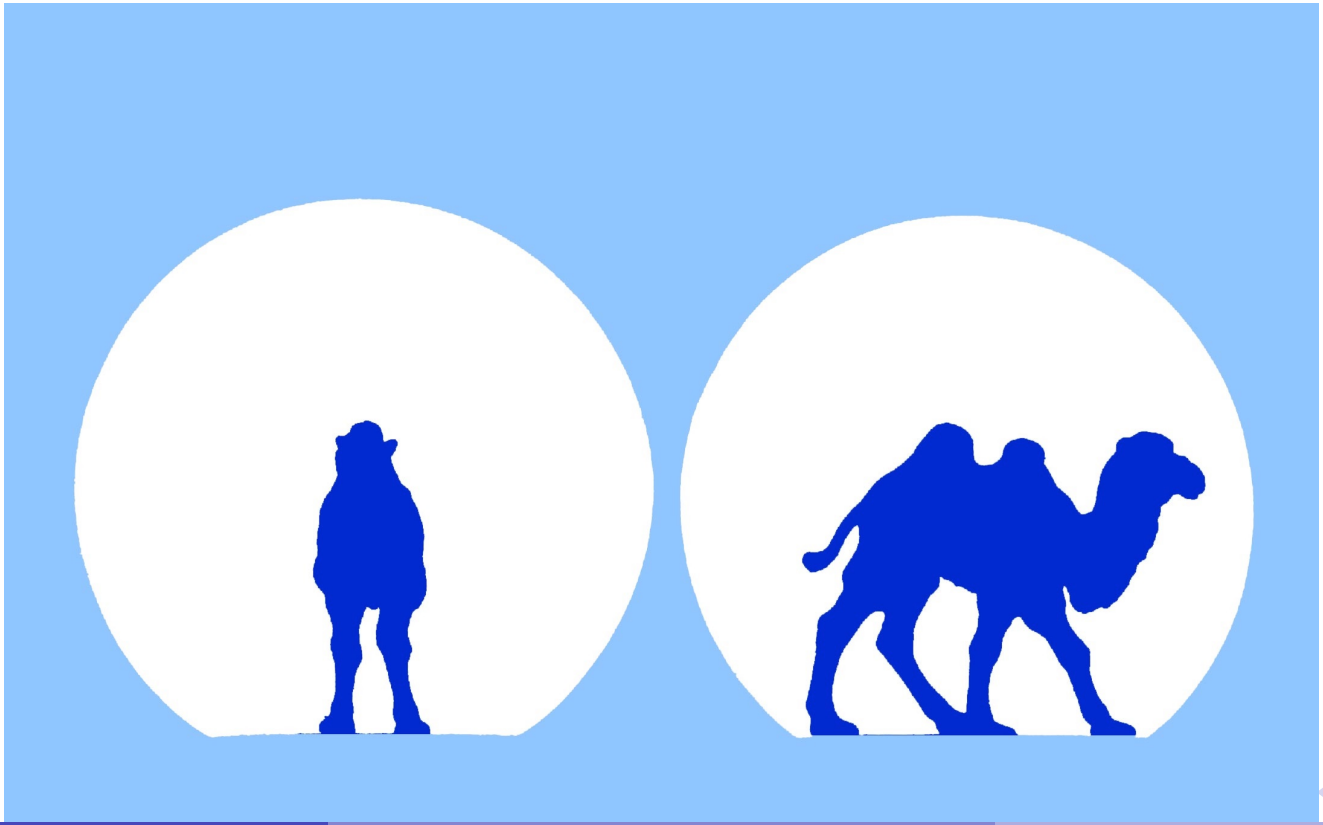
L'inertie des n observations $(x_1^1, \ldots, x_1^p), \ldots, (x_n^1, \ldots, x_n^p)$ *est définie par :*

$$I = \sum_{i=1}^{p} var(x^j) = \frac{1}{n} \sum_{k=1}^{n} d^2(I_k, \bar{I})$$

C'est donc à la fois la somme des variances des variables aléatoires observées et la moyenne des distance au carrée des individus à l'individu moyen. C'est donc bien une mesure de la dispersion autour de l'individu moyen dans $\mathbb{R}^k$.

Analyse en composante principale (ACP)

• Lorsqu'on dispose de plus de 3 variables observées par une série statistique, il n'est plus possible d'en donner une représentation graphique exacte. Un individu (ou une observation) est alors un vecteur de $\mathbb{R}^k$ qu'il nous faut essayer de projeter en dimension 2 avec un minimum de perte d'information. Mais les différentes projections ne font pas perdre autant d'information.



Analyse en composante principale (ACP)

- La représentation graphique du nuage de points n'est pas le seul objectif. Réduire la taille des données en remplaçant toutes les variables par quelques combinaisons linéaires les plus pertinentes de ces variables est fondamental pour réduire le temps de traitement des algorithmes d'analyse de données.
- L'ACP procède exactement ainsi, elle consiste à calculer des "variables-synthèse" appelées **composantes principales** qui sont des combinaisons linéaires des variables initiales et qui exprime le plus fidèlement les observations initiales.

Analyse en composante principale (ACP)

- Comme l'image la plus évocatrice est celle qui occupe le plus d'espace, les variables les plus fidèles seront celles de plus grande variance.
- Pour simplifier les calculs et pouvoir utiliser les fonctions de numpy, les variables x^j doivent tout d'abord être **centrées et réduites** en remplaçant x_i^j par

$$\frac{x_i^j - \bar{x^j}}{\sigma_j}.$$

Théorème

La matrice de corrélation étant symétrique est diagonalisable dans une base orthonormée de vecteurs propres $(u^1, u^2, \ldots, u^p)$. *On a donc* $Cor = P.D.{}^tP$

- *Les valeurs propres sont toutes positives. On les ordonne* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. *Elles constituent la diagonale de la matrice D.*
- *Les vecteur propres* $u^1, u^2, \ldots, u^p$ *sont deux à deux orthogonaux. Ils constituent les colonnes de la matrice P.*

- On peut alors définir p nouvelles variables aléatoires $c^1, c^2, \ldots, c^p$ par combinaison linéaires des variables initiales :

$$c^j = u_1^j x^1 + \ldots + u_p^j x^p$$

- On les appelle les **variables principales.**

Composantes principales

$$P = \begin{pmatrix} u_1^1 & \dots & u_1^p \\ \vdots & & \vdots \\ u_p^1 & \dots & u_p^p \end{pmatrix}$$

Donc $C = XP$, est une matrice à n lignes et p colonnes dont les colonnes sont les observations des variables c^j et les lignes sont les observations de ces variables sur les n individus.

Définition

Cette matrice $C = XP$ est appelée **matrice des composantes principales.**

Théorème

La matrice de covariance des variables principales est la matrice diagonale

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}$$

- *Les variables c^j sont donc deux à deux non corrélées et leur variance est $\text{var}(c^j) = \lambda_j$, elles sont donc de variance décroissante.*
- *Les deux premières composantes principales sont donc les deux directions orthogonales dans lesquelles la dispersion des données est la plus importante.*
- *Le plan qu'elles engendrent est appelé* **plan principal** *et c'est dans ce plan que nous représenterons les données.*

Cercle de corrélation

Afin de visualiser le rôle joué par chaque variable initiale dans la constitution des 2 composantes principales c^1 et c^2, on va représenter chaque variable x^j par le vecteur $V_j = (\rho(x^j, c^1), \rho(x^j, c^2))$.

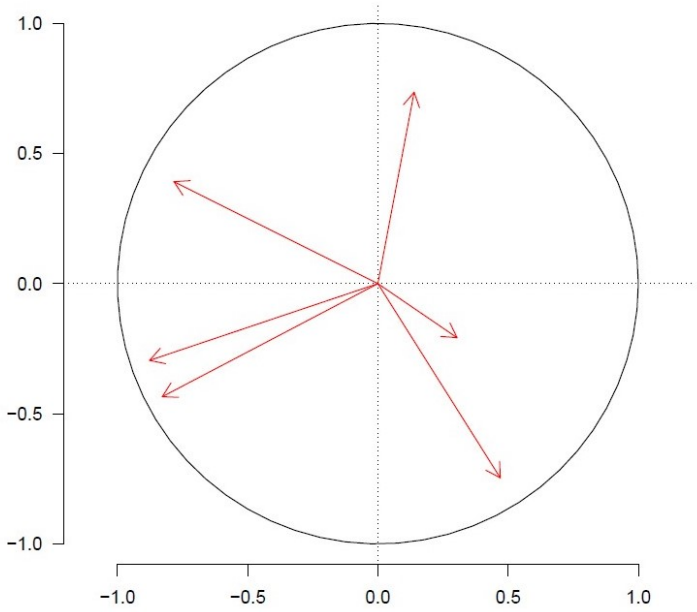


FIGURE 5.3 – Cercle des corrélations de $p = 6$ variables.

Cette représentation permet de visualiser en amplitude et en direction la corrélation de la variable initiale x^j avec les deux composantes principales.

Inertie

L'objectif initial était de donner la meilleure représentation des données initiales avec seulement 2 combinaisons linéaires. On peut utiliser l'inertie pour évaluer la qualité de notre choix.

Théorème

- *L'ACP conserve l'inertie initiale :*

$$I(x^1, \ldots, x^p) = \sum_{i=1}^{p} var(x^j) = tr(Cor)$$

$$= \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} var(c^j) = I(c^1, \ldots, c^p).$$

Inertie

Définition

1. *La contribution de chaque composante principale est*

$$\frac{var(c^j)}{I} = \frac{\lambda_j}{\lambda_1 + \ldots + \lambda_p}$$

2. *La contribution du plan principal est donc caractérisé par l'indicateur suivant appelé* **part d'inertie** *:*

$$r = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_p}$$

Inertie

On peut se faire une idée de cet indicateur à l'aide d'une représentation graphique de la décroissance des valeurs propre appelé **ébouli des valeurs propres** :

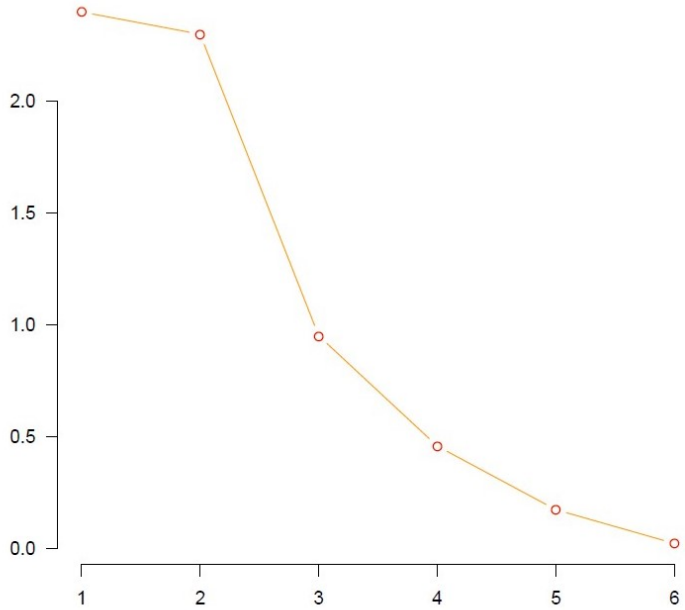


FIGURE 5.4 – Eboulis de $p = 6$ valeurs propres.