Maths pour l'I.A.

Thierry Montaut

Plan du cours

- Compléments d'algèbre linéaire
- Compléments d'analyse
- **Compléments de statistiques**
  - **Couple de vecteurs aléatoires, covariance**
  - **Régression linéaire**
  - **Vecteurs aléatoires, matrices de covariance et de corrélation**
  - **Analyse en composante principale**

Couples et vecteurs de variables aléatoires discrètes

- Il arrive fréquemment que les individus observés en statistiques dépendent de deux ou plusieurs caractéristiques. Une voiture peut être décrite par sa puissance, sa forme, son prix ; Un jus d'orange par son odeur, son goût, son acidité etc ; un lycéen par quelques-une de ses notes ou classements durant ses études secondaires.
- Ces caractéristiques peuvent être considérées comme des variables aléatoires et l'observation d'une série d'individus (quelques voitures, quelques jus d'oranges, quelques lycéens) comme des observations couplées de ces variables aléatoires.

Le but de ce chapitre est de répondre à quelques questions à l'issue de ces observations, telles que :

1. Quelles sont les caractéristiques qui sont liées ? qui agissent dans le même sens ? en sens contraire ?
2. Quelles sont celles qui expliquent le mieux une caractéristique d'un individu (le prix d'une voiture, le caractère agréable d'un jus d'orange, la réussite d'un étudiant ?)
3. En ce qui concerne les individus, quels sont ceux qui se ressemblent ?
4. Peut on les regrouper en plusieurs classes et si oui, en prenant en compte quelles caractéristiques ?

Un exemple de jeu de données

Modèle	Cyl	Puiss	Long	Larg	Poids	V max
Alfasud	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta	1 1570	109	428	162	1060	175
Princess	1798	82	445	172	1160	158
Datsun	1998	115	469	169	1370	160
Taunus	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140

Inutile de vous rappeler ce que sont espérance et variance, n'est-ce pas ?

Définition

Soit X une variable aléatoire discrète prenant les valeurs (finies) $x_1, \ldots, x_n$ avec des probabilités $p_1, \ldots, p_n$.

- *L'espérance de X est $E(X) = \sum_{i=1}^{n} p_i x_i$*
- *La variance de x est $var(X) = \sum_{i=1}^{n} p_i (x_i - E(x))^2 = E(X^2) - E(X)^2$*
- *L'écart-type de X est $\sigma(X) = \sqrt{var(X)}$.*

Je vous laisse établir ce que deviennent ces formules dans le cas (fréquent pour nous) où les valeurs sont équiprobables.

Et dans le cas d'une série statistique ?

Définition

Soit $X = (x_1, \ldots, x_n)$ *une série de données.*

- *La moyenne de* X *est* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- *La variance de* x *est* $var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\bar{X^2}) - \bar{X}^2$
- *L'écart-type de* X *est* $\sigma(X) = \sqrt{var(X)}$.

Couples de variables aléatoires discrètes.

Définition

On suppose que l'on dispose de deux variables aléatoires X *et* Y*. On parle de couple de variables aléatoires, lorsque ces deux variables sont observables à l'issue d'une même expérience.*
On peut donc considérer que l'issue ω *d'une expérience fournit un couple de valeurs pour* X *et* Y *:* $\omega = (a, b)$.

Loi conjointe et lois marginales

Connaître les lois des variables X et Y ne suffit pas en général pour connaître la probabilité que le couple (X, Y) prenne une valeur (a, b) donnée. Par exemple si les deux variables peuvent prendre les valeurs 0,1 et 2 avec les probabilités

$$P\{X = 0\} = P\{Y = 0\} = 0.3$$
$$P\{X = 1\} = P\{1 = 0\} = 0.5$$
$$P\{X = 2\} = P\{Y = 2\} = 0.2$$

On ne sait pas si les probabilités que le couple prenne les 9 valeurs possibles sont

P	$\{Y = 0\}$	$\{Y = 1\}$	$\{Y = 2\}$
$\{X = 0\}$	0.3	0	0
$\{X = 1\}$	0	0.5	0
$\{X = 2\}$	0	0	0.2

ou

P	$\{Y = 0\}$	$\{Y = 1\}$	$\{Y = 2\}$
$\{X = 0\}$	0.1	0	0.2
$\{X = 1\}$	0.1	0.4	0
$\{X = 2\}$	0.1	0.1	0

Loi conjointe et lois marginales

Définition

1. *On appelle* **loi conjointe** *la probabilité que le couple* (X, Y) *de variables prenne une valeur* (a, b)

$$P\{(X, Y) = (a, b)\} = P\{X = a, Y = b\} = P\{X = a\} \cap \{Y = b\}$$

2. *Les lois des variables aléatoires* X *et* Y *sont appelées les* **lois marginales**.

Loi conjointe et lois marginales

On peut en utilisant la formule des probabilités totales

$$P\{X=a\} = \sum_{b\in\Omega_2} P\{(X,Y)=(a,b)\}$$

obtenir les lois marginales à partir de la loi conjointe :

P	$\{Y=0\}$	$\{Y=1\}$	$\{Y=2\}$	$P(X)$
$\{X=0\}$	0.1	0	0.2	0.3
$\{X=1\}$	0.1	0.4	0	0.5
$\{X=2\}$	0.1	0.1	0	0.2
$P(Y)$	0.3	0.5	0.2	1

Loi conjointe et lois conditionnelle

Rappel : Il ne faut pas confondre la loi conjointe avec les lois conditionnelles (la probabilité de X sachant ce que vaut Y).

Définition

$$P\{X = a | Y = b\} = \frac{P\{(X, Y) = (a, b)\}}{P\{Y = b\}}$$

Indépendance

Définition

1. *On rappelle que deux événements A et B sont indépendants ssi* $P(A \cap B) = P(A).P(B)$.
2. *On définit de même : deux variables X et Y sont indépendantes ssi* $\forall (a,b) \in \Omega_1 \times \Omega_2$ *:*

$$P\{(X,Y) = (a,b)\} = P\{X = a\} \times P\{Y = b\}$$

Indépendance et loi conjointe

Dans le cas de deux variables indépendantes, on peut déduire la loi conjointe des lois marginales :

P	$\{Y = b_1\}$	$\{Y = b_2\}$	$\ldots$	$\{Y = b_j\}$	$P(X)$
$\{X = a_1\}$	$p_1 q_1$	$p_1 q_2$	$\ldots$	$p_1 q_j$	p_1
$\{X = a_2\}$	$p_2 q_1$	$p_2 q_2$	$\ldots$	$p_2 q_j$	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{X = a_i\}$	$p_i q_1$	$p_i q_2$	$\ldots$	$p_i q_j$	p_i
$P(Y)$	q_1	q_2	$\ldots$	q_j	1

Indépendance

On sait que sans conditions sur les variables aléatoires, l'espérance est linéaire. En particulier, on a toujours

$$E(X+Y) = E(X) + E(Y)$$

En revanche, ce n'est pas le cas pour le produit en général. C'est néanmoins vrai si les variables sont indépendantes

Théorème

Si x et y sont indépendantes alors

$$E(XY) = E(X) \times E(Y).$$

Démonstration

Covariance d'un couple de V.A.

Cette relation n'est pas vraie dans le cas général et la covariance mesure cette différence :

Définition

La covariance du couple de variables (X, Y) *est définie par :*

$$cov(X, Y) = E(XY) - E(X).E(Y) = E[(X - E(X)).(Y - E(Y)]$$

Covariance

Propriété

On en déduit que si X et X sont indépendantes, leur covariance est nulle.

Mais la réciproque est fausse en général.
Le signe de la covariance donne une information importante :

- Si $cov(X, Y) > 0$ les variables X et Y ont tendance à varier dans le même sens
- Si $cov(X, Y) <$ les variables X et Y ont tendance à varier en sens contraire

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

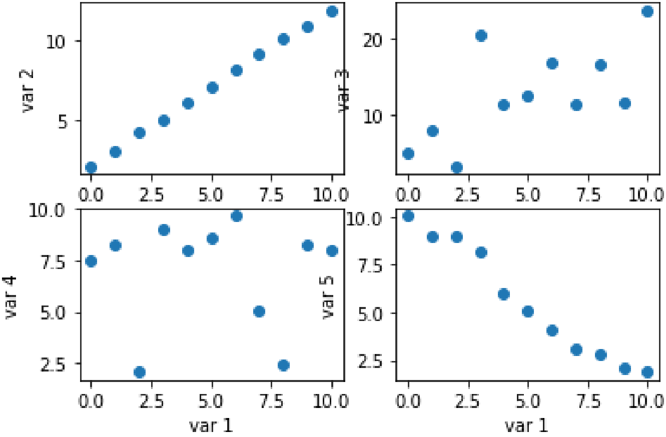
La covariance entre les observations x et y des variables X et Y est définie par :

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

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

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

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 !

Exercice :

Soit X une variable aléatoire de loi uniforme discrète sur $\{-1, 0, 1\}$ et $Y = X^2$.

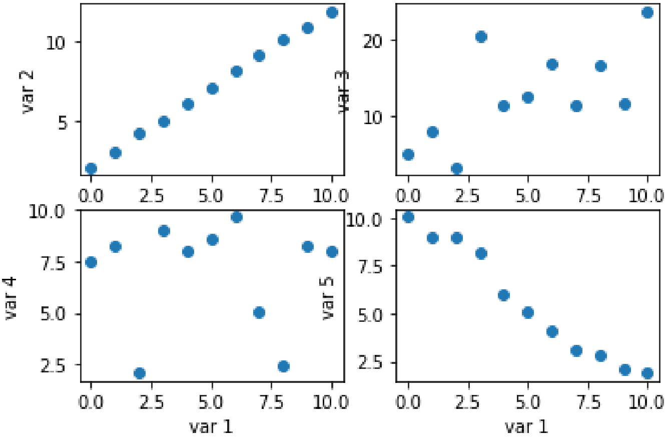
1°**)**Montrer que X et Y ne sont pas indépendantes.

2°**)**Calculer leur covariance.

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".

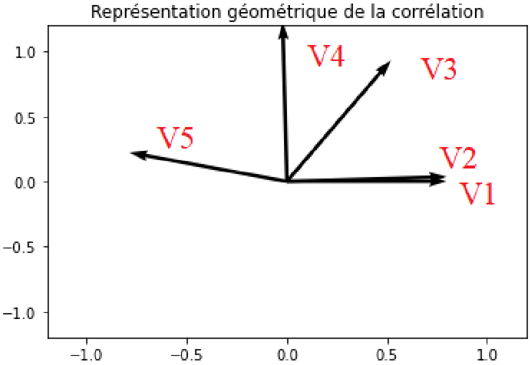
Exemples



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

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

Représentation graphique des corrélations



Régression linéaire

Définition

On dispose de deux séries statistiques de même taille $x = (x_1, \dots, x_n)$ *et* $y = (y_1, \dots, y_n)$, *et on cherche à déterminer la droite d'équation* $y = ax + b$ *donnant la meilleure approximation de cette observation.*

Ce procédé est appelé régression linéaire.

Erreur moyenne

Définition

Nous avons déjà vu que l'erreur commise en approchant les observations y_i par $a.x_i + b$ est

$$err(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a.x_i + b))^2$$

Notre but est de minimiser cette erreur, donc de minimiser une fonction de deux variables, ce qu'on sait assez bien faire depuis... la semaine dernière.

Régression linéaire

Théorème

L'erreur est minimale est obtenue pour la droite d'équation $y = a.x + b$ *avec :*

$$a = \frac{cov(x, y)}{var(x)^2}$$

et

$$b = \bar{y} - a\bar{x}$$

Ce qui prouve au passage que le point moyen $(\bar{x}, \bar{y})$ *appartient à la droite de meilleure approximation.*

Démonstration

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

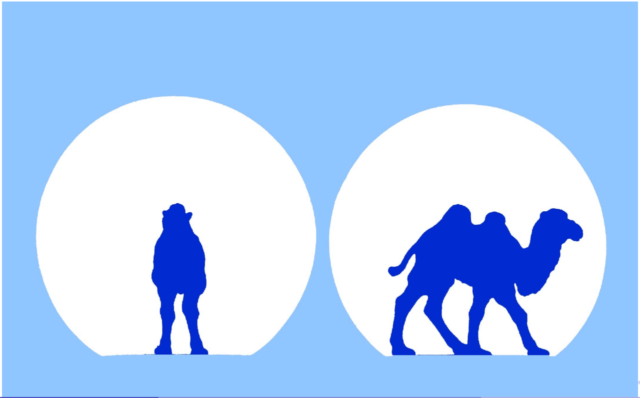
L'inertie des n observations $(x_1^1, \ldots, x_1^p), \ldots, (x_n^1, \ldots, x_n^p)$ *est définie par :*

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

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 $var(c^j) = \lambda_j$, 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 $V_j = (\rho(x^j, c^1), \rho(x^j, c^2))$.

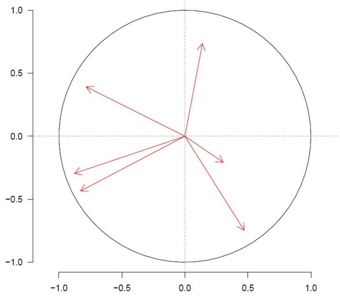


FIGURE 5.3 – Cercle des corrélations de $p = 6$ variables.

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 **éboulis des valeurs propres** :

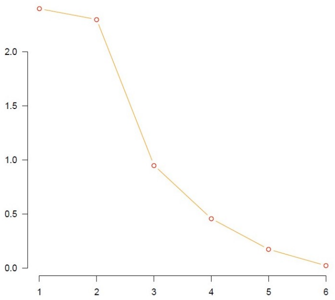


FIGURE 5.4 – Eboulis de $p = 6$ valeurs propres.

Exemple

Réalisez l'ACP de la série de données suivante :

Modèle	Cyl	Puiss	Long	Larg	Poids	V max
Alfasud	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta	1 1570	109	428	162	1060	175
Princess	1798	82	445	172	1160	158
Datsun	1998	115	469	169	1370	160
Taunus	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140