# Partie 1 - Réduction de la dimensionalité (Analyse en composantes principales)

Anne-Sophie Charest

École interdisciplinaire outils et méthodes
Cheminement 2 - Science des données

Apprentissage statistique (*machine learning*)

28 au 30 août 2024

# Plan

# Retour sur la corrélation

# Qu’en savez-vous ?

Retour sur le quiz

`https://forms.gle/nVtg5MR6JFRWoTv47`

# Coefficient de corrélation de Pearson

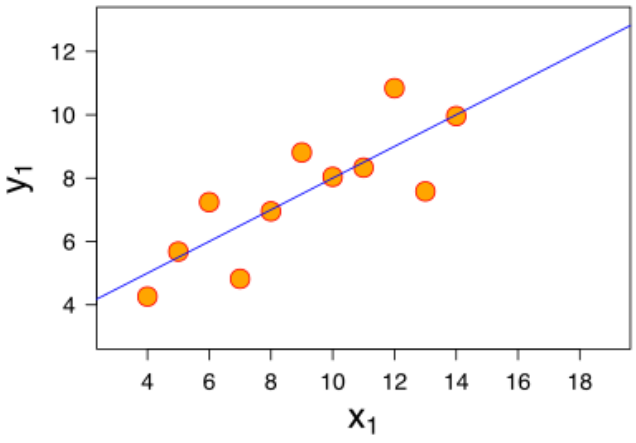
Voici la définition :

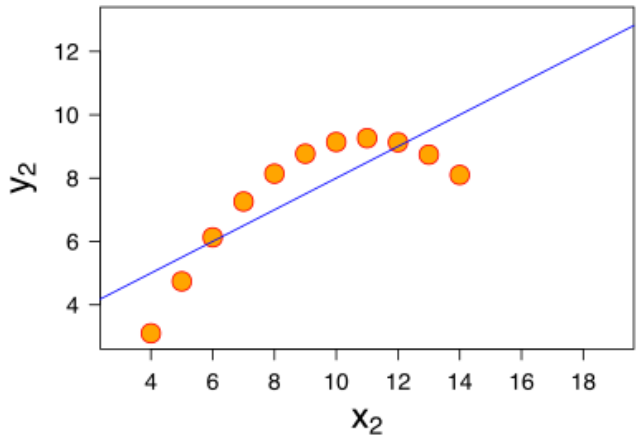
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{(n-1)s_x s_y}$$

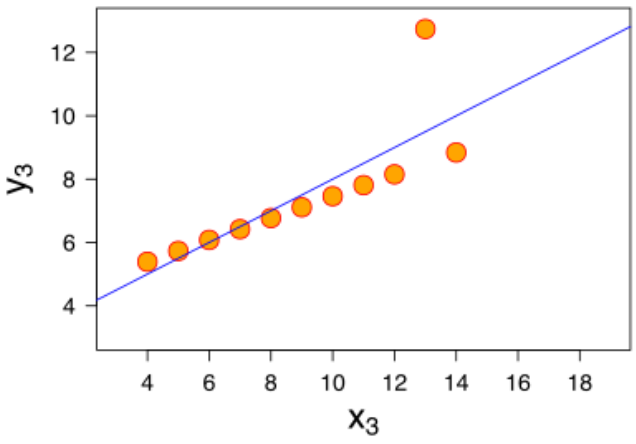
où $x_i$ et $y_i$ sont les valeurs des variables $X$ et $Y$ pour chacune des observations d'un jeu de données ($i = 1, \ldots, n$), et $\bar{x}$, $\bar{y}$, $s_x$ et $s_y$ sont les moyennes et écarts-types échantillonnaux des $x_i$ et $y_i$ respectivement.

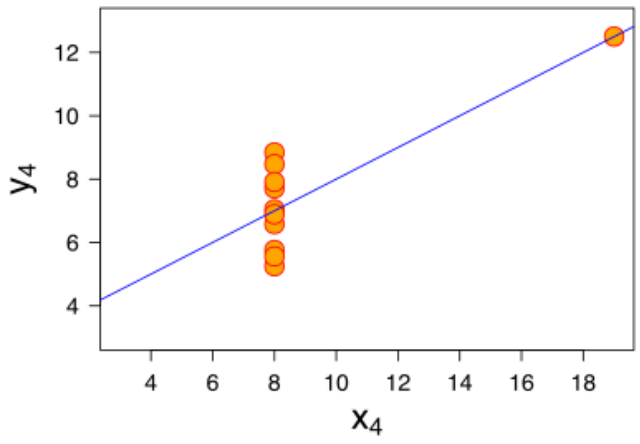
Le coefficient de corrélation $r$ peut prendre des valeurs entre -1 et 1. Plus la valeur de $r$ est proche de 1 en valeur absolue, plus la relation **linéaire** entre les variables est forte.

# Quelques exemples

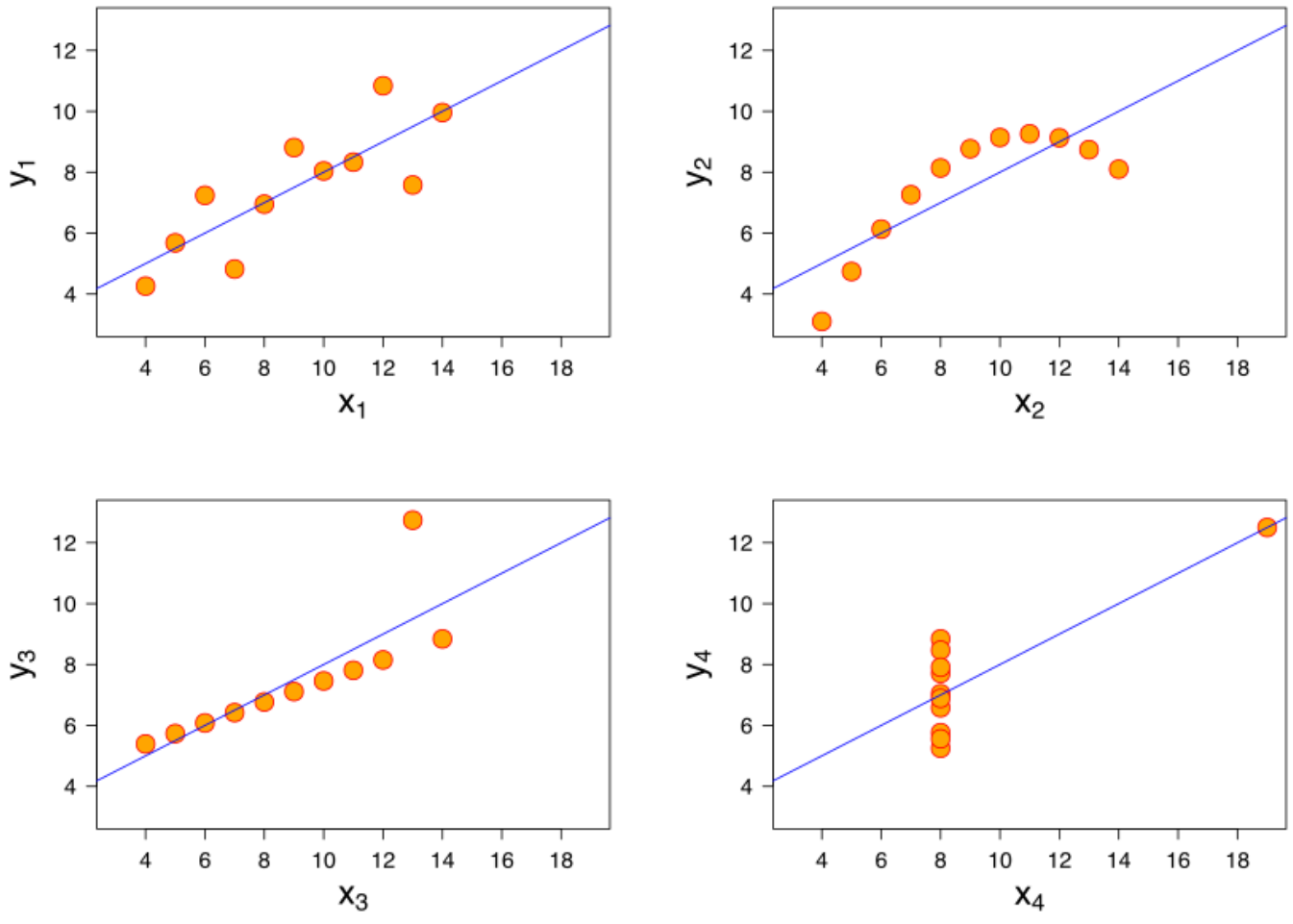


# Quelques exemples

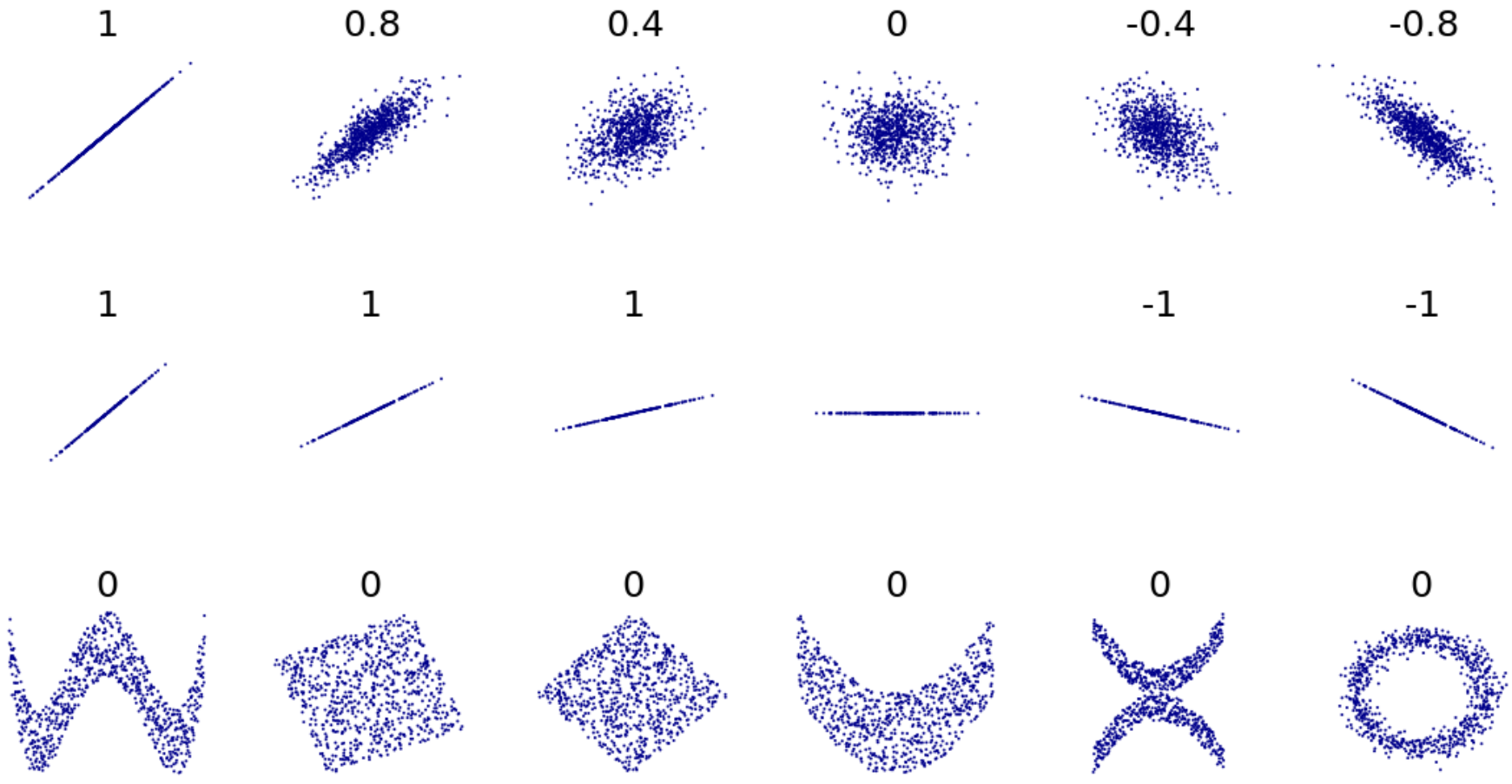



La corrélation est de 0.816 dans les quatre cas !
(voir Anscombe's quartet sur Wikipedia)

# Impact d'une transformation linéaire

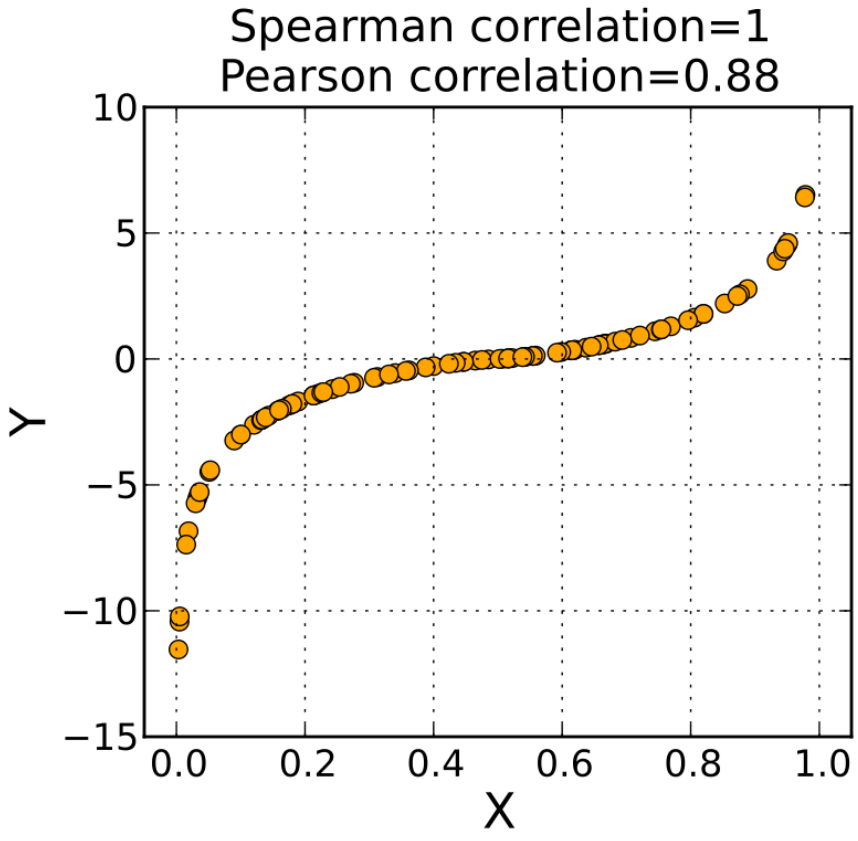
- Le coefficient de corrélation de Pearson reste inchangé lors de l'addition d'une constante, positive ou négative, à toutes les valeurs d'une variable ou même des deux variables.
- De même, la multiplication des valeurs par une constante positive n'affecte pas le coefficient.
- Par contre, si une seule des deux variables est multipliée par une constante négative, le coefficient changera de signe.

# Linéarité du coefficient de corrélation de Pearson




(source : Pearson correlation coefficient sur Wikipedia)

# Alternative : coefficient de corrélation de Spearman




Mesure si deux variables ont tendance à augmenter et diminuer simultanément, sans que le lien entre les deux variables ne soit nécessairement linéaire. On le calcule en utilisant la formule de Pearson, mais en remplaçant les observations par leur rang.

Et il existe d'autres mesures de corrélation, notamment pour des variables catégoriques.

# Transitivité de la corrélation

**Exemple : Paradoxes in Film Ratings**

https://www.tandfonline.com/doi/full/10.1080/10691898.2006.11910579

**Résultat théorique :**

Soit X,Y,Z des variables aléatoires telles que X et Y sont corrélées positivement avec coefficient de corrélation $\rho_{XY}$ et Y et Z sont corrélées positivement avec coefficient de corrélation $\rho_{XY}$. Si $\rho_{XY}^2 + \rho_{YZ}^2 > 1$ alors les variables X et Z sont également corrélées positivement.

Voir : Langford, Eric, Neil Schwertman, and Margaret Owens. *Is the Property of Being Positively Correlated Transitive ?* The American Statistician 55, no. 4 (2001) : 322–25.

http://www.jstor.org/stable/2685695.

# Analyse en composantes principales

# Motivation

Deux exemples en science politique :

- PCA and Brexit
  `https://lennybronner.com/post/2019/04/13/pca-brexit.html`

- Exploring the efficiency of Italian social cooperatives by descriptive and principal component analysis
  `https://link.springer.com/content/pdf/10.1007/s11628-011-0131-9.pdf`

Qu'avez-vous compris du deuxième article ?

# ACP en bref

Une façon d'extraire la structure d'un jeu de données pour en réduire la dimensionnalité.

Pourquoi réduire la dimensionnalité ?

- Visualiser les données
- Identifier des sous-groupes dans les données
- Compresser des images, vidéos
- Simplifier les analyses
  - pour simplifier l'interprétation
  - et/ou en réponse au fléau de la dimension

# Quelques détails

**Données**
Une matrice avec $n$ observations, chacune étant un vecteur de $p$ variables.

**Objectif**
Obtenir une représentation des données dans un espace plus restreint en conservant la plus grande quantité d'information possible.

**Origine de la méthode**
Harold Hotelling (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520.

# Plus précisément

**Espace restreint :** on considère les combinaisons linéaires des variables mesurées.

# Plus précisément

**Espace restreint :** on considère les combinaisons linéaires des variables mesurées.

**Information conservée :** on tente de maximiser la variabilité des données dans le nouvel espace.

# Description mathématiques

Il y a plusieurs façons d'écrire l'ACP mathématiquement. On ira ici pour une présentation plus simple. On considère l'extraction des composantes l'une après l'autre.

# Première composante principale

Soit un jeu de données

$$\mathbf{X} = (X_1, \ldots, X_p)^\top$$

avec matrice de covariance $\Sigma = \mathrm{var}(\mathbf{X})$.

On veut une première composante principale

$$Y_1 = \alpha_1^\top \mathbf{X} = \sum_{i=1}^{p} \alpha_{1i} X_i,$$

qui maximise $\mathrm{var}(Y_1)$.

# Première composante principale (suite)

Il s'agit d'un problème d'optimisation relativement simple.
Pour que $\mathbb{V}\text{ar}(Y_1)$ soit maximale, il faut prendre

(i) $\lambda = \lambda_1$, la plus grande valeur propre de $\Sigma$ ;

(ii) $\alpha_1$, le vecteur propre normé correspondant.

# Deuxième composante principale

On poursuit un objectif double :

(i) conserver le maximum de variation présente dans $\mathbf{X}$ ;

(ii) simplifier la structure de dépendance, pour faciliter l'interprétation.

# Deuxième composante principale (suite)

Étant donné $Y_1$, la deuxième composante principale

$$Y_2 = \alpha_2^\top \mathbf{X}$$

est définie telle que

(i) $\text{var}(Y_2) = \alpha_2^\top \Sigma \alpha_2$ est maximale ;

(ii) $\alpha_2^\top \alpha_2 = 1$

(iii) $\text{cov}(Y_1, Y_2) = 0$.

On peut montrer qu'il faut alors choisir le vecteur propre normé correspondant à la deuxième plus grande valeur propre de $\Sigma$

Procédant par maximisations successives, on conclut que

$$\begin{aligned} Y_k &= \lambda_k, \text{ la } k^{\text{e}} \text{ composante principale} \\ &= \alpha_k^\top \mathbf{X}, \end{aligned}$$

où $\alpha_k$ est le vecteur propre normé de $\Sigma$ associé à $\lambda_k$.

# Matrice des composantes principales

Pour définir simultanément et de façon plus compacte les composantes principales, on pose

$$\mathbf{Y} = \mathbf{A}^\top \mathbf{X},$$

où

$$\mathbf{A} = (\alpha_1, \ldots, \alpha_p) = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix}.$$

La matrice $\mathbf{A}$ a pour colonnes les vecteurs propres de $\Sigma$.

Note :

$$\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}_p, \quad \mathbf{A}^\top = \mathbf{A}^{-1}.$$

# Interprétation des composantes

En analysant les variables qui sont grandement corrélées avec chacune des composantes principales, on peut interpréter ces composantes.

# Scores des observations

La formule

$$\mathbf{Y}_i = \mathbf{A}^\top \mathbf{X}_i$$

donne les coordonnées de l'observation $\mathbf{X}_i$ dans le nouveau système d'axes.

On appelle

$$Y_{ij} = \mathbf{a}_j^\top \mathbf{X}_i = \sum_{k=1}^{p} a_{jk} X_{ik}$$

le score de $\mathbf{X}_i$ sur l'axe principal $j$.

Note : On peut montrer que la distance entre les observations dans le nouveau système d'axes est la même que celle entre les observations dans l'espace des données originales.

# Variation expliquée par chaque CP

La trace de Pillai

$$\text{trace}(\Sigma) = \text{trace}(\Lambda) = \sum_{i=1}^{p} \lambda_i,$$

est une mesure globale de variation.

Ainsi, la proportion de variation expliquée par $Y_i$ est

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

# Estimation de $\Sigma$

Dans la pratique, la matrice $\Sigma$ est inconnue.

Cependant, elle peut être estimée par

$$\hat{\Sigma} = \frac{\mathbf{S}}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^\top$$

à partir d'un échantillon aléatoire $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

# Un exemple

Voir le début du labo R sur l'ACP.

# Point crucial

On peut faire soit :

- l'ACP de la matrice des covariances ;
- l'ACP de la matrice des corrélations.

La seconde se fait à partir des variables standardisées.

# Point crucial

On peut faire soit :

- l'ACP de la matrice des covariances ;
- l'ACP de la matrice des corrélations.

La seconde se fait à partir des variables standardisées.

Elle est recommandée, à moins que les variables soient de variances semblables, ou que la différence de variabilité contienne de l'information d'intérêt.

# Choix du nombre de composantes

- Dépend de l'utilisation qu'on veut en faire.
- Pour la visualisation, toujours plus facile avec 2 ou 3.
- Je présente ici 3 règles parfois utilisées.
- Il existe aussi des règles plus avancées, notamment basées sur des méthodes de rééchantillonage.

# $1^{re}$ règle

Garder autant de composantes que nécessaire pour expliquer 80% de la variation.

Pourquoi 80% ? C'est purement arbitraire !

# $2^{e}$ règle : Kaiser (1960)

Si l'ACP est effectuée sur la matrice des corrélations,

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq 1.$$

Note : $l_k$ est le $k^{\text{i-ème}}$ vecteur propre de l'ACP de la matrice des corrélations empirique.

**Source** :

H. F. Kaiser (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.

# $2^{\text{e}}$ règle : Kaiser (1960)

## Autre formulation

Peu importe la matrice sur laquelle l'ACP est effectuée,

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq \bar{\ell},$$

où

$$\bar{\ell} = (\ell_1 + \cdots + \ell_p)/p.$$

On a $\bar{\ell} = 1$ pour une matrice des corrélations.

# $2^{e}$ règle : Kaiser (1960)

## Opinion divergente

Jolliffe (1972) recommande plutôt

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq 0.7.$$

**Source** :

I. T. Jolliffe (1972). Discarding variables in a principal component analysis I : Artificial data. *Applied Statistics*, 21, 160–173.

# $3^{e}$ règle : Cattell (1966)

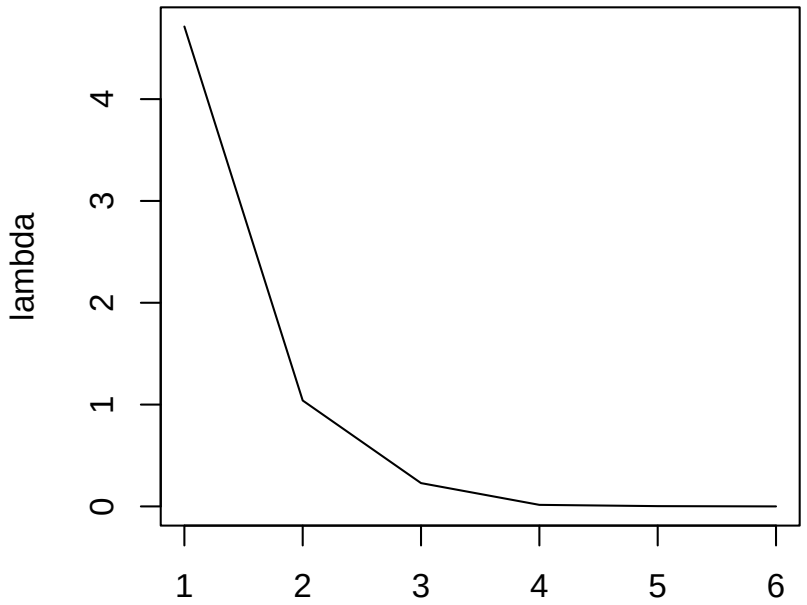
Dans le graphe des paires $(k, \ell_k)$,

garder les $\ell_k$ précédant le “pied de l’éboulis.”

**Source** :

R. B. Cattell (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.

Exemple :



# Extensions de l'ACP classique

- ACP avec noyaux
  Pour permettre de considérer autre chose que des combinaisons linéaires des variables.
- ACP parcimonieuse
  Offre une façon de limiter le nombre de variables incluses dans une composante principale.
- ACP avec données manquantes
  Plusieurs approches, avec ou sans imputation des données manquantes.
  Dans tous les cas il faut faire très attention aux hypothèses faites par ces méthodes.

# Autres approches

# Méthodes classiques

Il existe toute une panoplie de méthodes, similaires à l'ACP mais avec des objectifs et des stratégies un peu différentes :

- Analyse factorielle
- Analyse des coordonnées
  (ou positionnement multidimensionel classique)
- Analyse canonique des corrélations
- Analyse des correspondances
- Analyse conjointe
- etc.

Plusieurs de ces méthodes ont pour objectif de visualiser le jeu de données pour en faciliter l'interprétation. L'ACP est la méthode la plus fréquemment utilisée pour la réduction de la dimensionnalité dans le contexte d'une analyse supervisée (régression, classification).

# Positionnnement multidimensionel classique

Particulièrement utile si nos données ne sont pas sous la forme de variables pour diverses observations, mais directement d'une matrice de distances entre différentes observations.

Voir par exemple `https://webthesis.biblio.polito.it/9522/1/tesi.pdf`

R permet de faire du MDS classique (fonction `cmdscale()` ou du MDS non-métrique (fonction `isoMDS()` de la librairie `MASS`).