# Analyse en composantes principales

Julien JACQUES

Université de Lyon, Lyon 2 & ERIC EA 3083

# Un exemple de données

Données `USArrests` :

- nombres d'agressions, de meurtres et de viols (par 100 000 habitants)
- pourcentage de population urbaine
- pour chacun des 50 états des USA en 1973

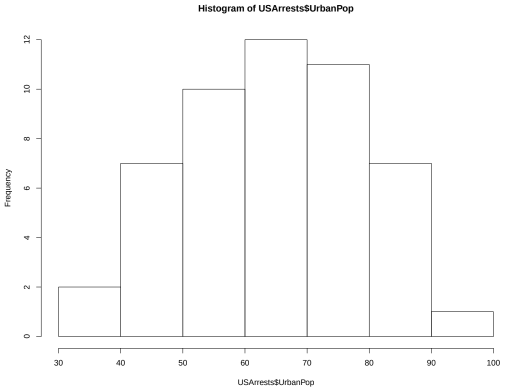
On dispose donc d'une matrice de 50 lignes (les 50 états) et 4 colonnes.

# Visualiser les données lorsque $p = 1$

- Question 1 : représenter le pourcentage de population urbaine.

# Visualiser les données lorsque $p = 1$

- Question 1 : représenter le pourcentage de population urbaine.
- Réponse : histogramme

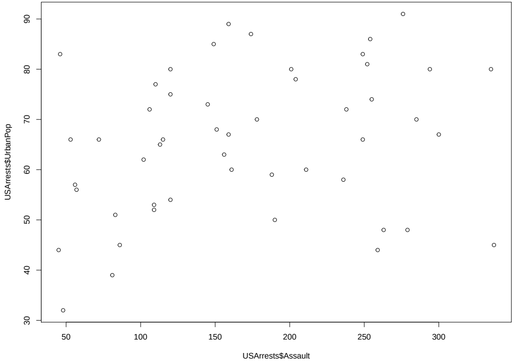


# Visualiser les données lorsque $p = 2$

- Question 2 : représenter le nombre d'agressions en fonction du pourcentage de population urbaine.

# Visualiser les données lorsque $p = 2$

- Question 2 : représenter le nombre d'agressions en fonction du pourcentage de population urbaine.
- Réponse : nuage de points



# Visualiser les données lorsque $p > 2$

- Question 3 : représenter le nombre d'agressions, de meurtres et de viols en fonction du pourcentage de population urbaine.

# Visualiser les données lorsque $p > 2$

- Question 3 : représenter le nombre d'agressions, de meurtres et de viols en fonction du pourcentage de population urbaine.
- Réponse : ???

# Visualiser les données lorsque $p > 2$

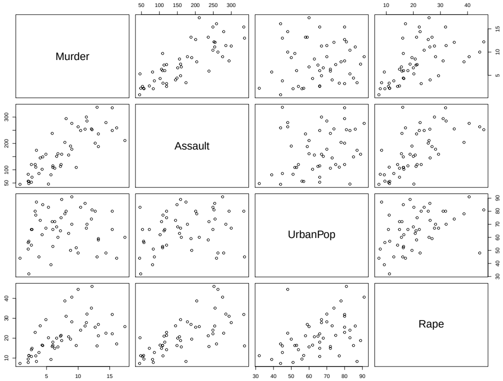
- Question 3 : représenter le nombre d'agressions, de meurtres et de viols en fonction du pourcentage de population urbaine.
- Réponse : ???

## Mathématiquement

- chaque observation est un point dans un espace de 4 dimensions : $\mathbb{R}^4$
- on ne peut visualiser un espace de dimension supérieur à 3 (et encore en dimension 3 ce n'est pas si facile...)
- ce que l'on sait bien visualiser est la dimension 2 !

# Visualiser les données lorsque $p > 2$

- Question 3 : représenter le nombre d’agressions, de meurtres et de viols en fonction du pourcentage de population urbaine.
- Réponse : une solution non optimale, le biplot



# Visualiser les données lorsque $p > 2$

- Question 3 : représenter le nombre d'agressions, de meurtres et de viols en fonction du pourcentage de population urbaine.
- Réponse : une solution optimale, l'analyse en composantes principales (ACP)

# Analyse en composantes principales

## Les data

On stocke les données sous forme d'une matrice

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

où $x_{ij}$ : valeur de la $j$-ème variable pour le $i$-ème individu.
On définit :

- la moyenne de la variable $j$ : $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$
- l'écart-type de la variable $j$ : $s_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$
- la distance entre deux individus : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$

# Analyse en composantes principales

On définit également:

- la covariance entre les variables $j$ et $k$:
  $cov_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$
- la corrélation entre les variables $j$ et $k$: $cor_{jk} = \frac{cov_{jk}}{s_j s_k} \in [-1, 1]$
- la matrice de variance-covariance des données $\mathbf{X}$:

$$\Sigma = \begin{pmatrix} s_1^2 & cov_{12} & \dots & cov_{1p} \\ cov_{21} & s_2^2 & \dots & cov_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ cov_{p1} & cov_{p2} & \dots & s_p^2 \end{pmatrix}$$

# Analyse en composantes principales

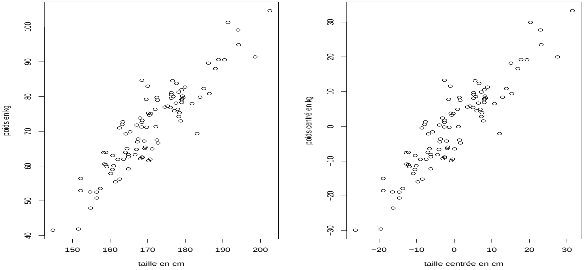
## Les objectifs

- visualiser les données (nuage de points en dimension $p$) dans un espace de faible dimension (2)
- identifier des individus semblables
- identifier des liens entre variables (liens linéaires, coefficient de corrélation)

# Analyse en composantes principales

## Pré-traitement : centrage réduction

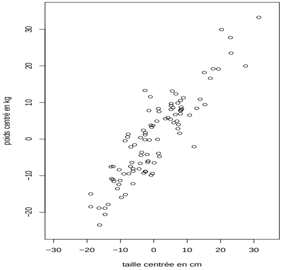
- centrer les données (retrancher à chaque variable sa moyenne) ne modifie pas la forme du nuage : on le fera toujours

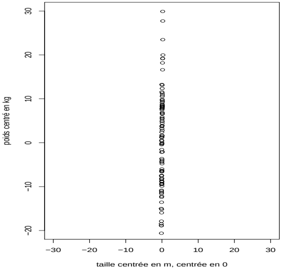


# Analyse en composantes principales

## Pré-traitement : centrage réduction

- centrer les données (retrancher à chaque variable sa moyenne) ne modifie pas la forme du nuage : on le fera toujours
- réduire les données (diviser chaque variable pas son écart-type) permet de s'affranchir des unités de mesures. En effet, le changement d'unité de mesure modifie la forme du nuage de points :



# Analyse en composantes principales

## Pré-traitement : centrage réduction

- centrer les données (retrancher à chaque variable sa moyenne) ne modifie pas la forme du nuage : on le fera toujours
- réduire les données (diviser chaque variable pas son écart-type) permet de s'affranchir des unités de mesures.
- Ainsi, on transformera chaque les observations comme suit :

$$x_{ij} \to \frac{x_{ij} - \bar{x}_j}{s_j}$$

# Analyse en composantes principales

## Recherche du sous-espace de projection optimal

- Visuellement

# Analyse en composantes principales

## Recherche du sous-espace de projection optimal

- Visuellement
- Caractérisation mathématique
  - projeter le nuage diminue nécessairement les distances entre individus
  - $\Rightarrow$ limiter la déformation = maximiser les distances entre individus
  - maximiser les distances entre individus $\Leftrightarrow$ maximiser la variance du nuage de points projeté

# Analyse en composantes principales

## Recherche du sous-espace de projection optimal

- Résolution
  - on recherche le premier axe (factoriel) $\mathbf{f}_1$ maximisant la variance des points projetés
  - on recherche ensuite un second $\mathbf{f}_2$ selon le même critère, mais orthogonal à $\mathbf{f}_1$ pour ne pas transcrire d'information redondante
  - $(\mathbf{f}_1, \mathbf{f}_2)$ forme le premier plan factoriel.
  - on peut continuer avec $\mathbf{f}_3, \mathbf{f}_4, \ldots$ suivant la quantité d'information qu'ils retranscrivent

# Analyse en composantes principales

## Recherche du sous-espace de projection optimal

- Résolution
  - on recherche le premier axe (factoriel) $\mathbf{f}_1$ maximisant la variance des points projetés
  - on recherche ensuite un second $\mathbf{f}_2$ selon le même critère, mais orthogonal à $\mathbf{f}_1$ pour ne pas transcrire d'information redondante
  - $(\mathbf{f}_1, \mathbf{f}_2)$ forme le premier plan factoriel.
  - on peut continuer avec $\mathbf{f}_3, \mathbf{f}_4, \ldots$ suivant la quantité d'information qu'ils retranscrivent
- Mathématiquement
  - chaque axe factoriel $\mathbf{f}_j$ est un axe dans l'espace $\mathbb{R}^p$ : il peut être vu comme une nouvelle variable, synthétique, définie comme une combinaison linéaire des variables initiales
  - $\mathbf{f}_1, \mathbf{f}_2, \ldots$ sont les vecteurs propres de la matrice de variance $\Sigma$ associés aux plus grande valeurs propres $\lambda_1 > \lambda_2 > \ldots$

# Analyse en composantes principales

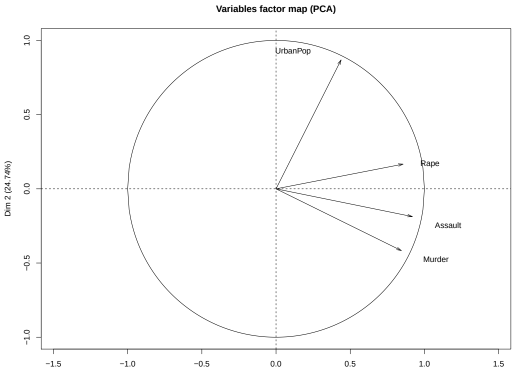
## ACP sous R

- `library(FactoMineR)`
- `res.pca <- PCA(USArrests, graph = FALSE)`

# Analyse en composantes principales

## Interprétation des axes factoriels

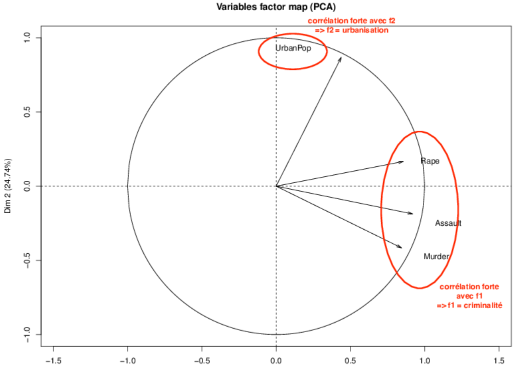
- on examine les corrélations entre les variables initiales et les axes (variables) factoriel(le)s : $r(\mathbf{x}_j, \mathbf{f}_\ell) \in [-1, 1]$
  `plot(res.pca,choix = "var")`



# Analyse en composantes principales

## Interprétation des axes factoriels

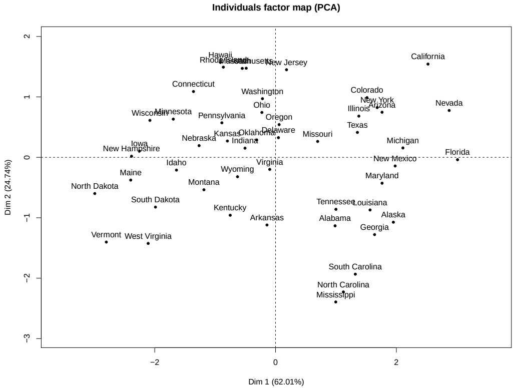
- on examine les corrélations entre les variables initiales et les axes (variables) factoriel(le)s : $r(\mathbf{x}_j, \mathbf{f}_\ell) \in [-1, 1]$
  `plot(res.pca,choix = "var")`



# Analyse en composantes principales

## Représentation (projection) des individus

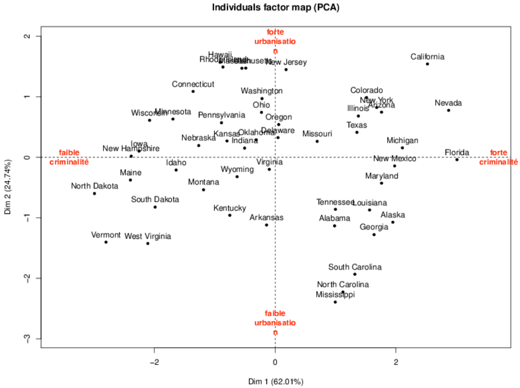
- on projette les individus sur les axes (variables) factoriel(le)s :
  `plot(res.pca,choix = "ind")`



# Analyse en composantes principales

## Représentation (projection) des individus

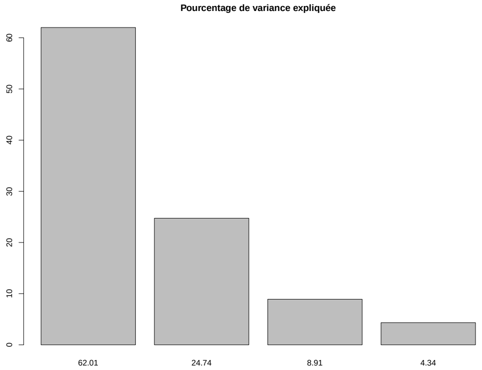
- on projette les individus sur les axes (variables) factoriel(le)s :
  `plot(res.pca,choix = "ind")`



# Analyse en composantes principales

## Choix du nombre d'axes factoriels

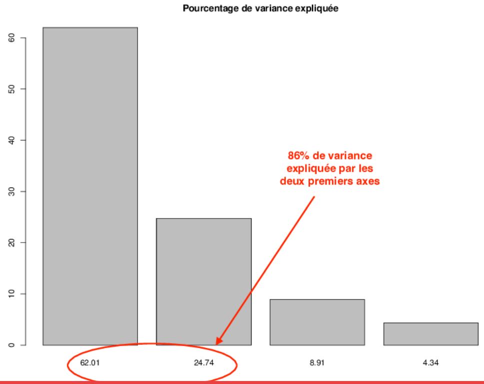
- on calcule le pourcentage d'information retranscrite par chaque axe :
  `barplot(res.pca$eig[,2],main="Pourcentage de variance expliquée")`



# Analyse en composantes principales

## Choix du nombre d'axes factoriels

- on calcule le pourcentage d'information retranscrite par chaque axe :
  `barplot(res.pca$eig[,2],main="Pourcentage de variance expliquée")`

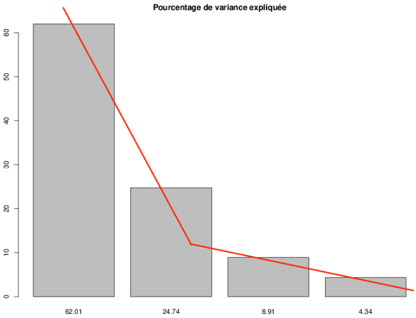


# Analyse en composantes principales

## Choix du nombre d'axes factoriels

Plusieurs façon de choisir

- retenir le nombre d'axes pour expliquer au moins 80% de la variance
- rechercher un coude dans l'ébouli des valeurs propres



# Analyse en composantes principales

## Aide à l'interprétation - variables

- corrélation variables/axes : `res.pca$var$coord`

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 |
|---|---|---|---|---|
| Murder | 0.84 | -0.41 | 0.20 | 0.27 |
| Assault | 0.91 | -0.18 | 0.16 | -0.30 |
| UrbanPop | 0.43 | 0.86 | 0.22 | 0.05 |
| Rape | 0.85 | 0.16 | -0.48 | 0.03 |

# Analyse en composantes principales

## Aide à l'interprétation - variables

- corrélation variables/axes : `res.pca$var$coord`

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 |
|---|---|---|---|---|
| Murder | 0.84 | -0.41 | 0.20 | 0.27 |
| Assault | 0.91 | -0.18 | 0.16 | -0.30 |
| UrbanPop | 0.43 | 0.86 | 0.22 | 0.05 |
| Rape | 0.85 | 0.16 | -0.48 | 0.03 |

- contributions des variables aux axes : `res.pca$var$contrib`

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | total |
|---|---|---|---|---|---|
| Murder | 28.71 | 17.48 | 11.64 | 42.14 | 100 |
| Assault | 34.01 | 3.53 | 7.19 | 55.26 | 100 |
| UrbanPop | 7.73 | 76.17 | 14.28 | 1.79 | 100 |
| Rape | 29.53 | 2.79 | 66.87 | 0.79 | 100 |

# Analyse en composantes principales

## Aide à l'interprétation - variables

- corrélation variables/axes : `res.pca$var$coord`

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 |
|---|---|---|---|---|
| Murder | 0.84 | -0.41 | 0.20 | 0.27 |
| Assault | 0.91 | -0.18 | 0.16 | -0.30 |
| UrbanPop | 0.43 | 0.86 | 0.22 | 0.05 |
| Rape | 0.85 | 0.16 | -0.48 | 0.03 |

- contributions des variables aux axes : `res.pca$var$contrib`

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | total |
|---|---|---|---|---|---|
| Murder | 28.71 | 17.48 | 11.64 | 42.14 | 100 |
| Assault | 34.01 | 3.53 | 7.19 | 55.26 | 100 |
| UrbanPop | 7.73 | 76.17 | 14.28 | 1.79 | 100 |
| Rape | 29.53 | 2.79 | 66.87 | 0.79 | 100 |

- qualité de la représentation des variables sur les axes : `res.pca$var$cos2`

# Analyse en composantes principales

## Aide à l’interprétation - individus

- coordonnées des individus sur les axes : `res.pca$ind$coord`
- contributions des individus aux axes : `res.pca$ind$contrib`
- qualité de la représentation des individus sur les axes : `res.pca$ind$cos2`

# Analyse en composantes principales

## Individus et variables supplémentaires

- il est possible de mettre des individus et/ou des variables (quantitatives ou qualitatives) en supplémentaire, ce qui signifie qu'ils n'interviennent pas dans le calcul des axes factoriels, mais ils seront représentés et il sera possible d'interpréter leur positions :
  `PCA(..., ind.sup = NULL, quanti.sup = NULL, quali.sup = NULL, ...)`

# Analyse en composantes principales

## Exercice d'application

- Réaliser une ACP du jeu de données `autos.xls`, en indiquant les variables `finition`, `prix` et `r-poid.puis` en supplémentaire.