Analyses descriptives Analyse en composantes principales (ACP)

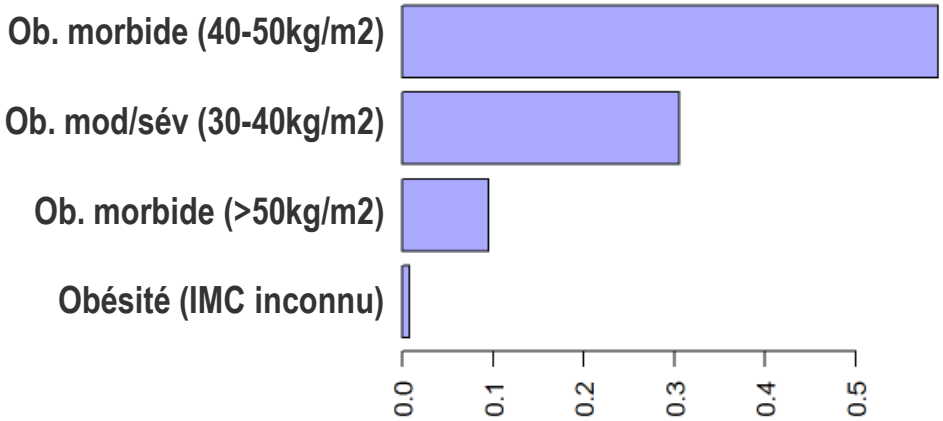
I. Analyses univariées - rappel
II. Analyses bivariées – rappel
III. Approche intuitive de l'ACP
IV. L'ACP
V. Exemples d'interprétation d'ACP
VI. L'ACM en bref

Rappel sur les analyses univariées

I. Variables qualitatives
II. Variables quantitatives discrètes
III. Variables quantitatives continues
IV. Variables binaires
V. Survie (variable censurée à droite)

Description univariée d'une variable qualitative

- Indice de masse corporelle de patients opérés en chirurgie bariatrique




```
Effectif : 215776
Valeurs manquantes : n= 10667 soit 4.71%.
```

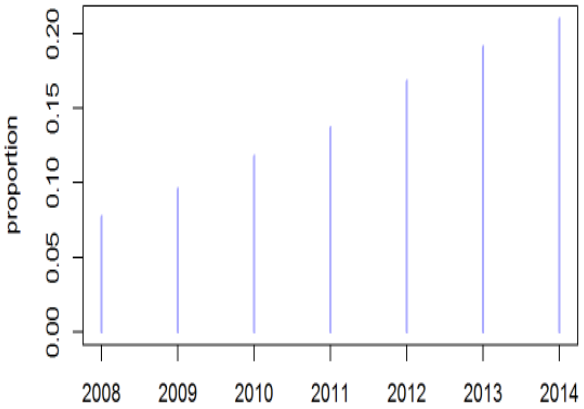
Modalite	Effectif	Proportion	IC95%
Ob. morbide (40-50kg/m2)	127497	59.09%	[58.88%;59.30%]
Ob. mod/sév (30-40kg/m2)	65994	30.58%	[30.39%;30.78%]
Ob. morbide (>50kg/m2)	20590	9.54%	[9.42%;9.67%]
Obésité (IMC inconnu)	1695	0.79%	[0.75%;0.82%]

- Graphique :
  - Diagramme en barres
  - Diagramme en secteurs (camembert)
- Métriques synthétiques :
  - Mode
- Métriques pour chaque modalité :
  - Nom
  - Effectif
  - Proportion
  - Intervalle de confiance

Description univariée d'une variable quantitative discrète

- Année de l'intervention chirurgicale (chirurgie bariatrique)

```
Effectif  226443
Min.        2008
1er Qu.     2010
Mediane     2012
Moyenne     2011.64
Ecart type     1.90
3ème Qu.    2013
Max.        2014
Aucune valeur manquante.
```

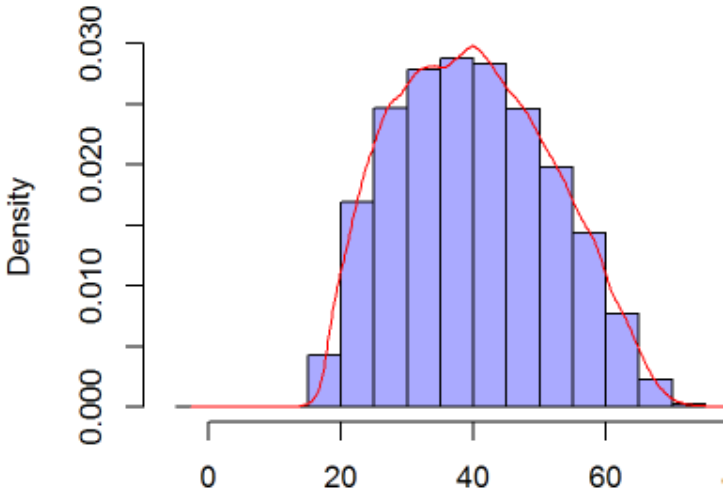


Modalite	Effectif	Proportion	IC95%
2008	17659	7.8%	[7.69%;7.91%]
2009	21778	9.62%	[9.5%;9.74%]
2010	26789	11.83%	[11.7%;11.96%]
2011	31160	13.76%	[13.62%;13.9%]
2012	38131	16.84%	[16.69%;16.99%]
2013	43382	19.16%	[19%;19.32%]
2014	47544	21%	[20.83%;21.16%]

- Graphique :
  - Diagramme en bâtons
- Métriques synthétiques :
  - Min, max, quartiles, médiane
  - Moyenne, écart type
- Métriques pour chaque modalité :
  - Nom
  - Effectif
  - Proportion
  - Intervalle de confiance

Description univariée d'une variable quantitative continue

- Âge des patients (chirurgie bariatrique)

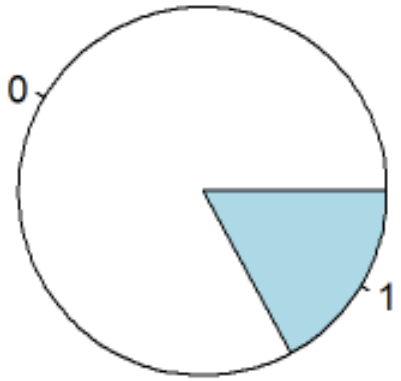



```
Effectif  226443
Min.          16
1er Qu.       30.49
Médiane       39.5
Moyenne       40.07
Ecart type    11.74
3ème Qu.      48.5
Max.          75
Aucune valeur manquante.
```

- Graphique :
  - Histogramme (estime densité de probabilité)
  - Fréquences cumulées (estime fonction de répartition)
  - Boxplot
- Métriques synthétiques :
  - Min, max, quartiles, médiane
  - Moyenne, écart type

Description univariée d'une variable binaire

- Sexe des patients (chirurgie bariatrique)



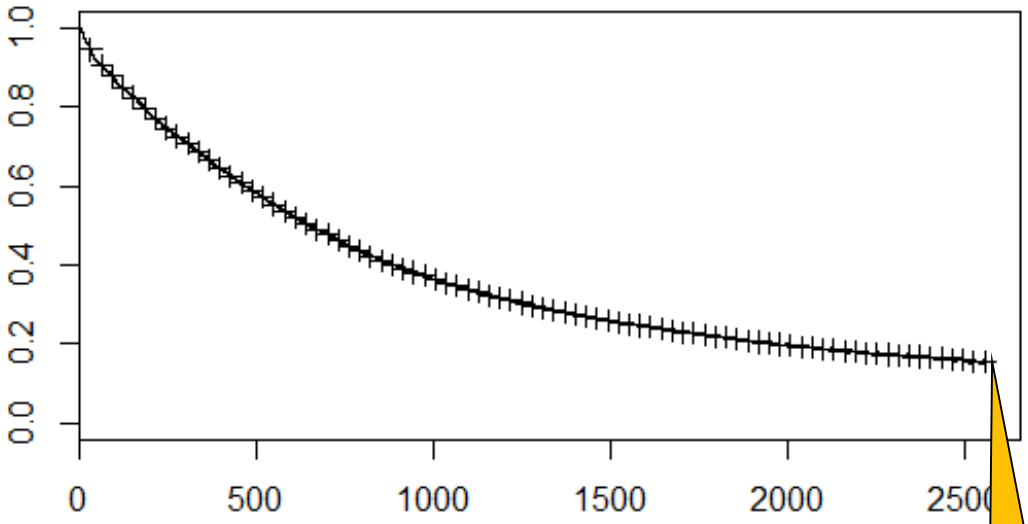

```
Effectif  226443
Moyenne          0.1709
Ecart type       0.376
Aucune valeur manquante
```

Modalite	Effectif	Proportion	IC95%
0 (F)	187736	82.91%	[82.75%;83.06%]
1 (H)	38707	17.09%	[16.94%;17.25%]

- Graphique :
  - Diagramme en secteurs
- Métriques synthétiques :
  - Moyenne =proportion_de_1=p
  - Ecart type $=\sqrt{p(1-p)}$
- Métriques pour chaque modalité :
  - Nom
  - Effectif
  - Proportion
  - Intervalle de confiance

Description univariée d'une variable de survie, censurée à droite

- Suite à une chirurgie bariatrique, risque dans le temps d'être ré-hospitalisé



Effectif départ	216766
Evénements	122811
Suivi médian (j)	646
Suivi max (j)	2555

~~56%~~ ≠82%

- Courbe de Kaplan-Meier : probabilité au fil du temps que l'événement ne soit toujours pas survenu
- Individus avec événement : font descendre la courbe (seul le premier événement est pris en compte)
- Individus suivis sans événement (exclus vivants ou perdus de vue) : ne font pas descendre la courbe, mais ne participent plus au dénominateur à partir d'une certaine date. Symbolisés par « + » sur la courbe.

Individus avec événement : généralement suivis moins longtemps que les individus censurés

Analyses bivariées

Mise en évidence d'une dépendance statistique entre deux variables

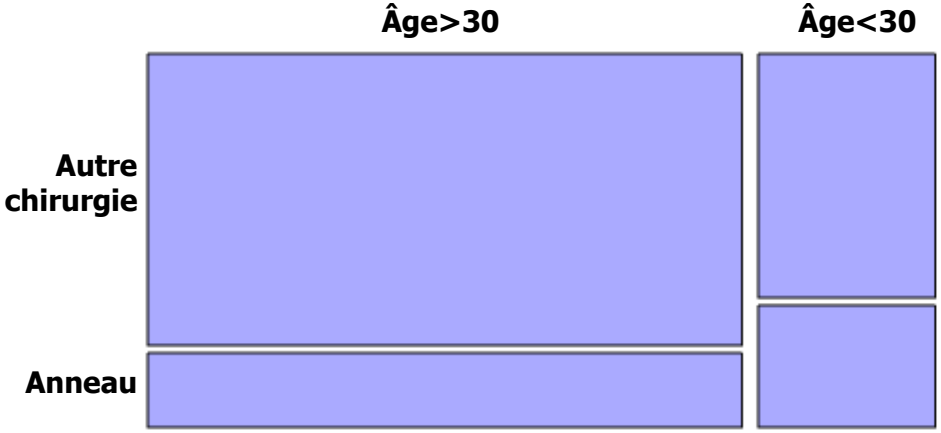
I. Qualitatif – qualitatif
II. Qualitatif – quantitatif
III. Quantitatif – quantitatif
IV. Survie – qualitatif

Notes préalables

- On peut faire plus que ce qui vous est montré. Ici : version simplifiée, 1 seul graphique à chaque fois
- Certaines variables pourront parfois être traitées comme quantitatives ou qualitatives, selon les préférences :
  - Variables quantitatives discrètes (avec ou sans seuil)
  - Variables binaires
  - Variables quantitatives, lorsqu'un seuil peut être appliqué

Entre deux variables qualitatives

- Parmi les chirurgies bariatriques, proportion d'anneau gastrique en fonction de l'âge




```
Test du Chi2 de Pearson
Chi2 = 3805.4
ddl = 1
p-valeur < 2.2e-16
```

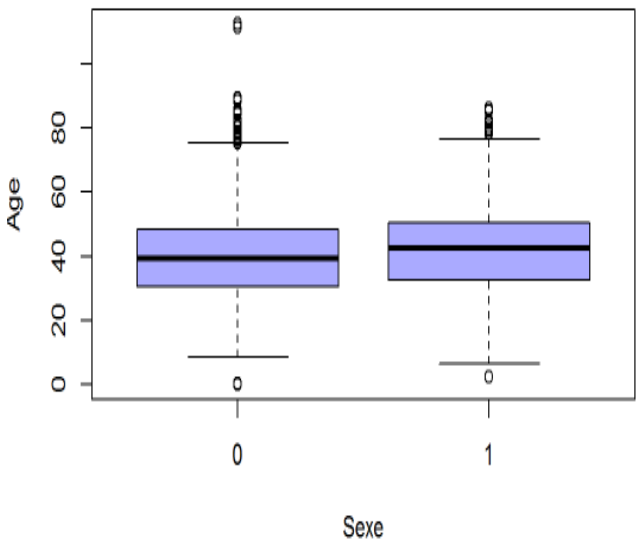
Tableau des effectifs :

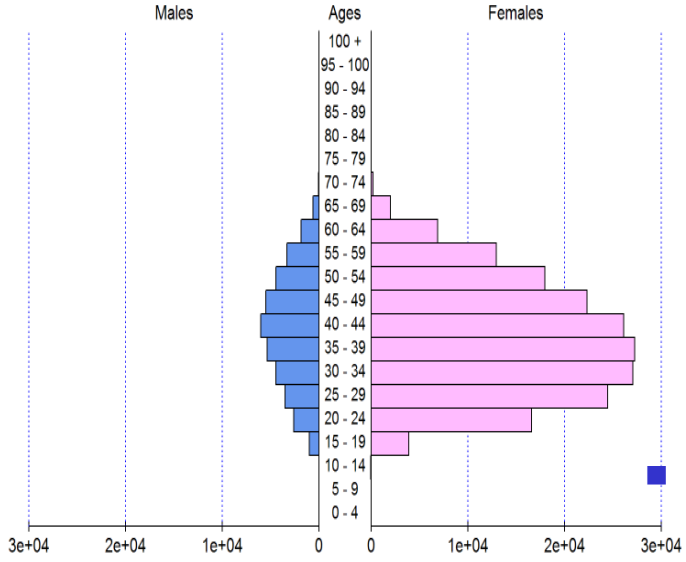
	age>30	age<30
Autre_chir	137027	34173
Anneau gas	34883	17147

- Graphique :
  - Mosaïque (proportionnalité entre surface et effectif. Ici, partition en colonnes puis en lignes)
- Test :
  - Khi²

Entre une variable qualitative et une variable quantitative

- Relation entre l'âge et le sexe (chirurgies bariatriques)





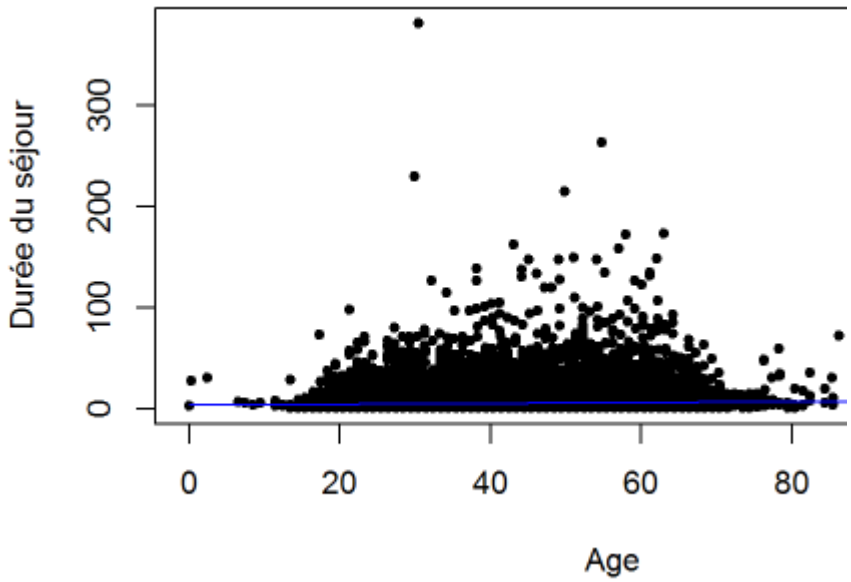

```
Test de Student
P-valeur : 4.825e-240
```

Sexe	Effectif	MoyenneAge	IC95%
0	187736	39.7	[39.6;39.7]
1	38707	41.9	[41.7;42.0]

- Graphique :
  - Boxplots : une boxplot de la variable quantitative par modalité de la variable qualitative, mises côte à côte
  - Ou cas particulier : ici pyramide des âges (un histogramme en effectif par modalité)
- Test :
  - Student si 2 modalités de la variable qualitative
  - ANOVA (analyse de la variance) si plus de modalités

Entre deux variables quantitatives

- Relation entre la durée de séjour et l'âge (chirurgies bariatriques)



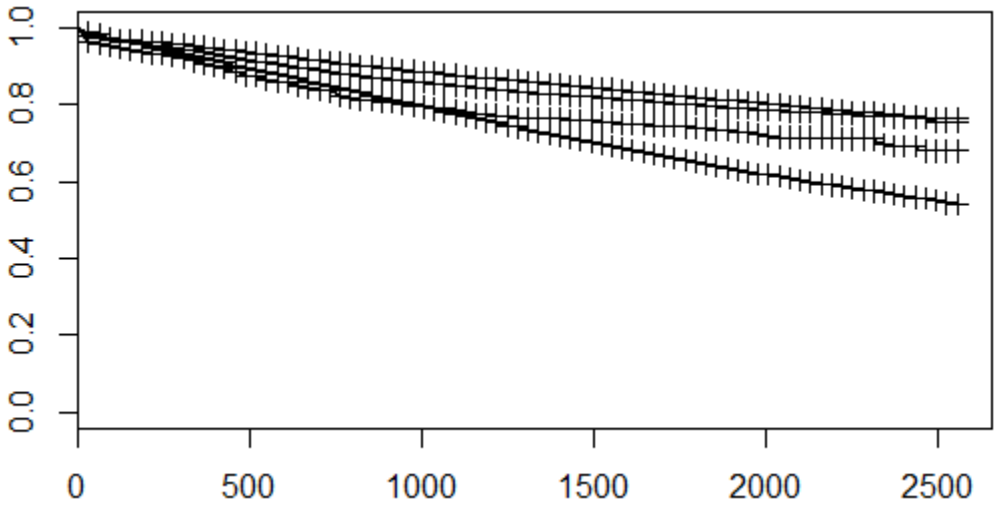

```
Coefficient de corrélation :
r = 0.149
r²= 0.02215
p-valeur = 0

Equation droite : y = 0.0444*x + 4.29 (+ bruit)
```

- Graphique :
  - Nuage de points : un point par individu, coordonnées = variables
  - Droite de régression surimprimée
- Test :
  - Test de nullité du coefficient de corrélation de Pearson ou de Spearman
  - Régression linéaire simple => équation de la droite de régression

Entre une survie et une variable qualitative

- Relation entre le type d'opération et le risque de réhospitalisation (chirurgies bariatriques)



- Graphique :
  - Une courbe de Kaplan-Meier pour chaque modalité de la variable qualitative
- Test (ne pas retenir) :
  - Test du Log-Rank
  - Modèle de Cox

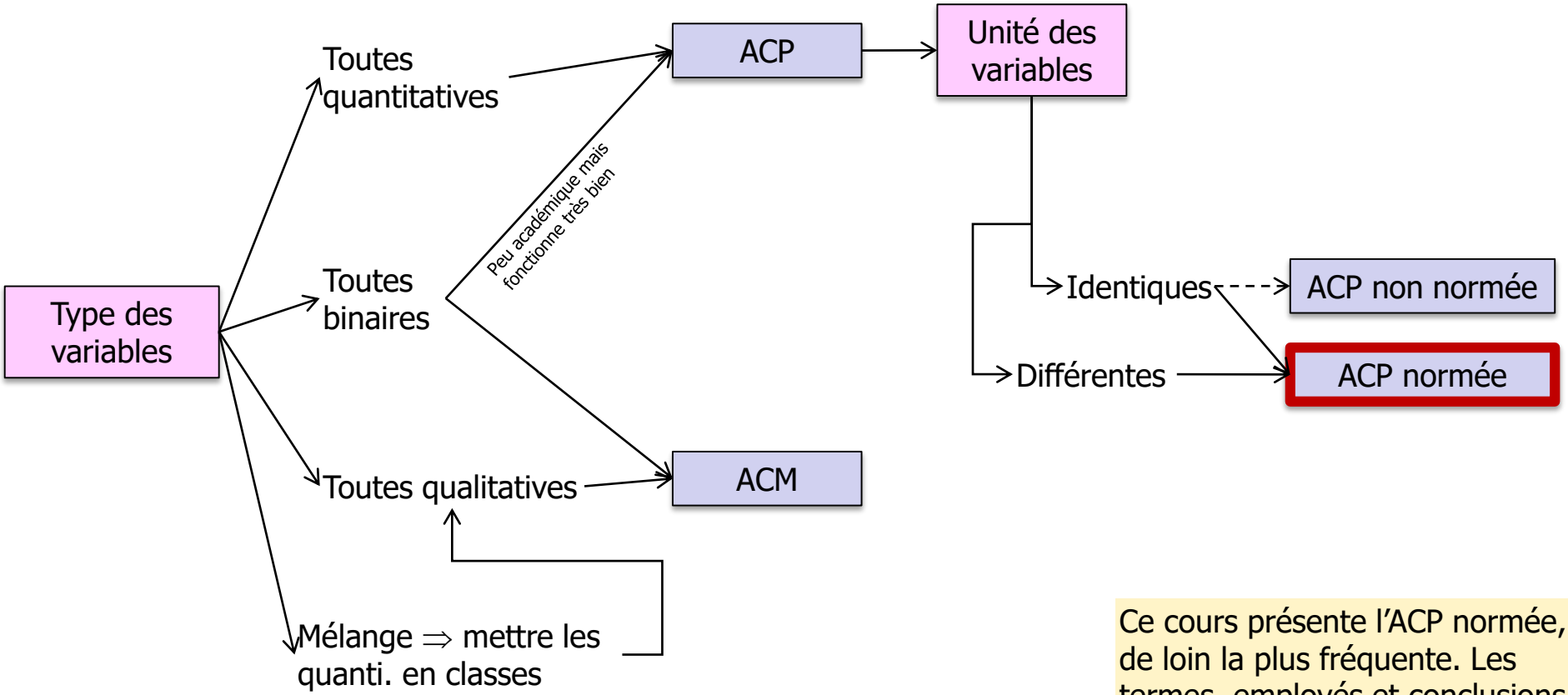
Analyses multivariées

Mise en évidence d’une dépendance statistique entre plus de deux variables

Analyses multivariées

- Mise en évidence de relations entre plus de 2 variables simultanément
- Analyses non-supervisées :
  - Relation sans parti pris entre $\{X_1, X_2, X_3, \ldots, X_n\}$
  - Ex :
    - analyse en composantes principales (objet de ce cours)
    - analyse des correspondances multiples
- Analyses supervisées :
  - Relation type $\{X_1, X_2, X_3, \ldots, X_n\} \rightarrow Y$
  - Y est la variable à expliquer
  - Ex :
    - régression multiple (vu dans le prochain cours)
    - arbres de décision (vu dans le cours « Pharmacovigilance »)

L'ACP, une sorte « d'analyse multiple de corrélations » entre variables quantitatives



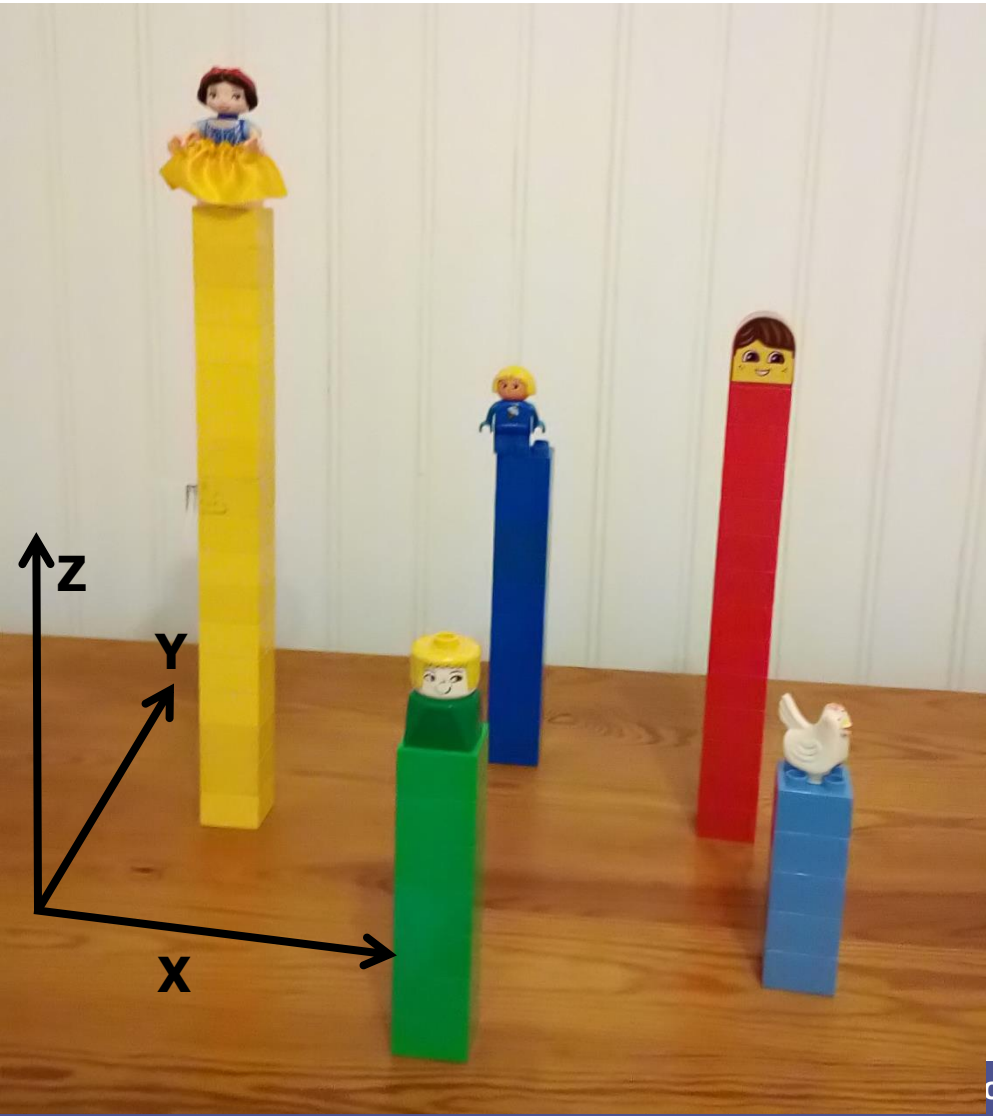
Ce cours présente l'ACP normée, de loin la plus fréquente. Les termes employés et conclusions ne sont pas strictement généralisables à toutes les ACP.

Approche intuitive : principes de la réduction de dimension

Les personnages

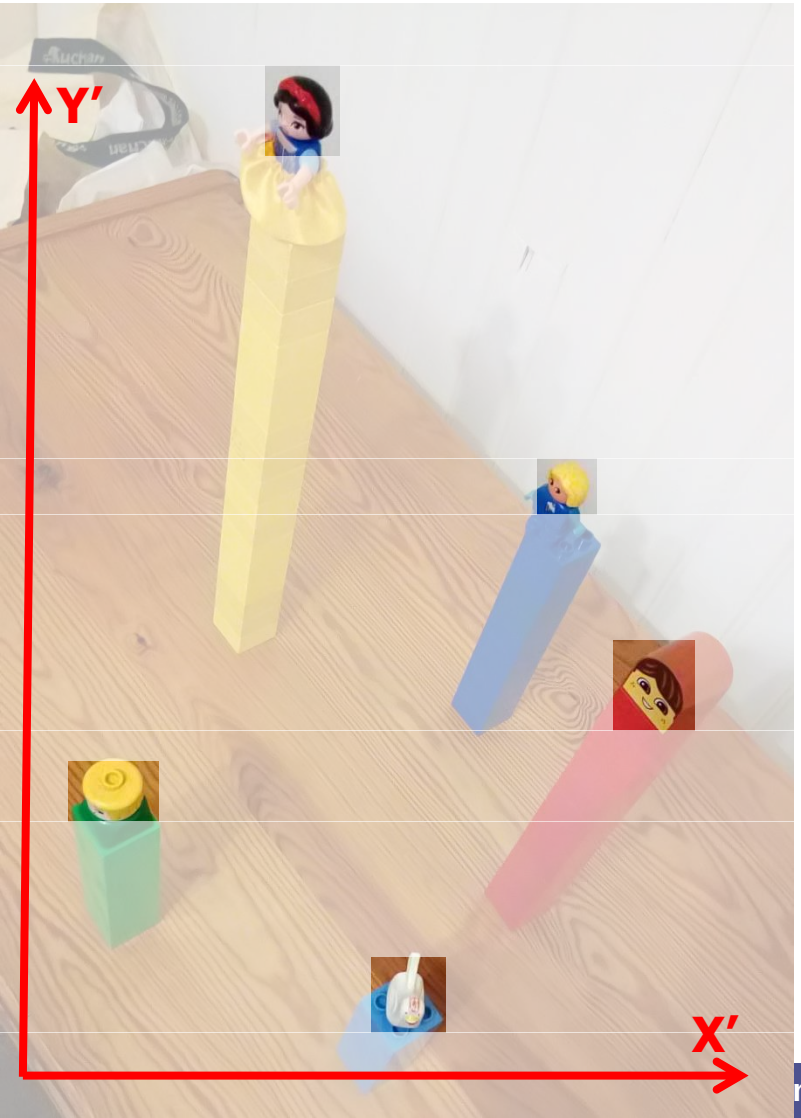
- Deux filles (à gauche)
- Trois garçons (à droite)

Position dans l'espace 3D



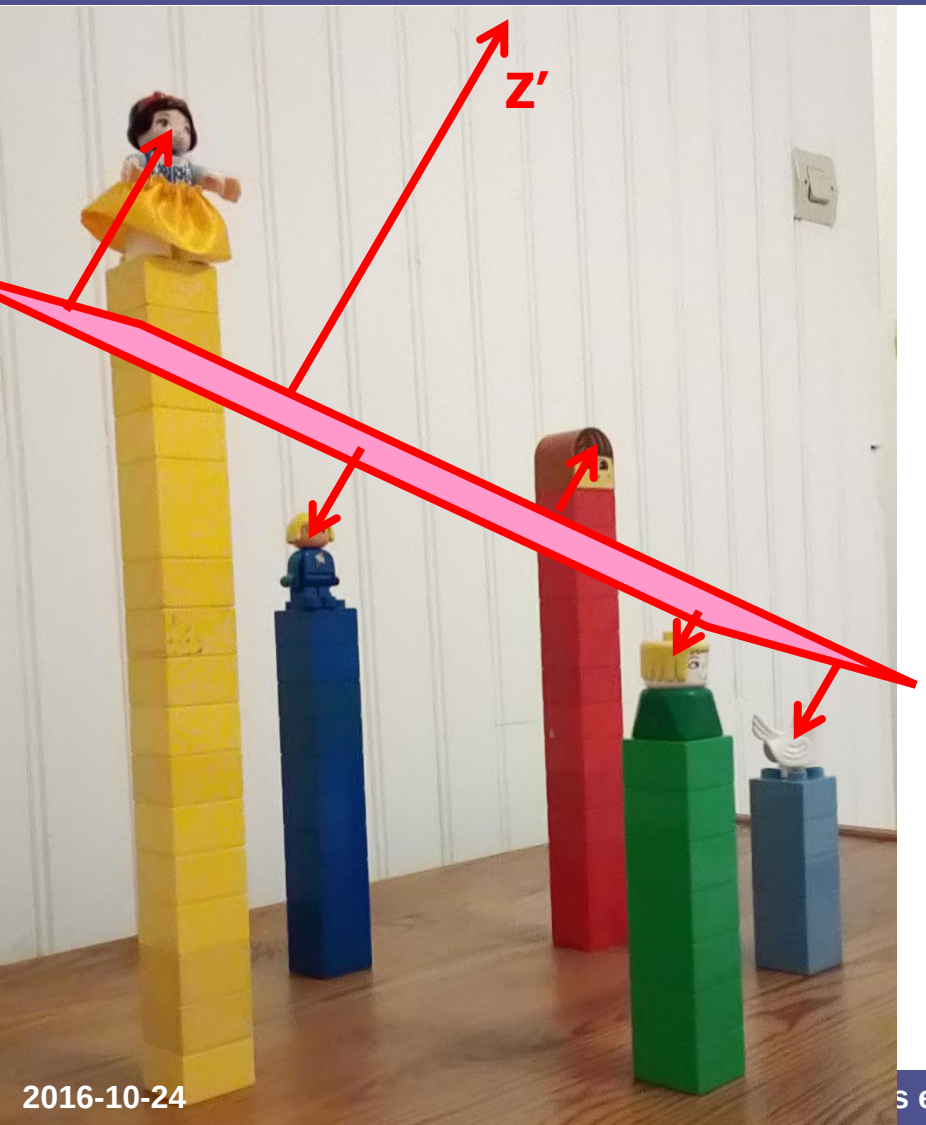
- Prise de photographie
  - = projection sur un plan 2D d'un espace 3D
  - donc réduction de dimension (3=>2)
  - donc perte d'information
- Inertie totale I
  - I = dispersion dans l'espace des personnages
  - I = var(X) + var(Y) + var(Z)
- Prise optimale
  - Quel angle aurait-on pu prendre pour perdre le moins possible d'inertie ?

Projection sur un plan 2D optimal (= minimisant la perte d'info)



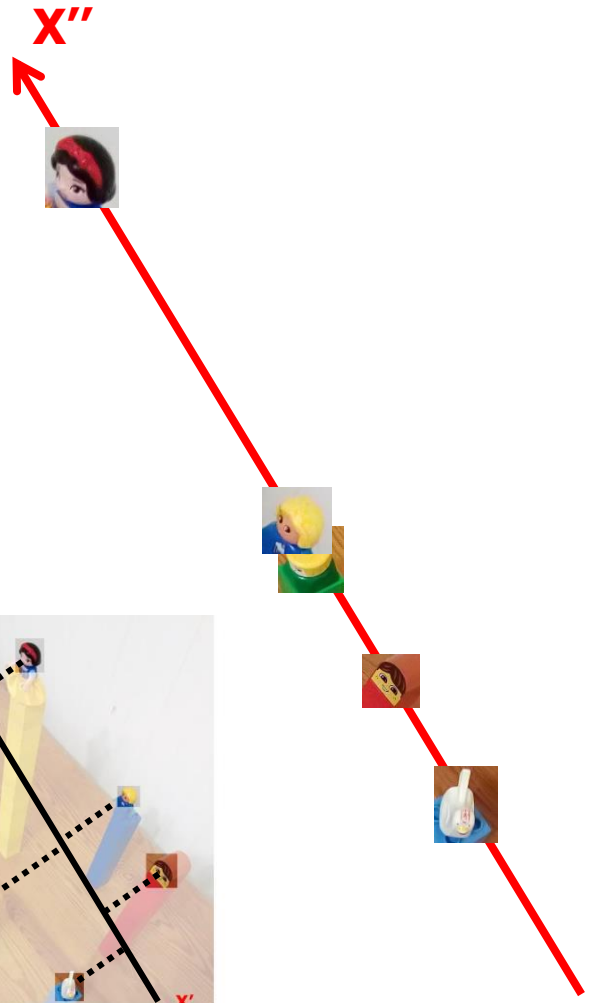
- Plan proposé ici :
  - X' est une combinaison linéaire de X, Y et Z : $X'=c_{1,1}.X + c_{1,2}.Y + c_{1,3}.Z$
  - Y' est une combinaison linéaire de X, Y et Z : $Y' = c_{2,1}.X + c_{2,2}.Y + c_{2,3}.Z$
  - $X \perp Y$
- Minimise la perte d'information :
  - $I=var(X')+var(Y')+I_{résiduelle}$
  - Avec $I_{résiduelle}$ la plus faible possible

Inertie résiduelle



- On a perdu de l'information (inertie) :
  - C'est l'inertie résiduelle
  - C'est la distance au plan (X', Y') choisi
- Cette information est indépendante (irrécupérable)
  - Troisième dimension, orthogonale au plan, donc orthogonale aux deux autres
  - Connaître les valeurs de X' et Y' n'aide nullement à connaître la valeur de Z'
- Cette perte est acceptable
  - Car moins importante que la perte d'une dimension initiale (X, Y ou Z)
  - On peut donc parfois ignorer Z'

Projection sur un axe 1D optimal (= minimisant la perte d'info)



- Choix de l'axe optimal X''
  - Sur le plan (X', Y')
  - Cet axe est une combinaison linéaire de X' et Y', donc finalement de X, Y et Z : $X''=c_{3,1}.X + c_{3,2}.Y + c_{3,3}.Z$
- Autant que possible :
  - Maximise l'écart entre individus =maximise l'inertie de X''
  - Minimise la perte d'information =minimise l'inertie résiduelle

Choix des projections ici, passage de 2D à 3D

- On aurait pu supprimer simplement une dimension
- Supprimer une variable

- On a choisi le plan qui minimisait la perte d'information, écartait le plus possible les individus
- Faire une analyse en composantes principales

- On aurait aussi pu choisir le plan qui séparait le mieux les garçons des filles (et les rapprochait entre eux)
- Faire une analyse discriminante

De la photographie à l'analyse en composantes principales

- Dimensions :
  - Pas que physiques, mais plus généralement ce sont des variables (…d'ailleurs il est habituel de représenter les individus sur un plan, ex X=taille Y=poids !)
  - Aisément plus de 3 dimensions, processus identique
  - Projection = combinaison linéaire
  - Orthogonalité géométrique = indépendance statistique
- L'analyse en composantes principales
  - Permet, de la même manière, des réductions de dimensions
  - Procède dans l'ordre inverse : identifie d'abord LE nouvel axe qui maximise l'inertie expliquée, puis le suivant (orthogonal au premier), puis le troisième (orthogonal aux deux autres), etc.

Les analyses en composantes principales normées
- principes généraux -

Finalités de l'ACP normée

- Tableau des données initial :
  - n lignes (1 par individu)
  - p colonnes (1 par variable)
- L'ACP permet d'ajouter au tableau de données des composantes *(facteurs, variables latentes)* :
  - p nouvelles variables (nouvelles colonnes)
  - Chaque composante est une combinaison linéaire des p variables initiales
  - Ces p composantes expliquent la totalité de l'inertie *(variance)* des p variables initiales
  - MAIS (propriétés majeures !) :
    - Elles sont totalement non-corrélées linéairement entre elles
    - Les premières composantes expliquent le maximum d'inertie possible
- L'ACP est utile lorsque les variables initiales sont assez fortement corrélées linéairement entre elles (multi-colinéarité)

ACP normée pas-à-pas (1)

- Toutes les p variables initiales sont d'abord centrées et réduites. X_1 à X_p désignent ces variables centrées réduites.
- Inertie totale = somme des variances des variables initiales centrées réduites :

$$I_{totale} = \sum_{i=1}^{p} var(X_i) = p$$

- Création d'une première composante C_1 :
  - Combinaison linéaire des X_i initiales avec des poids $u_{1,*}$

$$C_1 = u_{1,1}.X_1 + u_{1,2}.X_2 + \cdots + u_{1,p}.X_p$$

  - Avec des poids $u_{1,*}$ tels que

$$\sum_{i=1}^{p} (u_{1,i})^2 = 1$$

  - Et une inertie expliquée I_{c1} la plus élevée possible

$$I_{totale} = I_{C_1} + I_{résiduelle}$$

ACP normée pas-à-pas (2)

- Puis création d'une deuxième composante C_2 :
  - Combinaison linéaire des X_i initiales avec des poids u
    $$C_2 = u_{2,1}.X_1 + u_{2,2}.X_2 + \cdots + u_{2,p}.X_p$$
  - Avec des poids u tels que $\sum u^2 = 1$
  - Et C_2 est indépendante de C_1 (orthogonale)
  - Et une inertie expliquée I_{c2} la plus élevée possible
    $$I_{totale} = I_{C_1} + I_{C_2} + I_{résiduelle}$$
- Et ainsi de suite…
- Puis création de la dernière composante C_p :
  - Combinaison linéaire des X_i initiales avec des poids $u_{p,*}$
  - Avec des poids $u_{p,*}$ tels que $\sum u^2 = 1$
  - Et C_p est indépendante de C_1, de C_2, de C_3 … et de C_{p-1}
  - Elle explique l'inertie restante $I_{totale} = I_{C_1} + I_{C_2} + \cdots + I_{C_p}$
- Au final :
  - Pour chaque individu on dispose de p nouvelles variables <u>toutes indépendantes entre elles</u>, les composantes C_1 à C_p
  - Ces p nouvelles variables expliquent la totalité de l'inertie des p variables initiales.
  - On peut souvent se contenter des 2 ou 3 premières, dites « composantes principales »
  - On peut calculer les corrélations entre les variables initiales et les nouvelles composantes

Interprétation en pratique d’une ACP normée

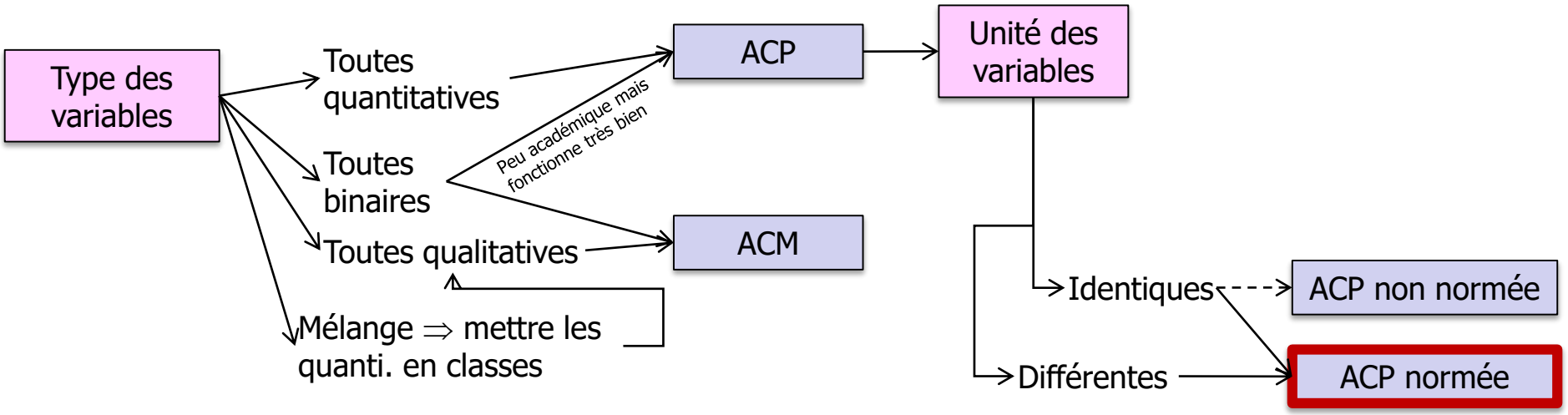
Etapes de l'interprétation d'une ACP normée

1. Les données et la technique
2. Part d'inertie expliquée par les CP
3. Plan factoriel des variables
4. Plan factoriel des individus (profils)

NB : les intitulés en italique sont des synonymes fréquemment employés, ne pas apprendre

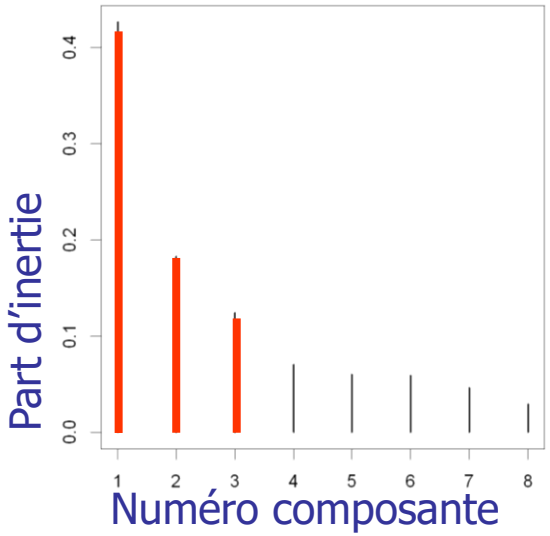
1- Les données et la technique

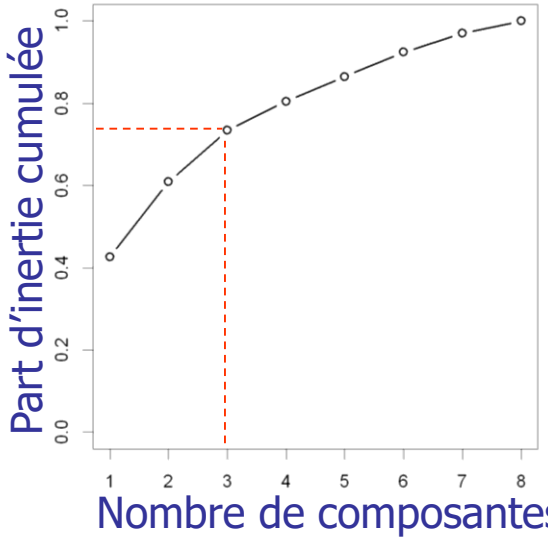
- Variables utilisées :
  - Nombre de variables (un des buts : réduction)
  - Type des variables (quantitatives, qualitatives, binaires)
  - Unité des variables (ex : années, euros, cm, etc.)
  - Rappel : technique « non supervisée », toute les variables au même niveau, pas de « variable à expliquer »



2- Part d'inertie expliquée

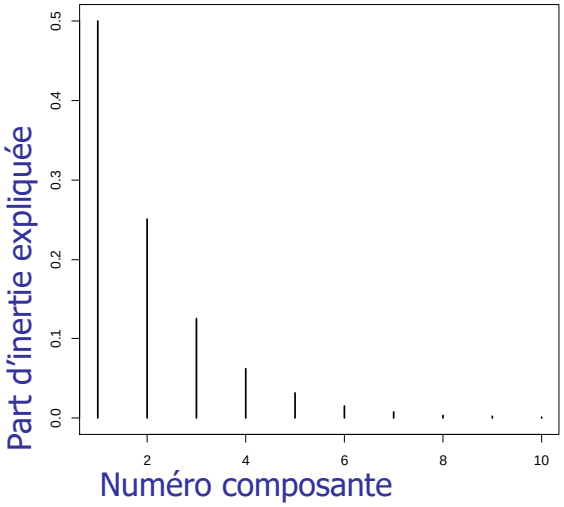
- Si p variables, on obtiendra p composantes *(facteurs, variables latentes)* expliquant 100% de l'inertie *(variance, information)* totale
- Les p composantes sont non-corrélées linéairement
  $\Rightarrow$ les inerties s'additionnent
- ACP intéressante seulement si les k premières composantes (composantes principales, avec k=1, 2 ou 3) expliquent une part d'inertie suffisante (donc si variables initiales assez corrélées ; nb : il existe de nombreux critères).
- Diagramme de la part d'inertie (proportionnelle aux Valeurs Propres) et diagramme cumulatif

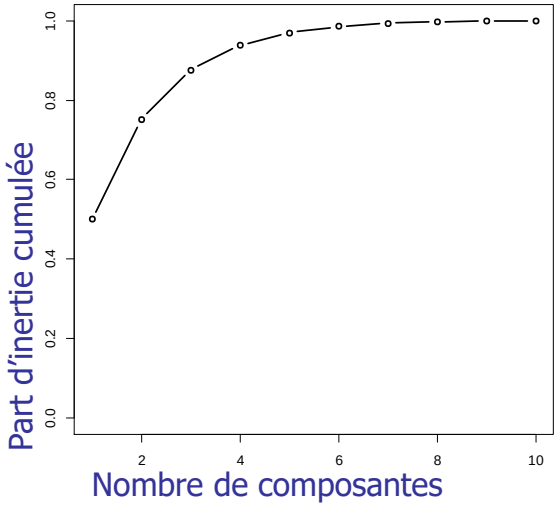




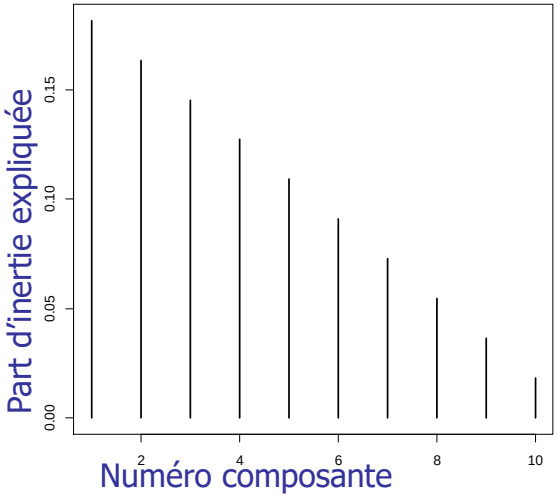
2- Part d'inertie expliquée

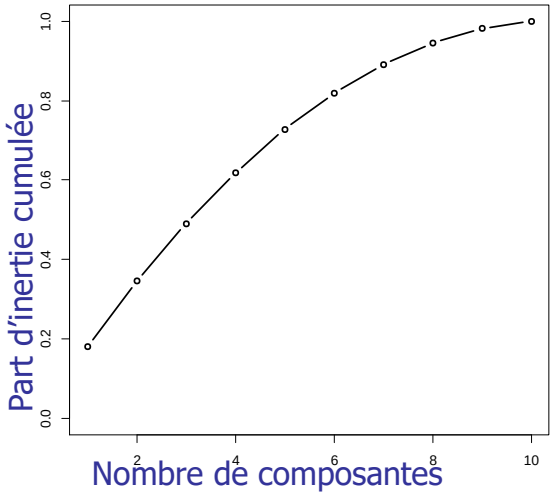
- Exemple d'ACP intéressante (8 variables initiales très corrélées linéairement)





- Exemple d'ACP sans intérêt (8 variables initiales peu corrélées linéairement)





3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3
CP2	0	1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4
CP3	0	0	1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39
CP4	0	0	0	1	0	0	0	0	0.04	-0	0.21	0.36	0.05	-0.3	0.34	0.74
CP5	0	0	0	0	1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22
CP6	0	0	0	0	0	1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07
CP7	0	0	0	0	0	0	1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0
CP8	0	0	0	0	0	0	0	1	0	-0	0.09	-0.5	0.39	-0	0.01	0
Var1	-0.9	-0.4	-0.4	0.04	0	0.01	-0	0	1	0.6	0.58	0.25	0.47	0.45	0.23	0.31
Var2	-0.9	0.38	0.14	-0	-0	0.02	0	-0	0.6	1	0.46	0.32	0.63	0.12	0.06	0.17
Var3	-0.6	-0.3	0.18	0.21	-0.1	-0.3	0.59	0.09	0.58	0.46	1	0.31	0.45	0.45	0.42	0.48
Var4	-0.3	0.13	-0	0.36	0.08	-0.7	-0.1	-0.5	0.25	0.32	0.31	1	0.44	-0.1	0.08	0.29
Var5	-0.7	0.07	0.19	0.05	0.14	-0.5	-0.3	0.39	0.47	0.63	0.45	0.44	1	0.25	0.09	0.29
Var6	-0.4	-0.8	0.38	-0.3	0.05	0	-0	-0	0.45	0.12	0.45	-0.1	0.25	1	0.37	0.33
Var7	-0.2	-0.4	0.31	0.34	-0.8	0	-0.1	0.01	0.23	0.06	0.42	0.08	0.09	0.37	1	0.43
Var8	-0.3	-0.4	0.39	0.74	0.22	0.07	-0	0	0.31	0.17	0.48	0.29	0.29	0.33	0.43	1

3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3
CP2	0	1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4
CP3	0	0	1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39
CP4	0	0	0	1	0	0	0	0	0.04	-0	0.21	0.36	0.05	-0.3	0.34	0.74
CP5	0	0	0	0	1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22
CP6	0	0	0	0	0	1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07
CP7	0	0	0	0	0	0	1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0
CP8	0	0	0	0	0	0	0	1	0	-0	0.09	-0.5	0.39	-0	0.01	0
Var1	-0.9	-0.4	-0.4	0.04	0	0.01	-0	0	1	0.6	0.58	0.25	0.47	0.45	0.23	0.31
Var2	-0.9	0.38	0.14	-0	-0	0.02	0	-0	0.6	1	0.46	0.32	0.63	0.12	0.06	0.17
Var3	-0.6	-0.3	0.18	0.21	-0.1	-0.3	0.59	0.09	0.58	0.46	1	0.31	0.45	0.45	0.42	0.48
Var4	-0.3	0.13	-0	0.36	0.08	-0.7	-0.1	-0.5	0.25	0.32	0.31	1	0.44	-0.1	0.08	0.29
Var5	-0.7	0.07	0.19	0.05	0.14	-0.5	-0.3	0.39	0.47	0.63	0.45	0.44	1	0.25	0.09	0.29
Var6	-0.4	-0.8	0.38	-0.3	0.05	0	-0	-0	0.45	0.12	0.45	-0.1	0.25	1	0.37	0.33
Var7	-0.2	-0.4	0.31	0.34	-0.8	0	-0.1	0.01	0.23	0.06	0.42	0.08	0.09	0.37	1	0.43
Var8	-0.3	-0.4	0.39	0.74	0.22	0.07	-0	0	0.31	0.17	0.48	0.29	0.29	0.33	0.43	1

- Toutes les cases de la diagonale valent 1 car corr(X,X)=1
- La matrice est symétrique par rapport à la diagonale car corr(X,Y)=corr(Y,X)

3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3	CP1
CP2		1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4	CP2
CP3			1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39	CP3
CP4				1	0	0	0	0	0.04	-0	0.21	0.36	0.05	-0.3	0.34	0.74	CP4
CP5					1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22	CP5
CP6						1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07	CP6
CP7							1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0	CP7
CP8								1	0	-0	0.09	-0.5	0.39	-0	0.01	0	CP8
Var1									1	0.6	0.58	0.25	0.47	0.45	0.23	0.31	Var1
Var2										1	0.46	0.32	0.63	0.12	0.06	0.17	Var2
Var3											1	0.31	0.45	0.45	0.42	0.48	Var3
Var4												1	0.44	-0.1	0.08	0.29	Var4
Var5													1	0.25	0.09	0.29	Var5
Var6														1	0.37	0.33	Var6
Var7															1	0.43	Var7
Var8																1	Var8

- Toutes les cases de la diagonale valent 1 car corr(X,X)=1
- La matrice est symétrique par rapport à la diagonale car corr(X,Y)=corr(Y,X)
- Les composantes sont non-corrélées linéairement
  $corr(CP_i,CP_j)=0$
  $corr(CP_3,CP_6)=0$

3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3	CP1
CP2		1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4	CP2
CP3			1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39	CP3
CP4				1	0	0	0	0	0.04	-0	0.21	0.36	0.05	-0.3	0.34	0.74	CP4
CP5					1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22	CP5
CP6						1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07	CP6
CP7							1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0	CP7
CP8								1	0	-0	0.09	-0.5	0.39	-0	0.01	0	CP8
Var1									1	0.6	0.58	0.25	0.47	0.45	0.23	0.31	Var1
Var2										1	0.46	0.32	0.63	0.12	0.06	0.17	Var2
Var3											1	0.31	0.45	0.45	0.42	0.48	Var3
Var4												1	0.44	-0.1	0.08	0.29	Var4
Var5													1	0.25	0.09	0.29	Var5
Var6														1	0.37	0.33	Var6
Var7															1	0.43	Var7
Var8																1	Var8

- Toutes les cases de la diagonale valent 1 car corr(X,X)=1
- La matrice est symétrique par rapport à la diagonale car corr(X,Y)=corr(Y,X)
- Les composantes sont non-corrélées linéairement
  $corr(CP_i,CP_j)=0$
  $corr(CP_3,CP_6)=0$
- Corrélation entre 2 variables initiales : corr(var3, var4)=0.31

3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3	CP1
CP2		1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4	CP2
CP3			1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39	CP3
CP4				1	0	0	0	0	0.04	-0	0.21	0.36	0.05	-0.3	0.34	0.74	CP4
CP5					1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22	CP5
CP6						1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07	CP6
CP7							1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0	CP7
CP8								1	0	-0	0.09	-0.5	0.39	-0	0.01	0	CP8
Var1									1	0.6	0.58	0.25	0.47	0.45	0.23	0.31	Var1
Var2										1	0.46	0.32	0.63	0.12	0.06	0.17	Var2
Var3											1	0.31	0.45	0.45	0.42	0.48	Var3
Var4												1	0.44	-0.1	0.08	0.29	Var4
Var5													1	0.25	0.09	0.29	Var5
Var6														1	0.37	0.33	Var6
Var7															1	0.43	Var7
Var8																1	Var8

- Toutes les cases de la diagonale valent 1 car corr(X,X)=1
- La matrice est symétrique par rapport à la diagonale car corr(X,Y)=corr(Y,X)
- Les composantes sont non-corrélées linéairement $corr(CP_i,CP_j)=0$ $corr(CP_3,CP_6)=0$
- Corrélation entre 2 variables initiales : corr(var3, var4)=0.31
- Corrélation entre CP et variable : corr(var2, CP3)=0.14

3- Plan factoriel des variables

3a – matrice de corrélation

- Tableau croisant les variables (et si on veut, les CP), donnant le coefficient de corrélation 2 à 2
- Information objective, complète mais complexe
- Plan factoriel des variables = synthèse visuelle

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	
CP1	1	0	0	0	0	0	0	0	-0.9	-0.9	-0.6	-0.3	-0.7	-0.4	-0.2	-0.3	CP1
CP2		1	0	0	0	0	0	0	-0.4	0.38	-0.3	0.13	0.07	-0.8	-0.4	-0.4	CP2
CP3			1	0	0	0	0	0	-0.4	0.14	0.18	-0	0.19	0.38	0.31	0.39	CP3
CP4				1	0	0	0	0	0.04	-0	0.21	0.06	0.05	-0.3	0.34	0.74	CP4
CP5					1	0	0	0	0	-0	-0.1	0.08	0.14	0.05	-0.8	0.22	CP5
CP6						1	0	0	0.01	0.02	-0.3	-0.7	-0.5	0	0	0.07	CP6
CP7							1	0	-0	0	0.59	-0.1	-0.3	-0	-0.1	-0	CP7
CP8								1	0	-0	0.09	-0.5	0.39	-0	0.01	0	CP8
Var1									1	0.6	0.58	0.25	0.47	0.45	0.23	0.31	Var1
Var2										1	0.46	0.32	0.63	0.12	0.06	0.17	Var2
Var3											1	0.31	0.45	0.45	0.42	0.48	Var3
Var4												1	0.44	-0.1	0.08	0.29	Var4
Var5													1	0.25	0.09	0.29	Var5
Var6														1	0.37	0.33	Var6
Var7															1	0.43	Var7
Var8																1	Var8

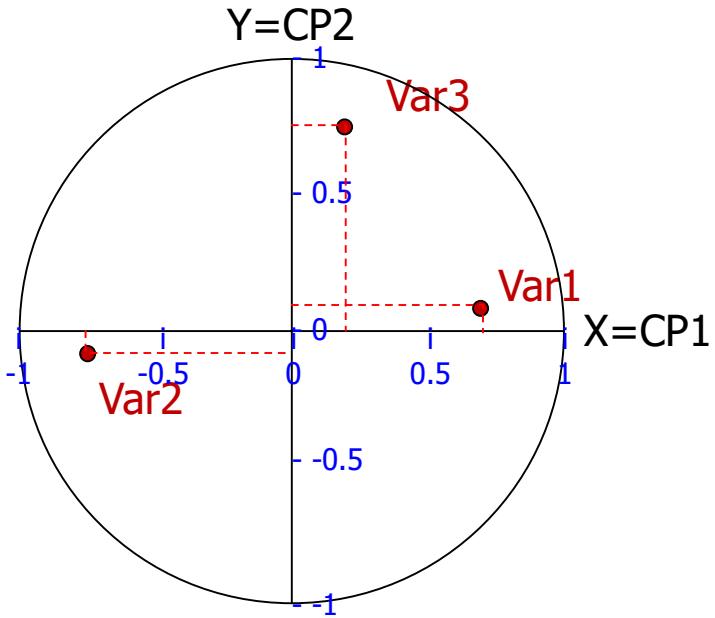
Information représentée par le premier plan factoriel des variables

- Toutes les cases de la diagonale valent 1 car corr(X,X)=1
- La matrice est symétrique par rapport à la diagonale car corr(X,Y)=corr(Y,X)
- Les composantes sont non-corrélées linéairement $corr(CP_i,CP_j)=0$ $corr(CP_3,CP_6)=0$
- Corrélation entre 2 variables initiales : corr(var3, var4)=0.31
- Corrélation entre CP et variable : corr(var2, CP3)=0.14

3- Plan factoriel des variables

3b - interpréter les composantes principales

- Plan factoriel des variables *(Cercle des corrélations)* :
  premier plan X=CP1, Y=CP2 (et souvent deuxième plan X=CP2, Y=CP3)
- Chaque variable initiale est représentée par un point
- Coordonnées d'une variable sur le plan
  = coefficients de corrélation entre la variable et les CP servant d'axes
- Corrélation(CP1, CP2)=0 ⇒ une variable ne peut être fortement corrélée aux deux en même temps ⇒ variables à l'intérieur du cercle

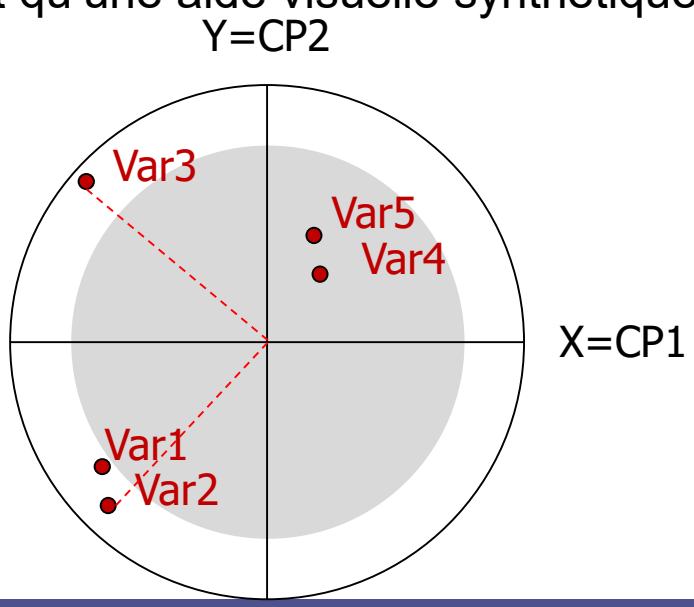


- Var1 : corrélée lin. positivement à CP1, quasi-indépendante de CP2
- Var2 : corrélée lin. négativement à CP1, quasi-indépendante de CP2
- Var 3 : corrélée lin. positivement à CP2, quasi-indépendante de CP1
- Donc CP1 représente assez bien Var1 et l'opposé de Var2
- Et CP2 représente assez bien Var3

3- Plan factoriel des variables

3c – corrélations entre variables

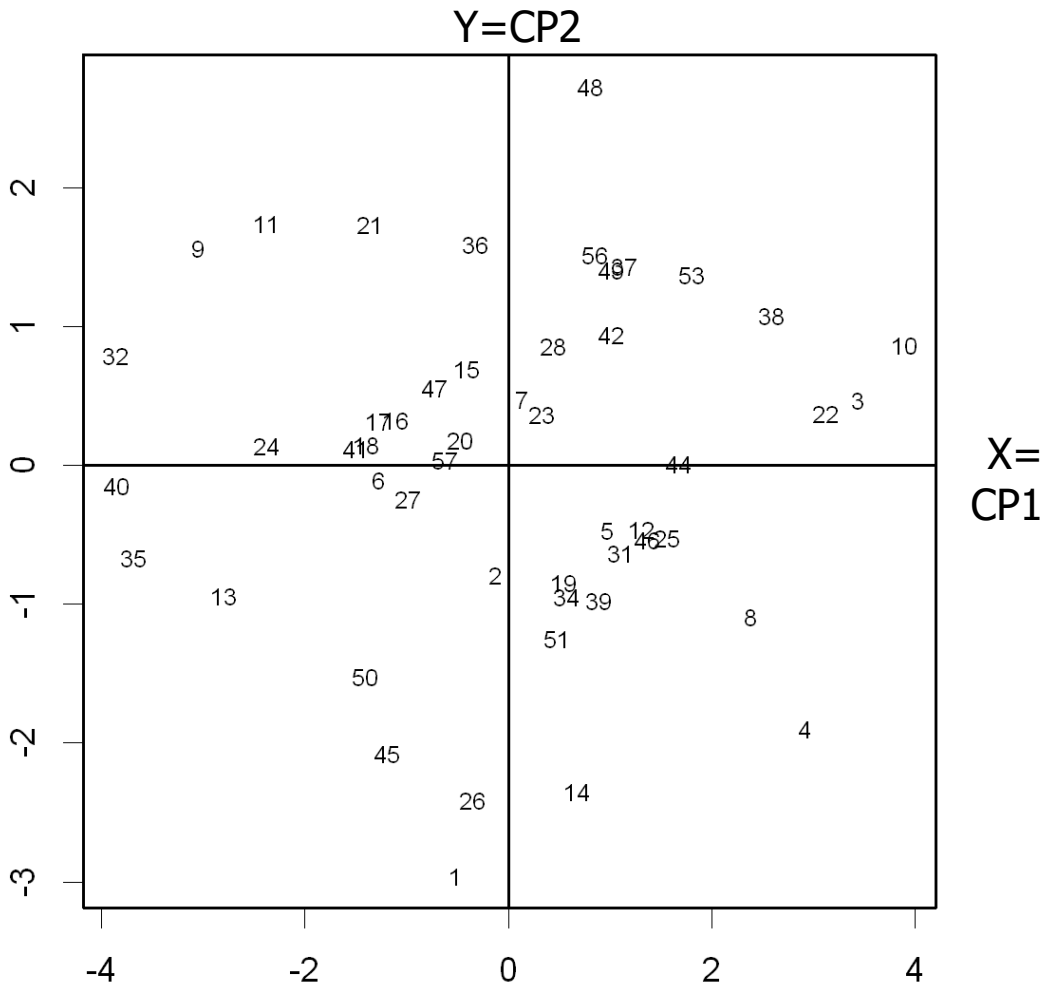
- Les variables proches du cercle sont essentiellement expliquées par les CP des axes, et donc seront très peu corrélées linéairement aux CP suivantes (4, 5…)
- Sur le plan factoriel des variables *(cercle des corrélations)* :
  - Deux variables proches l'une de l'autre (ou à 180°) ET proches du cercle : sont corrélées linéairement entre elles
  - Deux variables à 90° l'une de l'autre ET proches du cercle : sont non-corrélées linéairement
  - Deux variables éloignées du cercle : ON NE PEUT PAS INTERPRETER !
- Ce n'est qu'une aide visuelle synthétique : seule la matrice corrélation importe



- Var1 et Var2 sont corrélées linéairement entre elles
- Var2 et Var3 sont non-corrélées linéairement entre elles
- On ne sait pas si Var4 et Var5 sont corrélées linéairement entre elles
- Seule la matrice de corrélation donne un information certaine (mais moins synthétique)

4- Plan factoriel des individus

- Premier plan : X=CP1, Y=CP2
  (et souvent deuxième plan X=CP2, Y=CP3)
- Un point par individu : valeurs des CP pour cet individu
- Une fois qu'on « comprend » ce que représentent les composantes principales, on peut :
  - Décrire des groupes d'individus
  - Identifier des individus marginaux



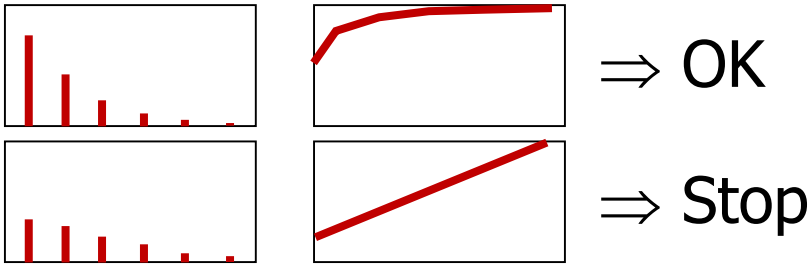
Analyse en Composantes Principales normée

Fiche d'interprétation

1- Données et technique

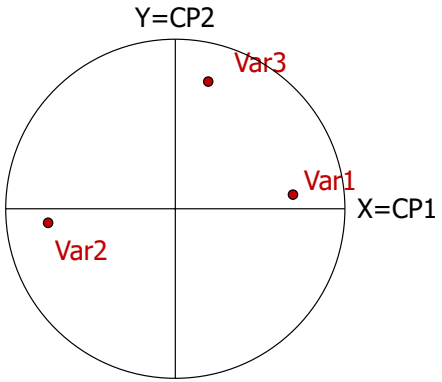
- Nb, type et unité des variables
- ACP normée, ACP non-normée (ou ACM)

2- Part d'inertie expliquée par les CP1, 2 et 3



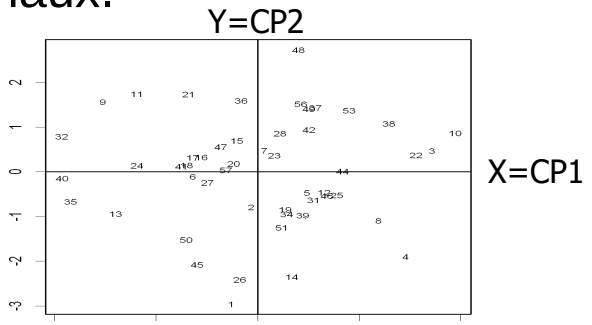
3- Plan factoriel des variables

- Que signifient les CP1, 2 et 3 ? Quelles variables initiales représentent-elles ?
- Eventuellement, quelles variable sont corrélées linéairement ?



4- Plan factoriel des individus

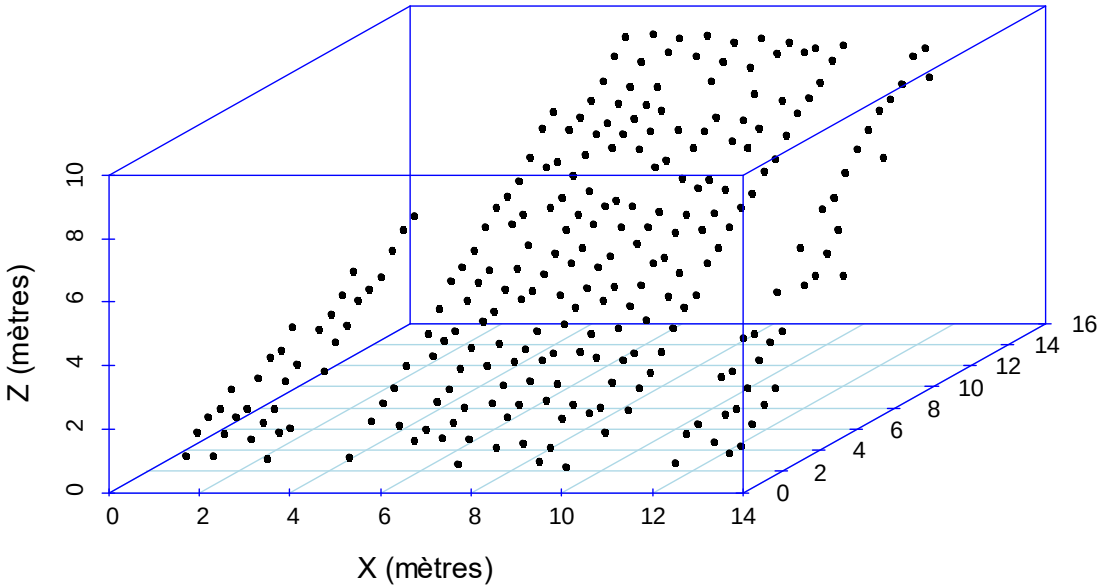
- Des groupes ? Des marginaux ?
- Si les CP ont une signification, interpréter ces groupes et marginaux.



Exemple n°1
Etudiants dans un amphithéâtre

Données étudiées = position des étudiants (tête) dans un amphi

- 256 étudiants
- Variables :
  - X= position G-D en mètres
  - Y=position devant-derrière en mètres
  - Z=hauteur de la tête en mètres (hauteur du banc de la rangée + distance bassin-tête)
- Technique employée : ACP normée



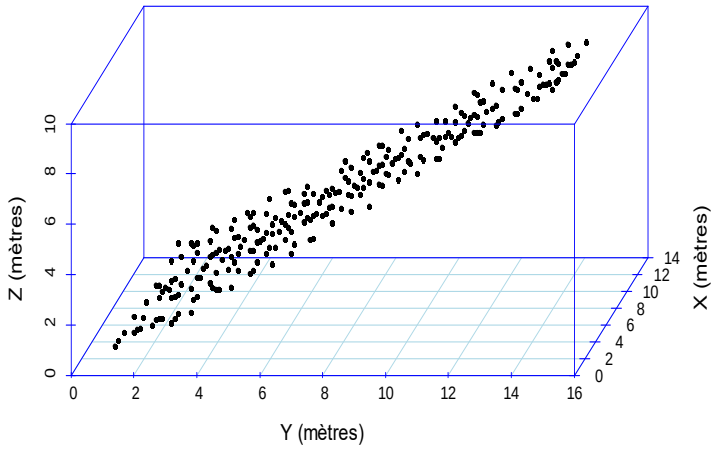
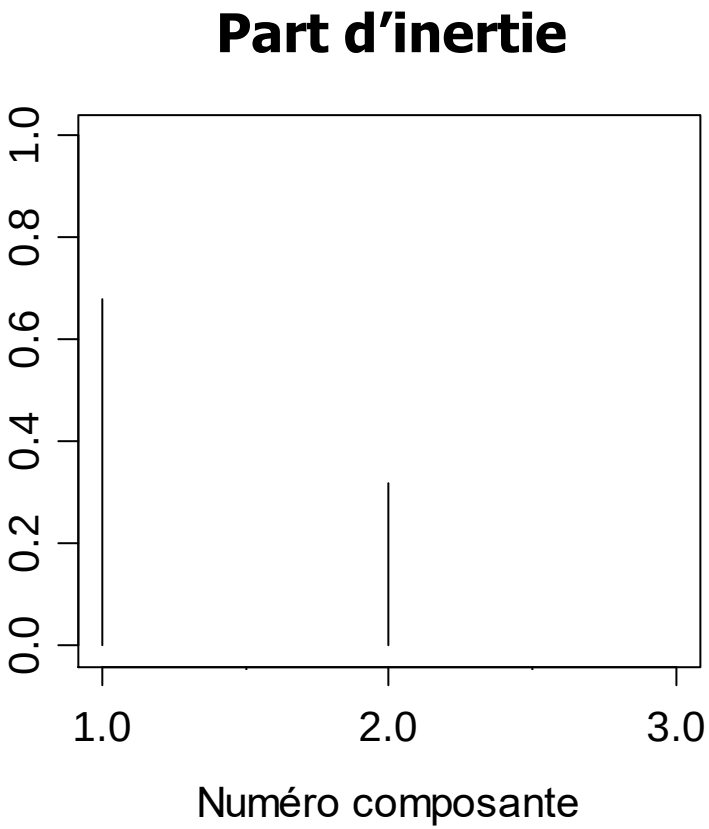
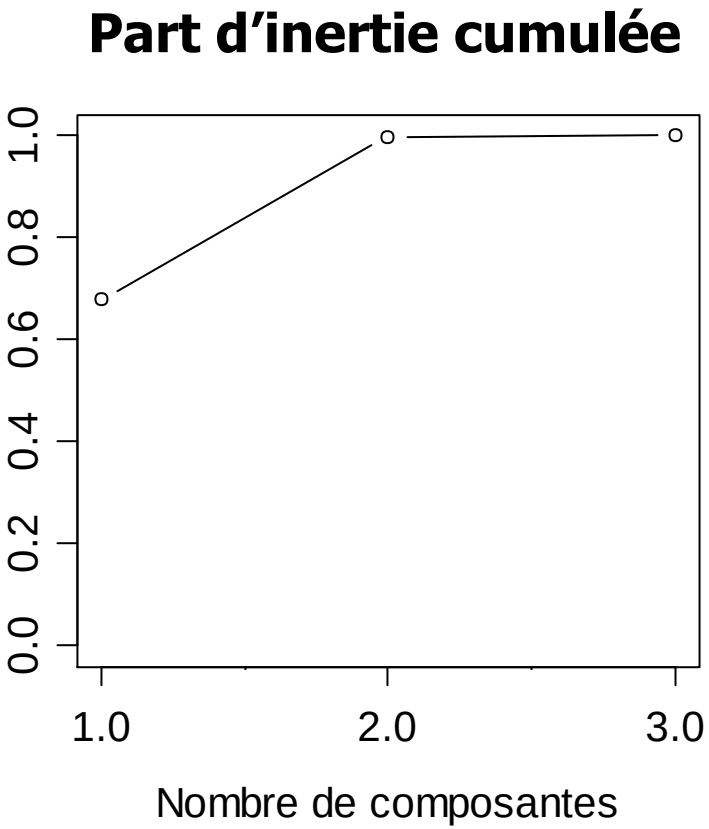
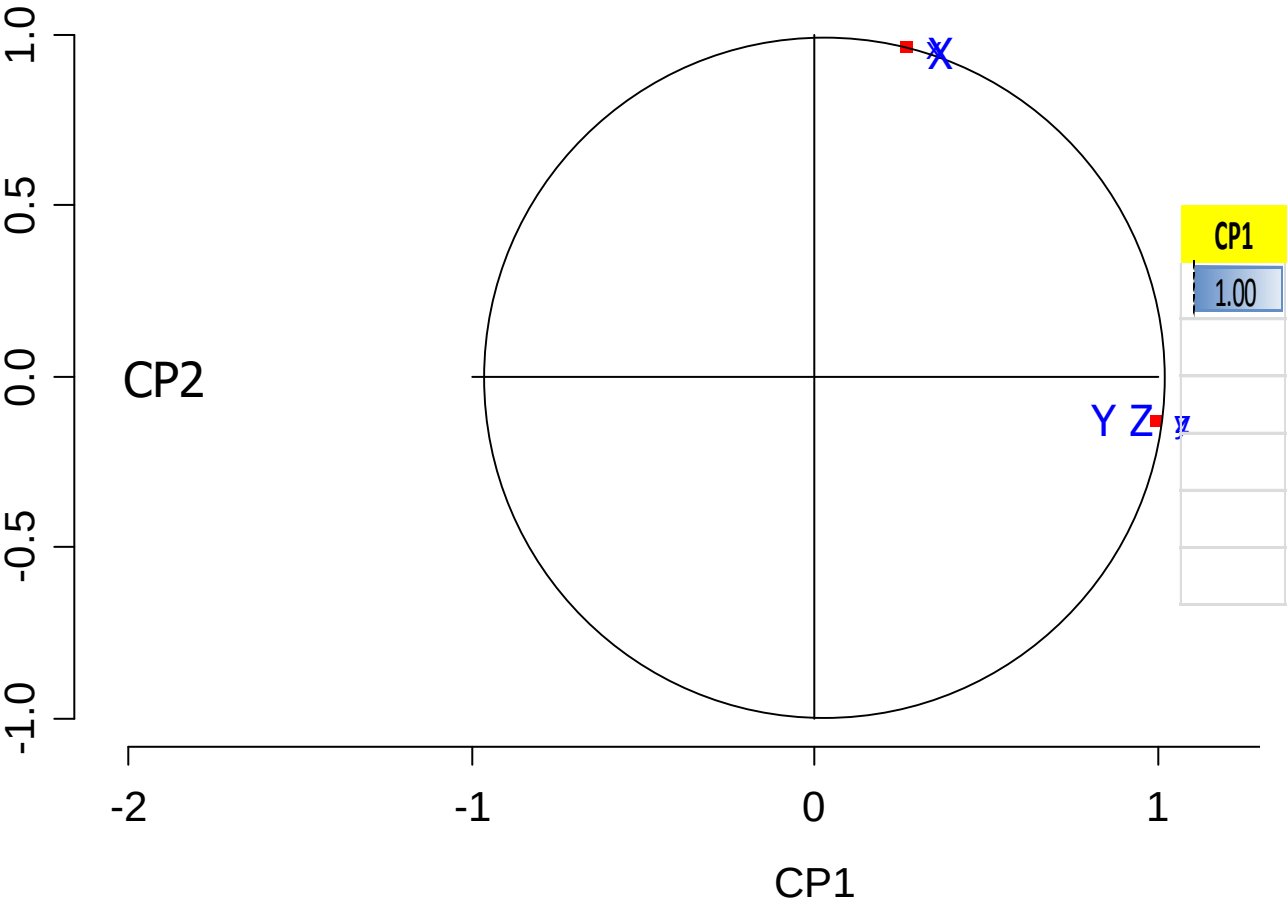


Diagramme des valeurs propres



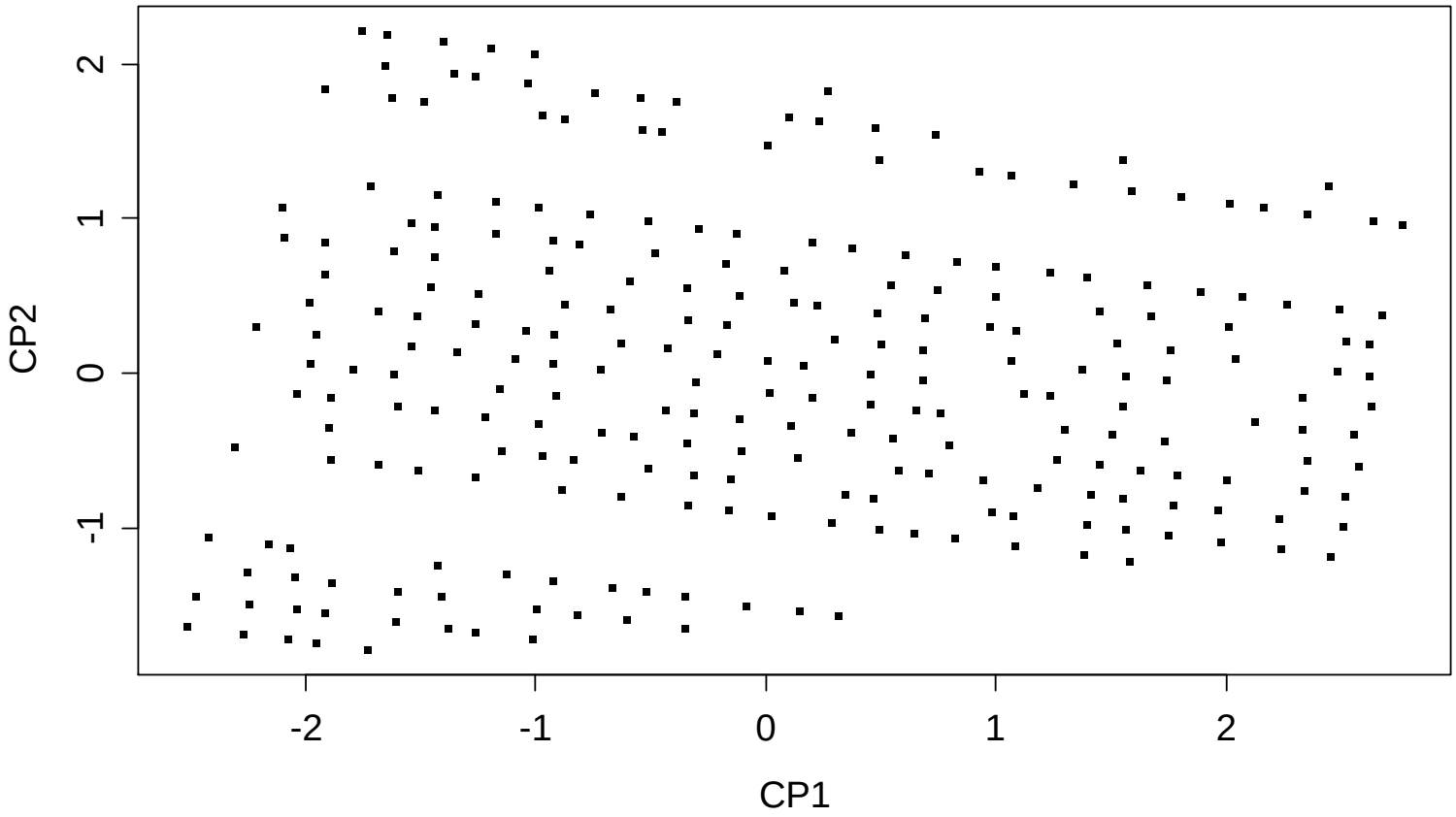


Plan factoriel des variables



	CP1	CP2	CP2	x	y	z
CP1	1.00	0.00	0.00	0.27	0.99	0.99
CP2		1.00	0.00	0.96	0.13	0.13
CP3			1.00	0.00	0.03	0.03
x				1.00	0.14	0.14
y					1.00	1.00
z						1.00

Plan factoriel des individus



Analyse du cas

1. Données et technique :
   - Données quanti, unité identique : on pourrait utiliser une ACP non-normée, et l'ACP normée est aussi valide
   - On est en quasi-2D : Z dépend presque uniquement de Y
   - Un changement de repère simplifierait les choses
2. Part d'inertie expliquée
   - CP1 : 68%. CP1+CP2 : 99.9%
   - Sans surprise, données en réalité 2D donc CP3 est inutile
3. Plan factoriel des variables
   - CP1 représente surtout Y et Z (très corrélées entre elles) : à droite les individus situés en arrière/en haut de l'amphi, à gauche les individus situés en avant/en bas de l'amphi
   - CP2 représente surtout X : en haut les individus situés à droite dans l'amphi (pour le prof), en bas ceux situés à gauche dans l'amphi (pour le prof)
4. Plan factoriel des individus
   - Dans ce cas, l'ACP a simplement retrouvé le plan incliné de l'amphithéâtre (photographié par en-bas), et a identifié l'axe principal des étudiants, légèrement dévié
   - En géométrie, changement de repère = création de nouveaux axes par combinaison linéaire des axes initiaux

Exemple n°2
Corpulence des sujets

Données étudiées = données anthropométriques

- 247 hommes
- Variables :
  - Taille : moy=171.1 cm
  - Poids : moy=69.1 kg
  - Tour de nombril : moy=85.6 cm
- Technique employée : ACP normée

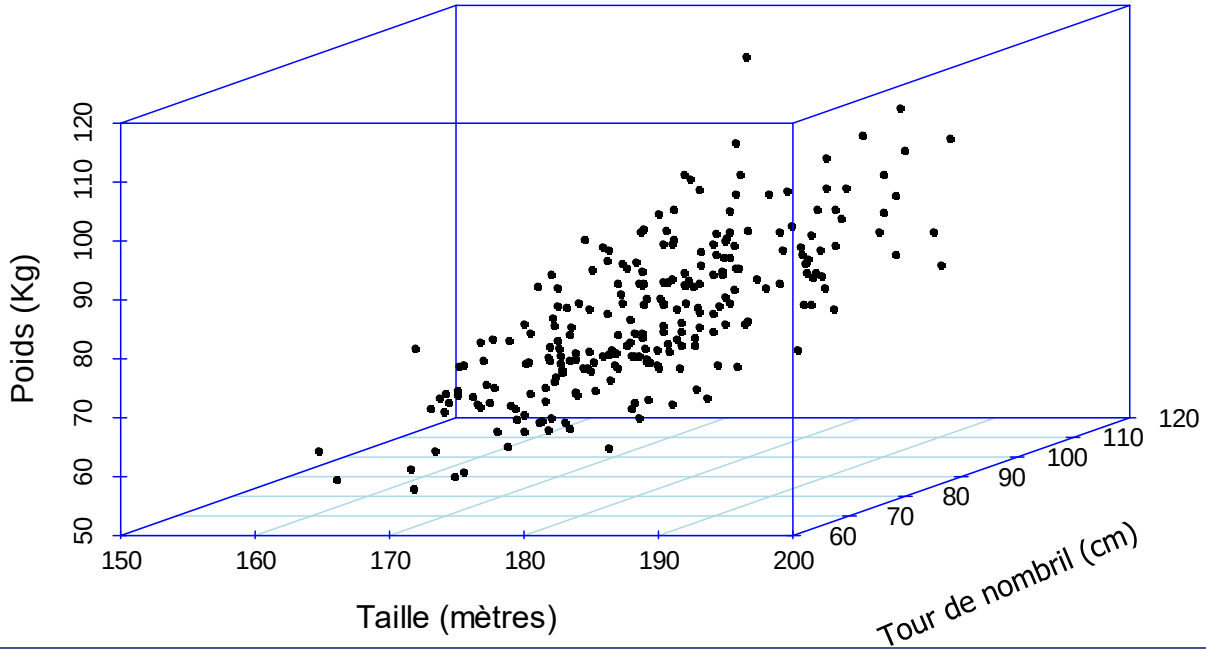
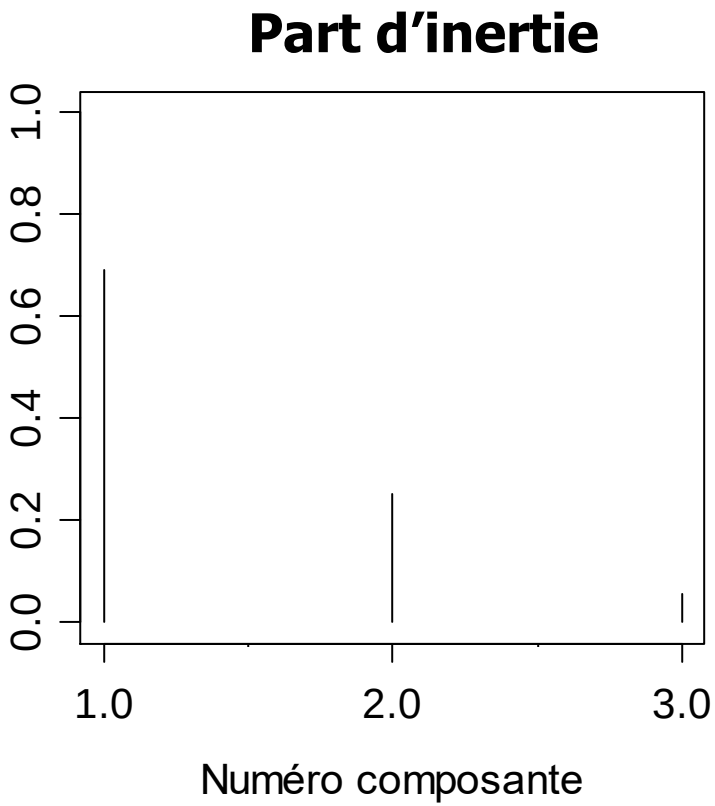
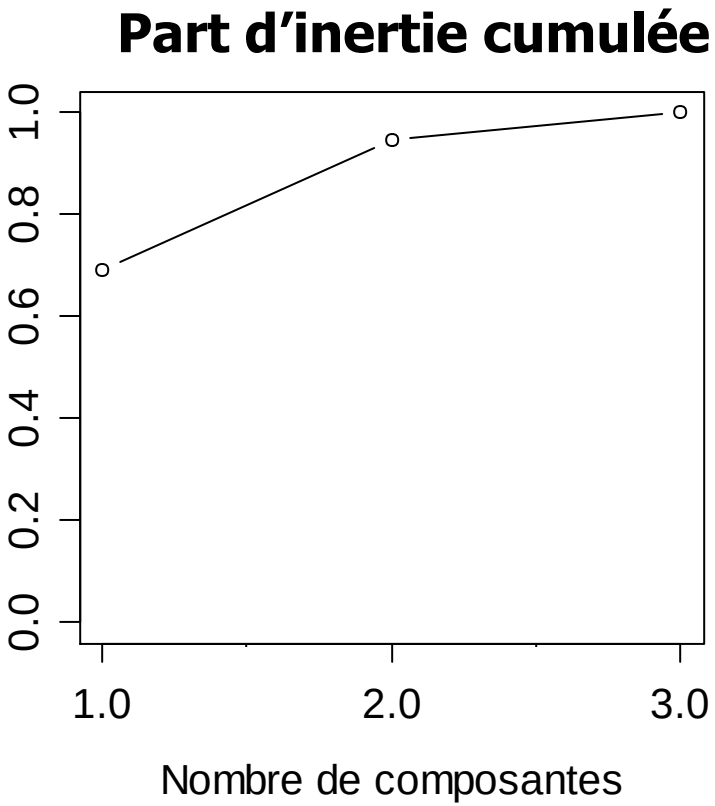
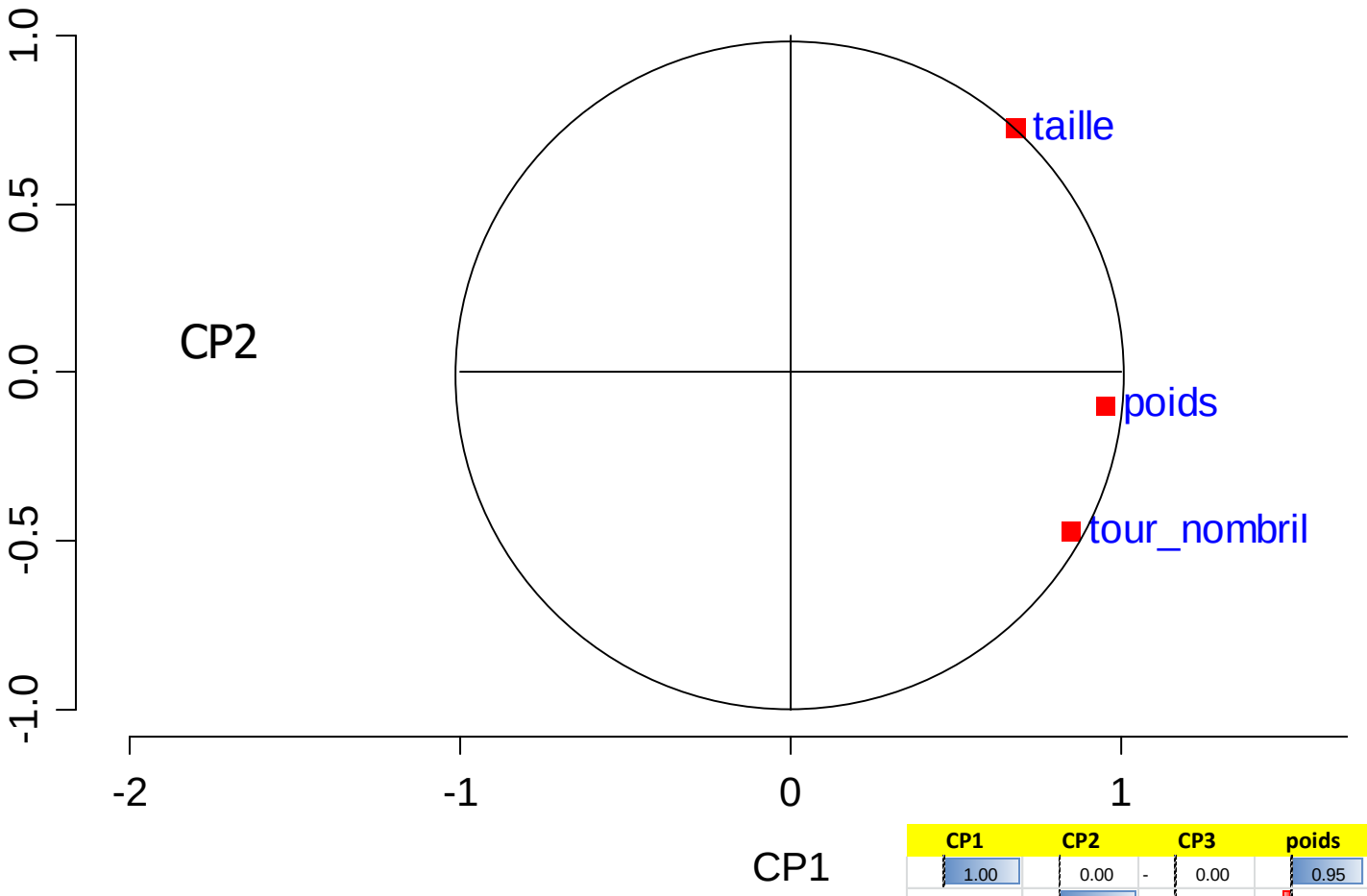


Diagramme des valeurs propres



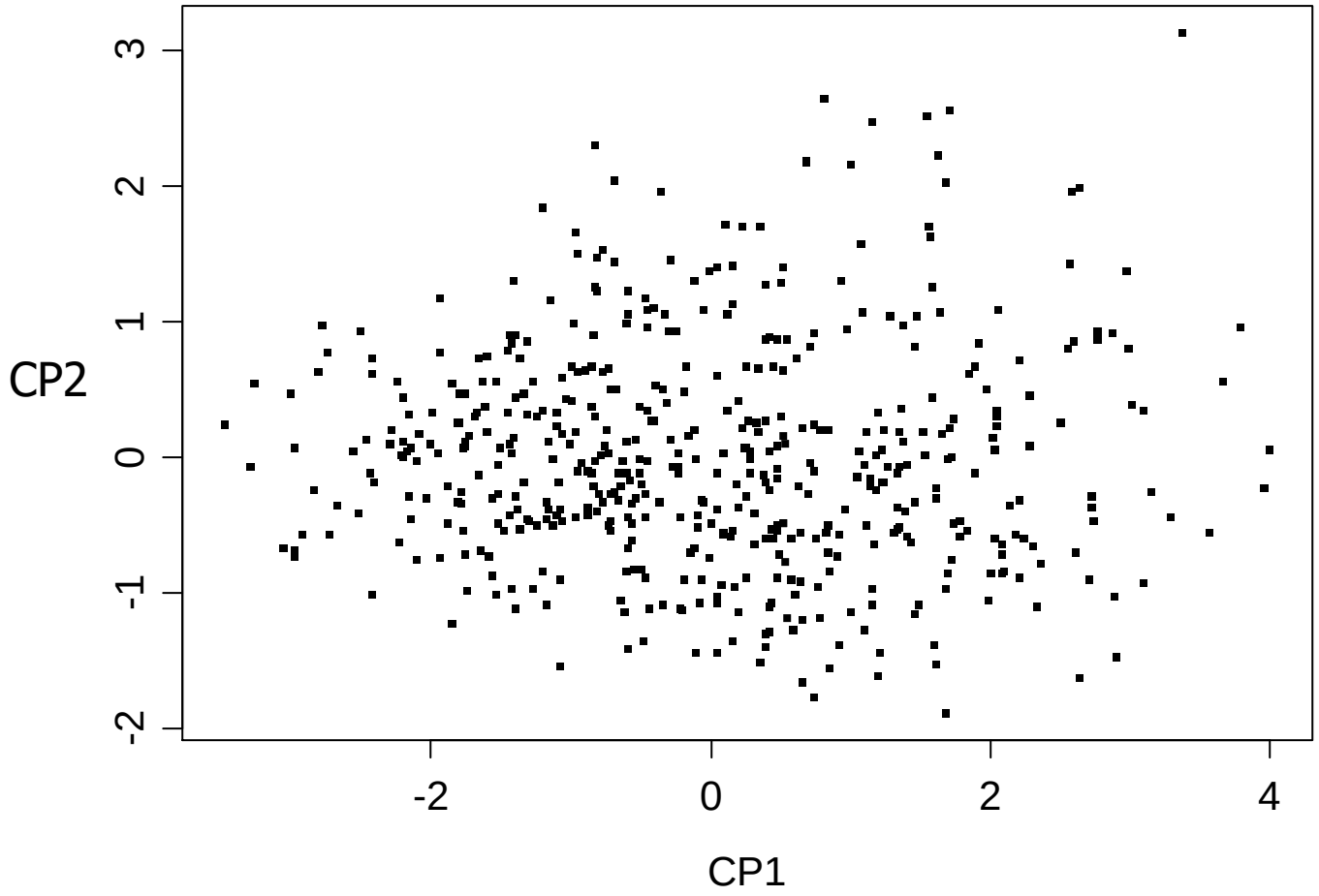


Plan factoriel des variables



	CP1	CP2	CP3	poids	taille	tournomb
CP1	1.00	0.00	0.00	0.95	0.68	0.85
CP2		1.00	0.00	0.10	0.73	0.47
CP3			1.00	0.30	0.12	0.24
poids				1.00	0.53	0.78
taille					1.00	0.26
tournomb						1.00

Plan factoriel des individus



Analyse du cas

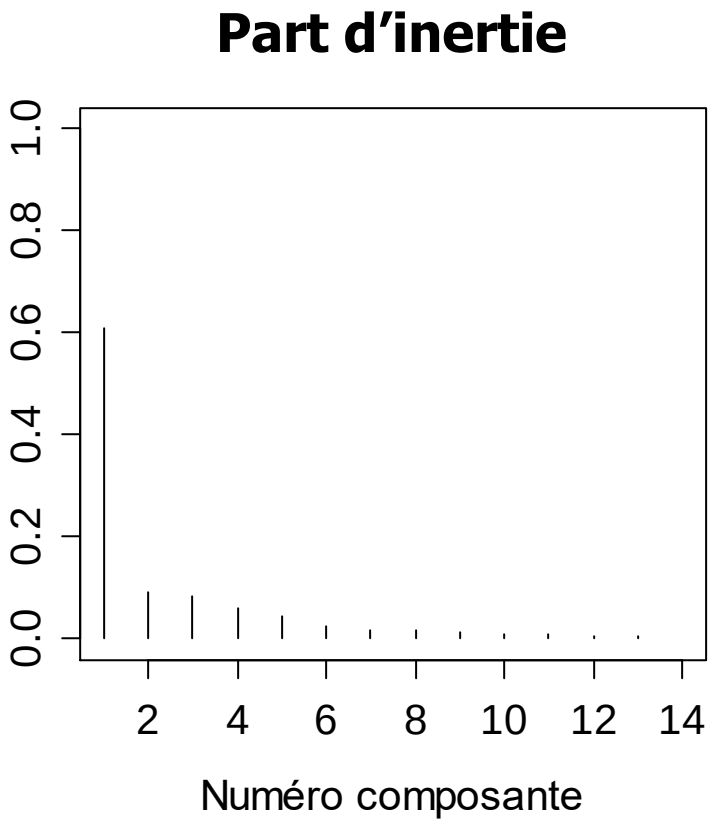
1. Données et technique :
   - Données quantitatives, unités différentes => ACP normée
   - Impression assez nette que le poids augmente lorsque la taille ou le tour de nombril augmente
2. Part d'inertie expliquée
   - CP1 : 69%. CP1+CP2 : 94%
   - Sans surprise, CP3 est inutile
3. Plan factoriel des variables
   - CP1 représente la corpulence (surtout poids, et aussi de manière moindre taille et au tour de nombril)
   - CP2 représente l'adiposité : à corpulence égale, cette composante sépare les grands et minces (valeurs positives) des petits et gras (valeurs négatives)
4. Plan factoriel des individus
   - A droite les corpulents (lourds), à gauche les peu corpulents (légers)
   - Pour une corpulence donnée, en haut les grands maigres en bas les petits gras
   - Moins de variété selon CP2 pour les valeurs faibles de CP1 (un individu peu corpulent ne peut être ni grand, ni gras)
   - Quelques individus extrêmes se distinguent

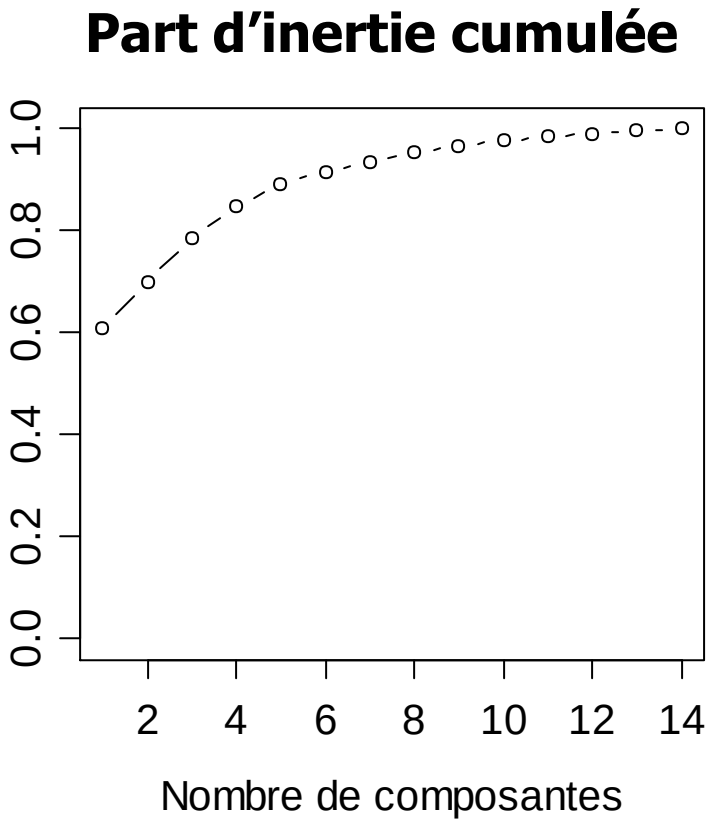
Exemple n°3
Plus de données anthropométriques

Données étudiées = données anthropométriques complétées

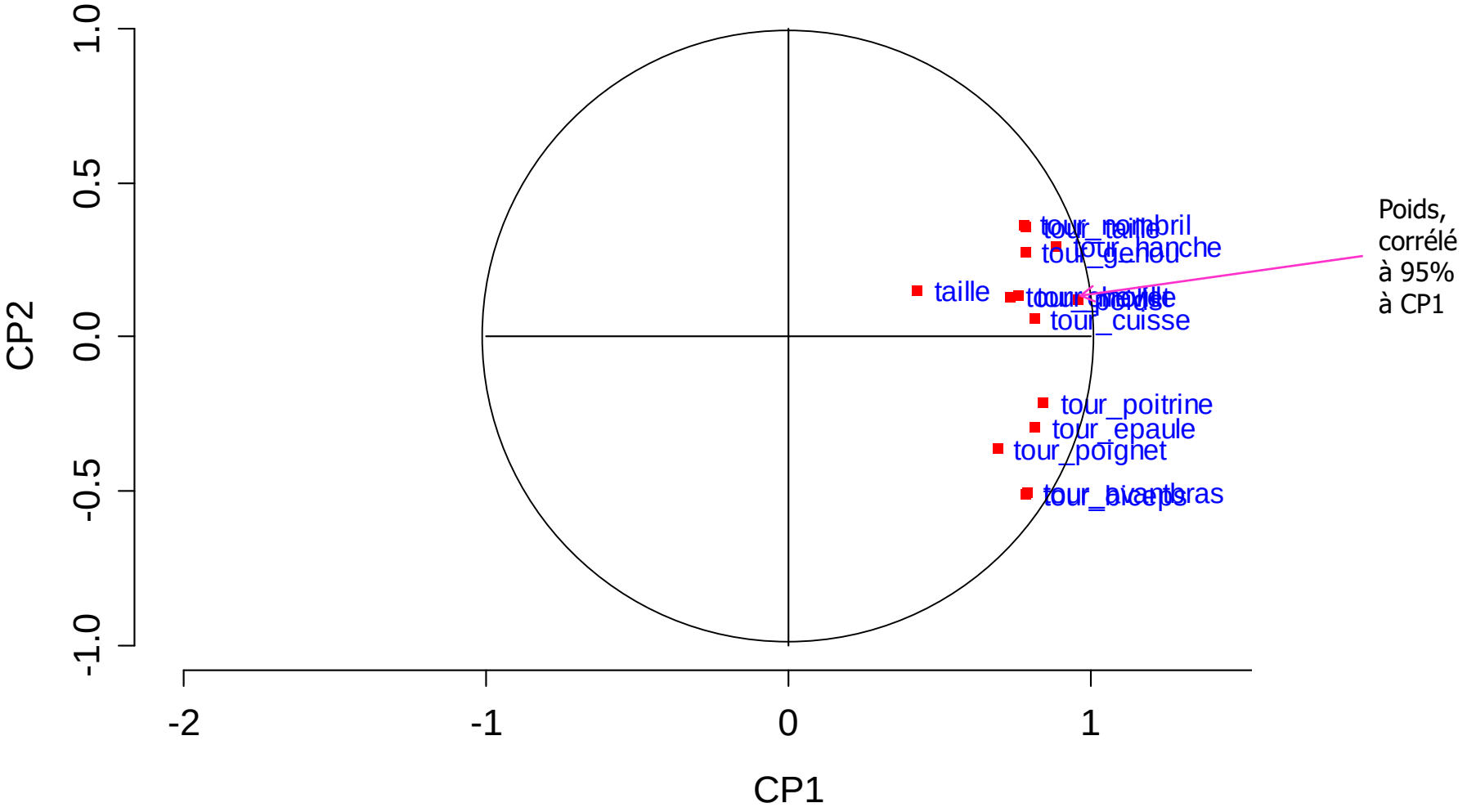
- 247 hommes. Même fichier que précédemment, plus complet
- 14 variables :
  - Poids en kg, taille en cm
  - Tours en cm : épaule, poitrine, taille, nombril, hanche, cuisse, biceps, avant-bras, genou, mollet, cheville, poignet
- Technique employée : ACP normée

Diagramme des valeurs propres

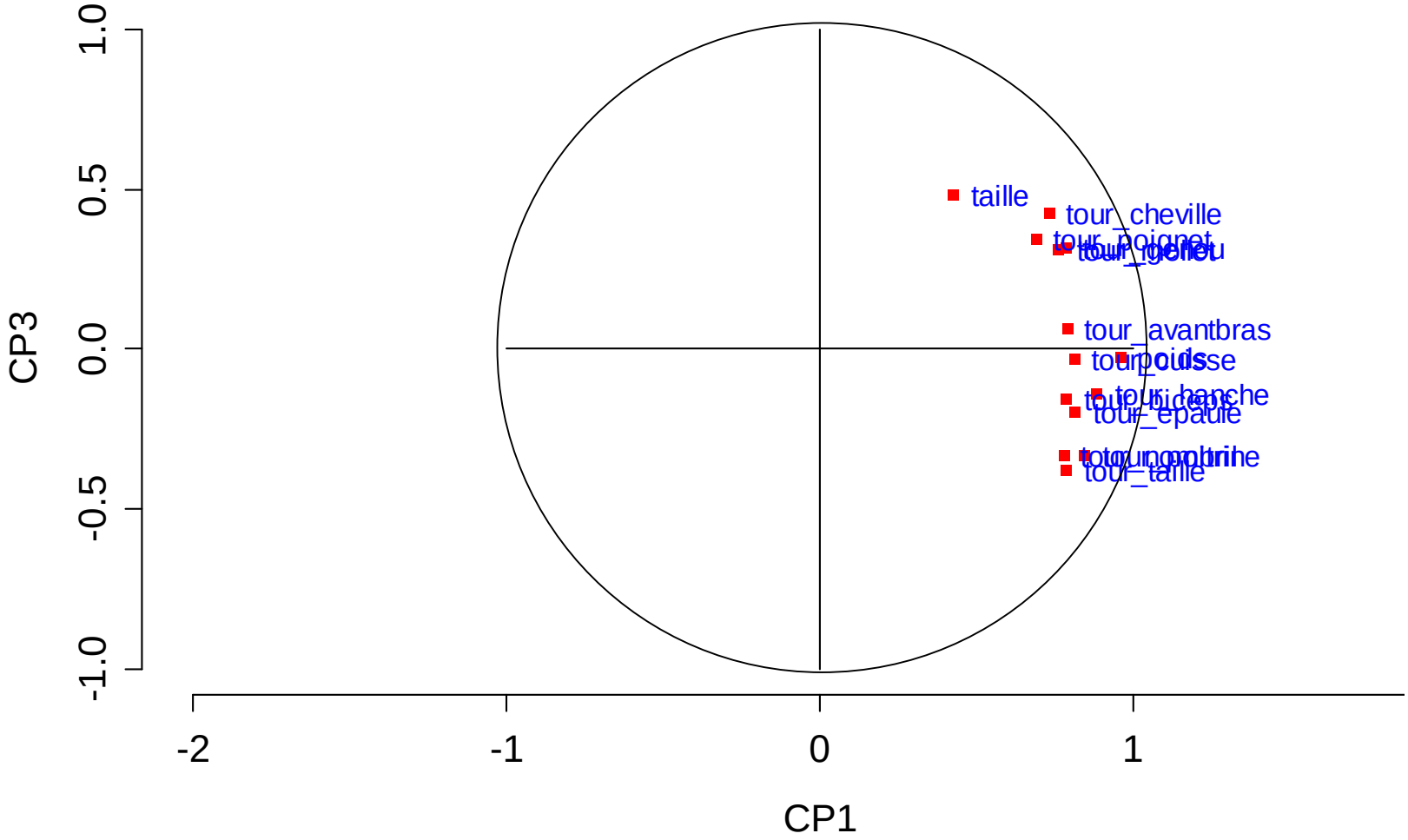




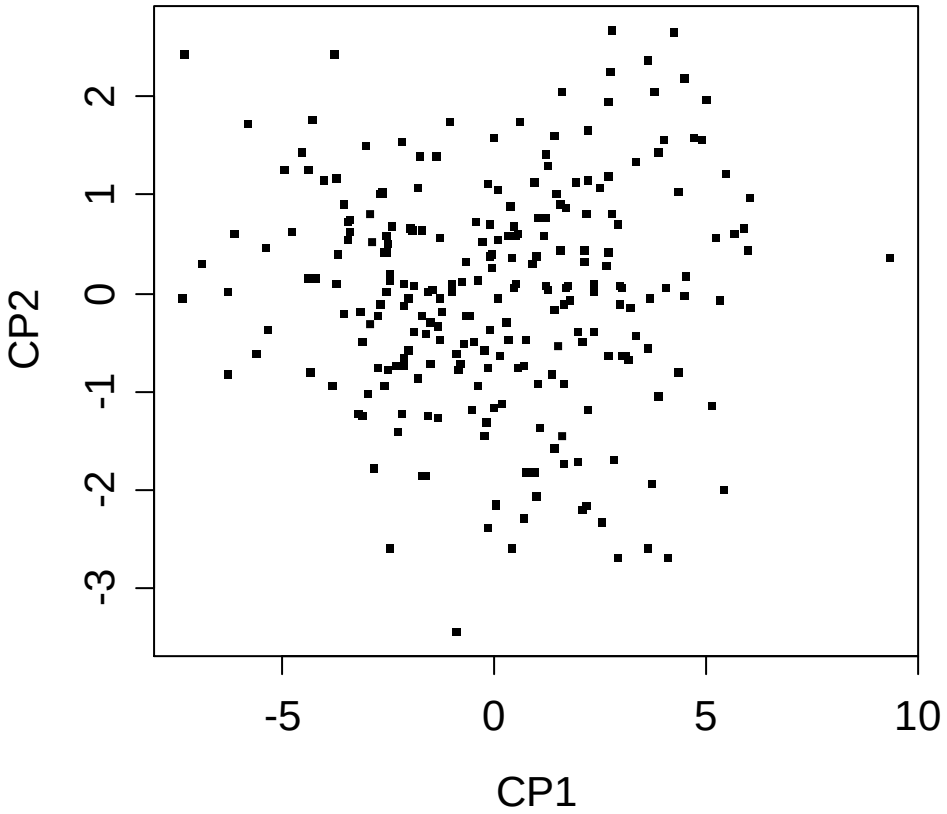
Plan factoriel des variables : CP1 et CP2

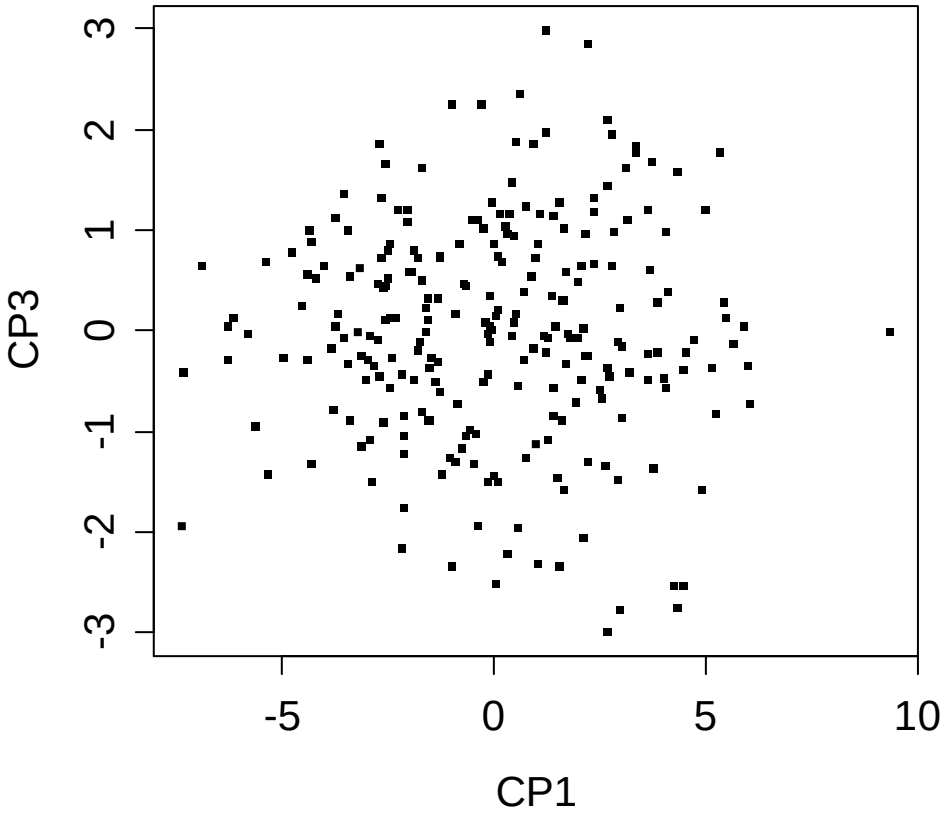


Plan factoriel des variables : CP1 et CP3



Plan factoriel des individus





Analyse du cas

1. Données et technique :
   - Données quantitatives, unités différentes => ACP normée
2. Part d'inertie expliquée
   - CP1 : 61%. CP1+CP2+CP3 : 78%
   - Très bon compte tenu qu'il y a 14 variables.
   - Clairement, la première CP est de loin la plus importante
3. Plan factoriel des variables
   - CP1 positionne à droite les individus de corpulence élevée (poids et tous les tours de…)
   - CP2 positionne vers le bas, à corpulence égale, les individus dont le membre supérieur est bien développé (biceps, avant-bras, poignet), et vers le haut ceux dont le tronc est imposant (nombril, taille).
   - CP3 positionne vers le haut, à corpulence égale, la corpulence liée au squelette (grande taille, tour de poignet et cheville élevés), et vers le bas la corpulence liée aux tissus mous (muscle et graisse).
4. Plan factoriel des individus
   - CP1 : à droite les lourds, à gauche les légers
   - CP2 : vers le bas les musclés du membre sup, vers le haut les gras
   - CP3 : vers le haut les grands, vers le bas les petits
   - Plus de variété dans les corpulences élevées (dispersion selon CP2)
   - Quelques individus extrêmes se distinguent
   - Possibilité de positionner de nouveaux individus pour les caractériser synthétiquement

Exemple n°4
Notes au tronc commun de PACES

Données étudiées = notes de chaque épreuve du tronc commun

- ~2500 étudiants ayant passé toutes ces épreuves
- 12 Variables :
  - 1 ou 2 notes sur 20 par UE du Tronc Commun PACES
  - Une variable qualitative : affectation finale
- Technique employée : ACP normée (sans la variable qualitative)

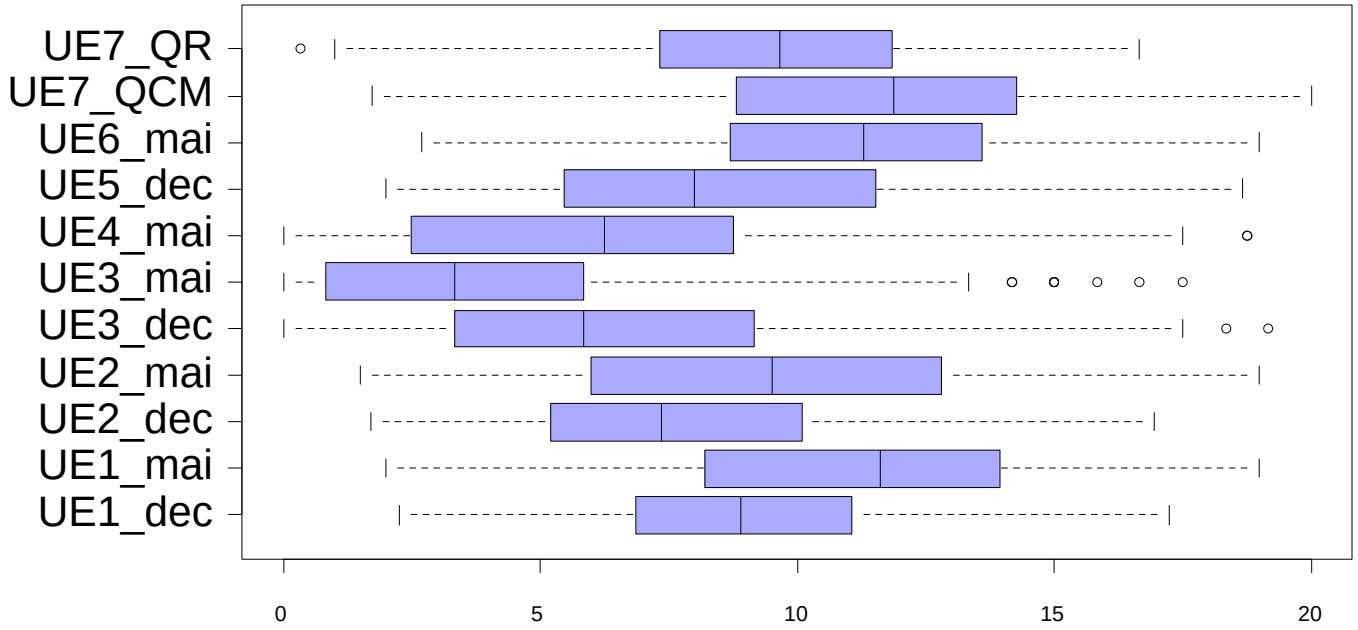
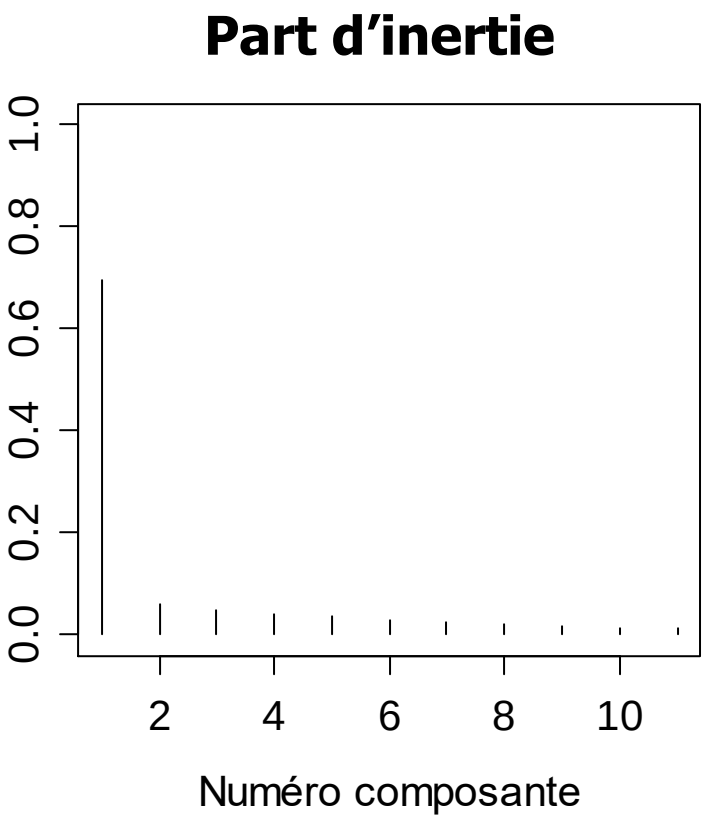
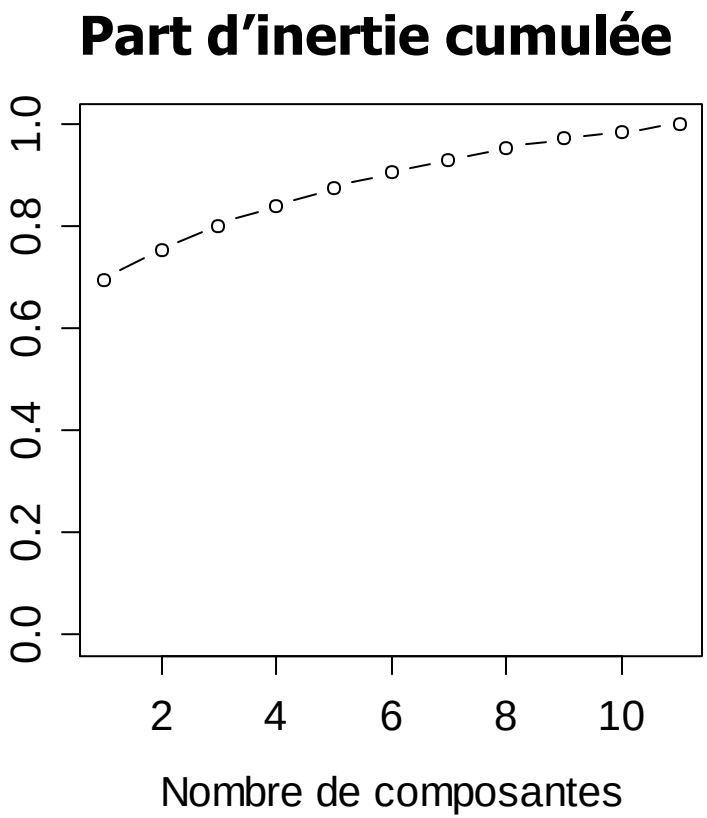
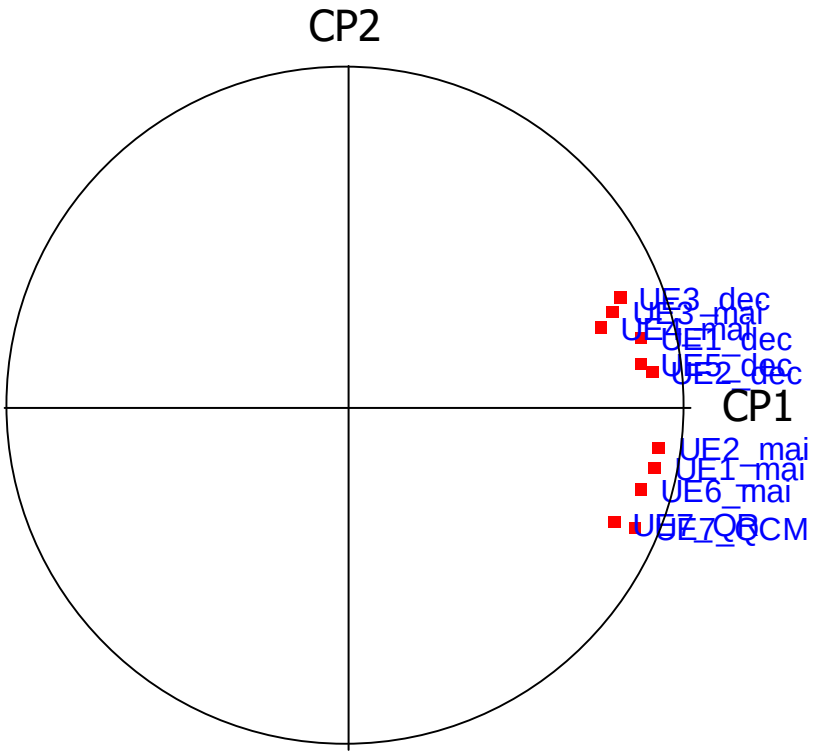


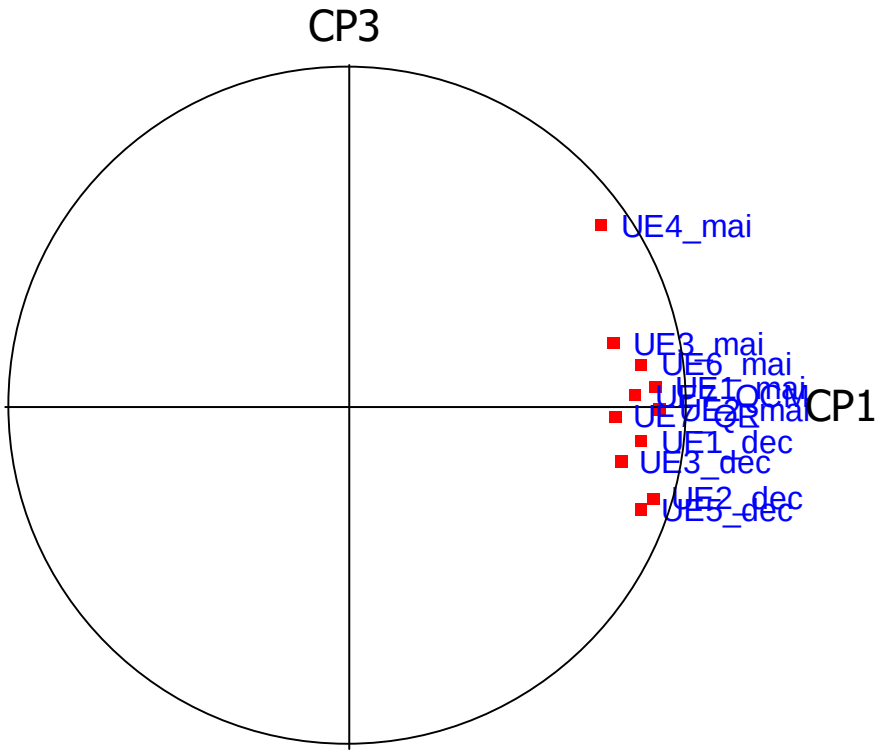
Diagramme des valeurs propres



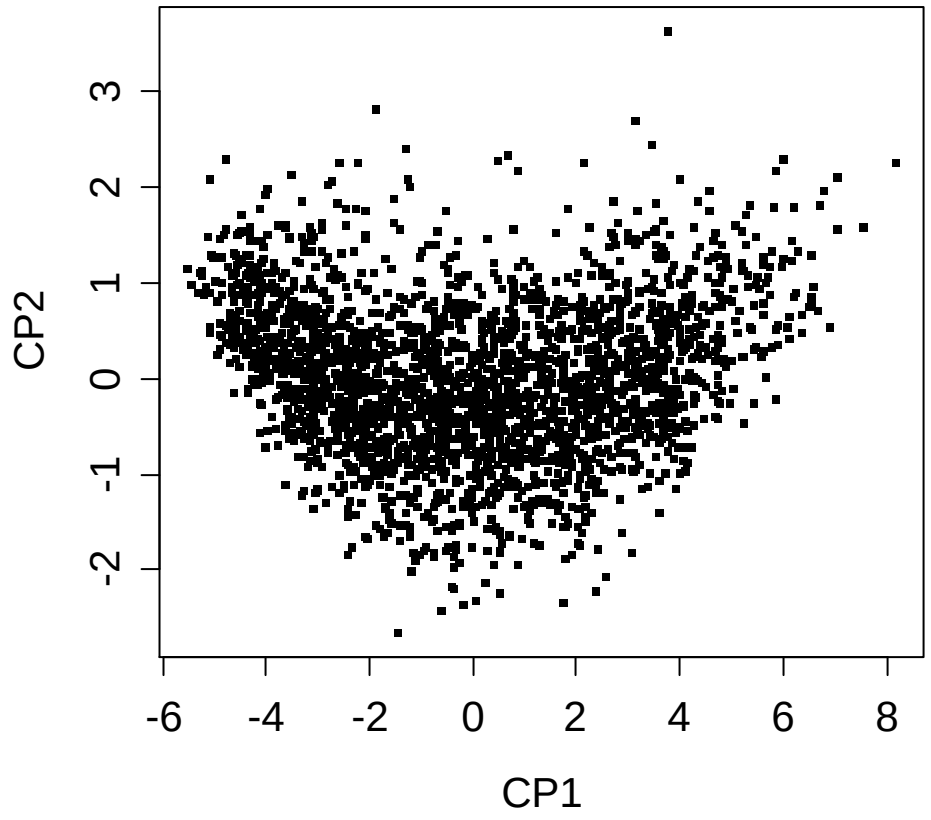


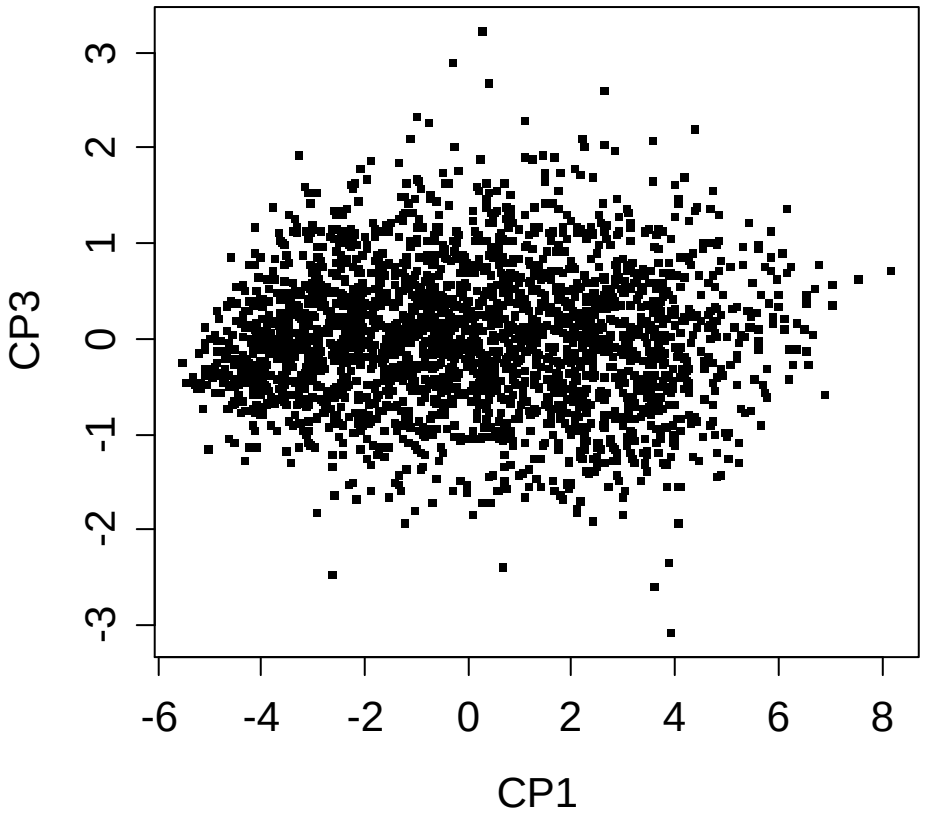
Plan factoriel des variables



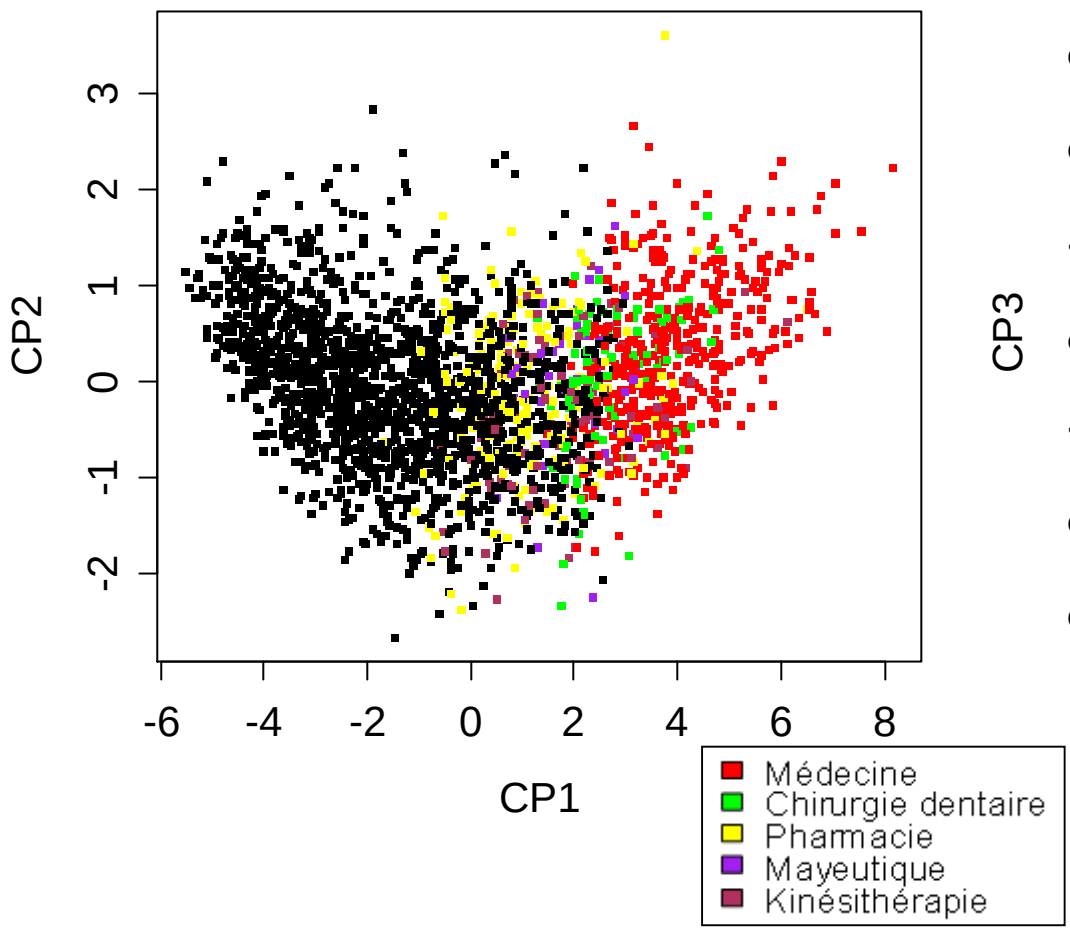


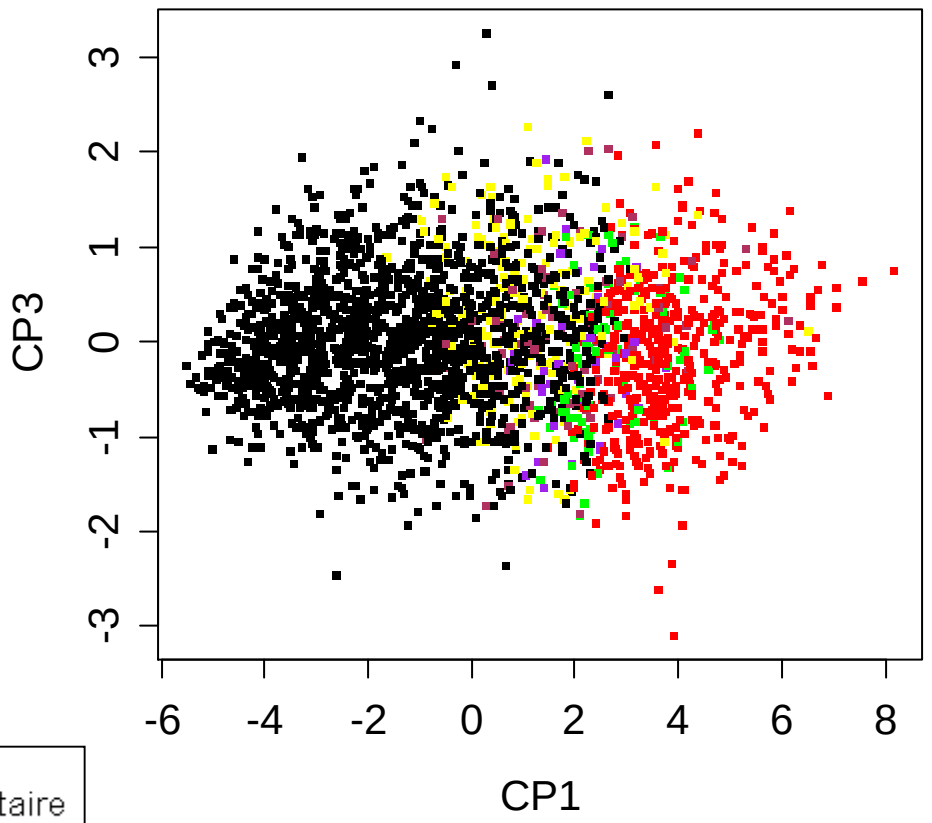
Plan factoriel des individus





Plan factoriel des individus
Couleur = affectation finale





Analyse du cas

1. Données et technique :
   - Données quantitatives, unités identiques => ACP non-normée également possible
   - L'affectation finale ne peut être utilisée pour construire l'ACP, mais pour illustrer
2. Part d'inertie expliquée
   - CP1 : 69%. CP1+CP2+CP3 : 80%
   - Impression générale : CP1 explique beaucoup, les suivantes n'apportent que des nuances qui sont toutes du même ordre de grandeur, et de manière général peu déterminantes. Peut-être même une CP par UE…
3. Plan factoriel des variables
   - CP1 représente le niveau général des étudiants. Ce niveau général semble se répercuter dans toutes les UE, ce qui est rassurant. Toutes les variables sont très corrélées à CP1.
   - CP2 est une composante complexe représentant notamment l'UE7 (bonnes notes : vers le bas), qui se distingue peut-être par le côté rédactionnel
   - CP3 est une composante complexe représentant notamment l'UE4 (bonnes notes : vers le haut), qui se distingue peut-être par les calculs à réaliser
   - A noter que les différentes notes d'une même UE (décembre et mai, QR et QCM) sont généralement très corrélées
4. Plan factoriel des individus
   - On a l'impression que les très bons et très mauvais sont nécessairement homogènes : les nuages se pincent à gauche et à droite
   - De nombreux étudiants semblent miser principalement sur l'UE7, et restent moyens
   - Affectation (ou choix de redoubler) en fonction du profil de l'étudiant

ACM : Analyse des correspondances multiples

Partie présentée pour votre culture.

Aucune question à l’examen.

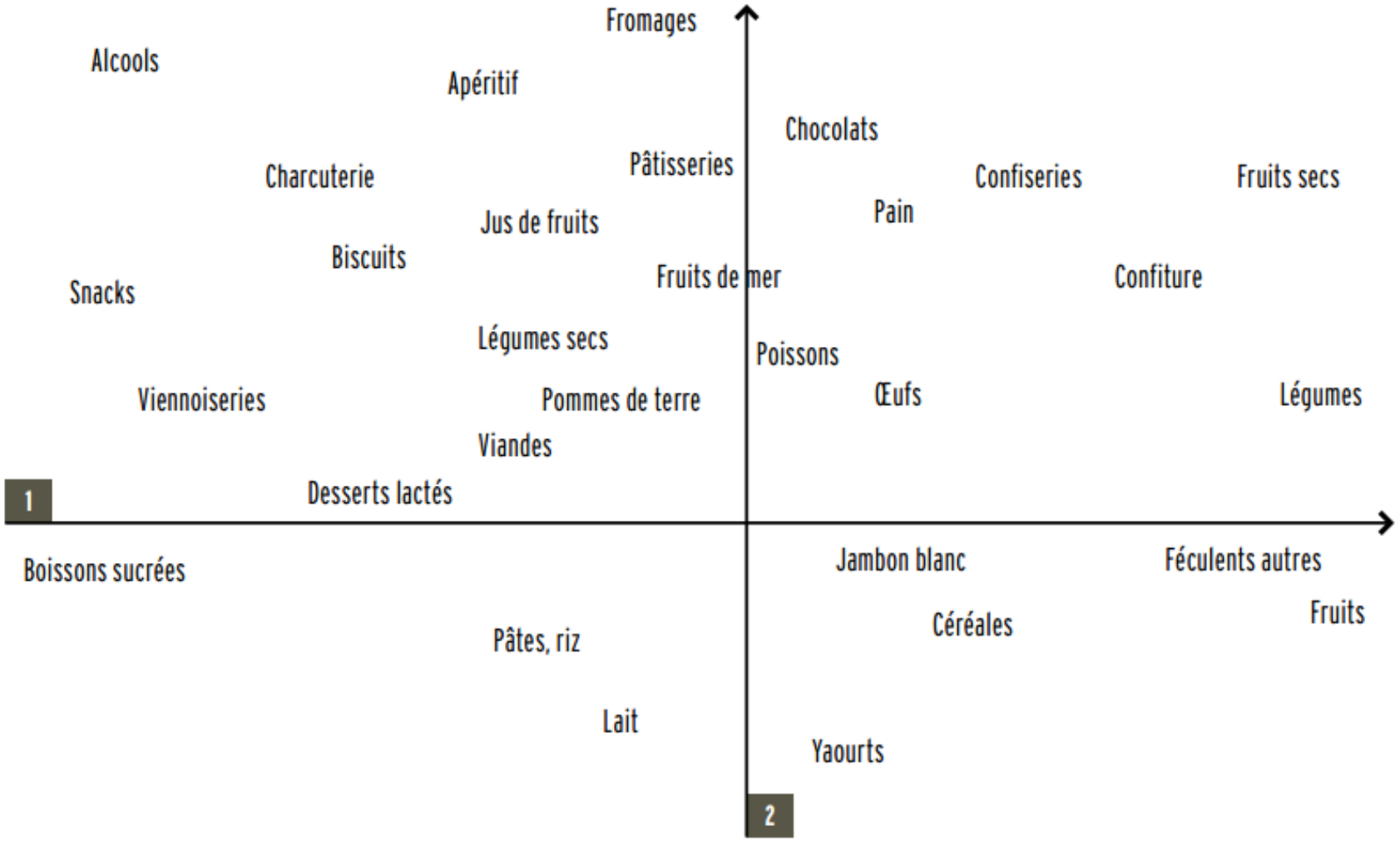
ACM en bref

- Même principe que l'ACP. Quelques différences :
- Données :
  - variables qualitatives, binaires, et quantitatives seulement si réduites en classes
- Part d'inertie expliquée par les CP :
  - Idem
- Plan factoriel des modalités :
  - Représente les modalités et non les variables
  - Les points ne sont pas circonscrits par un « cercle »
- Plan factoriel des individus (profils) :
  - Idem, mais par construction beaucoup d'ex-aequo, donc visualisation généralement peu intéressante

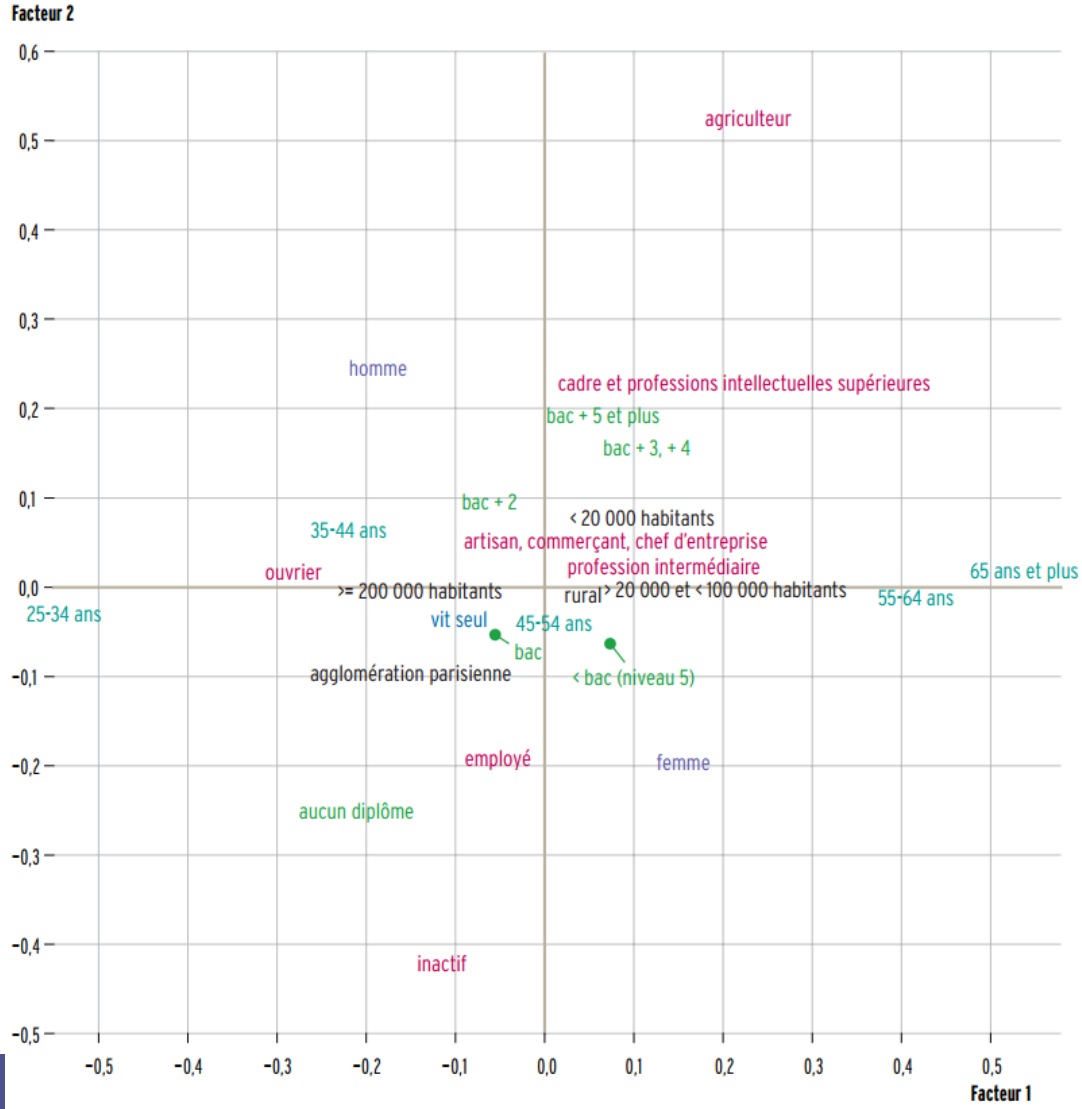
Exemple n°5
Baromètre nutrition santé de l'INPES

Plusieurs ACM réalisées depuis un questionnaire complet. Nous ne présenterons ici que les plans factoriels des modalités.

Plan factoriel des modalités
Aliment consommé le plus la veille



Plan factoriel des modalités
Profils sociodémographiques



Plan factoriel des modalités
Profils économiques