Compte-rendu du TP n°2 - Analyse en composantes principales partie 2

**Noms du groupe : AMINI Nada, ESSAYEGH Nour**

**----------------------------------------------------------------------------------------------------------------**

**Partie 1** - **l’ACP et étude de nuage de point dans .**

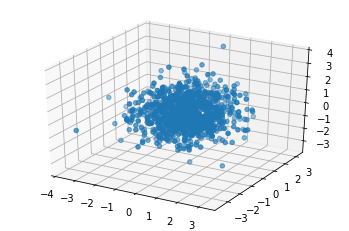
**Etude de la forme du nuage initiale et de sa répercussion sur la réduction de dimension**

* 1. Nuage isotrope

# On génère un nuage de données isotrope

# L’aperçu du nuage de points pour n=1000

On remarque que la distribution ne tend pas vers un des axes en particulier.



# La matrice de covariance

Mat\_cov=tp1.matrice\_covariance(V)

print(Mat\_cov)

Par construction de la data frame les variables X1, X2 et X3 sont indépendantes et on peut le visualiser sur la matrice de covariance ci-dessous :

Pour n=100

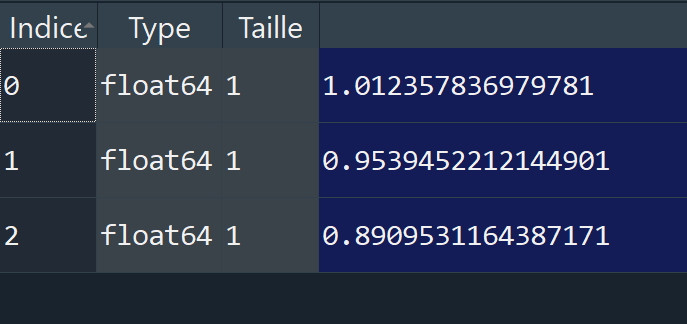


Pour n=10000



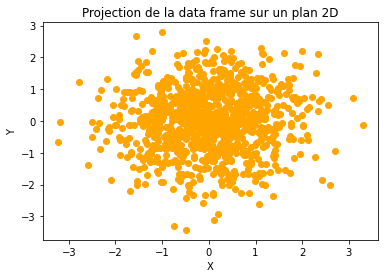
On remarque que pour un n plus grand la matrice de covariance tend vers la matrice identité. Ce résultat est assez intuitif.

# La cascade de valeurs propres



D'après le critère de **Karlis - Saporta - Spinaki**, on ne prend que les valeurs propres supérieurs ou égales à = 0.28 et donc toutes les valeurs propres. Elles sont toutes importantes sinon on aura une perte d’informations, mais on va quand même faire une projection sur les deux premiers hyperplans et voir la qualité de projection.

# Nouvelles coordonnées

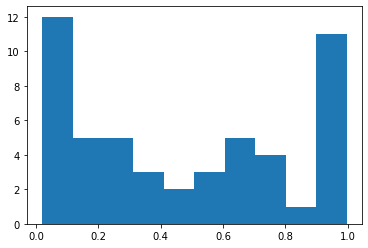
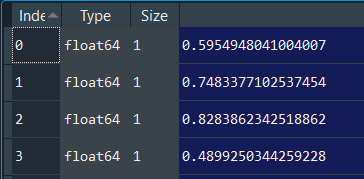


# La qualité de projection sur

nv\_coord=tp1.coord\_Rk(V,2).loc[:50,:]

Q = [tp1.quali\_representation(V,nv\_coord,i) for i in range(nv\_coord.shape[0])]

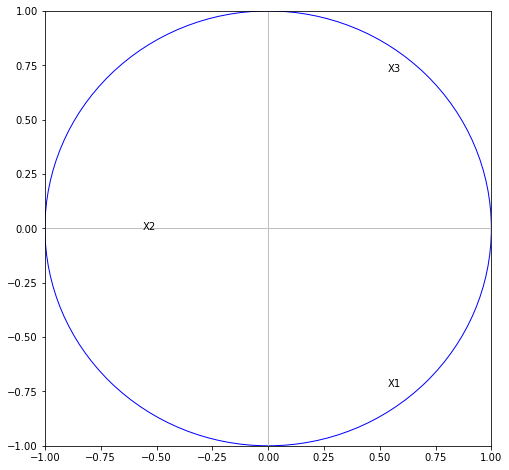
plt.hist(Q)

Quand on envisage de faire la projection sur les deux premiers hyperplans et qu’on visualise l’histogramme des qualités de projection, on remarque qu’il y a beaucoup d’individus dont la qualité de projection est entre 0 et 0.5 donc nous ne pouvons pas retenir une diminution de variables dans le cas de ce nuage.

Cela est prédictible vu la construction du nuage isotrope.

# Cercle des corrélations



Les variables sont indépendantes par construction et on voit bien que l'angle séparant X1 et X2 et l'angle séparant X1 et X3 sont proche de 90°, elles sont alors deux à deux indépendantes.

1.2 Nuage non isotrope

# On génère un nuage de données non isotrope avec une forte corrélation (relation de linéarité)

n=1000

X\_ni=X

Y\_ni=2\*X+1

Z\_ni=Z+X\_ni

X1\_ni=pd.DataFrame(data=X\_ni,columns=['X1'])

X2\_ni=pd.DataFrame(data=Y\_ni,columns=['X2'])

X3\_ni=pd.DataFrame(data=Z\_ni,columns=['X3'])

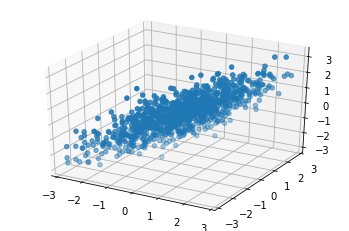
V\_ni=pd.concat([X1\_ni,X2\_ni,X3\_ni],axis=1)

# Il faut normer ce tableau

V\_ni=tp1.acp\_normee(V\_ni)

# L’aperçu du nuage de points après normalisation pour n=1000

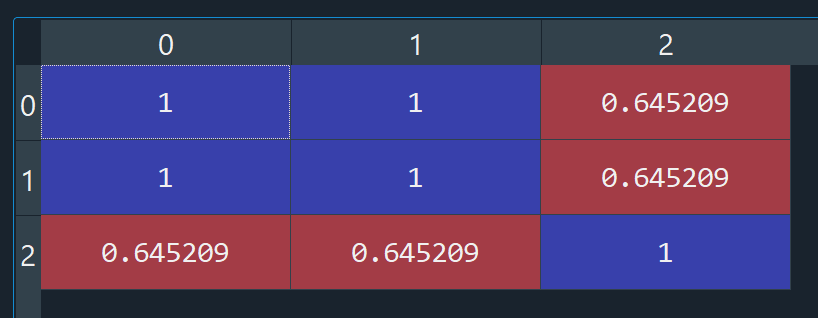
Ici on retrouve une forme elliptique allongée suivant un certain axe ce qui exprime une forte dépendance des variables Xi (notamment la relation de linéarité)



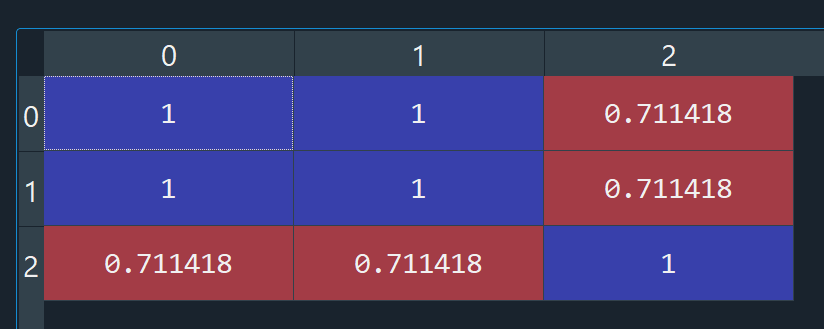
# La matrice de covariance

Par construction de la data frame, les variables X1 et X2 sont linéairement dépendante ce qui s’illustre dans la matrice de covariance par M\_cov[0,1] = 1. Il y a aussi une dépendance de X1 et X3 et cela revient à l’ajout de X1 à un échantillon de loi normale. (Bruit)

Pour n=100

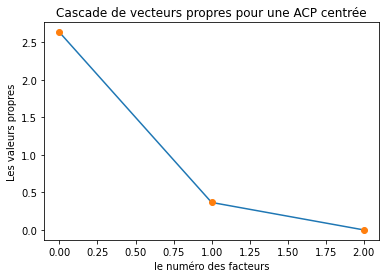
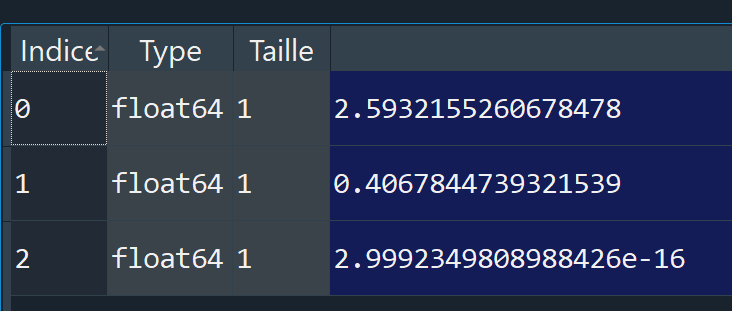


Pour n=10000



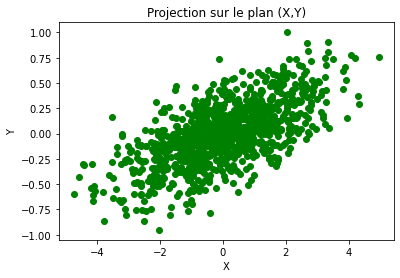
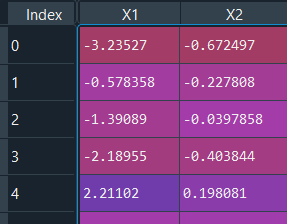
Les valeurs de la matrice de covariance s’affinent au fur et à mesure que n grandit.

# La cascade de valeurs propres

Sans avoir recours aux critères vus en cours, on voit ici que la troisième valeur propre est très négligeable devant les deux premières. Donc on peut envisager une projection sur les deux hyperplans correspondant aux deux premières valeurs propres.

# Nouvelles coordonnées

# La qualité de projection

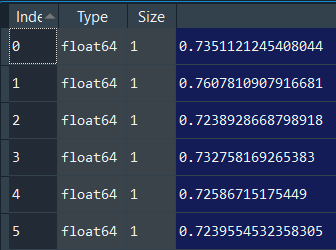
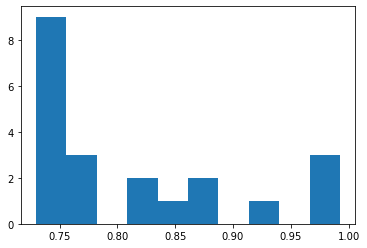
nv\_coord2=tp1.coord\_Rk(V\_ni,2).loc[:20,:]

Q2 = [tp1.quali\_representation(V\_ni,nv\_coord2,i) for i in range(nv\_coord2.shape[0])]

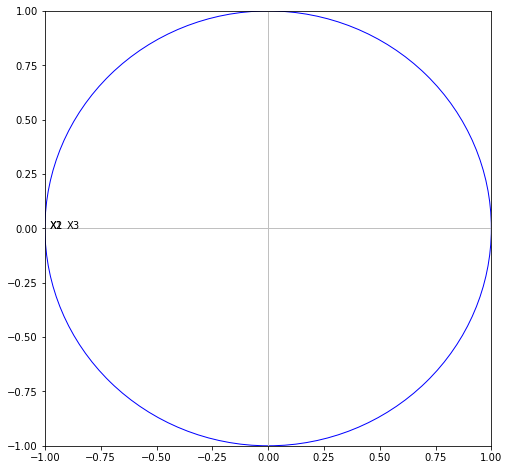
plt.hist(Q2)

plt.show()

La qualité de la projection est très élevée puisque pour tous les individus, elle est supérieure à 0.7. On le voit sur l’histogramme.

# Cercle des corrélations

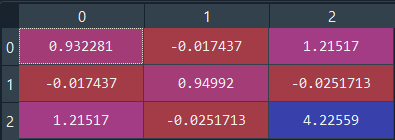


Cette fois-ci, les variables sont corrélées par construction et on voit clairement que X1 et X2 sont superposés, l'angle entre elles est donc 0 et elles sont positivement corrélés, X1 et X3 sont colinéaire et donc très fortement corrélés.

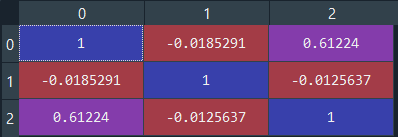
2) Cas du nuage proposé :*x <- sort(rnorm(1000)) y <- rnorm(1000) z <- rnorm(1000) + atan2(x, y)*

# On génère le nuage de données non isotrope proposé par l’énoncé

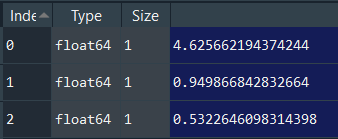
# Matrice de covariance non normé



# Matrice de covariance normé



# Les valeurs propres

On peut remarquer qu’entre les valeurs propres de la data frame normée et non normée, il peut y avoir une erreur d’interprétation dans la considération de l’hyperplan de la projection. En effet, avec la liste de droite des valeurs propres et d’après la règle de **Kaiser-Guttman**, on a envie de prendre une projection sur un seul axe factoriel correspondant à la première valeur propre.

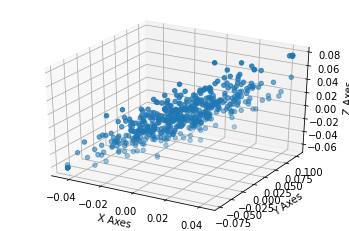
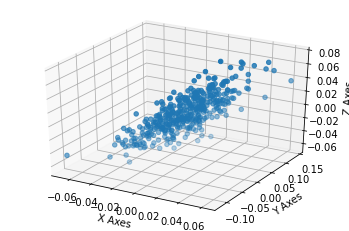
Cependant après la normalisation on voit que les deux premières valeurs propres sont importantes et obéissent à la règle de **Kaiser-Guttman**. On aurait eu une perte d’information si l’on avait considéré que la première.

**Conclusion :**

Il est essentiel d’utiliser l’ACP normé pour une bonne interprétation des résultats.

1.3 Points extrémaux

# Pour le nuage anisotrope

A droite, on a le nuage de points initial, et à gauche on a le nuage de points auquel on a ajouté une dizaine de points extrémaux. On voit bien qu’il y a un allongement dans une direction oblique.

En ajoutant des points extrémaux, on augmente la variance. Il est préférable de centrer et réduire la data frame pour que les grandes variances n’écrasent pas les plus petites et qu’on soit en mesure de bien interpréter les résultats.

**Partie 2. Etude de la forme du nuage initiale sur la réduction de dimension dans les deux espaces ou**

Vous avez développé l’ACP centrée et centrée réduite sur l’espaceen TP1 : on vous demande successivement de :

1. Mettre en œuvre l’ACP sur espace Rn sur les deux cas
   1. Nuage isotrope

Cette fois-ci on cherche à maximiser la matrice X\*X’ où X est la data frame initiale.

Dans toute la suite, on prend n = 50.

XX\_prime = D\_norm.dot(np.transpose(D\_norm))

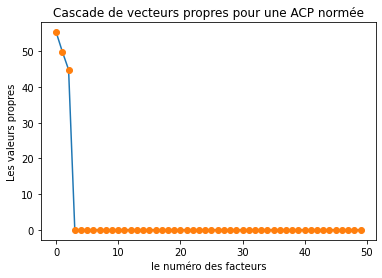
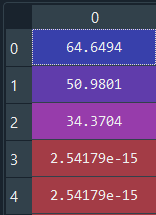
XX\_prime.columns = ["X{}".format(i+1) for i in range(XX\_prime.shape[0])]

val\_propre\_Rn, vec\_propre\_Rn = np.linalg.eig(XX\_prime)

vec\_propre\_Rn = np.transpose(vec\_propre\_Rn)

val\_propre\_Rn,vec\_propre\_Rn = classer(val\_propre\_Rn,vec\_propre\_Rn)

# Cascade des valeurs propres

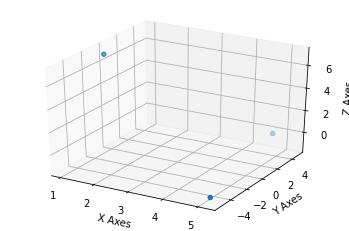
 

On voit que la cascade des valeurs propres s’écrase rapidement, après les 3 premières valeurs. D'après le critère de **Karlis - Saporta - Spinaki**, on ne prend que les valeurs propres supérieurs ou égales à = 9.89 et donc les trois premières valeurs propres.

# Nouvelle coordonnées dans l’espace

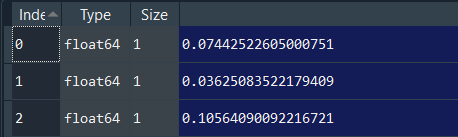
nouvelle\_coord = (np.transpose(D\_norm)).dot(np.transpose(vec\_propre\_Rn[:3]))

nouvelle\_coord.columns = ["X1","X2","X3"]



On a du mal à interpréter ce qui se passe dans ce « nuage » de points en dimension 50 car le nombre d’individus est très faible devant le nombre de variable (50 >> 3). Et ceci se reflète aussi dans la qualité de projection qui est très faible.

# Qualité de projection sur



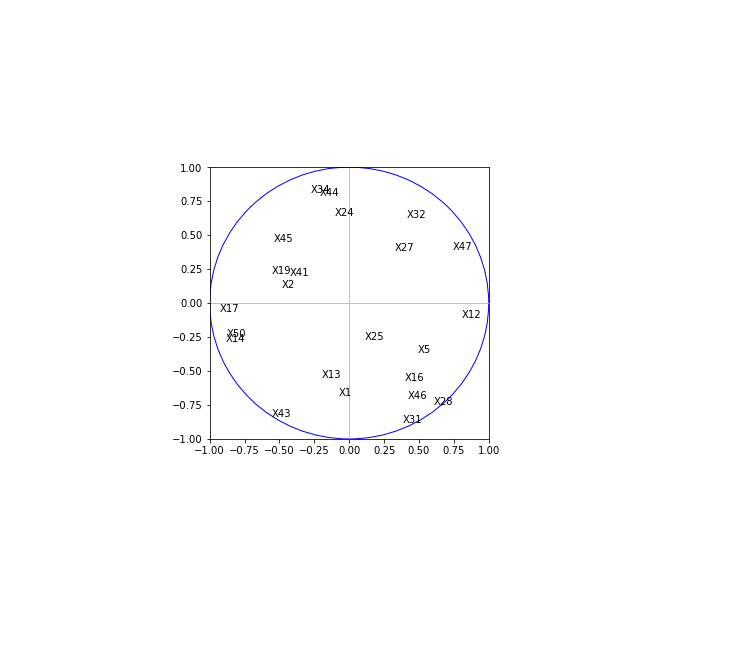
On doit retenir que dans le cas où dans le nombre de variable est très faible devant le nombre d’individus, (dans le cas de ce TP on a dans 3 variables X, Y, Z) Il ne faudrait pas passer à l’ACP dual sur car cela donne lieu à des aberrations.

# Cercle des corrélations

cor = np.zeros((m,m))

for i in range(m):

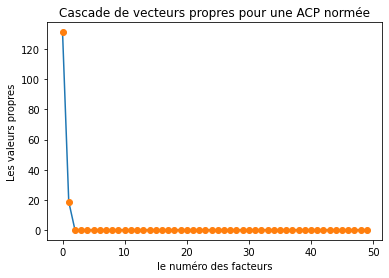
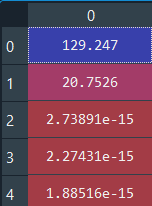
cor[:,i] = np.sqrt(abs(val\_propre\_Rn[i])) \* vec\_propre\_Rn[i]



On remarque que plusieurs variables sont corrélées : l’angle compris entre elles et proche de 0 (ou 180°) et donc le cosinus est proche de 1 (ou -1). Cependant, plusieurs variables sont indépendantes comme X2 et X13 ou encore X1 et X17 (angle de 90° entre elles). Cela pourrait -peut-être- en partie expliquer la faible qualité de projection vu qu’on a éliminé des axes factoriels indépendants de ceux choisis pour la projection.

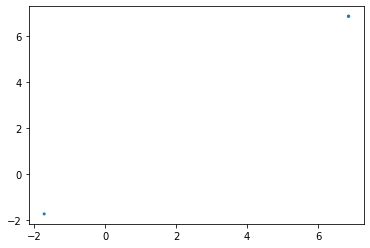
* 1. Nuage non isotrope

# Cascade des valeurs propres

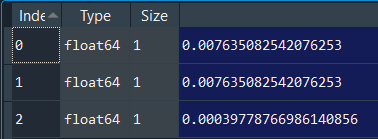
Encore une fois, la cascade des valeurs propres s’écrase rapidement. Cette fois-ci après les deux premières valeurs propres. Et d'après le critère de **Karlis - Saporta - Spinaki**, on ne prend que les valeurs propres supérieurs ou égales à = 9.89 et donc les deux premières valeurs propres.

# Nouvelle coordonnées dans

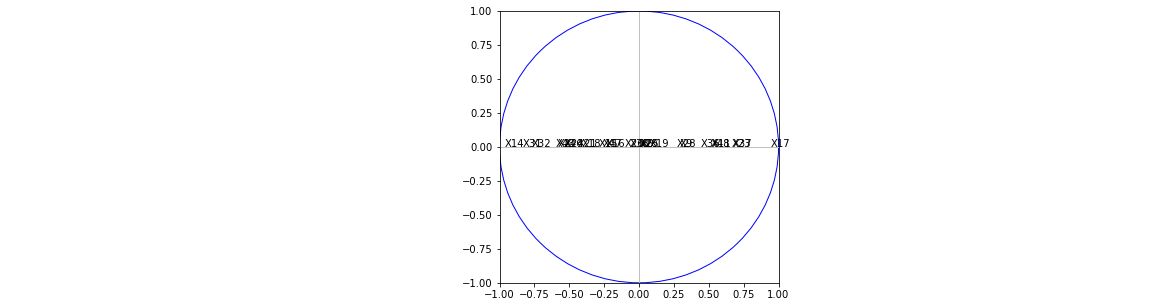
 

Comme dans dans le cas précédent de nuage isotrope, Il est difficile de visualiser et interpréter ce qui se passe en dimension 50 car le nombre d’individus est très inférieur au nombre de variables. Encore une fois, cela se reflète dans une qualité de projection encore plus faible que dans le cas de nuage isotrope.

# Qualité de projection dans



# Cercle des corrélations



Les variables sont toutes positivement ou négativement corrélées par construction et cela se reflète dans le cercle des corrélations.

1. Comparer avec les résultats dans sur les deux cas

Cas 1 : Nuage isotrope

Notre data frame initiale est où p = 3 (X, Y, Z) et n prend de grandes valeurs. L’ACP sur n’était pas nécessaire d’après les résultats obtenus. Elle est même déconseillée car sinon on a une perte d’informations et cela n’est pas plus mal car on peut déjà visualiser les résultats sur .

Cependant, quand on passe à l’ACP duale sur , bien qu’on arrive à réduire la dimension de n à 3 dans le respect des règles et théorèmes vu en cours, on n’a pas une visibilité sur ce qui se passe en dimension n car le nombre d’individus p est très faible (p=3) devant n.

Cas 2 : Nuage anisotrope

Dans ce cas pour l’ACP dans , on arrive à réduire l’espace de à sans perte d’informations.

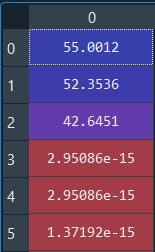
Similairement au cas du nuage isotrope, quand on effectue une ACP dual dans , on n’arrive pas à interpréter les résultats quand on réduit la dimension de n à 2 car encore une fois le nombre d’individus est très faible.

**Conclusion :**

L’ACP n’est pas efficace en espace duale si n >> p. Elle serait intéressante seulement quand n et p sont proches, sinon on n’arrive pas à interpréter les projections.

1. Retrouver les formules de passage à partir du cours et de vos résultats successifs
2. Valider les résultats sur vos exemples

Pour ne pas répéter deux fois les mêmes conclusions, on va se restreindre au cas du nuage isotrope et retrouver les formules du cours.

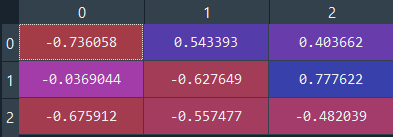
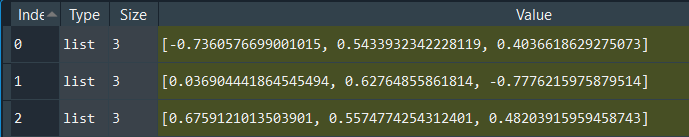
A droite, on a le tableau des valeurs propres de X\*X’ pour l‘ACP dans . On voit bien qu’à partie de la quatrième valeur propre, elles sont toutes nulles. Quand on normalise en divisant par n = 50 on retrouve successivement : 1.100024 et 1.047072 et 0.852902 qui sont exactement les valeurs propres de X’\*X dans .

->On retrouve la propriété du cours qui stipule que toutes les valeurs propres non nulles de X\*X’ et X’\*X sont égales.

u\_alpha = []

for i in range(vec\_propre.shape[0]):

u\_alpha.append(np.transpose(D\_norm).dot(np.transpose(vec\_propre\_Rn[i]))/np.sqrt(val\_propre\_Rn[i]))

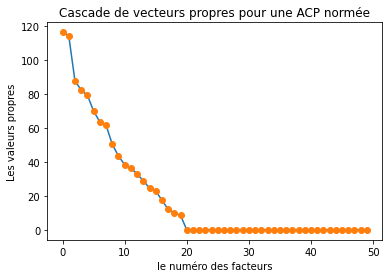
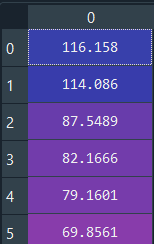
A gauche, on a la matrice des vecteurs propres de X’\*X et à droite on a les vecteurs propres obtenus avec la formule  . Si on arrondit à 3 chiffres après la virgule, ces deux matrices sont égales.

->On a donc retrouvé la formule du cours qui stipule que =

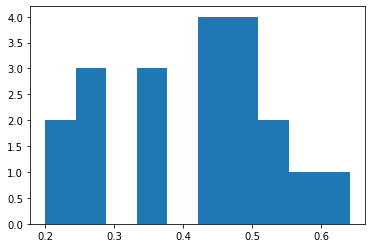
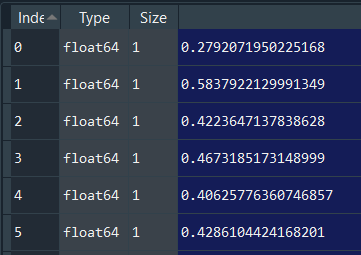
1. Tester sur les deux cas : l’effet de la dimension sur les deux décompositions
   1. Cas 1 : Nuage isotrope

Ici on augmente le p de 3 à 20, pour qu’il soit plus proche de n = 50 et on teste l’ACP dual.

# Cascade des valeurs propres

On remarque que les cascade s’écrase graduellement jusqu’au 20ème facteur contrairement aux cascades des questions précédentes. D'après le critère de **Karlis - Saporta - Spinaki**, on ne prend que les valeurs propres supérieurs ou égales à = 3,21 et donc les 20 premières valeurs propres.

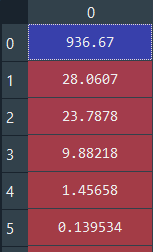
->La qualité de projection est déjà nettement supérieures à celle qu’on obtenait quand on avait p = 3 individus. Cela confirme notre conclusion, l’ACP duale est efficace quand n et p sont du même ordre de grandeur.

* 1. Cas 2 : Nuage anisotrope

On garde p = 20 et on génère une data frame ou les p variables sont corrélées.

# Cascade des valeurs propres

Une image contenant carré

Description générée automatiquement 

Dans ce cas aussi, la cascade s’écrase rapidement. D'après le critère de **Karlis - Saporta - Spinaki**, on ne prend que les valeurs propres supérieurs ou égales à = 3,21 et donc les 4 premières valeurs propres.

De même que pour le cas isotrope, l’histogramme des qualités de représentation s’améliore.

**Conclusion :**

L’augmentation de la dimension p a un impact positif sur l’ACP duale. En effet, comme démontré dans les questions précédentes, plus n et p sont proches l’un de l’autre plus l’ACP duale est efficace.

6) Proposer la projection de points supplémentaires (espace sur ) et de points colonnes (variables) dans ce nouvel espace . Pour cela vous devez proposer de générer soit des colonnes supplémentaires soit des lignes supplémentaires comme individus lignes :

Dans ce cas on augmente le n dans l’ACP sur . Pour n = 50 et n = 55, la qualité de projection ne change pas en moyenne. Et de même quand on augmente p dans l’ACP sur , de p = 3 à p = 5, la qualité de projection en moyenne augmente faiblement.

On en conclut donc que notre ACP est stable.