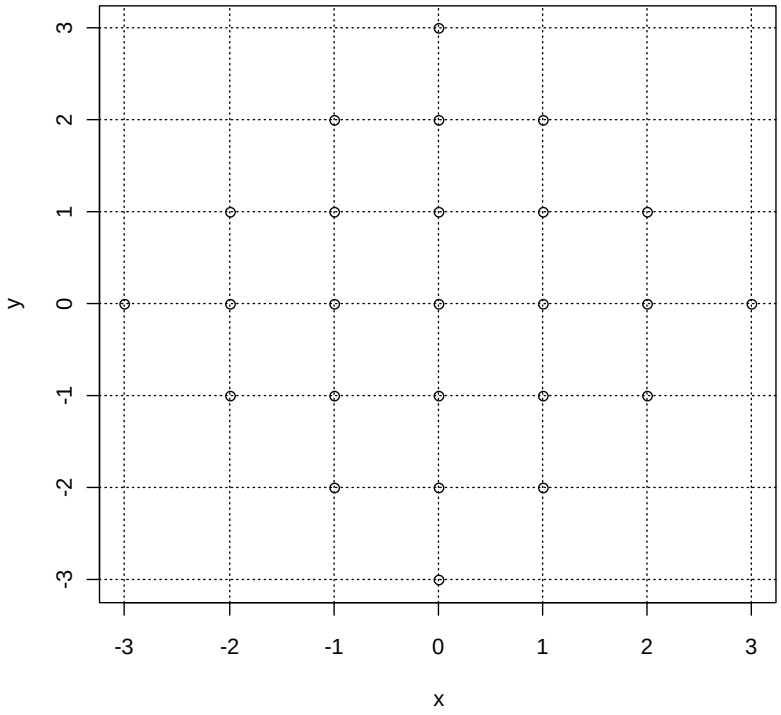
TD n°1*b* - On visse les boulons

Exercice n° 1.
Soit la matrice $\mathbf{X} = [\mathbf{x}, \mathbf{y}]$ constituée de 2 variables dont voici une représentation graphique :



1. Rappeler les métriques utilisées dans l'espace des variables et celui des individus lorsqu'on travaille avec a) avec une matrice de covariance, b) avec une matrice des corrélations.
2. Calculer la moyenne des variables. Que peut-on en déduire?
3. Calculer l'inertie totale a) de manière intelligente, b) en utilisant le nuage de points, c) en travaillant sur les variables.
4. Déterminer la matrice de variance $\mathbf{V}$ ainsi que la matrice des corrélations $\mathbf{R}$ a) en passant par le tableau centré réduit, b) par calcul explicite. Que remarque t-on?
5. Refaire l'exercice en tirant un nuage de points selon une loi normale de parametres $\boldsymbol{\mu} = (1, 2, 3)'$ et une matrice de corrélation de la forme
$$\mathbf{R} = \begin{pmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{pmatrix}$$
sachant que $var\,(X) = 1$, $var\,(Y) = 4$ et $var\,(Z) = 9$. On prendra $\rho = 0.5$.
6. Tracer les biplots des différentes variables du nuage. Tracer le graphique en 3D. Que remarque t-on?
7. Vérifier les paramètres des distributions marginales. Que remarque t-on? Pourquoi?

Correction

1. Voir le cours

2. L'axe des abscisses ainsi que celui des ordonnées sont axes de symétrie : le centre de gravité est donc $\mathbf{g} = (0,0)'$. Le nuage de points est centré : les variables sont de moyenne nulle.

3. Pour calculer l'inertie totale, nous allons procéder en utilisant les propriétés de symétrie du nuage. Considérons le quadrant supérieur gauche du graphique ainsi que les points situés sur la partie négative de l'axe des abscisses, soit en tout 6 points. La somme des carrés des distances au centre de gravité donne pour ces 6 points : $1+2+3+2+3+3$. Par symétrie, l'inertie totale correspond à 4 fois cette quantité divsée par le nombre total de points (25), soit
$$I = \frac{1}{25}\left(4 \times (1+4+9+5+5+2)\right) = \frac{4 \times 26}{25} = 4.16.$$
On peut également calculer cette inertie en utilisant les coordonnées des points $\mathbf{x}_i = (x_{i1}, x_{i2})'$, sans se poser de question, on aura ainsi :
$$I = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{g}\|^2 = \frac{1}{n}\sum_{i=1}^{n} x_{i_1}^2 + \frac{1}{n}\sum_{i=1}^{n} x_{i_2}^2 = \frac{1}{25}\left(52+52\right) = 4.16.$$
La dernière façon de procéder est de considérer la matrice de variance-covariance des deux variables qui constituent les coordonnées des individus. L'inertie totale est égale à la trace de cette matrice. Si on effectue les calculs :
$$tr(\mathbf{V}) = tr\begin{pmatrix} 2.08 & 0 \\ 0 & 2.08 \end{pmatrix} = 4.16.$$
Cette matrice nous renseigne également sur le fait que les deux variables sont orthogonales (la covariance est nulle).

4. Voir résultat ci-dessus. Soit $\mathbf{X} = \left(\mathbf{x}^1, \mathbf{x}^2\right)$ la matrice de données avec $\mathbf{x}^j = (x_{1j}, ..., x_{nj})'$ la $j$ième variable. On note la matrice centrée réduite $\mathbf{X}_{cr} = \left(\frac{1}{\sigma_1}\mathbf{x}^1, \frac{1}{\sigma_2}\mathbf{x}^2\right)$ où σ_j est l'écart-type de la $j$ème variable. Dans ce cas, le centre de gravité étant nul, il n'apparaît pas dans le centrage. La matrice de corrélation $\mathbf{R}$ est donnée par :
$$\mathbf{R} = \frac{1}{n}\mathbf{X}'_{cr}\mathbf{X}_{cr}.$$
La matrice de corrélation est équivalente à une matrice de variance calculée sur des données centrées réduites. D'un autre côté, on peut également la calculer à partir de $\mathbf{X}_c$, le tableau centré, en utilisant la matrice $\mathbf{M}^{1/2} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}$avec
$$\mathbf{R} = \frac{1}{n}\left(\mathbf{X}_c\mathbf{M}^{1/2}\right)'\left(\mathbf{X}_c\mathbf{M}^{1/2}\right) = \mathbf{M}^{1/2}\mathbf{V}\mathbf{M}^{1/2}.$$

5. Voit code R ci-dessous. Pour reconstituer la matrice de variance-covariance à partir de $\mathbf{R}$ on utilise
$$\mathbf{V} = \mathbf{M}^{-1/2}\mathbf{R}\mathbf{M}^{-1/2}.$$

6. Voir code R ci-dessous. Lors du tracé en 3D on se rend compte que les données sont regroupées sur un sous-espace de $\mathbb{R}^3$ (un plan). Ceci indique qu'il existe une relation linéaire entre deux variables du tableau initial.

7. On retrouve une estimation empirique des paramètres de départ (variance et moyenne des variables). En fait, plus on augmentera le nombre de valeurs tirées au hasard, plus les paramètres estimés empiriquement seront proches des peramètres de variance et de moyenne initiaux (lois des grands nombres).

```
########## CODE R ####################
#TD2 - tracé du nuage de points
x <- c(0,-1,0,1,-2,-1,0,1,2,-3,-2,-1,0,1,2,3,-2,-1,0,1,2,-1,0,1,0)
y <- c(-3,rep(-2,3),rep(-1,5),rep(0,7),rep(1,5),rep(2,3),3)
```

```
plot(x,y)
abline(v=-3:3,h=-3:3,lty=3)
#Le nuage de points est centré
xm = mean(x)
ym = mean(y)
#Calcul de l'inertie totale
#par symétrie
Itot =1/25 * 4 * (1+4+9+2+2+5)
#Par calcul direct
Itot = 1/25 * sum(x^2 + y^2)
#En calculant les variances des variables x et y
#ATTENTION var est sans biais (/(n-1)), il faut la corriger
varx = 24/25 * var(x)
vary = 24/25 * var(y)
covxy = 24/25 * cov(x,y)
covyx = 24/25 * cov(y,x)
#Creation matrice V
V = cbind(c(varx,covxy),c(covyx,vary))
#ou alors
X=cbind(x,y)
V=24/25 * t(X) %*% X
## Inertie totale, on retrouve les resultats precedents
## en passant par les variables
Itot = sum(diag(V))
##MAtrice de corrélation R
Mdemi = diag(c(1/sqrt(varx),1/sqrt(vary)))
R = Mdemi %*% V %*% Mdemi
## En passant par tableau centré réduit
Xcr=X %*% Mdemi
R = 1/25 * t(Xcr)%*%Xcr
##Refaire l'exercice en tirant un nuage de points en 3D
##selon une loi normale de parametres mu=c(1,2,3) et de matrice de
## corrélation R (voir enonce) sachant que varX, varY, varZ sont données (voir énoncé)
#Construction de la matrice de covariance
mu= c(1,2,3)
varX=1
varY=4
varZ=9
R=cbind(c(1,0.5,-0.5),c(0.5,1,0.5),c(-0.5,0.5,1))
#reconstitution variance a partir correlation
Mdeminv = diag(c(sqrt(varX),sqrt(varY),sqrt(varZ)))
V = Mdeminv %*% R %*% Mdeminv
##tirage aléatoire de 100 points selon une gaussienne
library(mvtnorm)
X = rmvnorm(100, mean = mu, sigma = V)
## faire représentation 2d puis 3d
##
Y=data.frame(X)
### Le biplot est directement donné par la fonction plot()
plot(Y)
##Représentation 3D : les données reposent sur un plan
library(rgl)
plot3d(X,size=5)
##Paramètre des distributions marginales (pour chaque variable)
apply(X,2, mean)
```

```
apply(X,2,var)
###On retrouve les valeurs approchées des paramètres de départ.
```