COURS « PROBABILITES » - UP1 : TP 2

Abdelhakim Benechehab - Younes Gueddari

Octobre 2019

1 Exercice 2

On reprend le jeu de données simulées exercice 1, partie 1, question 1. L'objectif de cet exercice est d'étudier empiriquement la loi conditionnelle de X2 sachant X1 = x1 où x1 est donné.

1.1 Pour h « petit » devant sigma1 = 1, extraire le jeu de données correspondant à X1 e [x1 − h, x1 + h] puis étudier la distribution empirique des valeurs associées de la variable X2. Qu'observez-vous ?

Taille de l'echantillon (les individus), a été augmentée pour une meilleure visualisation

```
n <- 500
```

Initialisation de la matrice des données simulées : n = 100 individus repérés par p = 2 deux variables en colonne –> c'est l'usage en Statistique

```
X <- matrix(0,nrow= n,ncol=2)

eps1 <- rnorm(n,mean=0,sd=1)
```

Bruit blanc N(0,1) de taille n (composantes sur l'axe x1)

```
X[,1] <- 1 + eps1
eps2 <- rnorm(n,mean=0,sd=1)
```

Bruit blanc N(0,1) de taille n (composantes sur l'axe x2)

```
X[,2] <- 2 + 0.8*eps1 + 0.6*eps2

x1 <- 1

h <- 0.1

indices <- which(X[,1]>=x1-h & X[,1]<=x1+h)

X2 <- X[indices, 2]
```

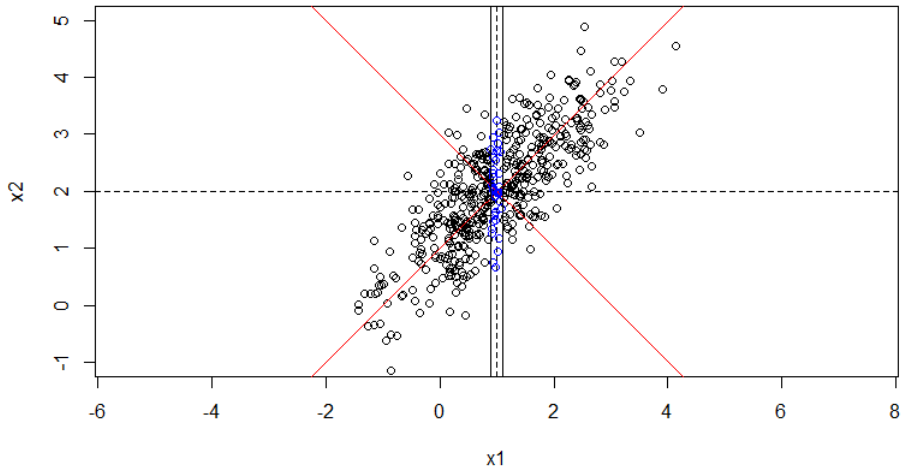

FIGURE 1 – distribution empirique des valeurs associées de la variable X2

```
plot(X[,1],X[,2],asp=1,xlim=c(1-3,1+3),ylim=c(2-3,2+3),
xlab="x1",ylab="x2")
abline(v=1,lty=2)
abline(h=2,lty=2)
abline(a=1,b=1,col="red")
abline(a=3,b=-1,col="red")
points(X[indices, 1], X[indices, 2], col="blue")
abline(v = c(x1-h, x1+h))
```

Afin d'étudier la distribution de X2 autour du point X1=x1 on va commencer par un test de normalité bien évidemment :

```
qqnorm(X2,main="Q-Q Plot x2")
qqline(X2,probs = c(0.1, 0.9),col="red")
```

On peut conclure depuis la figure 2 qu'il ne s'agit pas de loi normale car la distribution échappe de la droite indicatrice surtout aux éxtrémités.

On va maintenant calculer l'esperance et la variance de la loi de X2 sachant X1=x1 :

```
> mu <- mean(X2)
> print(mu)
[1] 2.0233
> vari <- var(X2)
> print(vari)
[1] 0.3930442
```

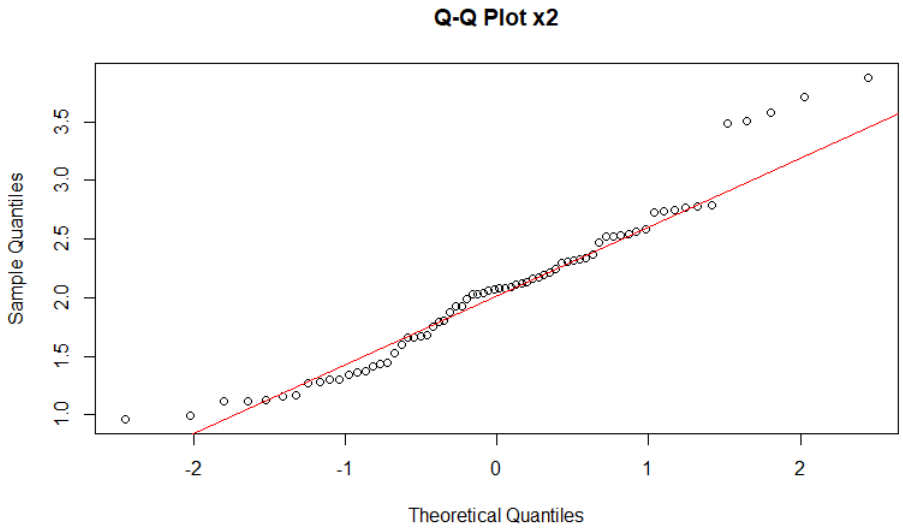

FIGURE 2 – test de normalité la distribution de X2 autour du point X1=x1

On remarque que l'ésperance de cette dernière est égale à l'ésperance de X2, or son écart type est bien plus petit que sigma2=1.

On va essayer d'augmenter la taille de l'échantillon n pour voir l'effet sur la distribution étudiée

```
n <- 5000
```

Dans la figure 3 on voit que la distribution ressemble plutôt bien à une distribution gaussienne, ainsi on pourrait conclure qu'elle est de loi gaussienne d'ésperance 2 et de variance 0.35 "valeurs approximatives"

1.2 Choisir d'autres valeurs de x1. Quelle est la nature de la fonction x1 -> E(X2 | X1 = x1) ? Que peut-on dire de Var(X2 | X1 = x1) ? Et de la loi conditionnelle de X2 sachant X1 = x1 ?

Pour répondre à cette question, nous avons décidé d'étudier la loi de X2 sachant X1=x1 pour chaque valeur x1 dans l'échantillon X1. pour faire ceci on a implementé une boucle sur la taille de l'échantillon X1 qui calcule et l'esperance et la variance de X2/X1=x1 pour chaque valeur de x1. Une dernière étape est de dessiner les deux courbes de l'ésperance et la variance en fonction de x1.

```
N <- length(X[,1])
```

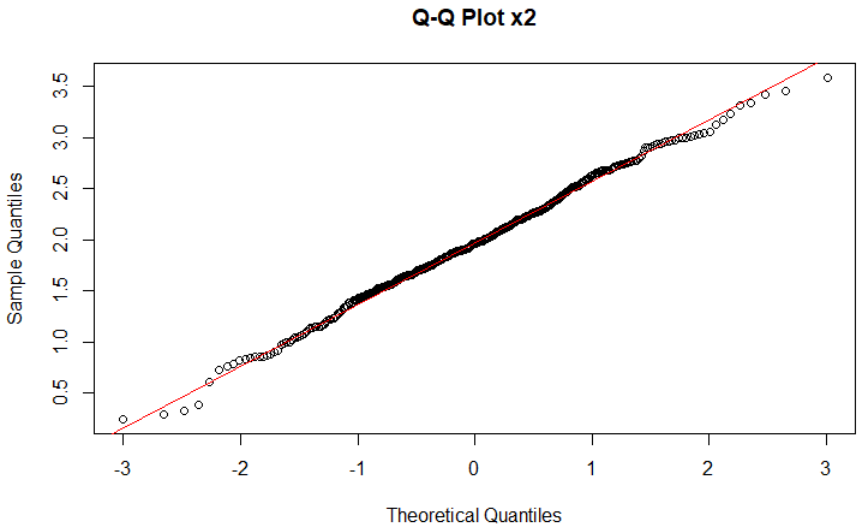

FIGURE 3 – test de normalité la distribution de X2 autour du point X1=x1, n=5000

```
esperance <- rep(0,N)
variance <- rep(0,N)

for (i in 1:N) {
  x1 <- X[i,1]

  h <- 0.1

  indices <- which(X[,1]>=x1-h & X[,1]<=x1+h)

  X2 <- X[indices, 2]

  esperance[i] <- mean(X2)

  variance[i] <- var(X2)
}

graphes <- data.frame(X[,1],esperance,variance)

p = ggplot() +
  geom_line(data = graphes, aes(x =X[,1], y =esperance ,
  color = "Esperance"))+
  geom_line(data = graphes, aes(x =X[,1], y =variance ,
```

```
  color = "Variance"))+
  scale_x_continuous(labels = scales::scientific)+
  scale_y_continuous(labels = scales::scientific)
plot(p)
```

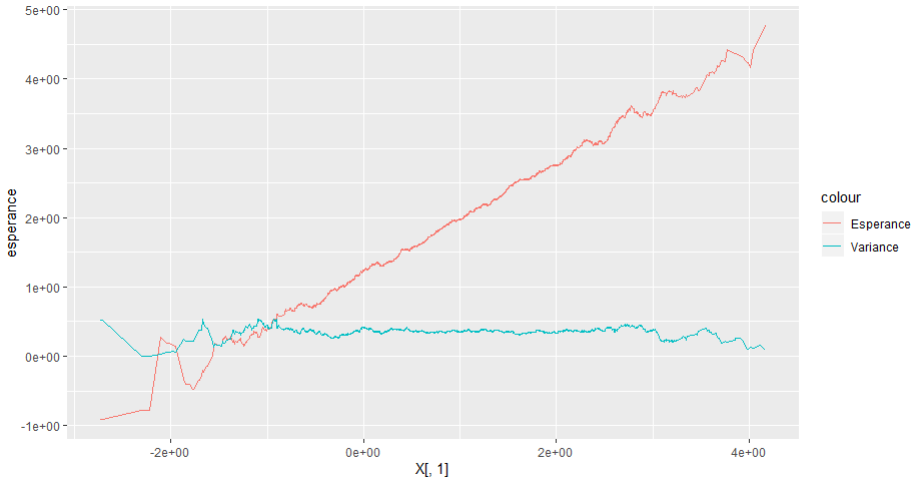

FIGURE 4 – Esperance et variance de la loi de X2/X1=x1 en fonction de x1

D'après la figure 4, si on néglige les effets de bords la nature de la fonction x1 -> E(X2 | X1 = x1) est 'croissante', ce qui est tout à fait logique car la distribution de X2 en fonction de X1 est orientée vers la première bissectrice. Et celle de le fonction Var(X2 | X1 = x1) est 'constante' autour de 0.36 un peu près.

Des tests de normalité sur des valeurs de x1 intermédiaires ont été réalisés et ont tous donné un résultat positif avec un nombre d'individus -n- important. Ainsi on pourrait dire que la loi de X2/X1=x1 est une loi normale de variance 0.36 et d'ésperance croissante en fonction de x1.

2 Exercice 3 : Un jeu de données financières

On met à disposition les données de marché du 24/09/2018 au 24/09/2019 de quarante actions cotées en continu sur le premier marché et qui constituent l'indice CAC40.

Les sociétés correspondantes, représentatives de différentes branches d'activités, reflètent la tendance globale de l'économie des grandes entreprises françaises et leur liste est revue régulièrement pour maintenir cette représentativité.

On s'intéressera au prix St à la clôture journalière (jour t) de ces actions, voir le code R fourni pour extraire par exemple l'évolution de l'action RENAULT.

Pour analyser l'évolution du prix St au jour le jour, on considère la série

des taux de hausse ou de baisse logarithmiques (ou taux de rendement logarithmiques) :

Rt = log(St/St - 1) = (St - St - 1/)St

L'hypothèse classique est de considérer que les variables R1, ..., Rn sont indépendantes (attendre le cours de « Séries Temporelles » pour une analyse de cette hypothèse).

Comme l'historique de marché est de 1 an, on supposera également que ces taux sont de même loi, en particulier de même volatilité (= écart-type).

2.1 En sélectionnant par exemple les deux actions SOCIETE GENERALE et BNP PARIBAS, tester empiriquement si la loi du vecteur bidimensionnel des taux de hausse ou de baisse de ces deux actions est un vecteur gaussien.

Une première étape serait d'importer les données des cours de la SOGE et de la BNP afin de pouvoir calculer les taux de hausse correspondants. Ensuite le graphe figure 5 montre les tracés de ces taux en fonction du temps.

```
SOGE_data <- read.csv("data CAC40/SOCIETEGENERALE_2019-
09-24.txt", header=T, sep="", dec=".")
SOGE <- SOGE_data$clot

BNP_data <- read.csv("data CAC40/BNPPARIBAS_2019-09-24.
txt", header=T, sep="", dec=".")
BNP <- BNP_data$clot

tx_SOGE <- diff(log(SOGE))
tx_BNP <- diff(log(BNP))
```

On voit deja que l'évolution des taux de hausse des deux banques est quasi la même (les courbes sont superposées).

Maintenant il faut tester si la loi du vecteur composé des deux variables décrivants les taux de hausse de la SOGE et de la BNP a une loi gaussienne ou non. Une première étape serait de visualiser l'éspace $[tx - SOGE, tx - BNP]$. Figure 6

```
plot(tx_SOGE,tx_BNP)
```

Depuis la Figure 6 le nuage de points dessiné ressemble bien à une cloche bi-gaussienne (distribution d'un vecteur gaussien bidimentionnel). Deja le test de normalité de chacune des composantes séparées montre qu'elles peuvent être qualifiées comme gaussiennes.

Or cela n'est pas suffisant pour conclure que c'est bien un vecteur gaussien.

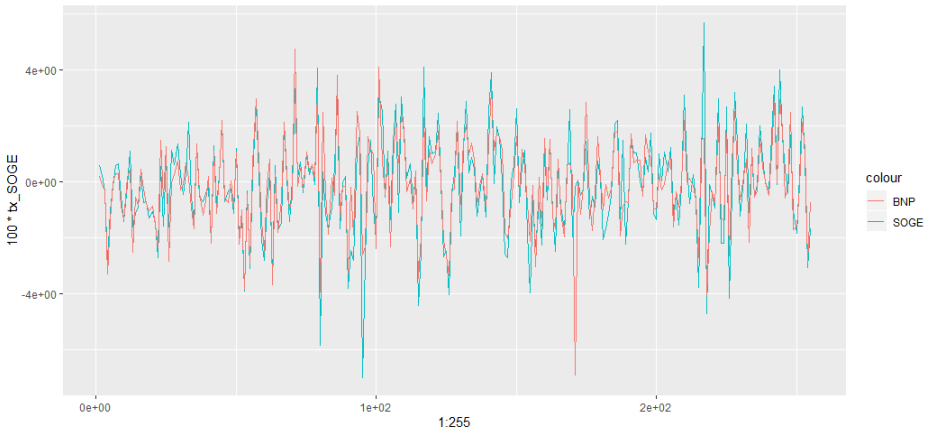

FIGURE 5 – Les taux de hausse des sociétés SOGE et BNP Paribas en fct du temps

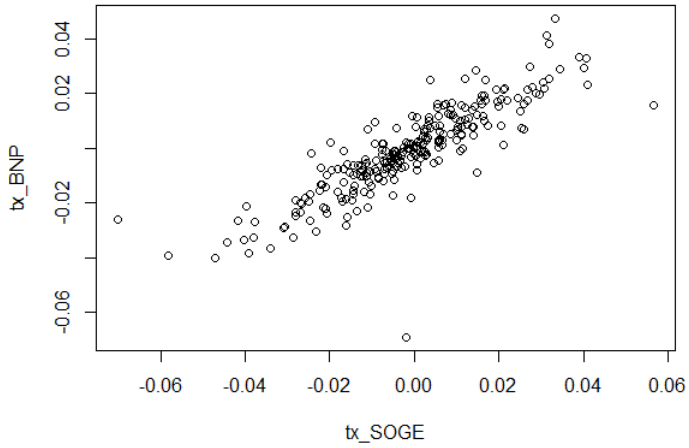

FIGURE 6 – L'éspace $[tx - SOGE, tx - BNP]$

Pour faire le test de normalité du vecteur qu'on veut étudier bous avons fait des recherches pour des méthodes éxistants sur R pour le réaliser (équivalent fonction qqnorm dans le cas unidimentionnel)

Nos recherches ont abouti à une bibliothèque de R peu connue mais très utile pour notre fin. C'est le package "RVAideMemoire" [1] qui contient deux fonctions qu'on va utiliser par la suite : mqqnorm et mshapiro.test

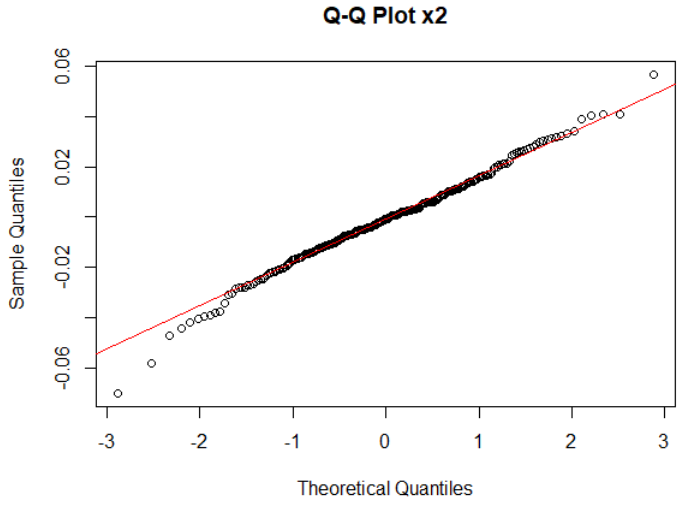

FIGURE 7 – Test de normalité de tx-SOGE

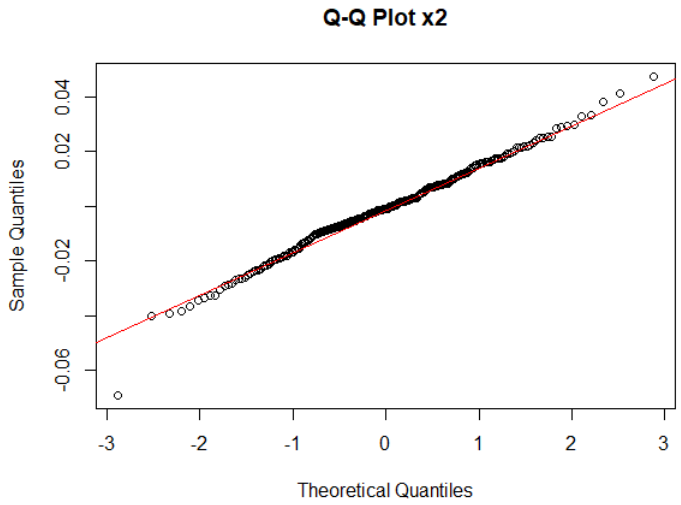

FIGURE 8 – Test de normalité de tx-BNP

```
install.packages("RVAideMemoire")
library(RVAideMemoire)
```

Dans ce qui suit un exemple d'utilisation de la fonction mshapiro.test

```
> x <- rnorm(30)
> y <- rnorm(30)
> mshapiro.test(cbind(x,y))

        Multivariate Shapiro-Wilk normality test

data:  (x,y)
W = 0.98757, p-value = 0.9726
```

Donc pour un vecteur gaussien le test de Shapiro-Wilk donne une p-value assez élevée, ce qui est confirmé dans [2].

Maintenant on va appliquer le test de Shapiro-Wilk sur notre vecteur.

```
> mshapiro.test(V)

        Multivariate Shapiro-Wilk normality test

data:  (tx_SOGE,tx_BNP)
W = 0.85134, p-value = 6.15e-15
```

La p-value qu'on a obtenu est assez proche de 0 ce qui permet de dire que le vecteur n'est **pas** un vecteur gaussien.

Confirmant cela par le test de normalité mqqnorm :

```
mqqnorm(V,main = "Multi-normal Q-Q Plot")
```

Dans la figure 9 On remarque que les quantiles de notre vecteur sortent très tôt du cadre des quantiles du vecteur gaussien de référence et s'éloignent considérablement. Ainsi on peut conclure avec une bonne approximation que le vecteur des taux des de hausse logarithmiques des société SOGE et BNP Paribas n'est pas un vecteur gaussien.

2.2 Tracer sur un même graphique l'évolution du cours St de ces deux actions et interpréter la corrélation entre les deux courbes en liaison avec la question précédente

La tracé des deux courbes est le suivant : Figure 10

```
graphes <- data.frame(1:256,SOGE,BNP,RENAULT,DASSAULT,
LOREAL,VEOLIA,VUITTON)

p = ggplot() +
  geom_line(data = graphes, aes(x =1:256, y =SOGE ,
  color = "SOGE"))+
  geom_line(data = graphes, aes(x =1:256, y =BNP ,
  color = "BNP"))+
```

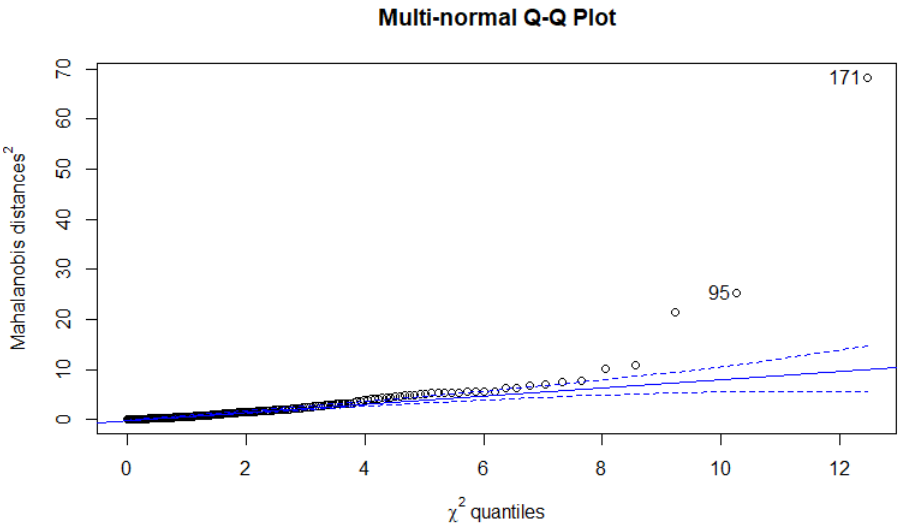

FIGURE 9 – Test de normalité du vecteur

```
  scale_x_continuous(labels = scales::scientific)+
  scale_y_continuous(labels = scales::scientific)
plot(p)
```

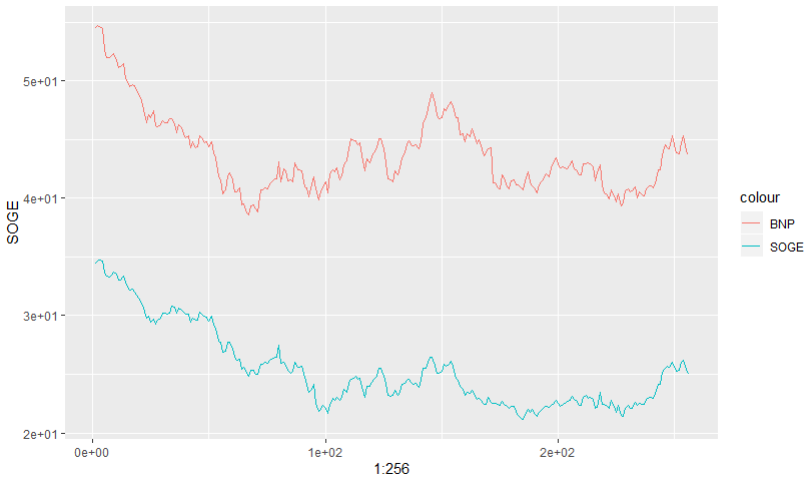

FIGURE 10 – Evolution des cours de la SOGE et la BNP Paribas

On peut voir depuis la figure 10 que les deux cours sont extremement corrélés.

Afin d'expliquer cela on va revenir au vecteur de l'évolution des taux de hausse logarithmiques étudié à la question précédente. Calculons sa matrice de covariance :

```
> mu <- c(mu_SOGE,mu_BNP)
> Gamma <- cbind(c(var_SOGE, covariance),c(covariance,var_
BNP))
> mu_SOGE <- mean(tx_SOGE)
> mu_BNP <- mean(tx_BNP)
> var_SOGE <- var(tx_SOGE)
> var_BNP <- var(tx_BNP)
> covariance <- cov(tx_BNP,tx_SOGE)
> mu <- c(mu_SOGE,mu_BNP)
> Gamma <- cbind(c(var_SOGE, covariance),c(covariance,var_
BNP))
> print(Gamma)
              [,1]          [,2]
[1,] 0.0003289405 0.0002491413
[2,] 0.0002491413 0.0002558383
> print(det(Gamma))
[1] 2.20842e-08
> correlation <- cor(tx_BNP,tx_SOGE)
> print(correlation)
[1] 0.8588241
```

Comme le détérminant de Gamma est presque nul (1E-08), la corrélation entre les taux de hausse des deux sociétés est assez élevée (mise en évidence par le calcul de corrélation : 0.85).

Cette corrélation entre les taux de hausse logarithmiques s'impose pour l'évolution des cours aussi ce qui est tout à fait logique. Et c'est ce qui explique le graphe Figure 10.

Qualitativement il ne faut pas oublier que ces deux sociétés font partie du même secteur (secteur bancaire) ce qui met en évidence cette dépendance entre les deux.

2.3 Choisir un panier d'actions qui vous intéresse formant ainsi un vecteur d-dimensionnel de taux de hausse ou de baisse (d <= 40). Faire l'analyse du vecteur correspondant et en dégager un certain nombre de conclusions qui peuvent être très utiles pour des investisseurs rationnels qui cherchent à limiter leur risque de perte par le biais de la diversification

Avant de choisir notre panier on a décidé de dessiner l'évolution du cours de plusieurs sociétés afin de faire le choix le plus diversifié possible dés le premier

coup :

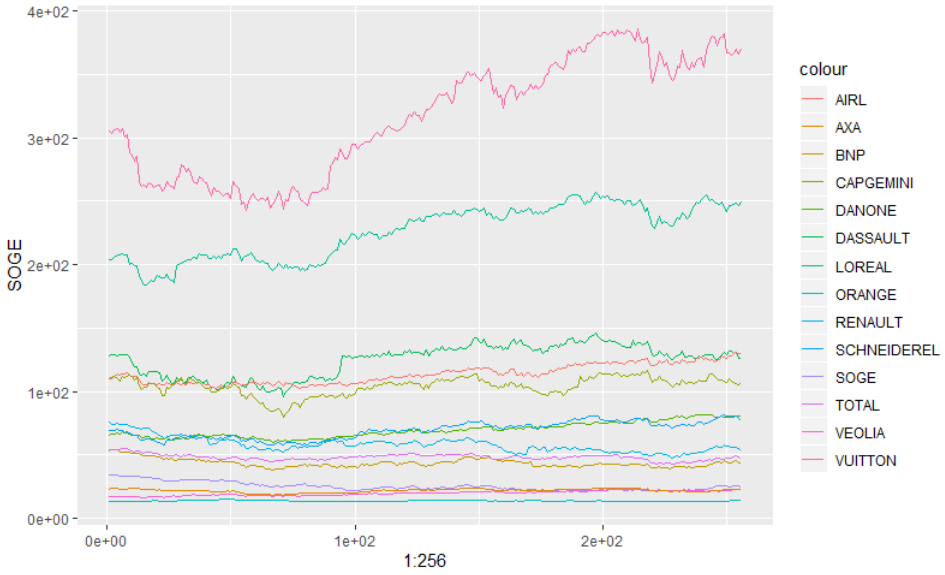

FIGURE 11 – Evolution des cours de plusieurs sociétés

Maintenant après analyse de ses cours choisis au hasard (en essayant de couvrire des secteurs différents) nous avons fait notre choix sur des cours qui nous paraissaient les plus indépendants possible

```
Vchoisi <- cbind(tx_VUITTON,tx_LOREAL,tx_RENAULT,tx_DANONE
,tx_TOTAL,tx_SOGE,tx_VEOLIA,tx_SCHNEIDEREL)

graphes <- data.frame(1:256,SOGE,RENAULT,LOREAL,VEOLIA,
VUITTON,DANONE,SCHNEIDEREL,TOTAL)

p = ggplot() +
  geom_line(data = graphes, aes(x =1:256, y =SOGE ,
  color = "SOGE"))+
  geom_line(data = graphes, aes(x =1:256, y =RENAULT ,
  color = "RENAULT"))+
  geom_line(data = graphes, aes(x =1:256, y =LOREAL ,
  color = "LOREAL"))+
  geom_line(data = graphes, aes(x =1:256, y =VEOLIA ,
  color = "VEOLIA"))+
  geom_line(data = graphes, aes(x =1:256, y =VUITTON ,
  color = "VUITTON"))+
  geom_line(data = graphes, aes(x =1:256, y =DANONE ,
```

```
  color = "DANONE"))+
  geom_line(data = graphes, aes(x =1:256, y =SCHNEIDEREL ,
  color = "SCHNEIDEREL"))+
  geom_line(data = graphes, aes(x =1:256, y =TOTAL ,
  color = "TOTAL"))+
  scale_x_continuous(labels = scales::scientific)+
  scale_y_continuous(labels = scales::scientific)
  #coord_cartesian(ylim =c(0, 100))
plot(p)
```

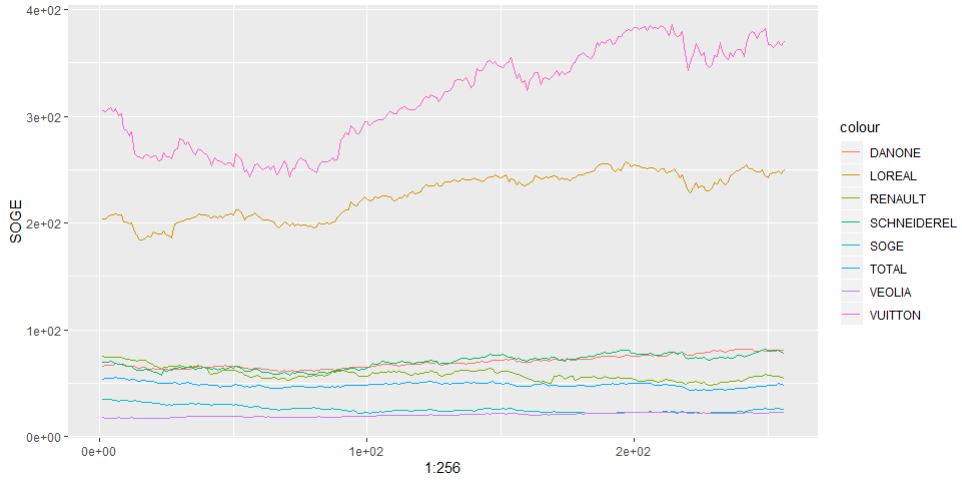

FIGURE 12 – Le vecteur choisi

Effectuons encore le test de normalité sur ce vecteur : voir Figure 13

```
> mshapiro.test(Vchoisi)

        Multivariate Shapiro-Wilk normality test

data:  (tx_VUITTON,tx_LOREAL,tx_RENAULT,tx_DANONE,tx_TOTAL,tx_SOGE,tx_VEOLIA,tx_S
W = 0.85395, p-value = 8.494e-15
```

Les deux tests montrent bien que ce n'est pas un vecteur gaussien, mais pour pouvoir faire l'analyse qui vient on va supposer que c'est un vecteur gaussien

La prochaine étape consiste à Calculer sa matrice de covariance Gamma, et faire la décomposition de Mahalanobis (comme en exercice 1) et en tirer les composantes principales. C'est le but des lignes de code suivantes :

```
> M <- cov(Vchoisi)
>  print(det(M))
```

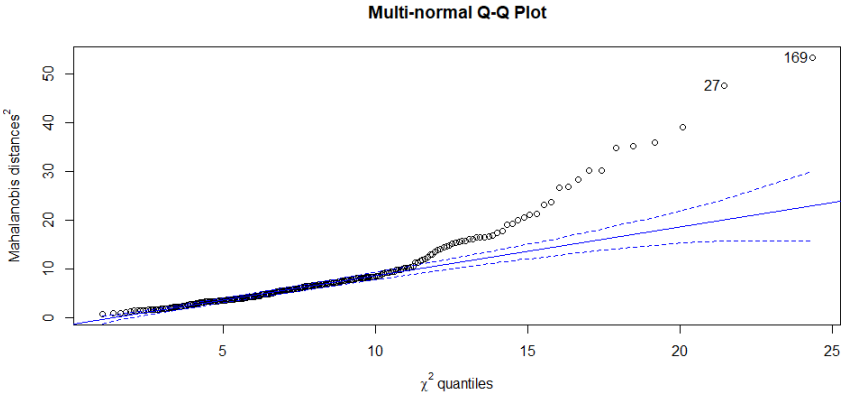

FIGURE 13 – fonction mqqnorm sur le vecteur obtenu

```
[1] 1.217829e-31
> decomp <- eigen(M, symmetric=TRUE)
> values <- decomp$values
> vectors <- decomp$vectors
> Vf <- matrix(0,nrow= 255,ncol=8)
> Vf[,1] <- Vchoisi[,1] - mean(Vchoisi[,1])
> Vf[,2] <- Vchoisi[,2] - mean(Vchoisi[,2])
> Vf[,3] <- Vchoisi[,3] - mean(Vchoisi[,3])
> Vf[,4] <- Vchoisi[,4] - mean(Vchoisi[,4])
> Vf[,5] <- Vchoisi[,5] - mean(Vchoisi[,5])
> Vf[,6] <- Vchoisi[,6] - mean(Vchoisi[,6])
> Vf[,7] <- Vchoisi[,7] - mean(Vchoisi[,7])
> Vf[,8] <- Vchoisi[,8] - mean(Vchoisi[,8])
> # Calcul des nouvelles coordonn es
> C1 <- Vf%*%vectors[,1]
> C2 <- Vf%*%vectors[,2]
> C3 <- Vf%*%vectors[,3]
> C4 <- Vf%*%vectors[,4]
> C5 <- Vf%*%vectors[,5]
> C6 <- Vf%*%vectors[,6]
> C7 <- Vf%*%vectors[,7]
> C8 <- Vf%*%vectors[,8]
> Vnew <- cbind(C1,C2,C3,C4,C5,C6,C7,C8)
```

La valeur du détérminant de la Matrice de covariance nous informe deja que notre vecteur est dégénéré. La figure 14 est la visualisation des nouvelles composantes :

Maintenant on va faire une analyse des valeurs singulières (les racines des valeurs propres de la matrice de covariance et au meme temps les écarts types

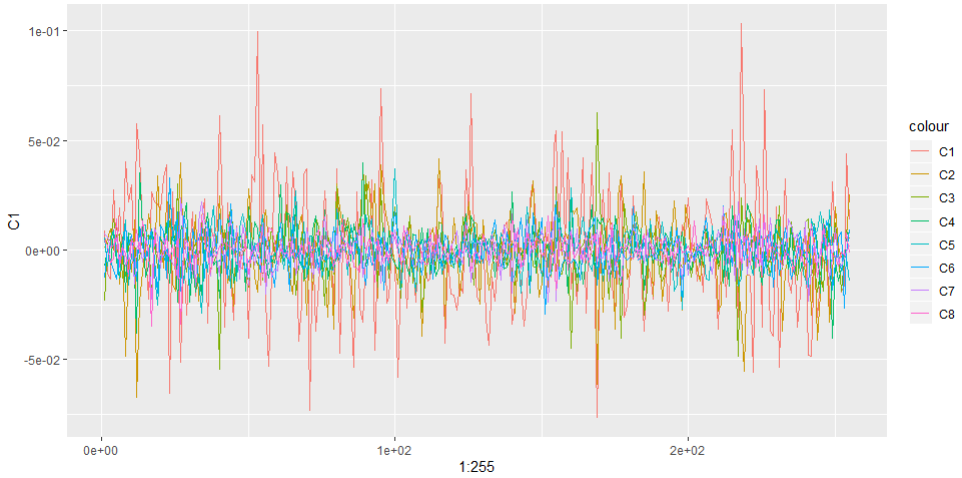

FIGURE 14 – Visualisation des nouvelles composantes principales

des nouvelles composantes) :

```
> sing <- sqrt(values)
> print(sing)
[1] 0.028640201 0.017168483 0.012608680 0.011235881
0.009733706 0.008875961 0.008461410
[8] 0.006852864
```

D'après les valeurs singulières qui représentent le poids de chaque composante principale, on peut dire qu'il n'existe pas de composante ou deux qui dominent toute l'information contenu dans notre vecteur vu que leurs valeurs sont assez proches :

```
> print(summary(sing))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.006853 0.008772 0.010485 0.012947 0.013749 0.028640
```

Ainsi en tenant compte de toutes les approximations qu'on a fait au passage, on peut considérer que le choix de sociétés qu'on a fait est judicieux.

2.4 Conclusion

Bien qu'elle manque de rigueur, nous recommandons aux investisseurs de faire une analyse pareille à celle qu'on vient de réaliser afin de choisir les bons cours pour investir. Merci !

Références

[1] Maxime Hervé. "Testing and Plotting Procedures for Biostatistics". In : *Package 'RVAideMemoire'* 0.9-73 (2018), p. 55-56. DOI : https://cran.r-project.org/web/packages/RVAideMemoire/RVAideMemoire.pdf.

[2] Unknown. *Shapiro–Wilk test*. URL : https://en.wikipedia.org/wiki/Shapiro-Wilk_test. (accessed : 06.10.2019).