Méthodes Statistiques pour la Finance

Indice : Nikkei

3A IF : MeQA

Benjamin Ruimy & Arnaud Ibakuyumcu & Kwan Voong

15/03/2017

Sommaire

Introduction

Ce TP a été effectué en trinôme. L'objectif est de travailler sur les rentabilités d'un indice (en l'occurrence le NIKKEI pour notre TP) sur la période 01/01/1990 - aujourd'hui.
En plus de ce rapport, vous pourrez trouver le code R dans l'archive.
Nous avons fait le choix de ne pas mettre le code directement dans le rapport afin de ne pas l'alourdir et de faire un rapport concis.

Question 1 : Étudiez la distribution empirique des rentabilités journalières, hebdomadaires, mensuelles de votre titre. Interprétez les états.

Question 2 : Étudiez la dépendance linéaire (modèle ARMA) des rentabilités journalières, hebdomadaires. Peut-on tirer profit de cette dépendance ? Comment ? Conséquences en termes d'efficience des marchés.

Question 3 : Étudiez le modèle à chaînes de Markov cachées (HMM) pour les rentabilités journalières de votre indices. Interprétez les états.

Question 4 : Sous des hypothèses standards, l'espérance des rentabilités journalières est-elle la même sur la période [1990-2002] et [2003-2016] ? En est-il de même pour la variance ?

Question 5 : Étudiez la corrélation entre les valeurs absolues de deux rentabilités journalières consécutives, entre leurs carrés. Peut-on profiter de ces dépendances ? Comment ?

Question 6 : Modélisez les rentabilités journalières et hebdomadaires à l'aide de modèle de type ARCH. Valider ces modèles en :

- Testant la signification des paramètres.
- Étudiant la nature du bruit (indépendance, normalité)

Question 7 :Pour le 'meilleur' modèle : ré-estimez les paramètres sur la période [1990, 2014]. A chaque date t $>$ 2014, calculez la variance h(t) prévue par le modèle, puis l'intervalle de confiance à 95% pour la rentabilité à la date t. Faire un graphique et ajoutez les rentabilités réalisées. Conclusions quant à la performance prédicative du modèle.

Question 8 : Pour le 'meilleur' modèle : ré-estimez les paramètres d'une part sur la période [1990, 2002], et d'autre part sur la période [2003, 2016]. Peut-on considérer que les deux jeux de paramètres sont les mêmes ? Conclusion sur la stabilité de ce modèle. Cela vous paraît-il prévisible ?

1 Question 1

Les rentabilités financières sont souvent statistiquement étudiées pour trouver des stratégies de gestion de portefeuille.
La première hypothèse de base est de supposer que les rentabilités suivent une loi normale (théorie de Markovitz). Vérifions cette hypothèse.

Tout d'abord voici les rentabilités au cours du temps pour l'indice `Nikkei` :

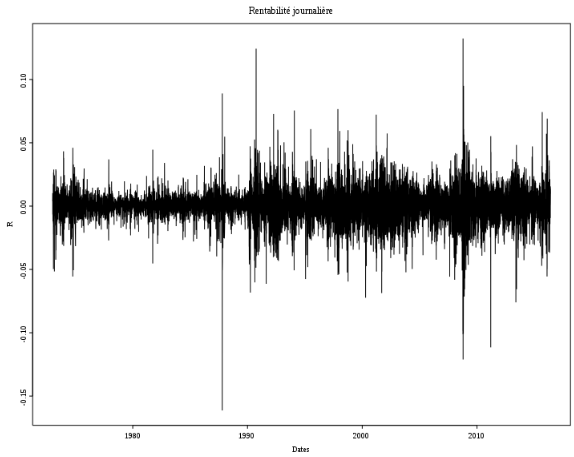

Figure 1 : Rentabilités Journalières

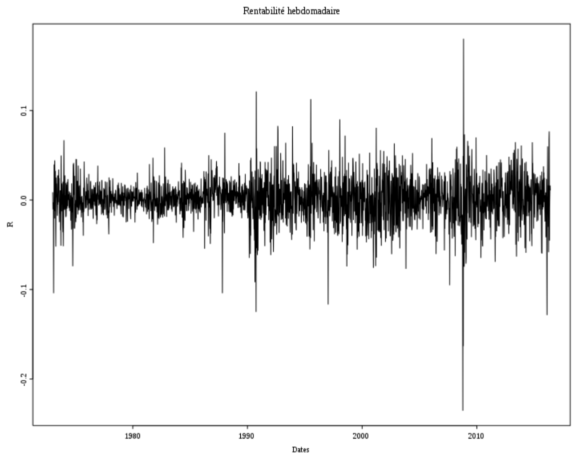

Figure 2 : Rentabilités Hebdomadaires

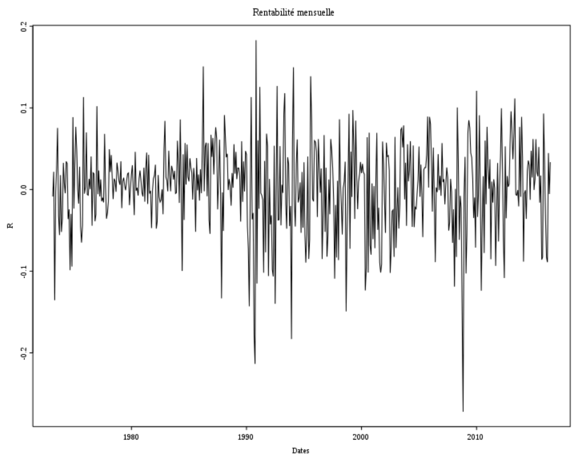

Figure 3 : Rentabilités Mensuelles

Dans ces graphiques, on remarque de nombreux pics ce qui nous donne un premier indice sur le fait que l'hypothèse de loi normale est a réfuter. Étudions la répartition des rentabilités avec en courbe rouge le témoin de la loi normale dans les graphiques suivants.

1.1 Rentabilités journalières

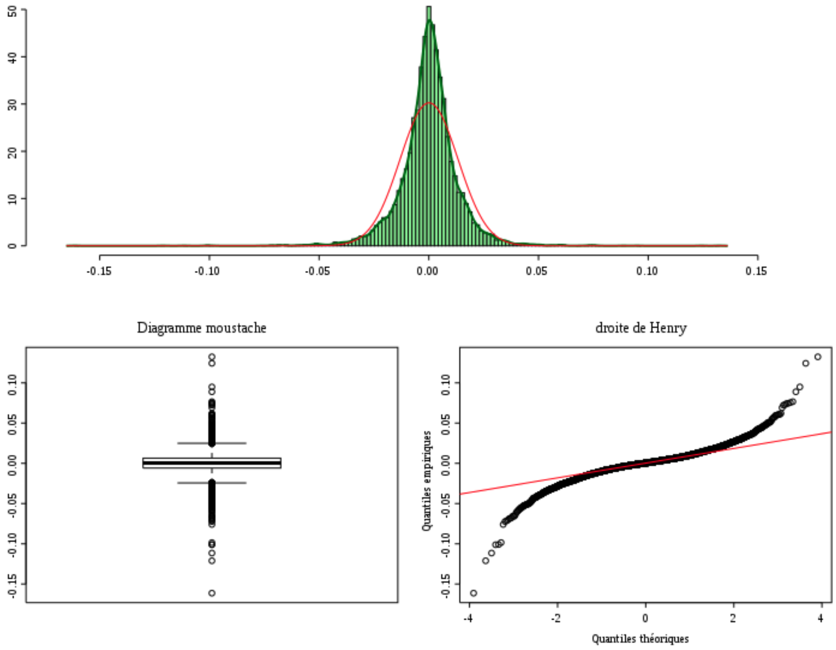

Figure 4 : Répartition Rentabilités Journalières

On peut réfuter l'idée d'une loi normale pour les répartitions journalières car la densité est très leptokurtique, et les quantiles sont très éloignées de la droite d'Henry.
De plus, le coefficient d'asymétrie (`skewness`) est de -0.3138 qui devrait être égale à 0 pour le cas d'une loi normale. La négativité de ce coefficient implique qu'il y a plus de poids sur la partie négative des rentabilités et donc plus de chance d'avoir une rentabilité négative.
Le coefficient d'applatissement (`kurtosys`) est de 12.097 qui est très élevé et spécifie le caractère leptokurtique de la densité. À titre d'information, il aurait du être de 3 pour le cas d'une loi normale.

1.2 Rentabilités hebdomadaires

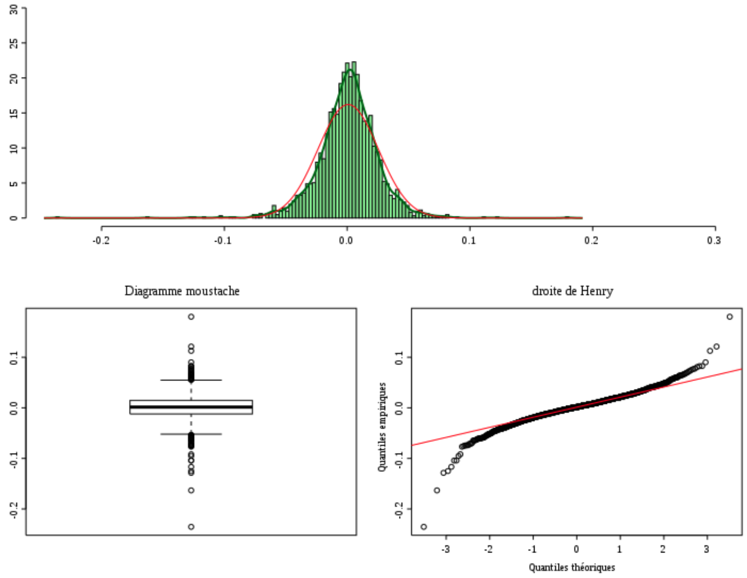

Figure 5 : Répartition Rentabilités Hebdomadaires

De la même façon, l'hypothèse d'une loi normale peut être réfutée puisque la densité est leptokurtique. Le coefficient d'asymétrie ici est de -0.5626 et le coefficient d'applatissement est de 10.428.

1.3 Rentabilités mensuelles

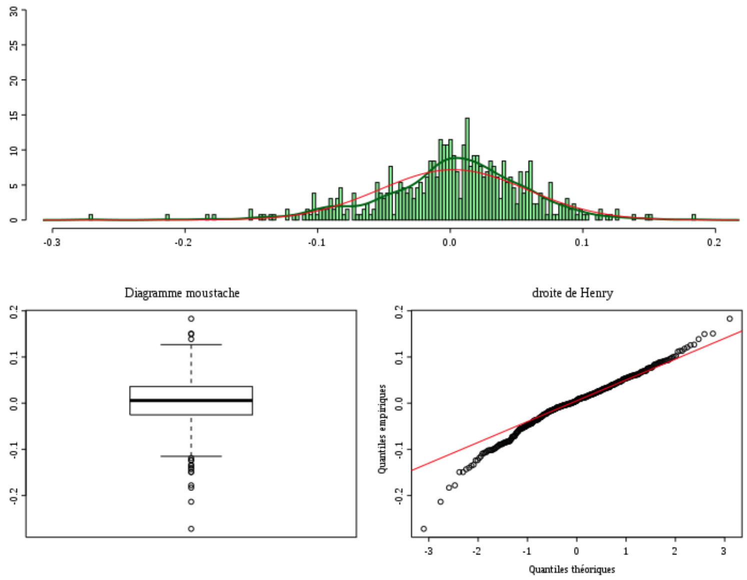

Figure 6 : Répartition Rentabilités Mensuelles

Encore une fois, l'hypothèse d'une loi normale peut être rejeté puisque la densité est très asymétrique à gauche cependant elle est quasiment applati comme la loi normale. Le coefficient d'asymétrie ici est de -0.5861 et le coefficient d'applatissement est de 4.70.
On va chercher à modéliser les rentabilités avec d'autres modèles dans les questions suivantes.

2 Question 2

Pour cette question, il est demandé de modéliser une dépendance linéaire (ARMA) pour les rentabilités.

⋆ La première étape est de trouver un ordre de grandeur pour le nombre de retard (p, q) en AR et MA. Pour cela, on utilise les fonctions d'auto-corrélation `acf` qui nous donne q^* pour un modèle MA et d'auto-corrélation partielle `pacf` qui nous donne (p^*) pour un modèle AR de `R`. Pour un modèle ARMA, on prend en général (p, q) tel que $p+q \leq p^*+q^*$.

⋆ La deuxième étape est d'utiliser la librairie `TSA` pour connaitre les combinaisons de p, q qui peuvent donner un bon modèle ARMA sur les rentabilités étudiées.

⋆ La troisième étape est d'étudier chaque 'bonne' combinaison de modèle ARMA et de choisir le meilleur modèle selon un ou plusieurs critère.
Nous voulions choisir le meilleur modèle en prenant celui qui minimise l'erreur de prédiction (méthode de split-validation) mais la simulation était complexe à gérer et nous avons alors utilisé le critère de minimum d'AIC et chercher à maximiser le rapport de détermination R^2 même s'il n'est pas toujours un bon critère de 'bon modèle'.

⋆ La quatrième étape consiste à étudier le modèle considéré avec le caractère significatif des coefficients estimés ainsi que la répartition des résidus.

2.1 Rentabilités Journalières

Les fonctions de corrélations et d'auto-corrélations partielles nous ont donné les graphiques suivants :

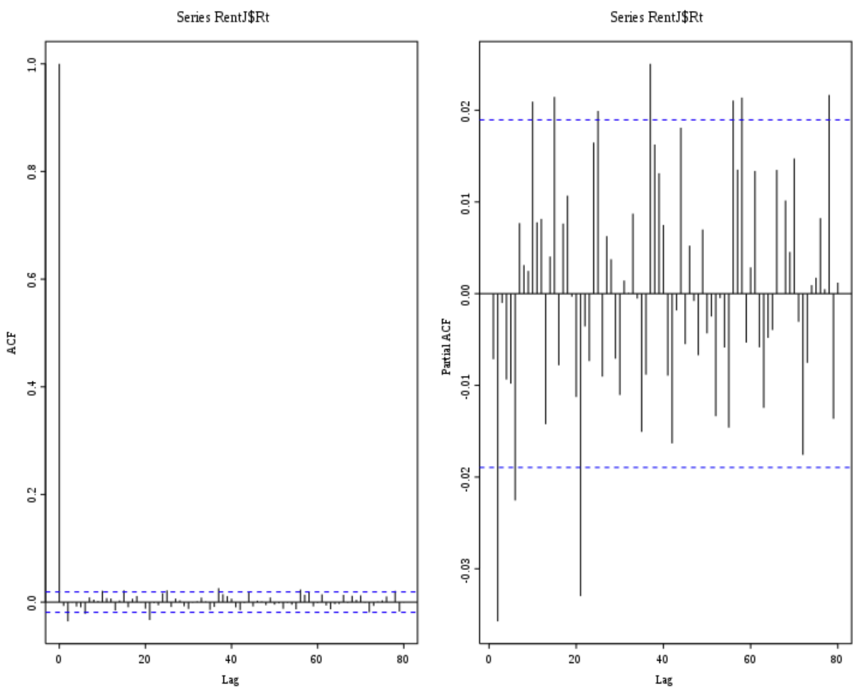

Figure : ACF et PACF des rentabilités journalières.

On remarque alors que $p + q < 6 + 1 = 7$, et la fonction `TSA::eacf` nous donne les combinaisons optimales suivantes (là où il y a les x) :

```
AR/MA
  0 1 2 3 4 5
0 o x o o o x
1 x x o o o x
2 x x o o o x
3 x x x o o x
4 x x x x o x
5 x x x x x x
```

En essayant chacun des couples (p, q), le modèle au meilleur AIC est `ARMA(2,0)` avec un AIC de -62235 et au meilleur R^2 est `ARMA(3,2)`.
En *fittant* les deux modèles précédent, celui ayant le plus de coefficients significatifs est le modèle `ARMA(2,0)` qu'on choisira pour la suite.

Dans ce modèle, le coefficient `AR[2]` est le seul étant très significatif (p-value de 0.000216) et les intervalles de confiances sont de l'ordre de 10^{-2} pour ces paramètres.

Les estimations sont : $\begin{cases} AR[1] & \approx -0.00738 \\ AR[2] & \approx -0.03575 \\ const & \approx 0.0001157 \end{cases}$

L'étude des résidus nous donne l'historique suivant :

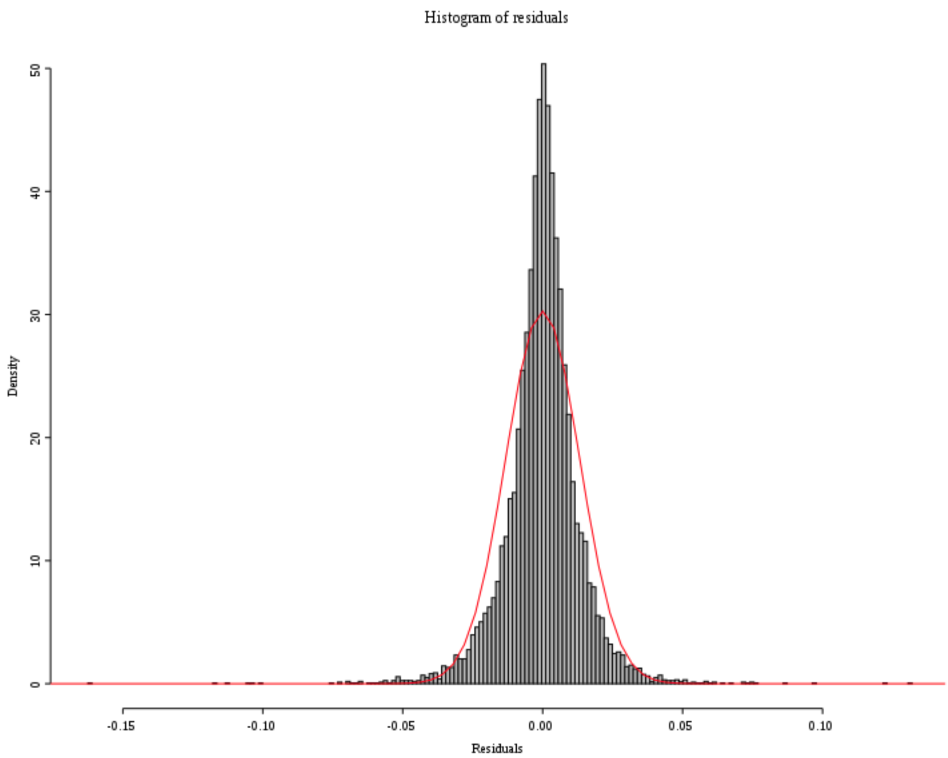

Figure : Histogramme des résidus contre densité de loi Normale.

On remarque que les résidus ne suivent pas du tout une loi normale, leur densité est très leptokurtique.
Ceci se confirme avec la répartition des résidus (voir graphique suivant). Le graphique quantile-quantile avec la droite de Henry montre bel et bien qu'on ne peut pas affirmer que les résidus suivent une loi normale.

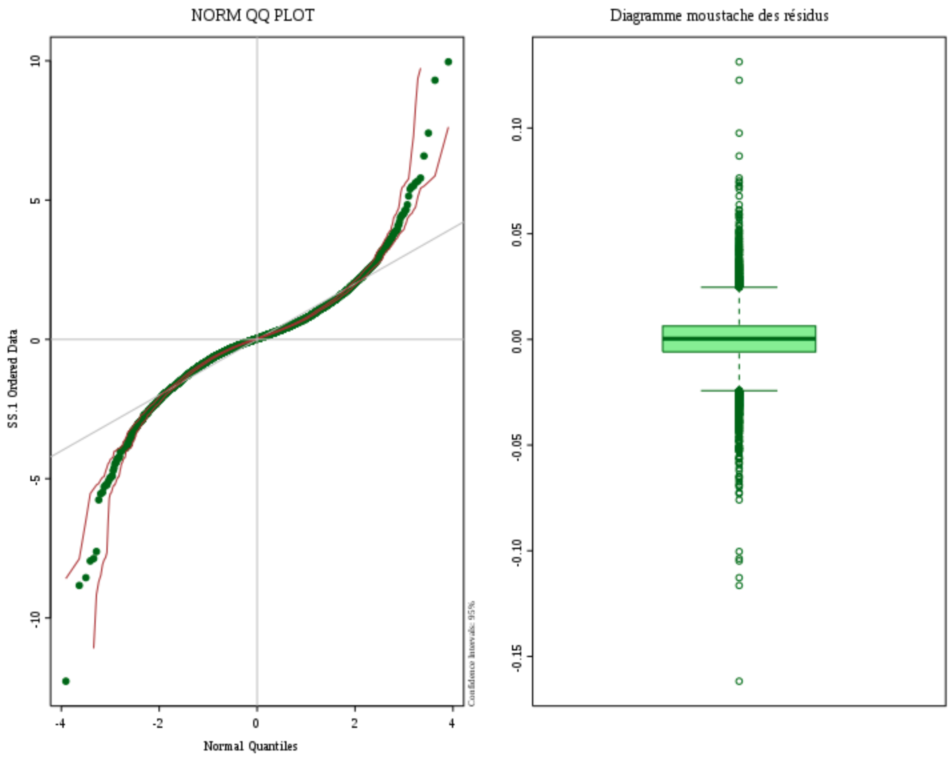

Figure Droite de Henry et Diagramme moustache des résidus.

En plus de ce constat, le coefficient de régression de ce modèle est $R^2 = 0.00116097$ qui est très faible et nous conforte dans l'idée qu'un modèle de dépendance linéaire n'est pas envisageable. D'autant plus que le *sumOfSquare* est de 1.86 qui est très élevé par rapport à la valeur des rentabilités qui est de l'ordre de $5E^{-3}$.

2.2 Rentabilités hebdomadaires

Les fonctions de corrélations et d'auto-corrélations partielles nous ont donné les graphiques suivants :

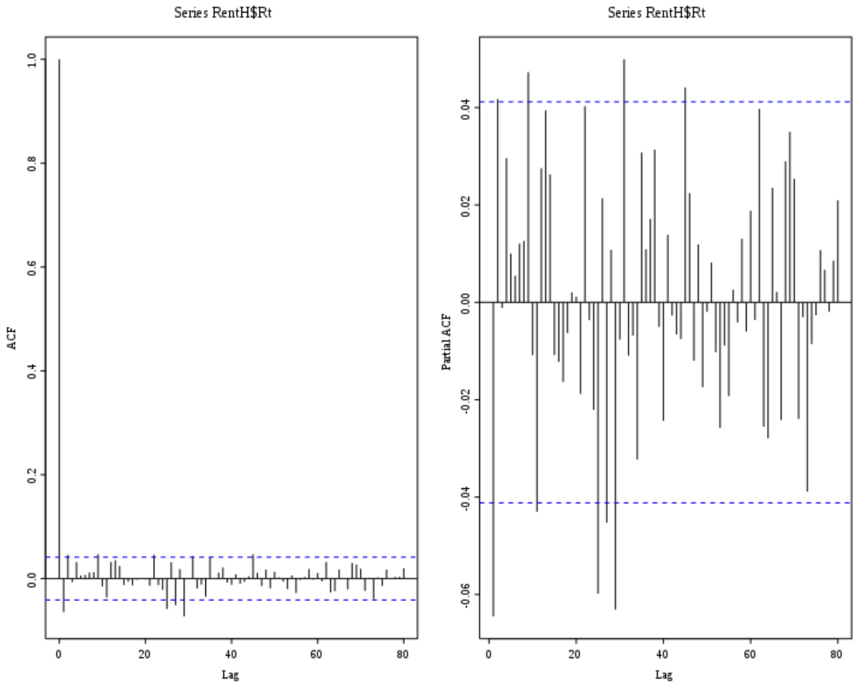

Figure : ACF et PACF des rentabilités hebdomadaires.

On remarque alors que $p+q < 8+1 = 9$, et la fonction `TSA::eacf` nous donne les combinaisons optimales suivantes (là où il y a les x) :

```
AR/MA
  0 1 2 3 4 5
0 x x o o o o
1 x o o o o o
2 o x o o o o
3 o x x o o o
4 x x x x o o
5 x x x x x o
6 x x x o x x
7 x x o o x x
8 x x o x x x
```

En essayant chacun des couples (p, q) parmi ceux qui convergent, le modèle au meilleur AIC et R^2 est `ARMA(2,1)` avec un AIC de -10358.

Dans ce modèle, tous les coefficients sont très significatif (p-value de l'ordre de 10^{-16} pour `AR[1]` et `MA[1]` et de 10^{-4} pour `AR[2]`) par contre les intervalles de confiances sont de l'ordre de 10^{-1} pour ces paramètres.

$$\text{Les estimations sont : } \begin{cases} AR[1] & \approx 0.75908 \\ AR[2] & \approx 0.081026 \\ MA[1] & \approx -0.823965 \\ const & \approx 0.000141 \end{cases}$$

L'étude des résidus nous donne l'historique suivant :

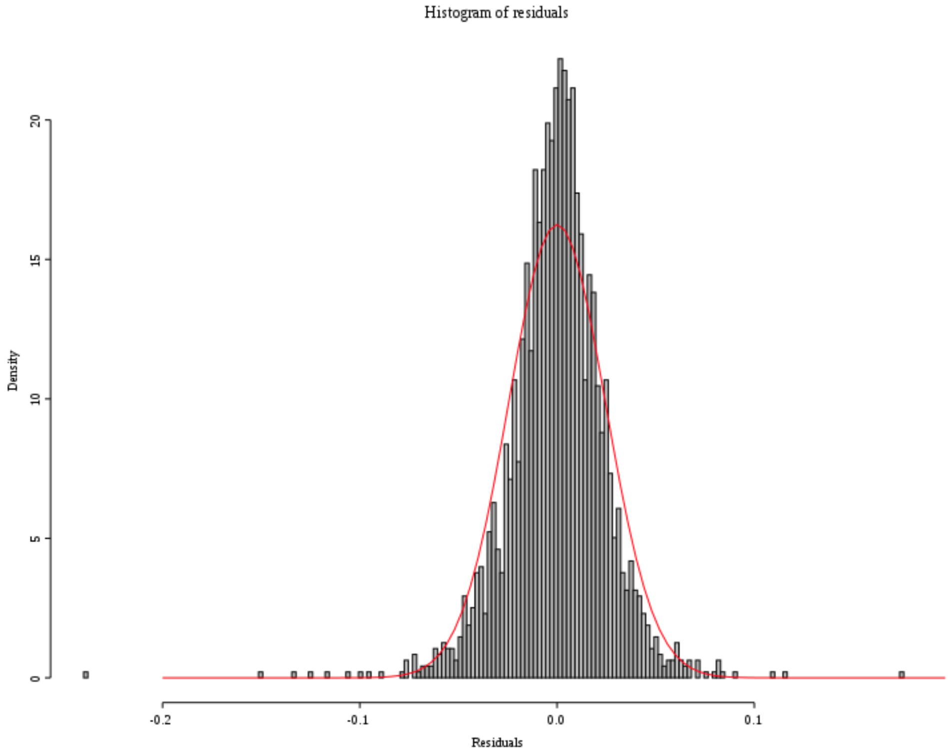

Figure : Histogramme des résidus contre densité de loi Normale.

On remarque que les résidus ne suivent pas du tout une loi normale, leur densité est toujours leptokurtique avec des queues lourdes et une légère asymétrie vers la droite. Ceci se confirme avec la répartition des résidus (voir graphique suivant). Le graphique quantile-quantile avec la droite de Henry montre bel et bien qu'on ne peut pas affirmer que les résidus suivent une loi normale même s'ils sont plus proche qu'avec les rentabilités journalières.

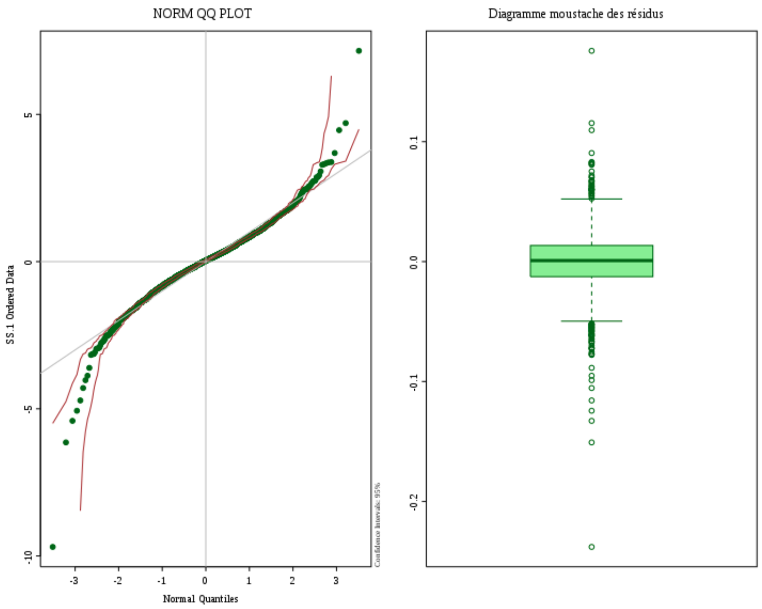

Figure Droite de Henry et Diagramme moustache des résidus.

En plus de ce constat, le coefficient de régression de ce modèle est $R^2 = 0.006$ qui est plus important que celui dse rentabilités journalières mais reste faible et nous conforte toujours dans l'idée qu'un modèle de dépendance linéaire n'est pas envisageable. De plus le *sumOfSquare* est de 1.36 qui est toujours élevé par rapport à la valeur des rentabilités qui est de l'ordre de E^{-2}.

2.3 Conclusion

S'il y avait un modèle de dépendance linéaire entre les rentabilités il serait possible de prédire les rentabilités d'actifs et donc de construire un portefeuille prédisant de manière quasi-certaine un gain.
Cependant, si de tels modèles existaient, ils seraient utilisés par tous et l'équilibre du marché ferait que ces modèles ne conviendraient plus (idem avec l'arbitrage, une opportunité d'arbitrage n'est jamais présente très longtemps car elle est de suite attrapé par des traders ce qui équilibre les prix et fait disparaitre l'opportunité).
On a pu remarquer dans les deux exemples précédents la médiocrité des modèles ARMA estimés.

3 Question 3

Nous devons afin de répondre à cette question, étudiez le modèle à chaînes de Markov cachées. Il convient alors d'utiliser la fonction *HMMFit*, en considérant 3 états (afin de tenter de retrouver l'état optimiste, l'état calme et l'état pessimiste). Puis nous utilisons **l'algorithme de Viterbi** afin de trouver les états cachés optimaux.

Lorsque nous affichons les densités estimées des trois états, nous pouvons nous apercevoir que les états ont des densités qui se rapprochent de la loi normale.

Néanmoins, l'état 1 est celui dont la densité colle le plus à la loi normale. L'état 2 à une densité leptokurtique, et l'état 3 a une densité leptokurtique asymétrique.

L'état 1 est volatile ($\sigma = 0.03$ et n'est pas centré (il a une moyenne négative estimée à -0.2%), il s'agit donc d'un **état de crise**. L'état 2 correspond à l'**état de marché classique**, il est centré en $1.337 * 10^{-4}$, et est assez volatile ($\sigma = 0.012$). L'état 3 quant à lui correspond à un **état de marché calme**, il est asymétrique et possède une queue plus épaisse à droite, ce qui signifie qu'il y a plus de chance d'avoir des rentabilités positives, la densité est centrée en une valeur positive ($7,5 * 10^{-4}$). De plus la volatilité de l'état 3 est faible ($\sigma = 0.005$).

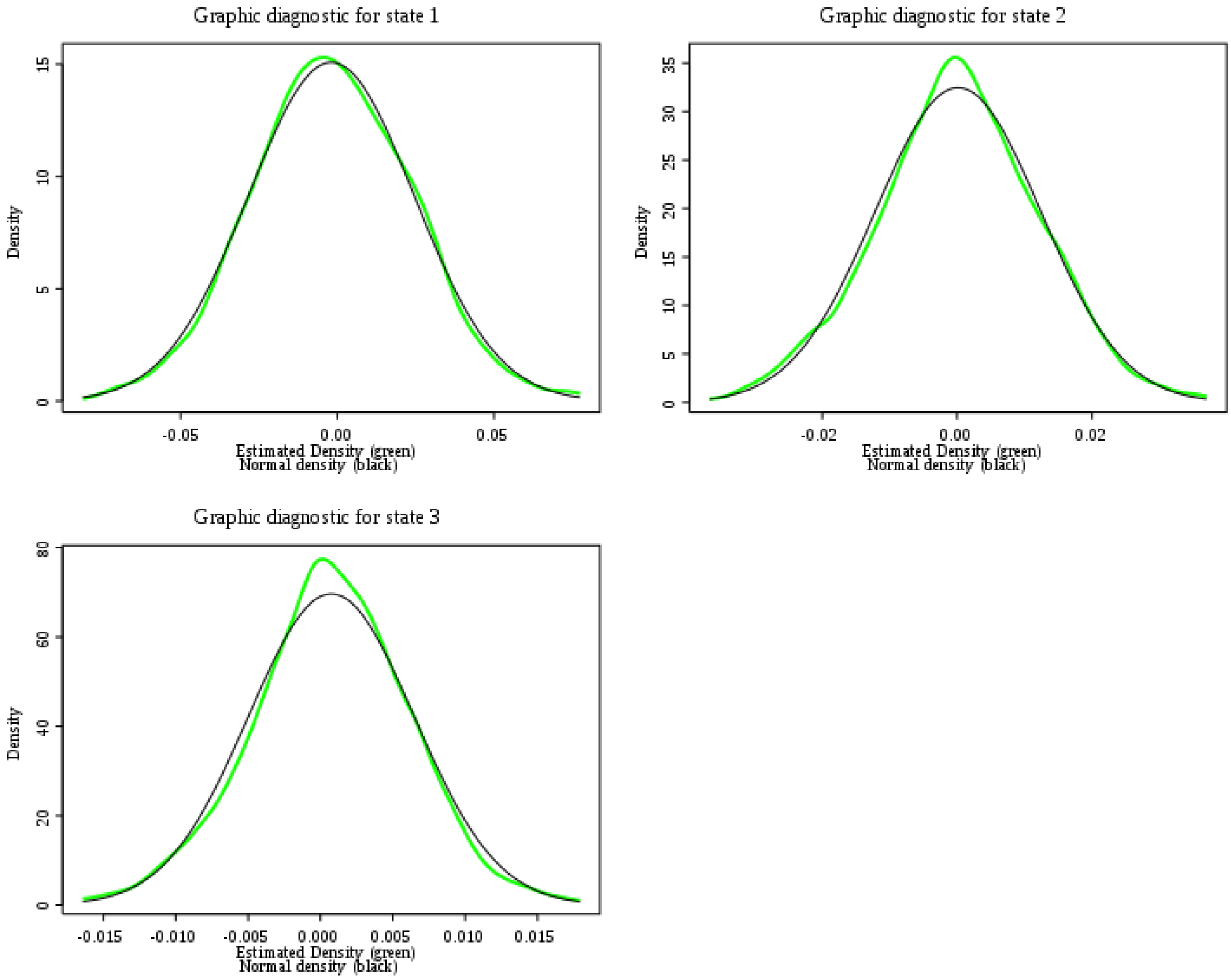

Figure 1: HMM Graphic

Nous pouvons remarquer que sur la période d'étude, il n'y a pas vraiment d'état de marché optimiste.

Nous pouvons analyser les probabilités de se trouver dans les différents états.
Nous obtenons :

Proba état 1	$4,94 \times 10^{-89}$
Proba état 2	$8,64 \times 10^{-47}$
Proba état 3	1.00

Un tracé pertinent est le tracé des rentabilités en fonction du temps, en faisant distinguer clairement l'état dans lequel nous nous trouvons.

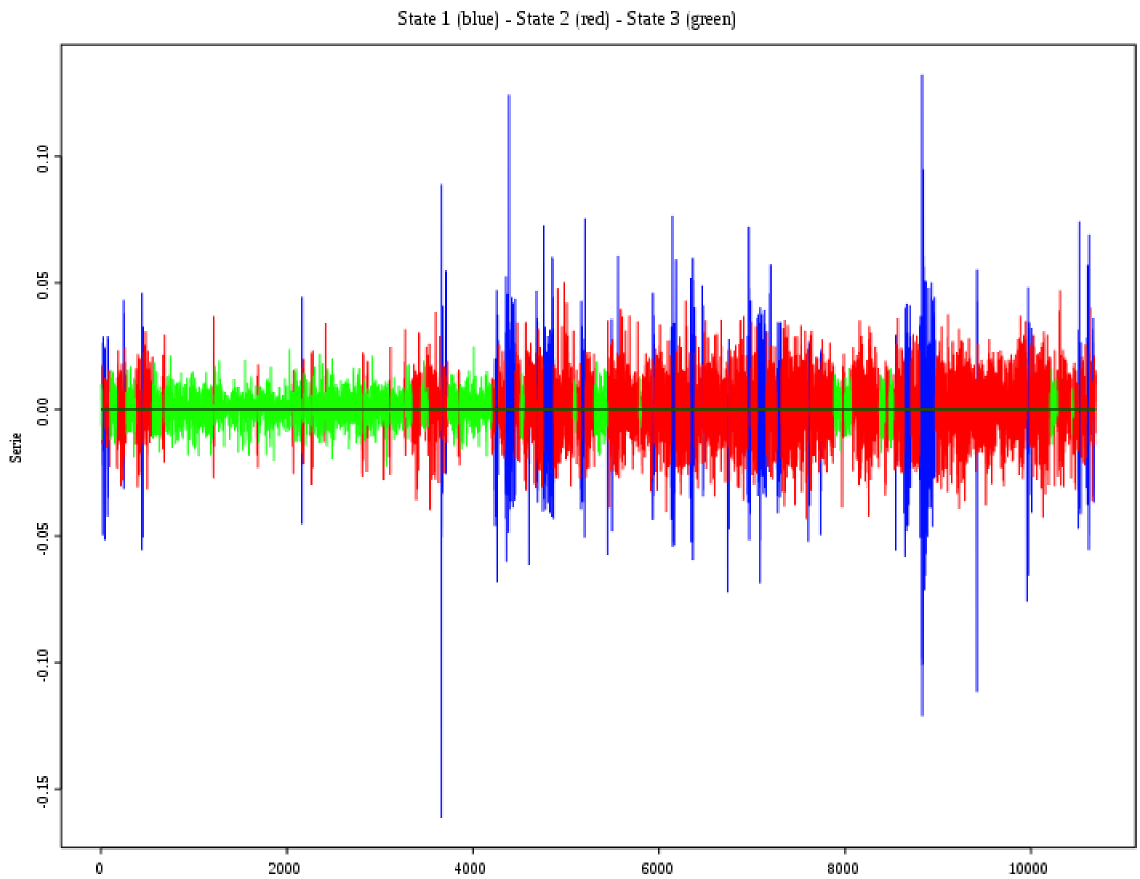

Figure 2: HMM PlotSeries (Rentabilités en Fonction du temps)

Lorsqu'on interprète ce graphe, nous pouvons constater que l'état le plus volatile est l'état 1 (représenté en bleu), d'ailleurs, il semble y avoir plus de rentabilité positive (les pics vers le bas sont plus nombreux et descendent plus bas). Néanmoins, ce état n'est pas très présent.
L'état 2 (représenté en rouge), semble très présent sur la seconde période de temps et très peu sur la première période. On a donc l'impression que les états 3 et 2 sont quasiment equi-répartis sur la période d'étude.
Néanmoins, cette conclusion est en contradiction avec ce que nous avons trouvé lors de la première analyse (qui disait que l'état le plus calme (l'état 3) est celui le plus représenté). En revanche, nous n'avons pas pu obtenir les p-valeurs de ces probas donc nous ne pouvons dire si les valeurs trouvées sont significatives.

4 Question 4

Cette question a pour objectif de comparer l'espérance et la variance des rentabilités journalières sur les périodes [1990-2002] et [2003-2016].

Traçons dans un premier temps l'allure des densités des rentabilités journalières sur les deux périodes.

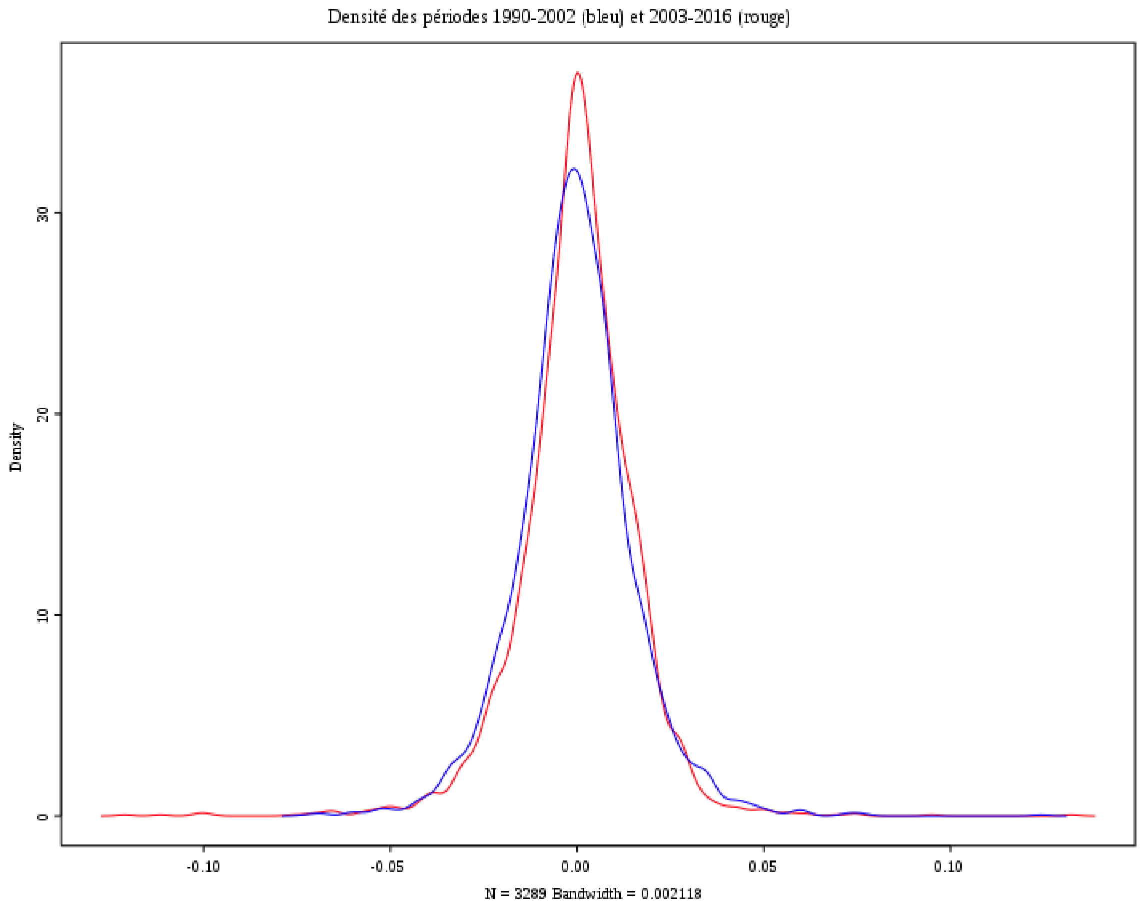

Figure 3: Densité des rentabilités journalières pour les périodes [1990-2002] et [2003-2016]

Nous trouvons les valeurs suivantes :

Moyenne sur la première période	$-4,717911 \times 10^{-4}$
Moyenne sur la seconde période	$2,121099 \times 10^{-4}$
Variation relative entre les périodes de la moyenne	-145%
Variance de la première période	$2,404016 \times 10^{-4}$
Variance de la seconde période	$2,365182 \times 10^{-4}$
Variation relative entre les périodes de la variance	-1.62%

Les variations relatives nous permettent d'avoir une analyse naïve. La première période est une période défavorable (moyenne négative) alors que la seconde période est une période favorable, il semble donc que les moyennes ne soient pas identiques. Les variances semblent être quasiment identiques.

Afin de pouvoir efficacement comparer les moyennes et les variances sur les deux périodes, nous utilisons un t test (test de Student) et un F test (test de Fisher) respectivement.

Le test de Student nous donne une p-valeur de $0.07443 > 0.05$, mais la moyenne des rentas sur la deuxième période ne se trouve pas dans l'intervalle de confiance à 95% (2% à droite de l'intervalle) . Ainsi, nous ne sommes pas vraiment en mesure de rejeter l'hypothèse H0 (le fait que les valeurs des moyennes soient significativement différentes) , même si nous sommes très tentés de le faire. Nous nous trouvons dans une sorte de cas limite (p-valeur légèrement trop élevé et une des valeurs n'est pas dans l'intervalle de confiance à 95% (mais déborde de seulement 2%).

Le test de Fisher quant à lui nous donne une p-valeur de 0.6426 et les variances sur les deux périodes se trouvent largement dans l'intervalle de confiance. Par conséquent on ne peut pas rejeter H0 (le fait que le ratio des variances soit égale à 1), et la position dans l'intervalle de confiance nous dit qu'elle ne sont pas significativement différentes.

5 Question 5

L'objectif est d'étudier la corrélation entre les valeurs absolues et le carré de deux rentabilités journalières successives.

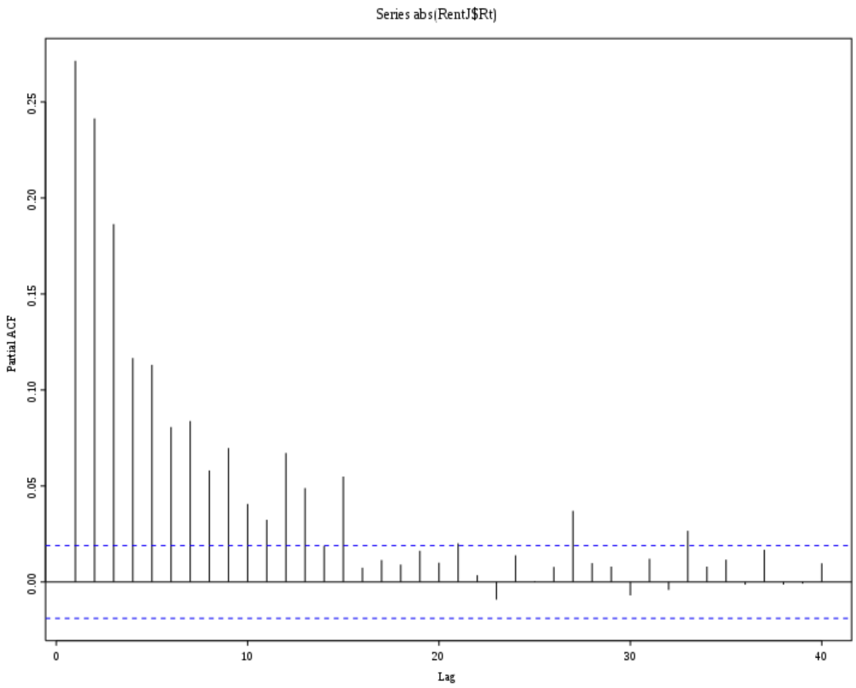

Figure 4: PACF sur la valeur absolue des rentas

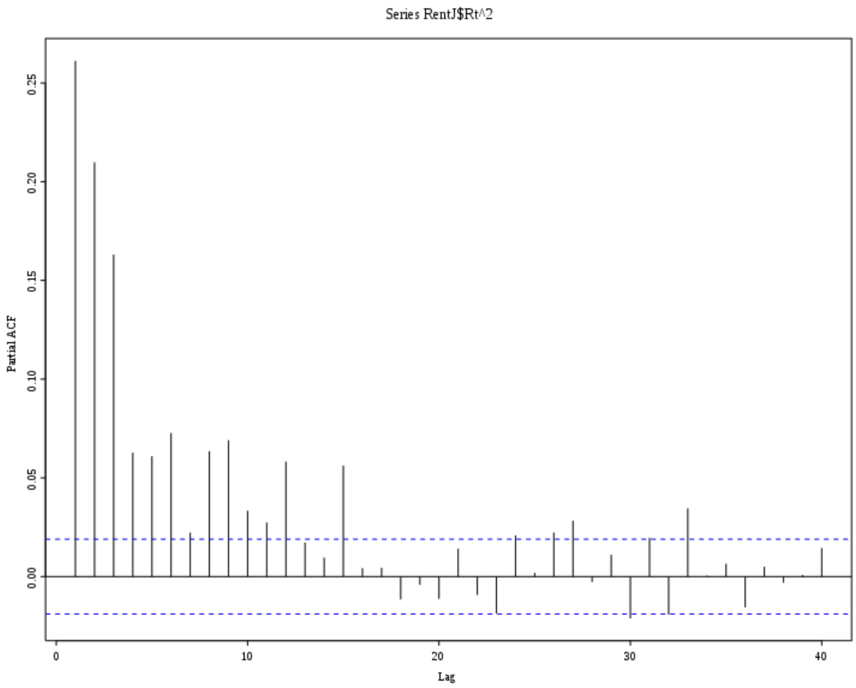

Figure 5: PACF sur le carré des rentas

Nous pouvons accéder aux valeurs des auto-corrélations partielles avec un lag1 facilement, néanmoins, si on le fait manuellement, nous obtenons des valeurs plus précises.

Nous obtenons une corrélation sur les rentabilités en valeur absolue successive de 0.2714215. Ainsi les rentabilités successives en valeur absolue sont positivement corrélés. La corrélation est faible mais nous indique néanmoins que d'un jour sur l'autre, les rentabilités peuvent être liées.

De plus, nous obtenons une corrélation sur le carré des rentabilités successives de 0.2612237, elles sont donc aussi positivement corrélés. L'information de cette corrélation peut concerner la variance des rentabilités.

Pourrions nous vraiment profiter de ces dépendances ?

En réalité, si nous considérons uniquement les actifs, nous ne pourrions pas réellement profiter de ces dépendances puisque nous n'avons pas d'information sur le fait que la rentabilité soit positive ou négative.
Cependant nous pouvons savoir si elle va varier ou non.
L'information sur la variance est assez intéressante, elle permet d'estimer le risque du lendemain par rapport à jour courant.

En revanche, si nous avons des produits sur les performances du NIKKEI, nous pourrions monter une stratégie. Il conviendrait de monter un portefeuille δ neutre et s'assurer du fait que notre γ soit positif. Dès lors, la valeur du portefeuille augmentera de la même manière que la rentabilité variera que ce soit une variation positive ou négative.

6 Question 6

Comme on a pu le voir dans la **Question 5**, il existe une corrélation pour les variances des rentabilités. C'est la raison pour laquelle on va mettre en place un modèle ARCH pour essayer d'expliquer cette corrélation.

Tout d'abord nous allons essayer de valider les différents modèles en fonction de l'auto-corrélation des résidus : si la famille de résidus est un bruit blanc, alors les coefficients d'auto-corrélation sont nuls. Nous allons donc analyser la fonction d'auto-corrélation empirique des résidus et nous assurer que les coefficients ne sont pas significativement non nuls.

Puis il viendra de considérer plusieurs critères pour choisir quel est le meilleur modèle parmi les modèles qui ont passé la phase de vérification. Pour ce TP, nous allons utiliser les critères de :

- pouvoir de prédiction : R^2
- quantité d'information : le critère d'Akaike AIC

6.1 Rentabilités journalières

Nous avons testé les modèles ARMA(i,j) + GARCH(k,l) et ARMA(i,j) + APARCH(k,l) pour $i, j \in [1, 7]$ et $k, l \in [1, 4]$.

Nous remarquons que la plupart des modèles présentent une qualité faible. Tout d'abord, ils ne convergent pas tous et même pour les différents algorithmes d'optimisation (nlminb, lbfgsb, nlminb+nm, lbfgsb+nm). En effet, le R^2 est très souvent faible. De plus, grâce aux graphes ACP, nous remarquons que les résidus sont souvent corrélés et sur la `qqnormplot` nous n'obtenons pas une droite, donc les résidus ne sont pas gaussiens.

Ainsi, nous trouvons comme meilleur (des mauvais) modèle le modèle : ARMA(1,1) + GARCH(1,2) avec R^2 = 0,30% et AIC = -6,2260.

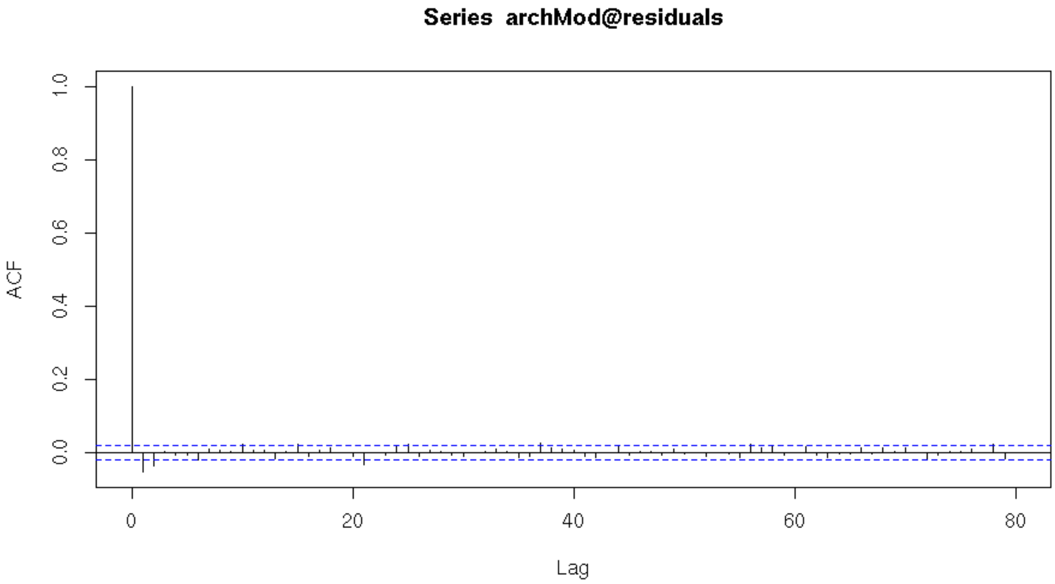

Figure 6: PACF sur la valeur absolue des rentas

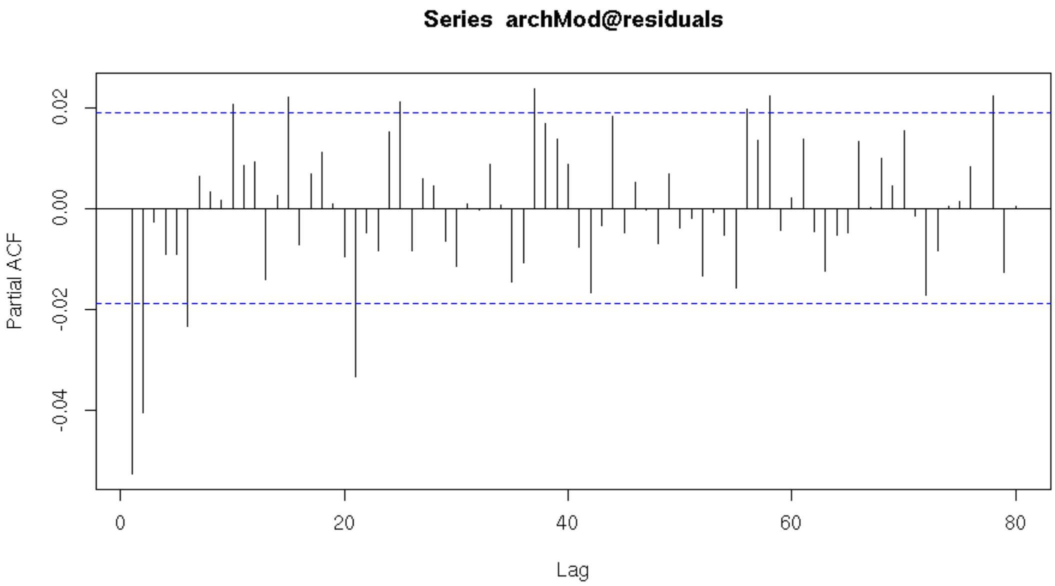

Figure 7: PACF sur la valeur absolue des rentas

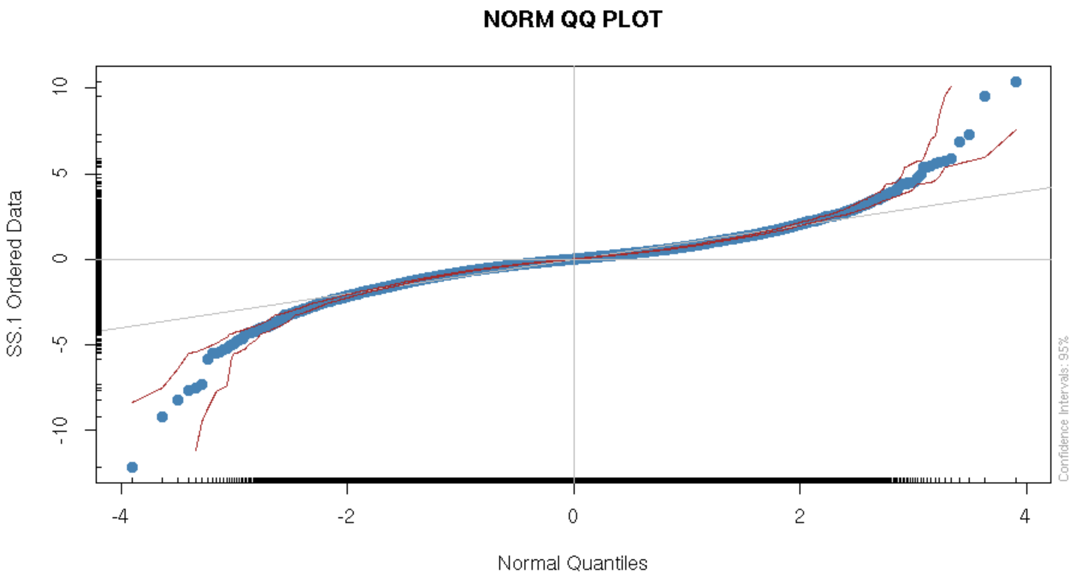

Figure 8: PACF sur la valeur absolue des rentas

6.2 Rentabilités hebdomadaires

Nous avons testé les modèles ARMA(i,j) + GARCH(k,l) et ARMA(i,j) + APARCH(k,l) pour $i, j \in [1, 4]$ et $k, l \in [1, 4]$.

Encore une fois, nous faisons face aux mêmes problèmes : les modèles ne convergent pas tous et même pour les différents algorithmes (nlminb, lbfgsb, nlminb+nm, lbfgsb+nm). En effet, le R^2 est souvent très faible. De plus, grâce aux graphes ACP, nous remarquons que les résidus sont souvent corrélés et sur la `qqnormplot` nous n'obtenons pas une droite, donc les résidus ne sont pas gaussiens. Cette dernière remarque est à nuancer car nous obtenons presque une droite.

Ainsi, nous trouvons comme meilleur (des mauvais) modèle le modèle : ARMA(2,1) + GARCH(1,3) avec R^2 = 0,01% et AIC = -4,7967.

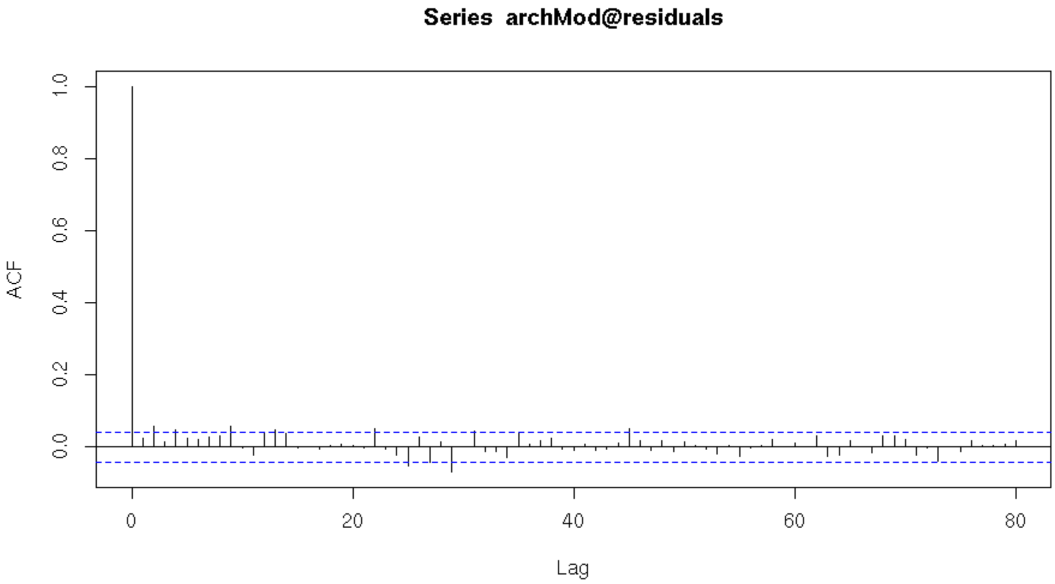

Figure 9: Corrélation des résidus

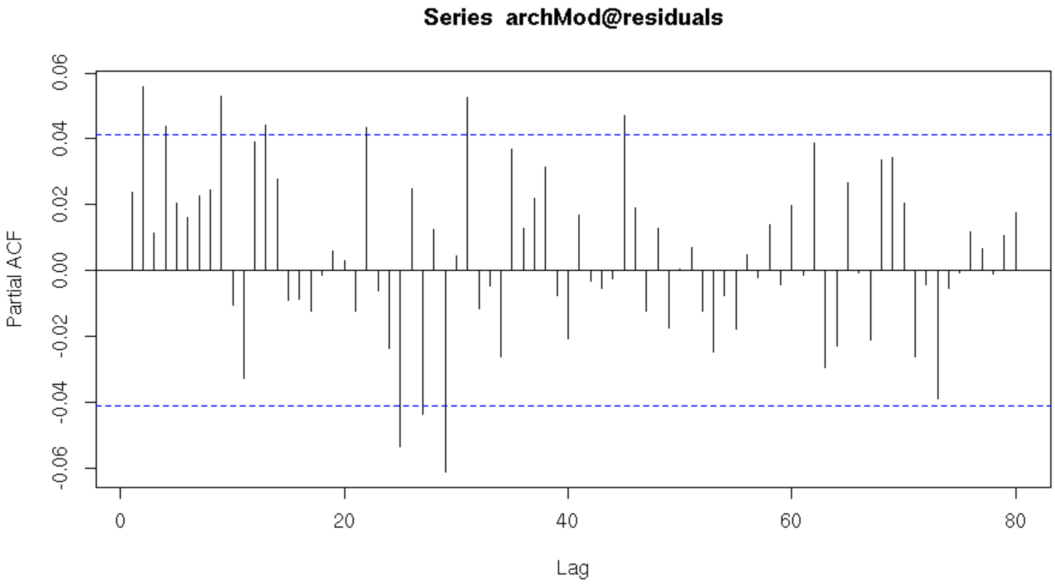

Figure 10: Corrélation partielle des résidus

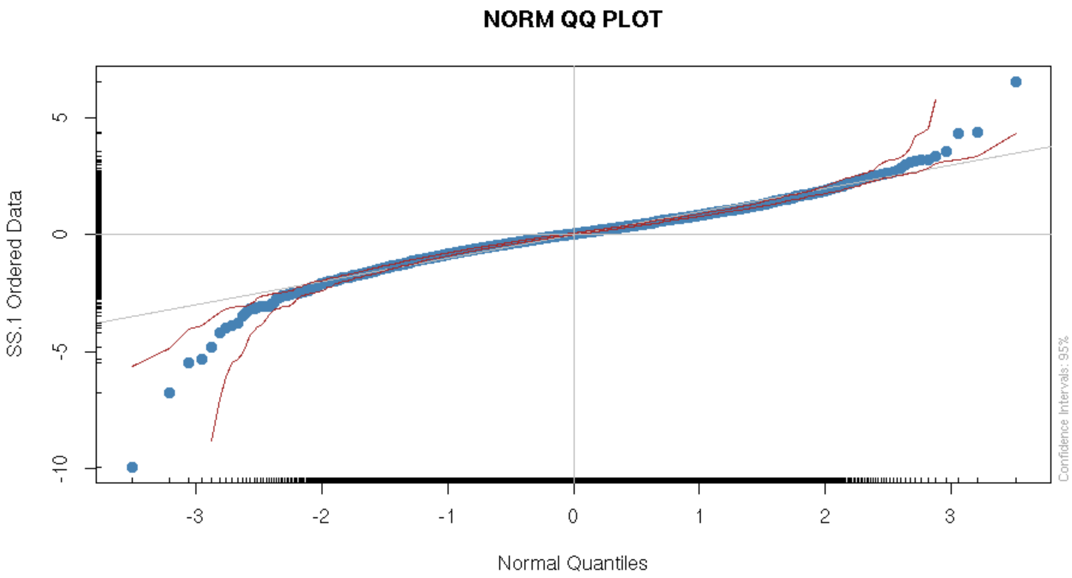

Figure 11: Qqnorm des résidus

6.3 Conclusion

Dans les deux cas, rentabilités hebdomadaires et journalières, nous observons que les modèles sont mauvais. Mais nous avons envie de dire que le modèle pour les rentabilités hebdomadaires semble meilleur que pour les rentabilités journalières.

7 Question 7

Avec le meilleur modèle que nous avons trouvé (pour les rentas hebdomadaires), nous ré-estimons les paramètres sur la période [1990-2014].

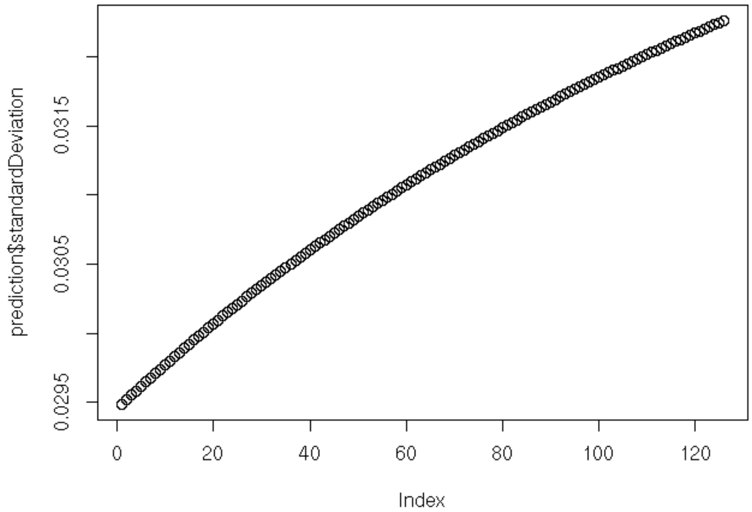

Figure 12: Standard Deviation du modèle

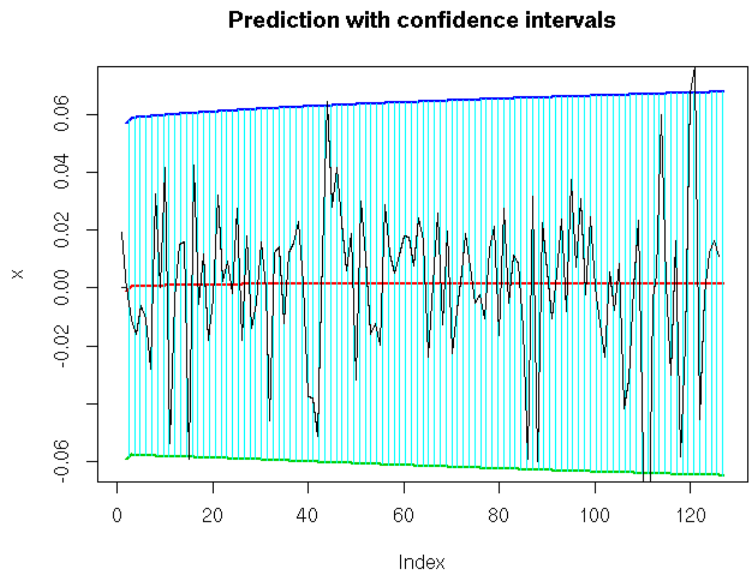

Figure 13: Qqnorm des résidus

Nous remarquons sur le graphe des rentabilités réalisées par notre modèle sont globalement dans l'intervalle de confiance, nous avons donc envie de dire que le modèle semble bien choisi. Or, avec le R^2 très faible que nous avions trouvé, finalement notre modèle n'est que très peu explicatif, nous ne pouvons pas nous fier à cet intervalle de confiance. Notre modèle n'est en réalité que du bruit.

8 Question 8

Pour vérifier que deux jeux de paramètres sont les mêmes, nous allons faire un test de Kolmogorov-Smirnov sur deux échantillons simulés avec les deux modèles. Nous trouvons une p-value inférieure à 2.2×10^{-16}, ainsi nous pouvons rejeter l'hypothèse que les deux modèles sont identiques.

Nous pouvons donc conclure que ce modèle n'est pas stable : nous ne pouvons pas prédire des rentabilités journalières avec un modèle ARMA/GARCH.

Nous faisons de même pour les rentabilités hebdomadaires et nous obtenons le même ordre de p-value. Ainsi, nous concluons que les modèles de rentabilité sont instables.

9 Conclusion

Lors de ce TP, nous avons testé plusieurs modèles pour les rentabilités journalières et hebdomadaires du NIKKEI 225. Cependant, nous n'avons pas réussi à trouver un modèle correct pour les rentabilités journalières. Les modèles étaient de mauvaises qualité (R^2 très faible). Il est donc peu probable d'avoir une espérance de gain avec ces modèles.
Les modèles ne sont pas réellement appropriés pour prédire les rentabilités futures notamment à cause de l'instabilité des modèles au fil du temps (Voir question 4).

Pour aller plus loin dans l'étude, nous pourrions essayer d'estimer sur de plus courtes périodes (1journée), afin de tenter de prédire la journée du lendemain. En effet nous avons remarqué que les R^2 des modèles pour les rentabilités hebdomadaires sont plus élevés que pour les rentabilités journalières. Ainsi, si nous arrivions à avoir plus de données sur une journée (par exemple en regardant les quotations toutes les secondes), nous pourrions peut être avoir des résultats plus concluants. Malheureusement, l'accès à ces données est souvent très coûteux.