Universite Claude Bernard Lyon 1 **Institut de Science Financière et d'Assurances**

✧✧✧

ESTIMATION DE COPULES

✧✧✧

Authors **: NGUYEN Tuan Anh**

Lyon, Avril 2023

Contents

INTRODUCTION DES DONNÉES

Sous la connaissance de la matière "Estimation de copule", les données que nous avons recueillies comprennent les cours de bourse de deux grandes entreprises technologiques: Apple et Microsoft. Pour chaque entreprise, nous avons enregistré le cours de clôture quotidien sur une période de 10 ans. Les données nous sont fournies sous forme de tableaux, contenant les dates et les prix de clôture pour chaque action, aussi *log de prix de clôture* pour chaque action.

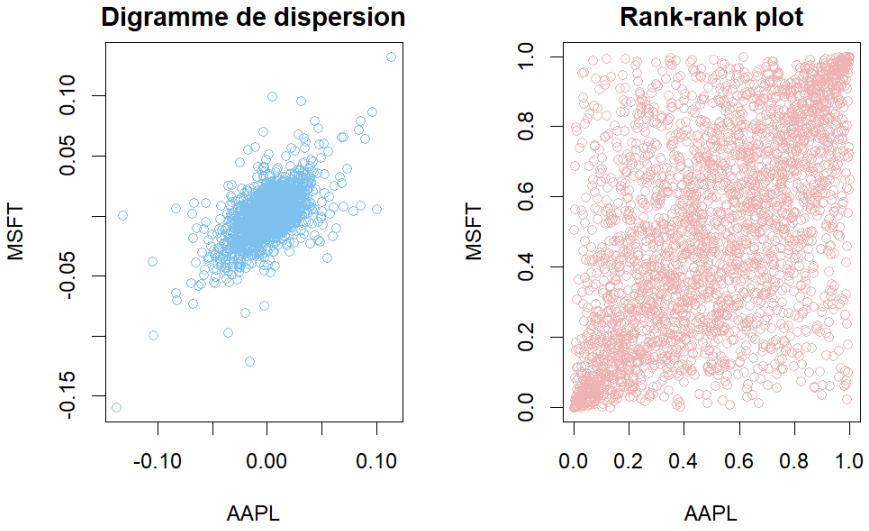
Nous allons utiliser des méthodes d'estimer de copules pour explorer la relation entre les deux variables, à savoir les cours de bourse d'Apple et de Microsoft. Pour ce faire, nous allons estimer des copules bivariées à partir des données, qui nous permettra d'examiner la structure de dépendance entre les deux actions. Ensuite, nous allons évaluer la qualité de l'ajustement de la copule et utiliser des mesures de dépendance pour évaluer la force de la relation entre les deux variables. En somme, notre objectif est de comprendre comment les fluctuations des cours de bourse de l'une des entreprises sont liées aux fluctuations des cours de bourse de l'autre entreprise.

DÉTECTION DE LA DÉPENDENCE

Dans cet étape, nous utiliserons les test de dépendance (*Test de Spearman*, *Test de Kendall* and *Test de Van der Waerden*) pour examiner de manière générale la corrélation entre deux variables. Les résultats:

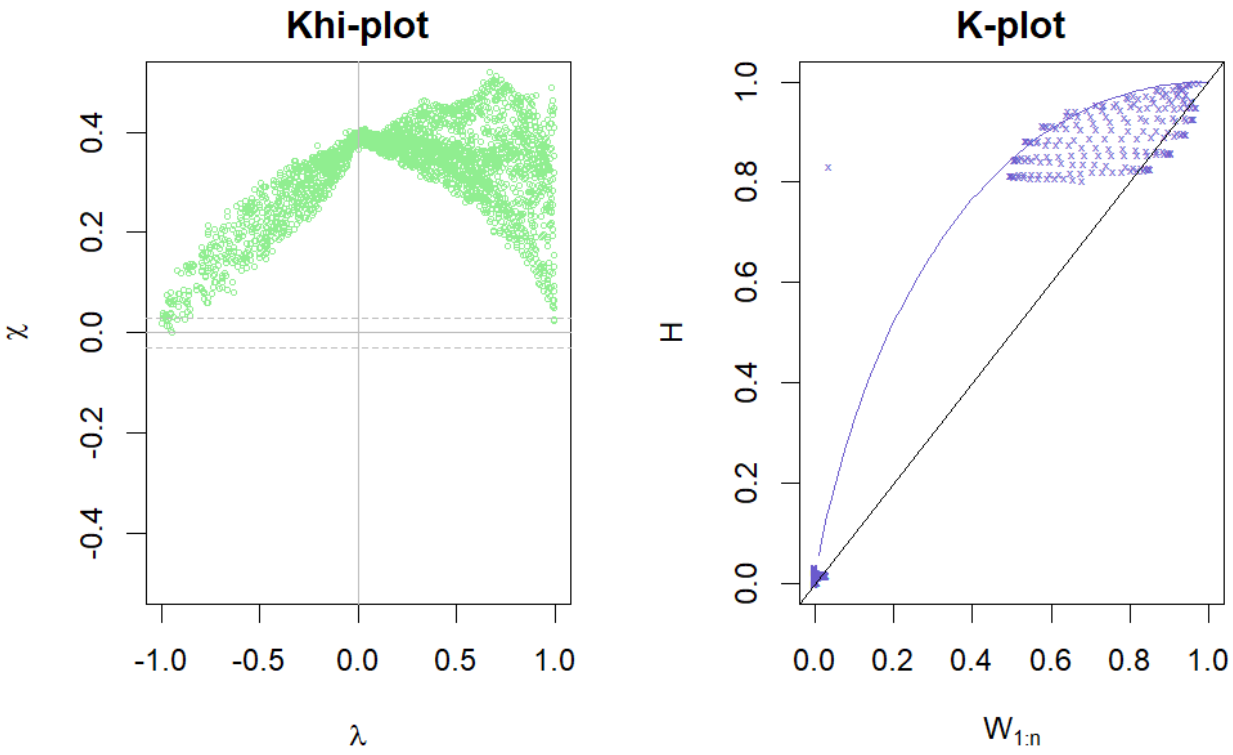
```
        Spearman's rank correlation rho

data:  X and Y
S = 1.584e+09, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.551385


        Kendall's rank correlation tau

data:  X and Y
z = 31.559, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.4002386
```

De plus, le résultat du *Test de Van der Waerden* a une valeur p-value = 0. Donc, tous les 3 tests montrent que les 2 variables sont dépendentes significative les uns des autres ($p\text{-}value < 0.05$).
Dans l'étapes suivantes, nous allons montrer la relation entre les cours de bourse de Apple et Microsoft par des graphiques.



Le graphique de nuage de points et aussi le *Rank-rank plot* peuvent être utilisés pour identifier toute différence

dans l'ordre relatif des deux ensembles de données, ainsi que toute tendance ou motif dans leur relation. En fait, les graphiques montrent une tendance linéaire claire à la hausse, cela suggère une forte corrélation entre les deux ensembles de données. De plus, nous pouvons aussi voir la dépendance légère de la queue. En outre, le Rank-Rank Plot peut suggérer que les données suivent une distribution bivariée symétrique, caractéristique d'une copule Student, si celui-ci montre une relation linéaire évidente entre les deux variables. Cela pourrait indiquer que l'utilisation d'une copule Student serait peut-êtr appropriée pour modéliser la dépendance entre les deux variables. Toutefois, ce sera encore une question.

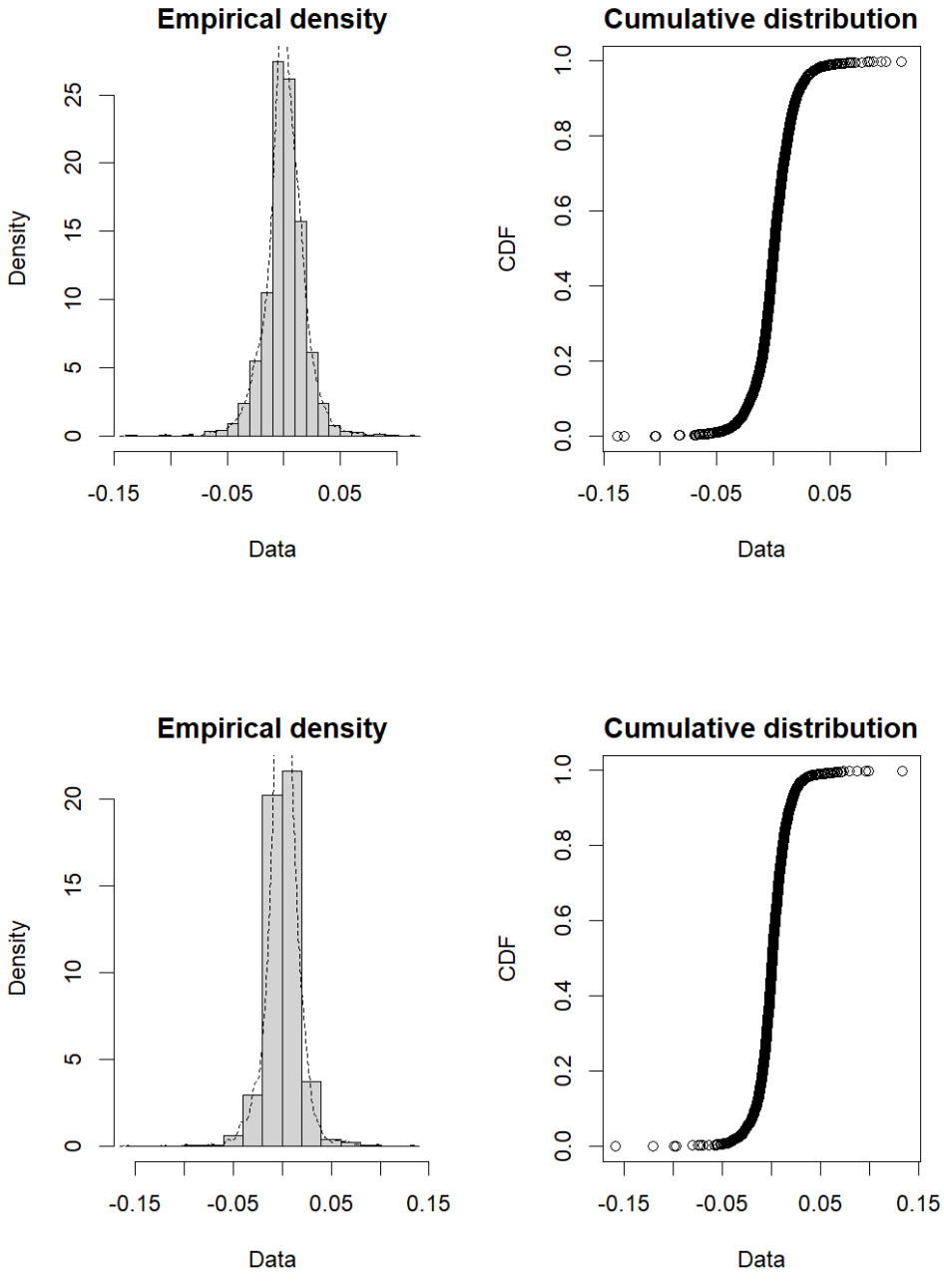


Le Khi-plot et la K-plot sont deux outils graphiques utilisés pour l'analyse des copules. Lorsque le Khi-plot permet de visualiser la différence entre la copule bivariée empirique et la copule théorique associée (ici, la copule Student), la K-plot est utilisée pour évaluer la force de la dépendance entre les deux variables. Bien que nous puissions facilement dire que ces deux variables sont dépendantes l'une de l'autre, nous ne pouvons pas conclure si elle est positive ou négative car il existe des valeurs lambda à la fois supérieures à 0 et inférieures à 0. Pourtant, nous pouvons constater une corrélation positive entre les deux variables, indiquée par la position des points sur le graphique K au-dessus de la ligne diagonale.

ESTIMATION DE COPULE

3.1 L'approach paramétrique

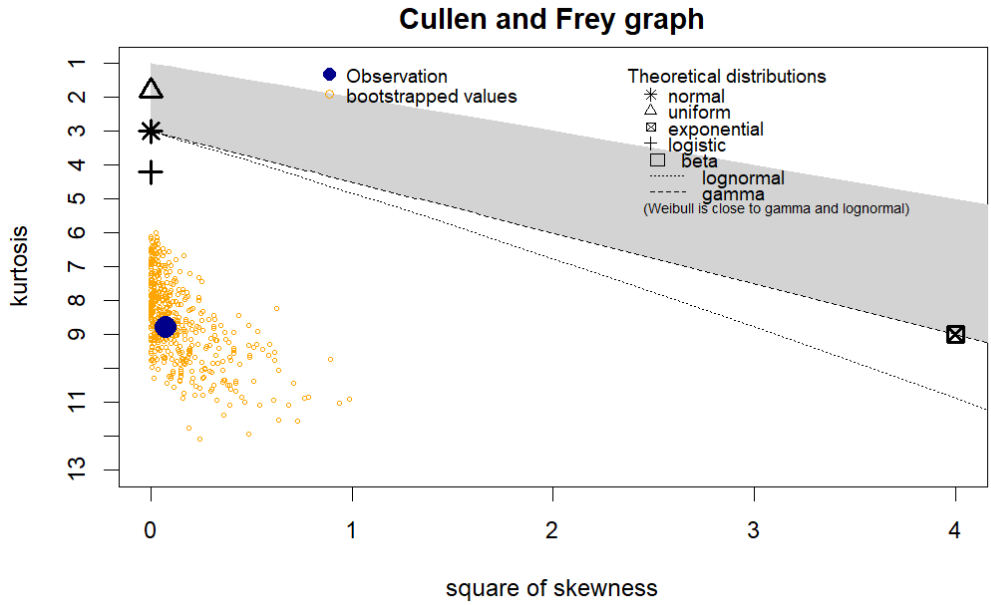
Tout d'abord, nous étudions toute de suite deux variables séparément.

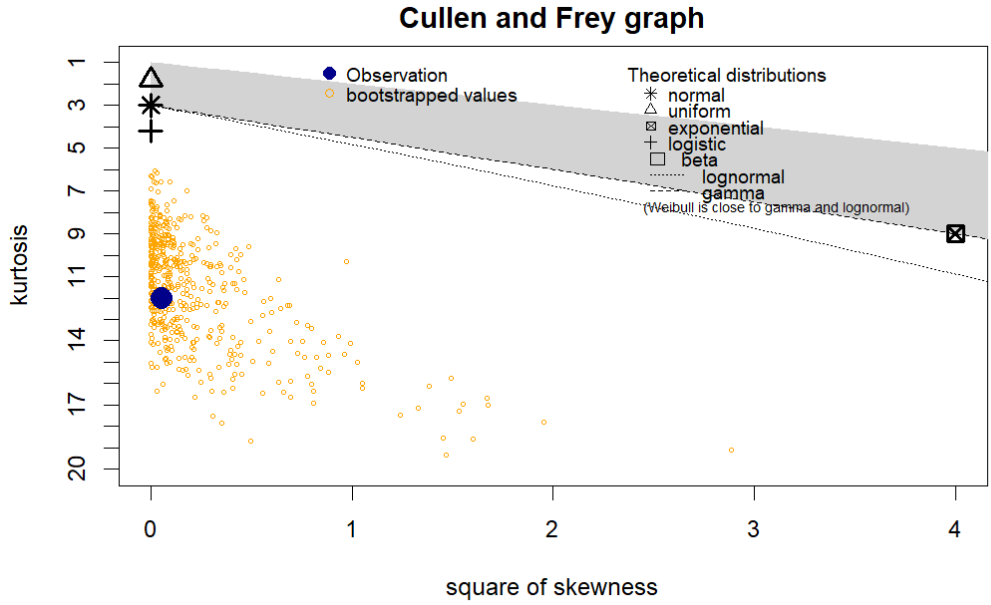


D'après les histogrammes des cours de bourses actions de l'Apple et aussi de Microsoft, nous supposons qu'ils

peuvent suivre la loi de Log-Normal (log des prix suivre la loi de Normal) comme les données sont normalement distribuées, le plot de nuage aura une forme de cloche centrée.

Mais nous allons appiquer le graphique *Cullen and Frey Graph* pour comparer quelques distributions populaires avec notre observation.





Nous concluons que nos variables ne suivent pas les distributions popualaires. Ensuite, parce qu'en général, les actions suivent la distribution log-normale, je vais utiliser le test de normalité pour revérifier.
La test de normalité de Shapiro-Wilk, appliquée aux données X et Y aussi, suggère que les données ne suivent pas une distribution normale. En effet, la valeur de D (test statistic) est de 0,07752 et de 0.097145, rétrospectivement, le p-value est très petit (7,216e-15 et 2.2e-16), ce qui indique que l'on peut rejeter l'hypothèse nulle de normalité des données. L'hypothèse alternative est donc acceptée, ce qui signifie que les données ne suivent probablement pas une distribution normale.

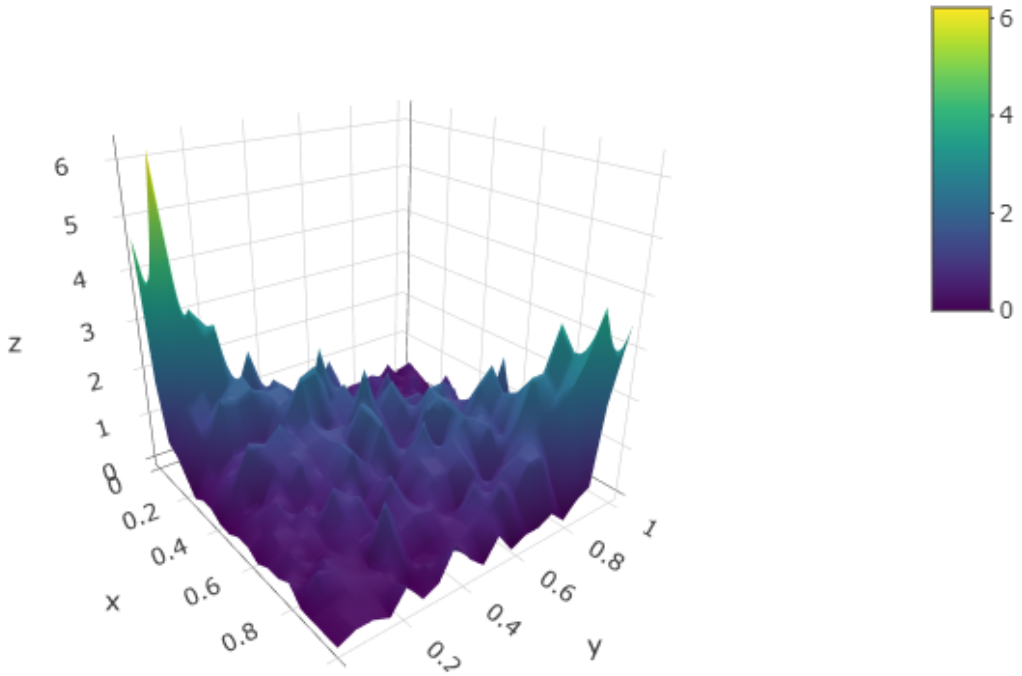
```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  X
D = 0.07752, p-value = 7.216e-15
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Y
D = 0.097145, p-value < 2.2e-16
alternative hypothesis: two-sided
```

En conclusion, parce que nous ne pouvons pas conclure la distribution des 2 variables, donc, faute de certitude, nous n'utiliserons pas cette méthode pour modéliser la dépendance entre eux et passer à l'approche suivante.

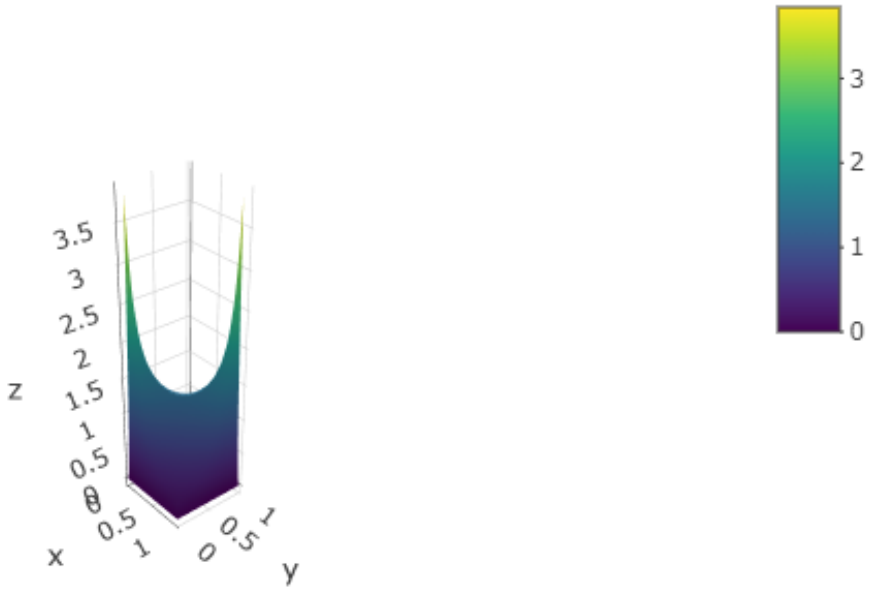
3.2 L'approach semi-paramétrique

Premièrement, je vais visualiser les données à l'aide de modèles 3D à partir de données empiriques



En utilisant le package `BiCopSelect`, nous pouvons choisir la copule appropriée pour l'ensemble de données. Alors, la copule Student est adaptée avec $\rho = 0.59$.

En utilisant la méthode CML (Canonical Maximum Likelihood), le paramètre estimé de la copule est 0.5918, il ne diffère pas trop du choix de paramètre d'origine. On peut donc supposer que la copule donne sa meilleure estimation. La valeur maximale de la fonction de log-vraisemblance est extrêmement grande et négative, ce qui signifie généralement que l'algorithme d'optimisation n'a pas convergé vers un résultat significatif. Cependant, dans ce cas, le message "Optimisation converge" indique que l'algorithme a bien convergé, et la grande valeur de vraisemblance log négative peut être due à la grande taille de l'ensemble de données.



Ensuite, nous utilisons le test d'adéquation paramétrique basé sur le bootstrap, qui est un test statistique qui évalue l'adéquation d'un modèle paramétrique ajusté aux données.

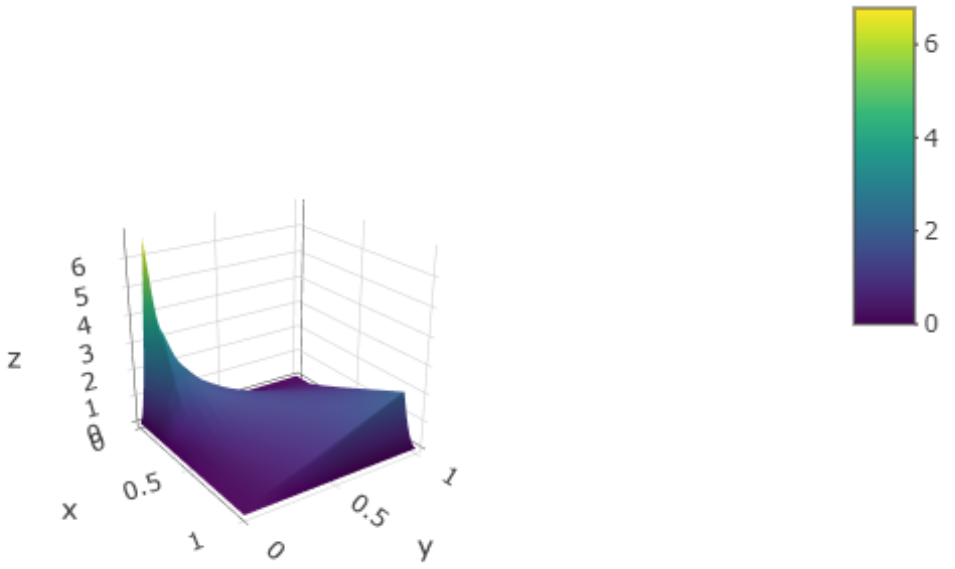
```
        Parametric bootstrap-based goodness-of-fit test of t-copula, dim. d = 2, with
        'method'="Sn", 'estim.method'="mpl":

data:  x
statistic = 0.077476, parameter = 0.61355, p-value = 0.0004995
```

Une *p-value* faible indique qu'il est peu probable que les données observées aient été générées à partir du modèle ajusté et que, par conséquent, le modèle n'est peut-être pas bien adapté aux données.
De plus, comme nous percevons au moins lower tail dependency, nous allons essayer la copule Clayton pour modéliser la dépendance entre ces 2 variables.

```
Call: fitCopula(claytonCopula(), data = data_rank, ... = pairlist(method = "mpl"))
Fit based on "maximum pseudo-likelihood" and 2767 2-dimensional observations.
Copula: claytonCopula
alpha
1.335
The maximized loglikelihood is 540.8
Optimization converged
Clayton copula, dim. d = 2
Dimension:  2
Parameters:
  alpha   = 1.335
```

En termes de description, le paramètre alpha de la copule est estimé à 1,335, ce qui indique le niveau de dépendance

entre les deux variables. Le log-vraisemblance maximisée est également fournie, indiquant l'adéquation de la copule aux données. L'optimisation a également convergé, ce qui signifie que le processus d'estimation de la copule a réussi à trouver un maximum pour la fonction de vraisemblance. En résumé, ces informations indiquent que la copule Clayton a été ajustée avec succès aux données et peut être utilisée pour modéliser la dépendance entre ces deux variables. Toutefois, l'histogramme tridimensionnel montre: Comme nous pouvons l'observer à travers



l'histogramme 3D de la copule étudiante et de la copule de Clayton et les comparer à la copule empirique 3D, nous pouvons déduire que ces copules ne peuvent pas montrer correctement nos données.

```
        Parametric bootstrap-based goodness-of-fit test of Clayton copula, dim. d = 2, with
        'method'="Sn", 'estim.method'="mpl":

data:  x
statistic = 0.41974, parameter = 1.3347, p-value = 0.0004995
```

Nous pouvons également conclure à la non-validation de clayton sur notre ensemble de données. C'est la raison que nous allons passer l'approach suivante.

3.3 L'approach non paramétrique

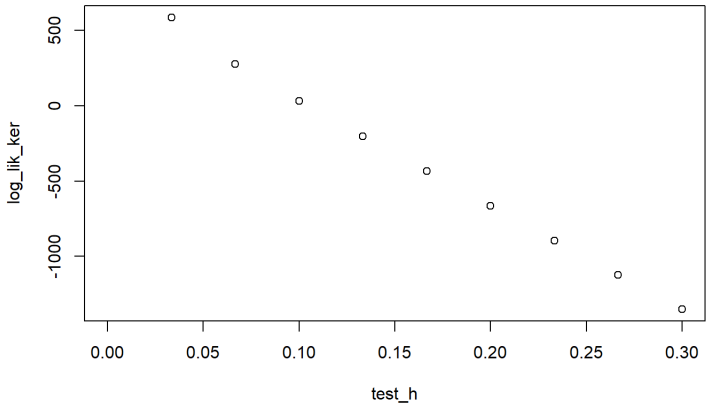
Quant à la dernière approche, nous choisissons l'estimateur à noyau pour estimer la densité c mais nous le modifierons en fonction de certaines méthodes d'estimation à noyau.

3.3.1 Noyau Gaussien Estimation

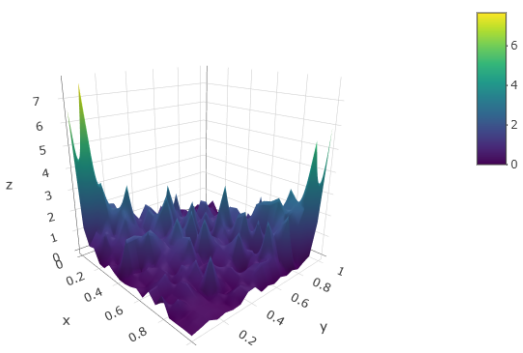
Noyau gaussien sans méthode de réflexion

Il s'agit d'une méthode d'estimation non paramétrique utilisée pour estimer la densité de probabilité d'une variable aléatoire à partir d'un échantillon de données. Cette méthode est généralement utilisée parce qu'elle est facile non seulement à lisser les données, mais aussi à impliquer. Cependent, le support de la copule se situe dans l'intervalle [0,1], de sorte que la densité dans la frontière est facilement biaisée, ce qui entraîne des phénomènes de sous-estimation dans la frontière.

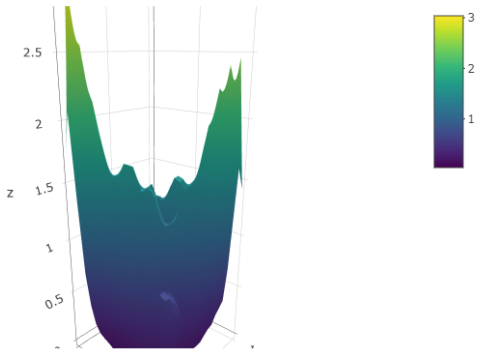
Avec l'estimation par noyau, la tâche la plus importante est de trouver la largeur de bande optimale (interprétation : la largeur des données prises pour estimer, plus les données choisies sont grandes, plus nos données sont lisses. Cependant, le niveau de certitude va être réduit).



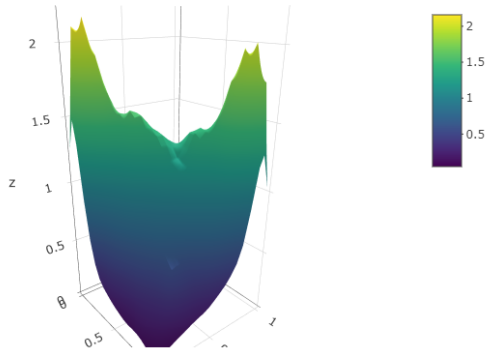
Nous pouvons observer que plus le paramètre de la largeur de bande est grand, plus la log-vraisemblance du modèle est élevée.



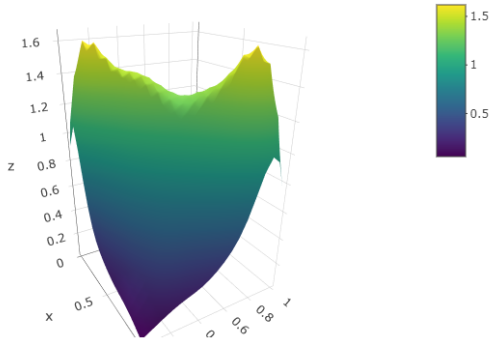
bandwidth = 0.01



bandwidth = 0.05



bandwidth = 0.07



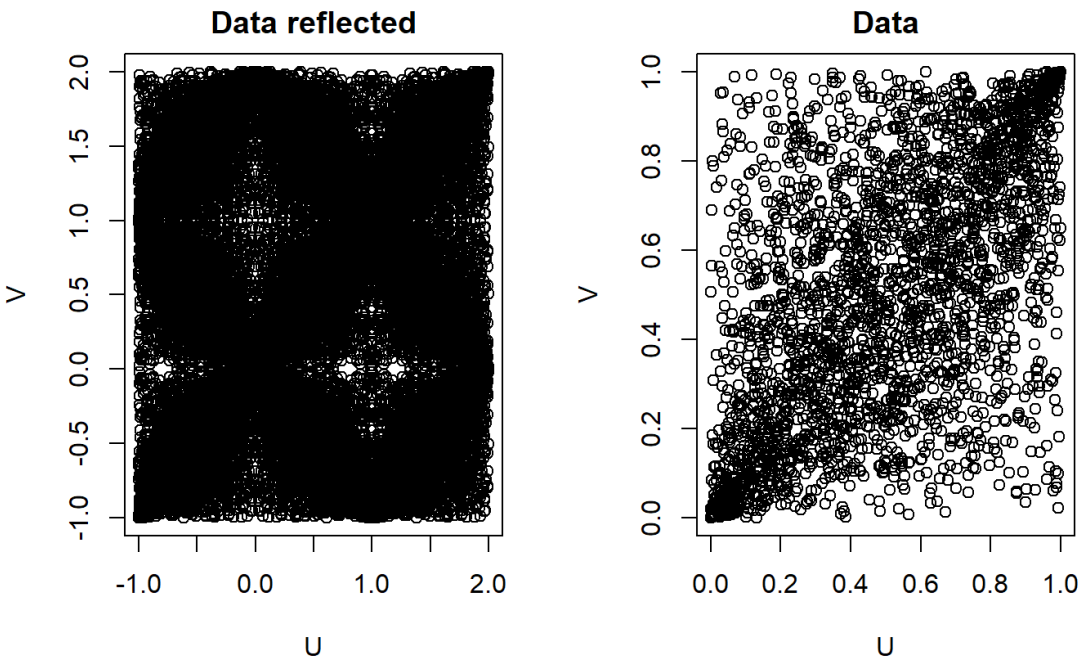
bandwidth = 0.1

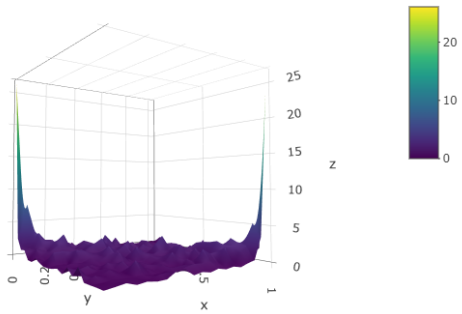
Lorsque nous observons le graphique de la log-vraisemblance avec chaque largeur de bande, nous pouvons remarquer que plus la largeur de bande est faible, plus la log-vraisemblance est élevée. Toutefois, nous allons choisir *bandwidth = 0.07*. Ensuite, nous utilisons aussi le calcul de la log-vraisemblance pour mesurer l'adéquation entre la fonction copule estimée et les données observées. Dans ce cas, la valeur de -5514,131 suggère que l'ajustement n'est pas très bon et qu'il est possible d'améliorer l'estimation de la fonction copule.

Noyau gaussien avec méthode de réflexion

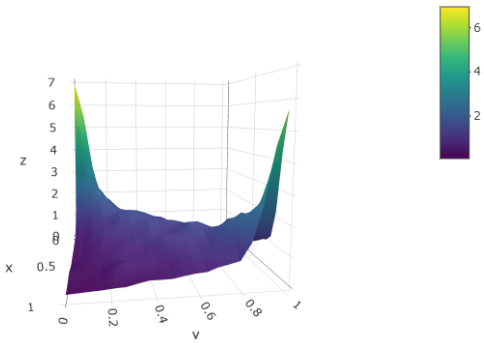
La méthode des données réfléchies est une technique utilisée dans l'estimation par noyau pour améliorer la performance de l'estimation de la densité près des frontières. Dans le contexte de l'estimation des copules, les frontières sont les bords de l'hypercube unitaire $[0,1]^d$, où d est la dimension de la copule. La méthode des données réfléchies permet de résoudre le problème de biais aux frontières qui peut survenir lors de l'utilisation de l'estimation de la densité par noyau standard près des frontières.

Pour résoudre le problème de la non représentation de la densité correcte dans la frontière, nous avons tendance à utiliser la méthode du noyau gaussien avec réflexion.

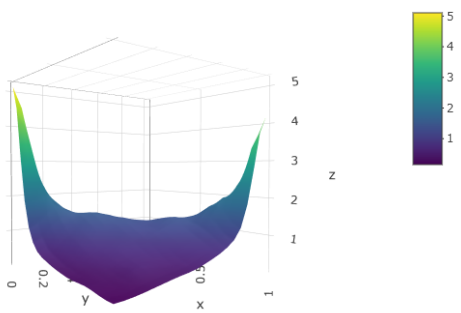




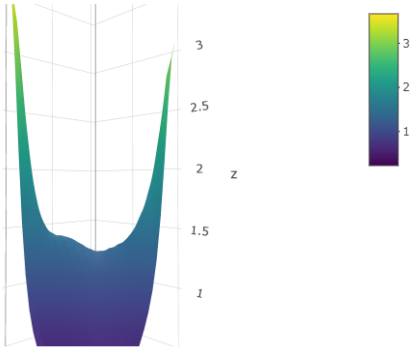
bandwidth = 0.01



bandwidth = 0.05



bandwidth = 0.7

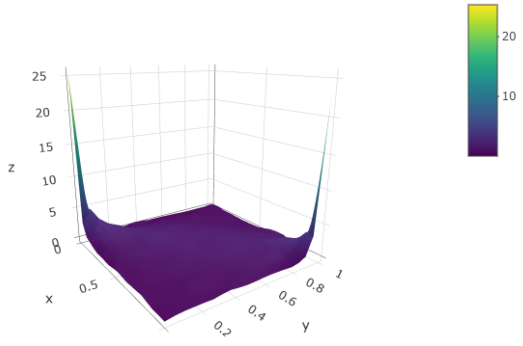


bandwidth = 0.1

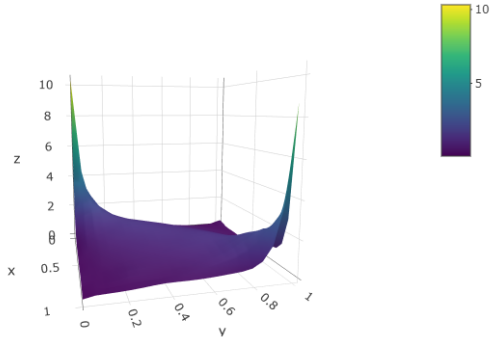
Nous verrons avec une *bandwitdth* = 0.05, c'est mieux par rapport au graphique empirique. La pseudo-vraisemblance logarithmique maximale pour l'estimation de la densité du noyau réfléchi égal 565.5899. Elle suggère que l'estimation de la densité du noyau par la méthode de réflexion a fourni un bon ajustement aux données.

3.3.2 Noyau Beta Estimation

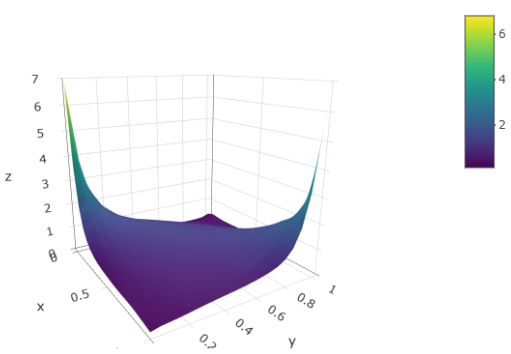
Le noyau gaussien est symétrique, ce qui signifie qu'il est plus sensible aux observations proches du point central et moins sensible aux observations éloignées. En revanche, le noyau beta est un noyau asymétrique qui assigne des poids en fonction de la position relative de chaque observation par rapport aux autres observations. Donc, l'enjeu de frontière n'existe pas avec ce noyau.



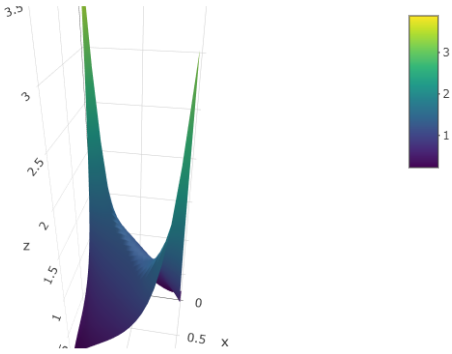
bandwidth = 0.01



bandwidth = 0.03



bandwidth = 0.05



bandwidth = 0.1

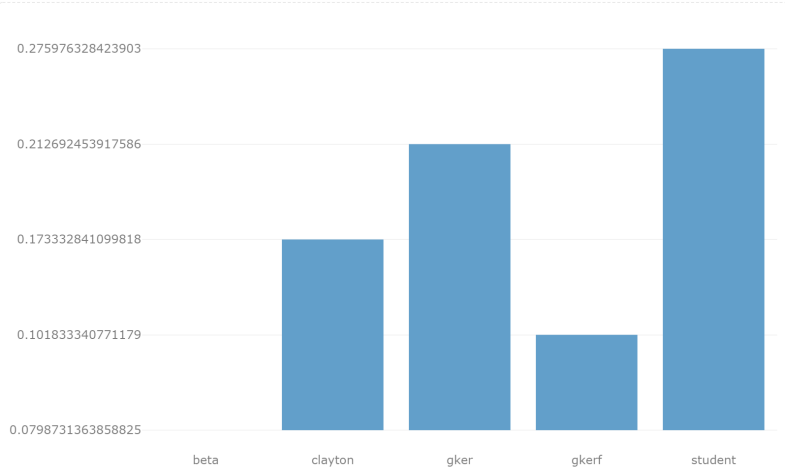
Comme nous pouvons l'observer à travers les graphiques, les graphiques entre bandwidth = 0.01 et 0.03 n'existent pas l'énorme différence. Nous avons donc tendance à choisir la bandwidth = 0.02. La somme du logarithme de la densité du noyau bêta bivarié estimée à l'aide de la largeur de bande donnée (0,02) est de 713,6121, ce qui est meilleur que le résultat du noyau de Gaussien.

CONCLUSION

Tableau récapitulatif des paramètres:

Copule	Student	Clayton	Noyau Gaussien sans réflection	Noyau Gaussien avec réflection	Noyau Beta
Paramètre	$\rho = 0.59$	$\alpha = 1.33$	bandwidth = 0.07	bandwidth = 0.05	bandwidth = 0.02

Nous utilisons *the Mean Integrated Squared Error (MISE)* pour les différentes méthodes d'estimation des copules. Le MISE mesure la distance entre la densité de copule estimée et la vraie densité de copule en élevant au carré la différence entre les deux densités et en intégrant sur l'unité carrée.



Comme nous pouvons l'observer, la MISE des modèles non paramétriques semble être plus faible que celle des modèles semi-paramétriques (les copules de Student et de Clayton en l'occurrence). Cela indique qu'en ajustant les paramètres de la bande passante pour obtenir une copule à la fois lisse et précise, il est possible d'obtenir un modèle non paramétrique offrant de meilleures performances que les modèles paramétriques ou semi-paramétriques classiques. Dans notre cas, la MISE du modèle à noyau gaussien normal est assez élevée, même en comparaison avec les modèles de Student et de Clayton. Ce phénomène peut être attribué à un biais engendré par les problèmes de bord. Pour pallier ce problème, nous utilisons le noyau gaussien avec la méthode de réflexion et le noyau bêta afin de confiner toute la densité à l'intérieur de l'hypercube (et ainsi limiter le biais). Finalement, c'est l'estimation par noyau bêta qui offre les meilleurs résultats, sa MISE étant nettement inférieure à celle des autres approches. En conclusion, lorsque les données présentent une complexité élevée et ne correspondent pas à un modèle de copule courant, la méthode non-paramétrique d'estimation des copules s'avère être une solution plus efficace. Cette approche permet d'obtenir un graphique plus clair et plus compréhensible pour l'analyse des résultats.