ISUP M2 Actuariat

Théorie des valeurs extrêmes

Application sur le S&P500

Lucas Jullia - Corentin Boyeau
02/11/2020

Table des matières

Introduction

A faire

I- Description et préparation des données

A) Description des données

Le S&P500 est un indice boursier basé sur 500 grandes sociétés cotées sur les bourses américaines. Cet indice est considéré comme un indice de référence qui fait office de « baromètre » de l'économie mondiale, étant donné qu'il regroupe un grand nombre de sociétés (pas forcément américaines) qui sont côtés sur les marchés américains.

Pour l'étude nous nous baserons sur une base de données regroupant l'ensemble des données journalières de l'indice de 1988 à nos jours, extraites depuis le site Yahoo Finance. La base de données contient 5 variables :

- « **Date** » : date à laquelle l'indice a été mesuré
- « **Open** » : valeur de l'indice à l'ouverture du marché américain pour une date donnée
- « **High** » : valeur du maximum atteint par l'indice pour une date donnée
- « **Low** » : valeur du minimum atteint par l'indice pour une date donnée
- « **Close** » : valeur de l'indice à la fermeture du marché américain pour une date donnée

De manière arbitraire nous considérerons pour la suite la variable « Open » pour le calcul des rendements.

B) Préparation des données

Plutôt que de travailler directement sur les indices, nous allons nous concentrer sur les rendements quotidiens. On peut les calculer de la manière suivante :

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Avec $P_t = valeur\ de\ l'indice$

On peut également s'intéresser à la rentabilité continue (« log returns ») qui se calcule de la manière suivante :

$$R_t = \log(1 + r_t) = \log\left(\frac{P_t}{P_{t-1}}\right)$$

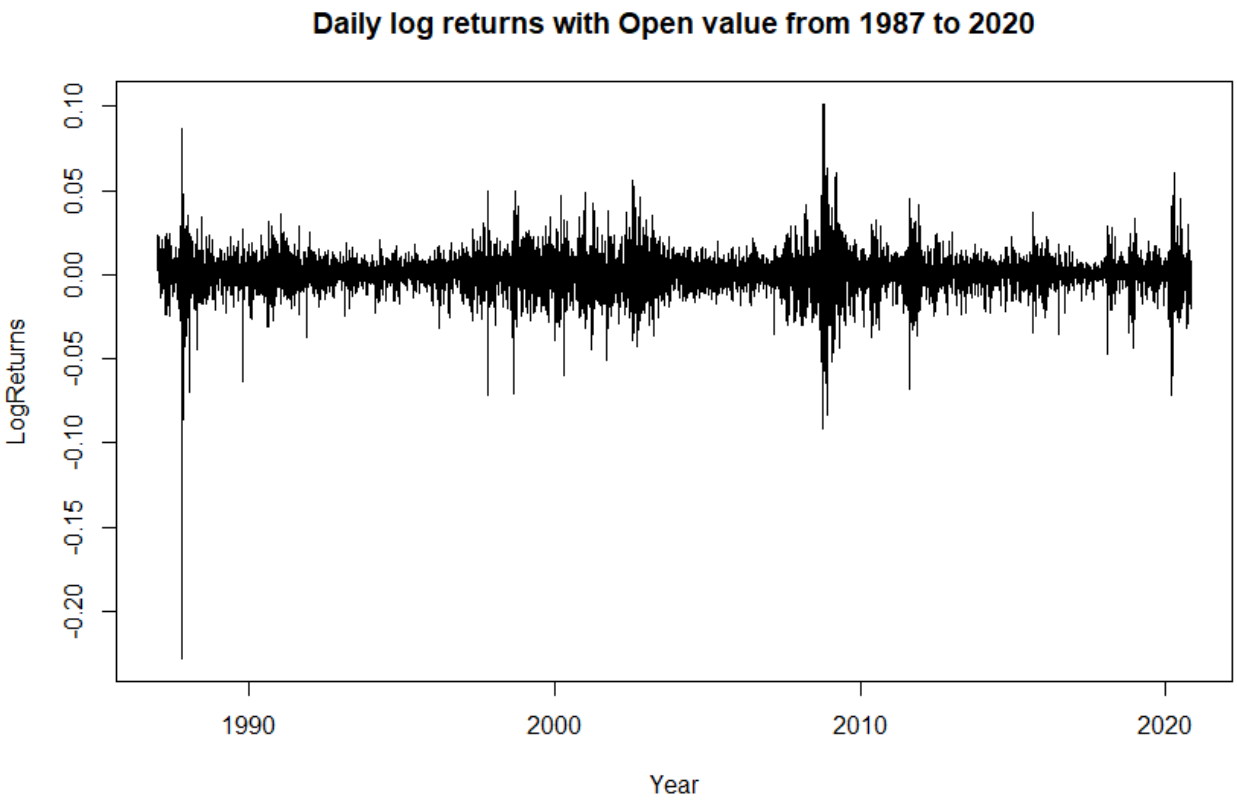
Nous utiliserons plutôt cette dernière formule pour nos futurs calculs.

C) Statistiques et premières visualisations des données

	S&P LogReturns
Moyenne	0.031%
Variance	0.013%
Minimum	-22.802%
1er quartile	-0.429%
médiane	0.069%
3eme quartile	0.552%
maximum	10.139%
Skewness	-1.312
Kurtosis	27.393

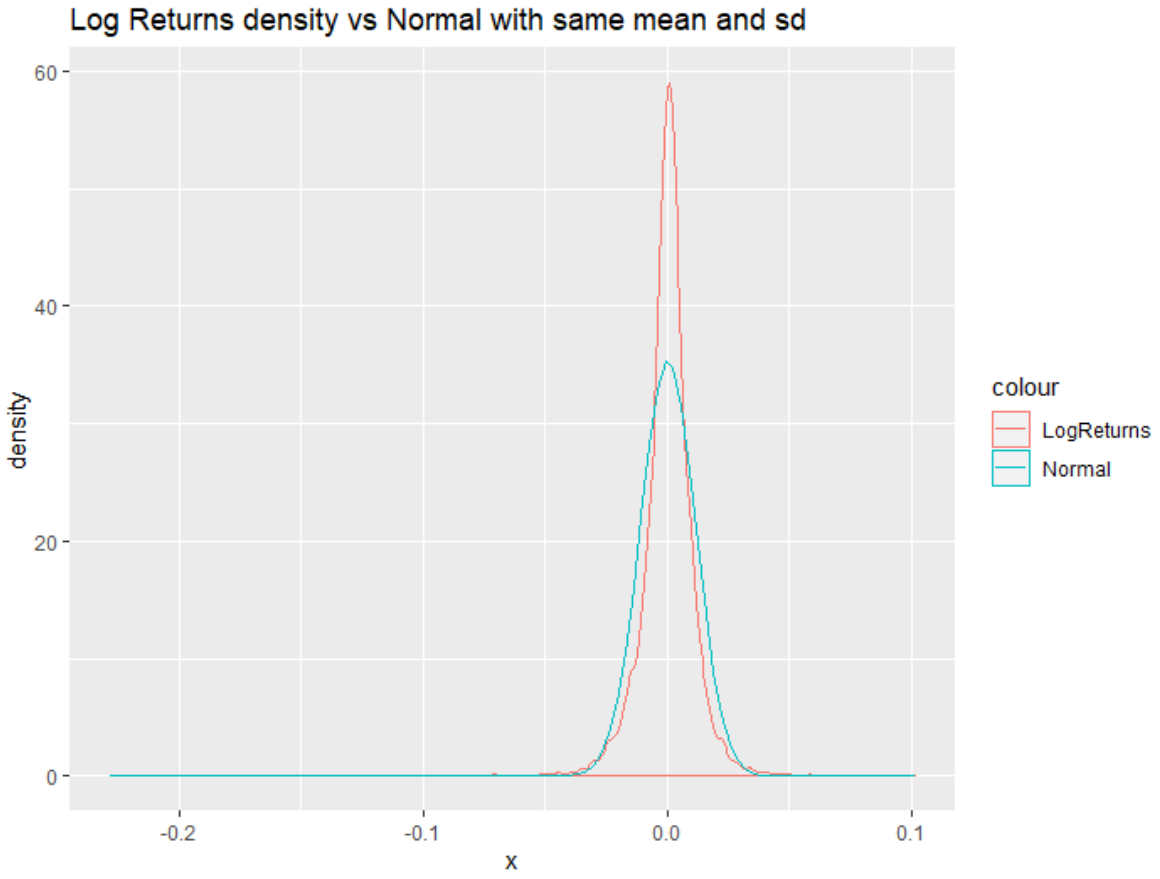
Cette table nous présente l'ensemble des statistiques qui pourraient nous intéresser. Tout d'abord, il semble que nos données soient concentrées autour de 0, vu que 75% d'entre elles se trouvent entre -0.42% et 0.55% de rendements, avec des rendements en moyenne légèrement positifs et une variance assez faible.

Cependant il apparaît que le kurtosis normalisé est légèrement positif, indiquant ainsi une distribution dont les queues sont plus épaisses qu'une loi normale. Pour rappel le kurtosis normalisée désigne le coefficient d'aplatissement auquel on retire le coefficient d'aplatissement d'une loi normale c'est-à-dire 3. De plus le coefficient d'asymétrie (skewness) est négatif, indiquant que notre distribution est décalée vers la droite de la médiane et donc que notre queue de distribution est étalée vers la gauche, indiquant ainsi une concentration de pertes extrêmes plus importante que de gains extrêmes.



La représentation des log rendements au cours du temps nous permet d'identifier très facilement les grands chocs qui ont marqués la bourse depuis 1987 :

- Black Monday le 19 octobre 1987 : une des plus importantes baisses journalières jamais enregistrée.
- Terrorisme et bulle internet : de 1998 à 2003 période marquée par de fortes variations causées notamment par l'éclatement de la bulle internet et par l'instabilité politique durant cette période.
- Crise des subprimes de 2007 provocant une crise mondiale en 2008.
- Tempêtes boursières de l'été 2011 marqué entre autres par les difficultés économiques de la Grèce.
- Crise économique de 2020 liée à la pandémie de Covid-19.



Il apparaît que la loi normale avec les mêmes paramètres que nos observations semble assez mal correspondre à la densité empirique. Nos rendements sont beaucoup plus concentrés au niveau de la moyenne avec des queues qui semblent plus épaisse que la loi normale.

L'objectif des parties suivantes est donc double. En priorité, nous souhaitons mieux comprendre nos risques extrêmes, c'est-à-dire les pertes extrêmes, et donc mieux modéliser cette queue de distribution en particulier (partie II et III). Dans la partie IV nous prendrons un peu de recul, pour tenter de modéliser l'ensemble de notre distribution, le comportement moyen, ainsi que nos 2 queues de distribution.

II – Block Maxima method

Afin de faciliter l'application des méthodes nous nous intéresserons pour la suite de l'étude à la fonction des pertes c'est-à-dire $Perte(x) = -LogReturns(x)$. Les valeurs positives sont donc des pertes et les valeurs négatives des gains.

A) Théorie

Afin d'établir quelles sont les valeurs extrêmes de notre série statistique, la méthode des Block Maxima se base sur une idée assez simple. En effet, le but est de diviser notre série statistique sur sa période complète en plusieurs sous-périodes de temps égal. Les valeurs extrêmes sont constituées alors de la valeur maximale de chaque sous-période.

Le théorème de Fisher-Tippett-Gnedenko assure ensuite le fait que les maxima d'un échantillon suivant une variable aléatoire de façon i.i.d. ne peuvent converger que vers 3 types de distribution différentes : Gumbel, Fréchet ou Weibull.

Ainsi, notre nouvelle série constituée des valeurs extrêmes selon la méthode des Block Maxima suit une loi Generalized Extreme Value, plus communément abrégée en GEV, avec la fonction de distribution suivante :

$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)_+^{-\frac{1}{\gamma}}\right) & si\ \gamma \neq 0 \\ \exp(-e^{-x}) & si\ \gamma = 0 \end{cases}$$

Où γ est nommé le paramètre de forme ou encore l' « extreme value index » et détermine le type de distribution dans laquelle nous nous trouvons :

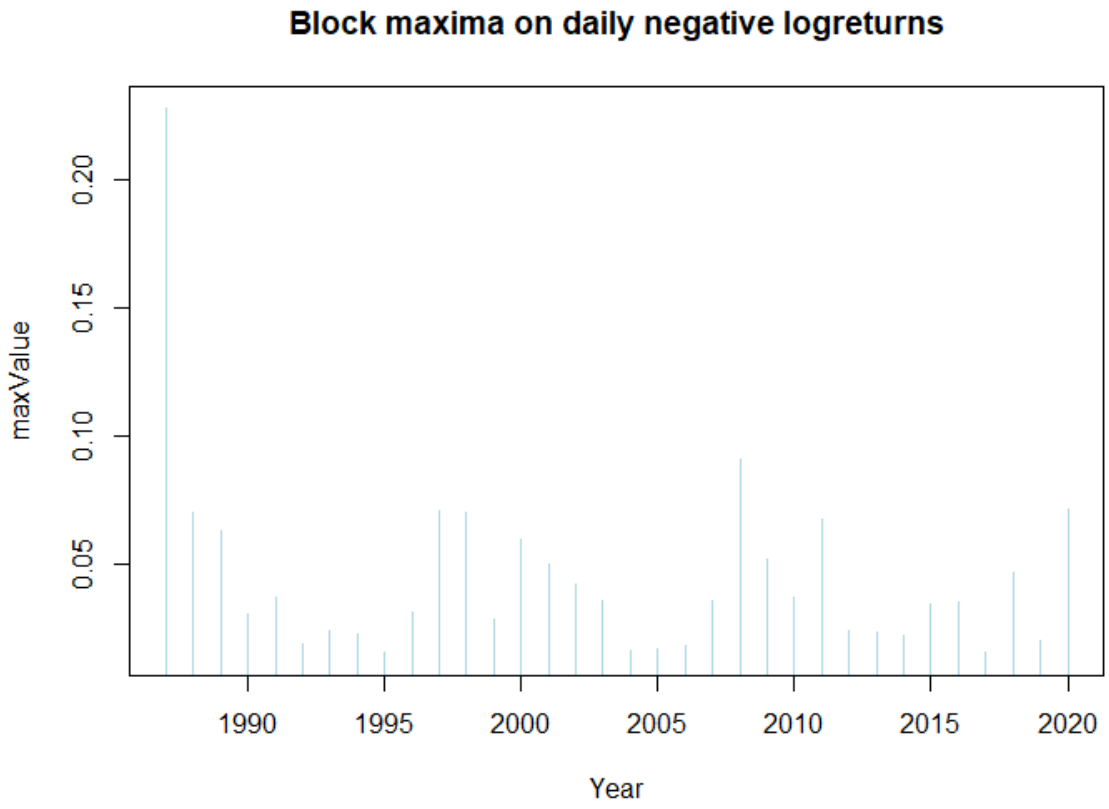
- $\gamma = 0$: Distribution de Gumbel (queue légère)
- $\gamma < 0$: Distribution de Weibull (queue lourde)
- $\gamma > 0$: Distribution de Fréchet (queue légère)

B) Résultats

Nous décidons ici de diviser notre période totale qui va de 1987 à 2020 en 34 sous-périodes, soit une période par an. Si chaque année ne comporte pas exactement le même nombre de jours travaillés qu'une autre, la différence est négligeable et nous considérons donc que nos sous-périodes sont de temps égal.

De plus, bien que l'année 2020 ne soit pas complète au moment où nous avons effectué cette méthode (en novembre), nous avons décidé de la garder étant donné que la valeur maximale correspondant à 2020 est d'ores et déjà assez significative comparée aux autres années. Cela n'est pas surprenant compte tenu de la crise induite par le covid-19.

Le graphique ci-dessous décrit la perte maximale de notre série par année :



Nous pouvons à nouveau remarquer sur ce graphique les grandes périodes où le marché a connu des fortes pertes telles que le Black Monday en 1987, la crise des subprimes en 2008 ou encore la crise liée au covid-19 en 2020.

Nous ajustons à présent notre loi GEV sur nos pertes journalières.

Nous obtenons les valeurs suivantes :

- Paramètre de forme :

$$\gamma = 0{,}538$$

- L'erreur standard associée au paramètre de forme :

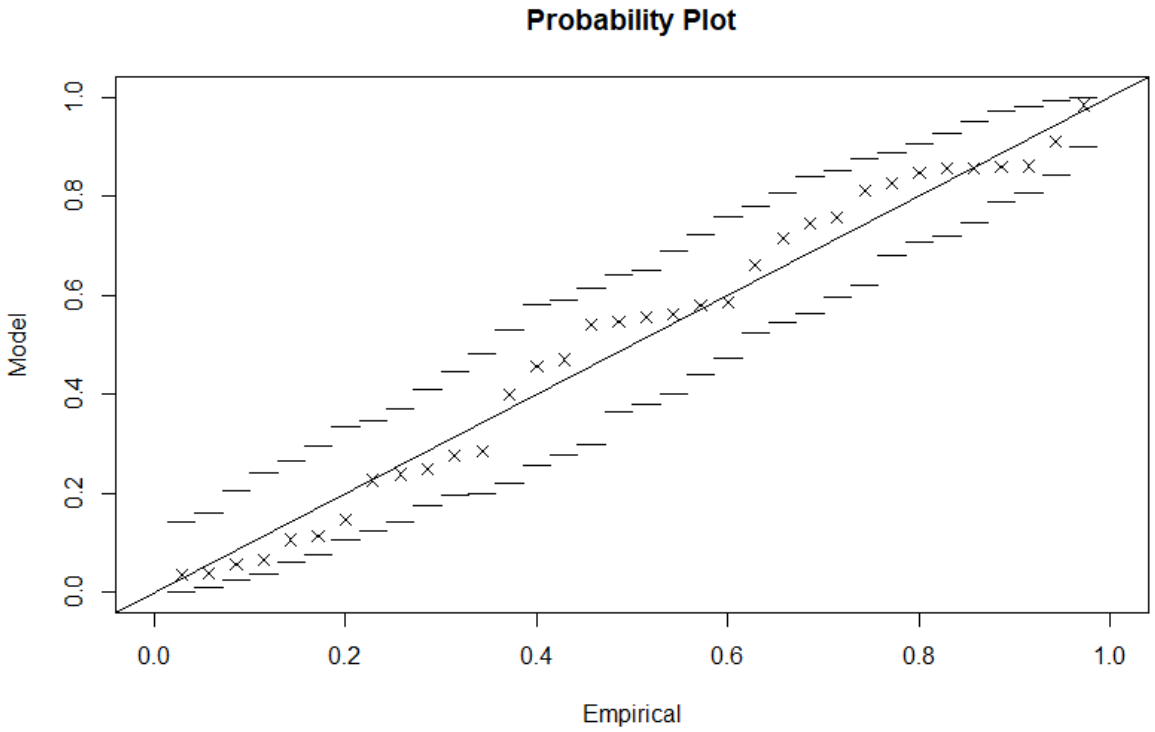
$$\sigma = 0{,}247$$

On en déduit un intervalle de confiance à 95% :

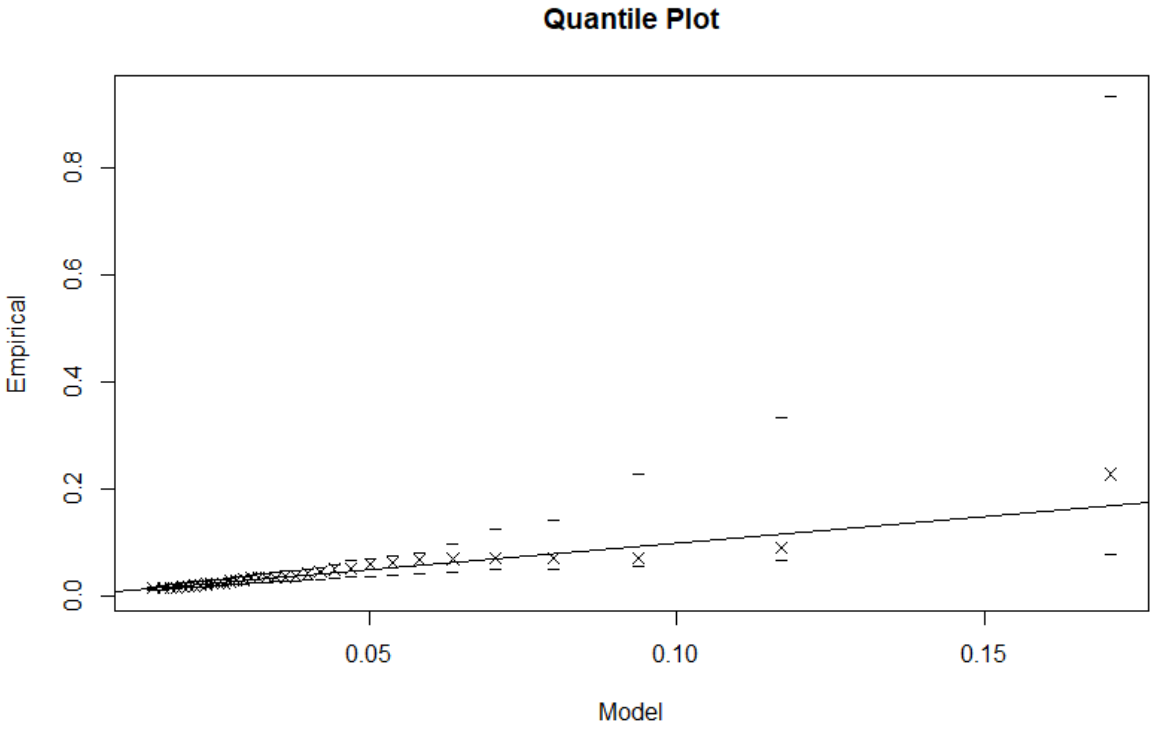
$$\gamma \in [\,0{,}0545\,;\ 1.022]$$

On peut donc conclure avec assez de confiance que le paramètre de forme est strictement positif et donc que les pertes extrêmes du S&P 500 suivent une loi de Fréchet. La queue de distribution est donc épaisse.

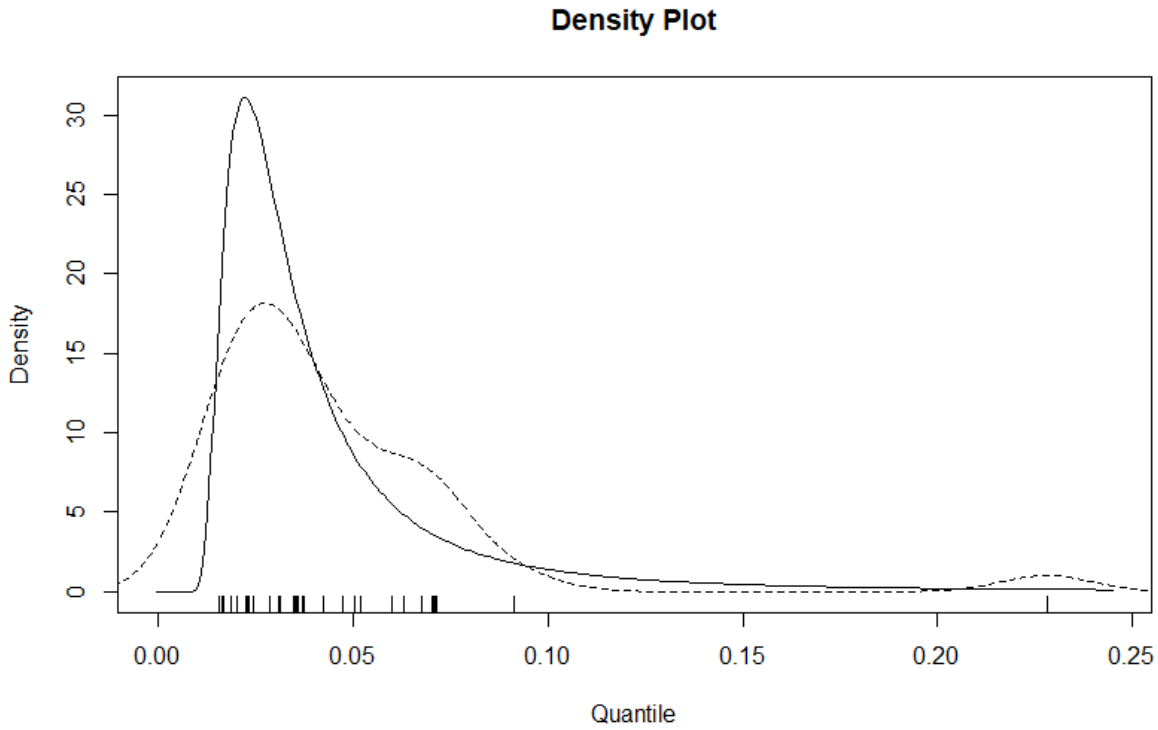
Nous pouvons à présent analyser graphiquement la précision de notre ajustement de notre queue de distribution par une loi de Fréchet.



Premièrement, les valeurs de notre modèle semblent fluctuer autour de notre série empirique sans trop s'en éloigner. De plus, elles restent toujours à l'intérieur de l'intervalle de confiance ce qui est un bon signe.



A nouveau, les valeurs sur ce QQ-plot restent à l'intérieur de l'intervalle de confiance, ce qui est cohérent avec le graphique de probabilité au-dessus.



Ce graphique est sans doute le plus intéressant à analyser. Notre loi de Fréchet ajustée – représentée par le trait plein – semble globalement suivre les variations de notre distribution empirique. Surtout, l'ajustement semble précis au niveau de la queue de distribution, bien que celle de notre loi de Fréchet semble légèrement surestimer l'épaisseur de celle-ci.
L'ajustement de la queue semble donc globalement satisfaisant mais pourrait sans doute être amélioré.

Concernant cette méthode, nous pouvons conclure que le nombre de sous-périodes choisi est prépondérant. En effet, la modification de ce nombre de sous-périodes pourrait grandement influer nos résultats. De plus, l'intervalle de confiance du paramètre de forme est très large, ce qui n'est pas optimal. Enfin, il est évident que nous n'étudions pas l'ensemble des valeurs les plus extrêmes de notre distribution des pertes mais seulement des maximums locaux. Cela pourrait par exemple avoir pour effet de ne pas tenir compte d'une grande valeur si une autre supérieure se trouve dans son voisinage.

III – Peak Over Threshold (POT) method

Contrairement à la méthode des Block Maxima, les méthodes appelées “Peak Over Threshold” se basent sur les plus grandes statistiques d’ordre d’une distribution statistique. En effet, cela permet de réaliser une étude sur les valeurs les plus extrêmes de la distribution. La problèmatique centrale de ces méthodes est d’établir un seuil à partir duquel nous pouvons considérer les valeurs comme étant extrêmes et donc pertinentes pour une analyse de la queue de distribution, et sur laquelle nous pouvons fitter une loi GPD comme l’indique le théorème de Pickands. Nous verrons trois d’entre elles dans cette partie

A) Mean Excess Plot

La méthode du Mean Excess Plot se base sur la probabilité conditionnelle suivante :

$$F_u(y) = \mathbb{P}(X - u \geq y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}$$

Où :

- F est la fonction de répartition de notre série de données
- X représente la variable aléatoire que prend nos pertes
- u est le seuil choisi
- F_u est la fonction de répartition conditionnelle au seuil u
- y > 0

L’objectif est d’établir l’équation permettant de calculer F_u :

$$\overline{F_u\,(y)} = 1 - F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}$$

A l’instar de la méthode des Block Maxima, nous pouvons supposer que les valeurs extrêmes de nos pertes suivent une loi GEV de fonction de répartition :

$$G(x) = \mathbb{P}(\max X_i \leq x) = F^n(x) = \exp\left(-\left(1 + \frac{\gamma(x-\mu)}{\sigma}\right)^{-\frac{1}{\gamma}}\right)$$

Or, étant donné que nous étudions des valeurs extrêmes nous pouvons émettre deux hypothèses :

- Nos valeurs étudiées sont très grandes et donc : $\ln\big(F(x)\big) \approx -(1 - F(x))$
- Notre seuil u est également très grand et donc : $u \to \infty$

En utilisant les expressions précédentes et en appliquant nos deux hypothèses après quelques manipulations d’équations, nous obtenons la formule suivante :

$$F_u(y) \approx G_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 + \frac{\gamma\mu}{\sigma}\right)^{-\frac{1}{\gamma}}, & \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right), & \gamma = 0 \end{cases}$$

Il s'agit d'une loi « Generalized Pareto Distribution », plus communément abrégé en GPD. C'est cette loi que nous allons ici ajuster en essayant d'obtenir le paramètre de forme γ le plus adapté à nos données.

Afin d'évaluer le γ optimal, il ne reste plus qu'à définir le seuil idéal. Celui-ci est déterminé graphiquement à partir de la fonction du Mean Excess définie comme telle :

$$e(u) = \mathbb{E}[X - u \mid X > u] = \frac{\sigma + \gamma u}{1 - \gamma}$$

Ou alternativement à partir de l'estimateur de cette dernière :

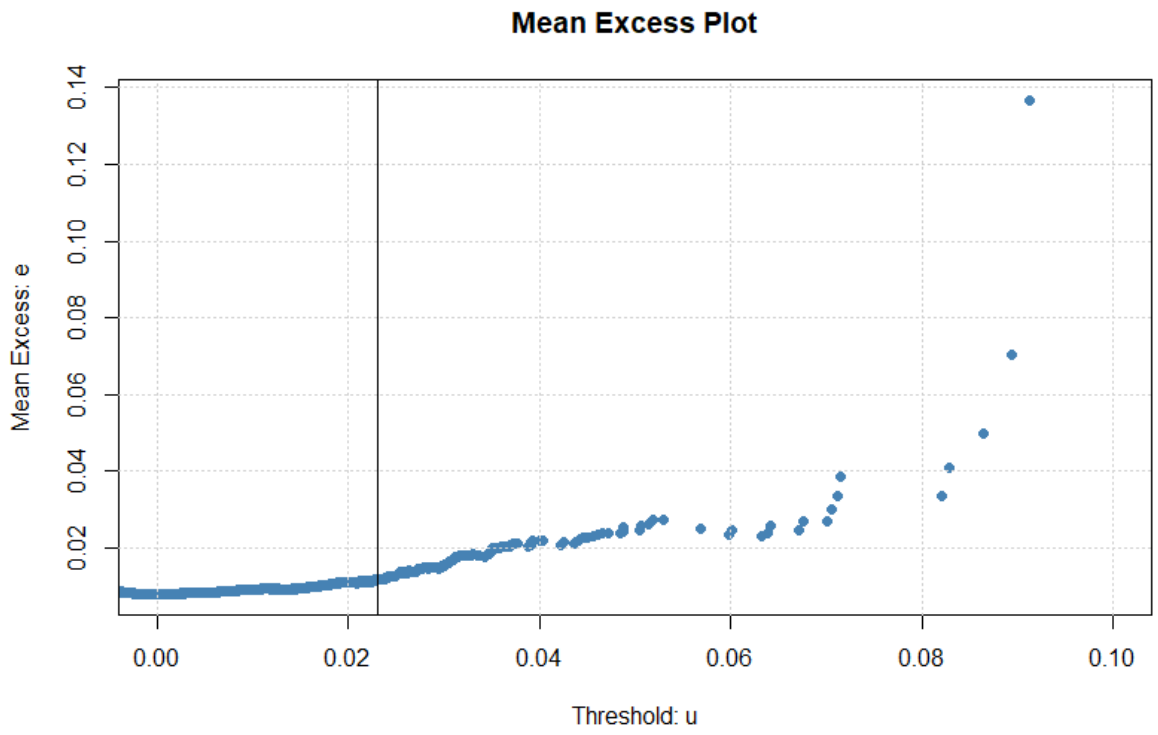
$$\widehat{e_n(u)} = \frac{1}{N_u} \sum_{i=1}^{N_u} (X_{i,n} - u)$$

Où N_u correspond au nombre d'observations de nos données étant supérieures au seuil u.

Le principal objectif de cette méthode est par conséquent de choisir un bon seuil, ni trop faible, ni trop élevé. Pour cela, il convient de tracer le graphique de la fonction du Mean Excess, en d'autres termes le Mean Excess Plot.

Il faut ensuite choisir un point à la suite duquel la fonction semble se stabiliser de façon linéaire. Ce point sera notre seuil.

Voici notre Mean Excess Plot :



Nous choisissons de positionner notre seuil à $u = 0{,}023$ puisque les points semblent former une fonction linéaire à la droite de celui-ci, bien que cela soit moins évident pour u supérieur ou égal à 0,05.

Ce seuil correspond au quantile d'ordre 97,50% de la distribution de nos pertes journalières. Cela signifie que nous travaillons avec les 214 valeurs de pertes les plus extrêmes.

Maintenant que nous avons notre seuil, nous pouvons calculer nos estimateurs des paramètres de la loi GPD par la méthode des moments sur nos données définies comme tel : $Y = X - u$

Pour cela, nous utilisons d'abord les formules de la moyenne et de la variance d'une loi GPD :

$$E[Y] = \frac{\sigma}{1-\gamma} \quad \text{et} \quad Var(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}$$

On a également :

$$\bar{Y} = \frac{\sum_{i=1}^{N_u} Y_i}{N_u} \quad \text{et} \quad S_Y^2 = \frac{\sum_{i=1}^{N_u} (Y_i - \bar{Y})^2}{N_u - 1}$$

On obtient donc les formules des estimateurs des moments de la loi GPD :

$$\hat{\gamma} = \frac{1}{2}\left(1 - \frac{\bar{Y}^2}{S_Y^2}\right)^2 \quad \text{et} \quad \hat{\sigma} = \bar{Y}\left(\frac{1}{2} + \frac{\bar{Y}^2}{S_Y^2}\right)$$

Et nous obtenons donc :

$$\hat{\gamma} = 0{,}313 \quad \text{et} \quad \hat{\sigma} = 0{,}0101$$

De plus, nous pouvons également estimer ces paramètres par la méthode du maximum de vraisemblance. En utilisant cette dernière, les paramètres estimés sont :

$$\hat{\gamma} = 0{,}382 \quad \text{et} \quad \hat{\sigma} = 0{,}00726$$

Ainsi, pour les 2 méthodes d'estimation des paramètres de la loi GPD, le paramètre de forme est inférieur à celui obtenu grâce à la méthode des Block Maxima. De ce fait, notre queue de distribution semble légèrement moins épaisse qu'au premier abord, bien que nous restions toujours dans le domaine des queues lourdes.

B) Hill estimator

La méthode de l'estimateur de Hill peut s'appliquer dans l'hypothèse où la queue de distribution suit une distribution de Pareto ; en d'autres termes que le paramètre de forme soit supérieur à 0. Comme cela a été le cas pour les deux précédentes méthodes, nous pouvons utiliser avec assez de confiance cet estimateur.

Soient $Z_{1,n} \leq \cdots \leq Z_{n,n}$ des statistiques d'ordre. L'estimateur de Hill est basé sur les k plus grandes statistiques d'ordre et se définit ainsi :

$$H_{k,n} = \frac{1}{k}\sum_{i=0}^{k-1} \ln\left(\frac{Z_{n-i,n}}{Z_{n-k,n}}\right)$$

Pour $k = k(n) \to \infty$ et $\lim_{n\to\infty} \frac{k}{n} = 0$; $H_{k,n} \to \alpha^{-1}$ en probabilité lorsque $n \to \infty$.

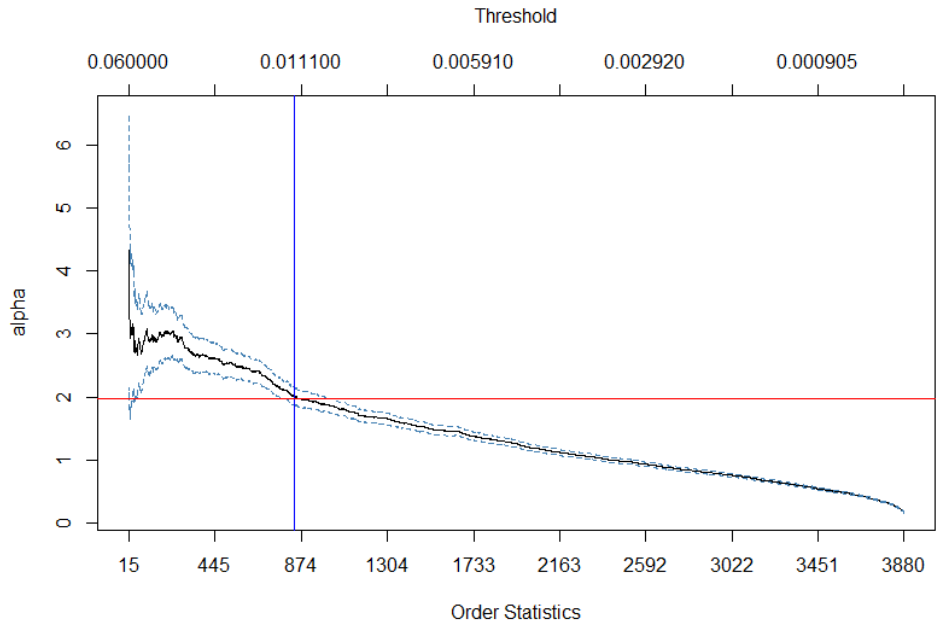
En appliquant la condition 2RV, nous pouvons montrer que l'estimateur de Hill est asymptotiquement normal :

$$\sqrt{k}\left(H_{k,n} - \alpha^{-1}\right) \to N(0, \alpha^{-2})$$

A nouveau, la difficulté de l'estimateur de Hill consiste à choisir un seuil. Pour cela, il faut ici tracer un Hill Plot via R. Ce dernier prend en abscisses les statistiques d'ordre de nos « log returns » négatifs et en ordonnées le coefficient alpha. En d'autres termes, nous visualisons différentes valeurs de alpha suivant le nombre de plus grandes valeurs que nous utilisons.

Ce nombre doit être choisi juste avant que la courbe du Hill Plot se stabilise, souvent à la suite d'une chute importante de celle-ci. Théoriquement, cela permet d'obtenir un alpha optimal puisque nous nous intéressons alors qu'aux statistiques d'ordre réellement pertinentes pour une analyse de valeurs extrêmes.

Voici le Hill Plot que nous obtenons :



Compte tenu de la courbe et des recommandations ci-dessus, nous choisissons un seuil à 840. Cela signifie que nous travaillons alors avec les 840 plus grandes valeurs de notre série statistique, ce qui correspond au quantile d'ordre 90,15% de notre distribution. Il est donc possible que nous prenons trop d'observations. Cependant, nous décidons de garder ce seuil étant donné que c'est celui qui nous semble le plus en accord avec les recommandations. Cela démontre que le choix graphique du seuil est loin d'être une évidence.

Nous prenons le alpha correspondant à notre seuil et nous obtenons :

$$\hat{\alpha} = 1{,}980959$$

Duquel nous déduisons notre paramètre de forme :

$$\hat{\gamma} = \frac{1}{\hat{\alpha}} = 0{,}504806$$

Nous pouvons également calculer un intervalle de confiance à 95%, également visible sur le Hill plot de part et d'autre de la courbe :

$$\hat{\gamma} \in [0{,}3708408\ ;\ 0{,}6387712]$$

Ainsi, nous obtenons un paramètre de forme qui se rapproche sensiblement plus de celui que nous obtenons avec la méthode des Block Maxima que de la méthode du Mean Excess Plot. La principale différence avec la méthode des Block Maxima réside dans le fait que l'intervalle de confiance est bien plus fin. Cela nous conforte dans notre hypothèse selon laquelle nous nous trouvons avec une queue épaisse. Nous pouvons également d'ores et déjà remarquer l'impact de la méthode et du seuil utilisé, puisque la valeur de notre paramètre de forme est ici assez différente de celle estimée à partir de la méthode du Mean Excess Plot.

C) QQ estimator

Enfin, nous présentons ici une dernière méthode liée au POT nommée QQ estimator. Son nom est dû au fait que nous utilisons un QQ-plot afin de comparer les quantiles théoriques d'une loi de Pareto et les quantiles empiriques de notre distribution pour déterminer α^{-1}.

Considérons $U_{1,n} \leq \cdots \leq U_{n,n}$ des variables aléatoires i.i.d. distribuées uniformément sur [0,1].

On a : $\mathbb{E}[U_{i,n}] = \frac{i}{n+1}$

Ainsi, le graphique $\left\{\left(\frac{i}{n+1}, U_{i,n}\right), 1 \leq i \leq n\right\}$ devrait présenter une courbe qui s'approche d'une fonction linéaire.

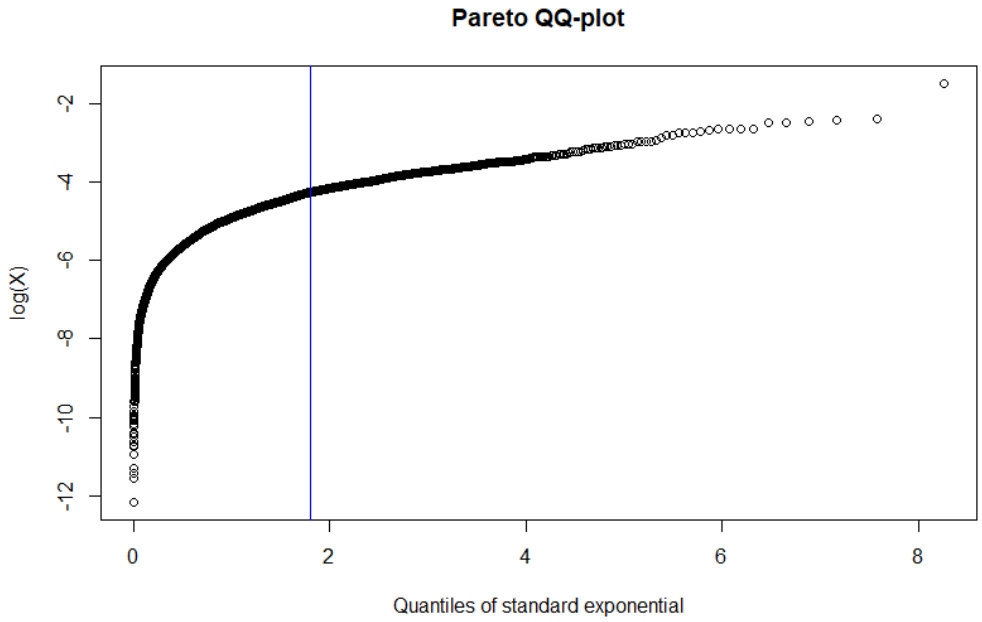
Ainsi, si l'on considère maintenant des variables aléatoires $X_{1,n} \leq \cdots \leq X_{n,n}$ i.i.d. que l'on suspecte de suivre une distribution continue G, le graphique $\left\{\left(\frac{i}{n+1}, G(X_{i,n})\right), 1 \leq i \leq n\right\}$ devrait donc lui aussi afficher une courbe approchant une fonction linéaire, tout comme le graphique $\left\{\left(G^{-1}\left(\frac{i}{n+1}\right), X_{i,n}\right), 1 \leq i \leq n\right\}$. Ce dernier graphique est en réalité un QQ-plot puisque $G^{-1}\left(\frac{i}{n+1}\right)$ sont les quantiles théoriques de la distribution G alors que les $X_{i,n}$ en sont les quantiles empiriques.

A présent, si l'on prend comme hypothèse que la distribution G est une fonction de distribution $G_{\mu,\sigma}$ telle que $G_{\mu,\sigma}(x) = G_{0,1}\left(\frac{x-\mu}{\sigma}\right)$; nous pouvons alors tracer le graphique $\left\{\left(G_{0,1}^{-1}\left(\frac{i}{n+1}\right), X_{i,n}\right), 1 \leq i \leq n\right\}$ pour vérifier si la courbe approche bien une fonction linéaire. Si tel est le cas, nous pouvons alors estimer le coefficient directeur σ et l'ordonnée à l'origine μ de la courbe.

Afin de ramener tout ceci à la méthode du QQ estimator, nous considérons ici des variables aléatoires $Z_{1,n} \leq \cdots \leq Z_{n,n}$ i.i.d. dont nous supposons qu'elles suivent une fonction de distribution de Pareto de paramètre α. Ainsi, en appliquant la méthode décrite ci-dessus aux variables aléatoires $\ln(Z_{i,n})$, nous pouvons tracer le graphique suivant qui est assimilé au QQ-plot de Pareto : $\left\{\left(-ln\left(1-\frac{i}{n+1}\right), \ln(Z_{i,n})\right), 1 \leq i \leq n\right\}$.

Le coefficient directeur de ce QQ-plot de Pareto vaut donc $\frac{1}{\alpha}$ ce qui est égale à notre paramètre de forme γ.

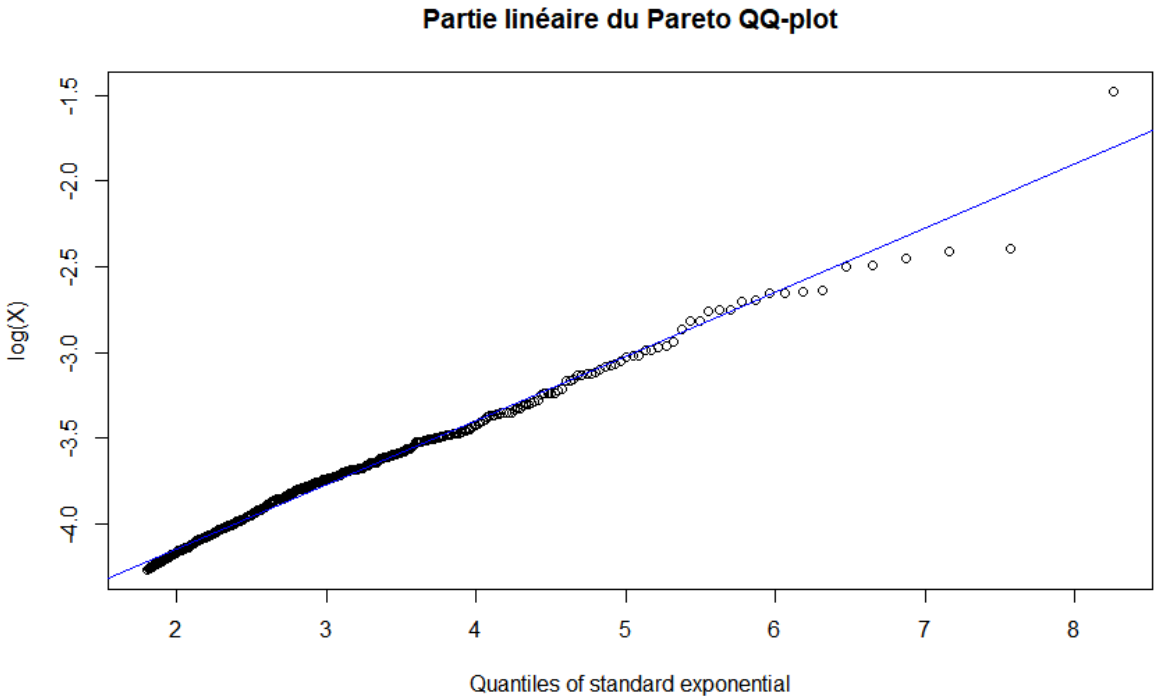
Voici le Pareto QQ-plot que nous obtenons, en excluant les valeurs négatives de nos log-returns étant donné que la fonction logarithme n'est définie que pour des valeurs supérieures à 0 :



Nous pouvons remarquer que nous n'obtenons pas un QQ-plot linéaire, mais il semble le devenir à partir d'un certain quantile. Ainsi, la difficulté de cette méthode réside dans le fait de trouver le bon quantile à partir duquel le QQ-plot semble devenir linéaire. Ici, nous choisissons de nous positionner pour les valeurs supérieures au quantile théorique 1,8 d'une distribution exponentielle standard.

Nous faisons donc l'hypothèse que nous sommes dans le cas d'une distribution de Pareto en excluant les valeurs inférieures au quantile théorique 1,8. Cela signifie ici que nous travaillons avec les valeurs supérieures au quantile d'ordre 92,48% de notre distribution, c'est-à-dire les 641 plus grandes statistiques d'ordre de nos pertes.

L'objectif est maintenant d'estimer le coefficient directeur de cette pente afin d'obtenir le QQ estimator. Pour cela, nous réalisons une régression linéaire :



Le QQ estimator correspondant au coefficient directeur estimé de la pente est :

$$Q_{k,n} = \hat{\gamma} = \frac{1}{\hat{\alpha}} = 0{,}3746871$$

Nous pouvons également calculer un intervalle de confiance à 95% :

$$\hat{\gamma} \in [0{,}3721244;\ 0{,}3772497]$$

Ainsi, nous nous trouvons à nouveau en présence d'une queue lourde. Le paramètre de forme estimé est très similaire à celui obtenu par la méthode du Means Excess Plot. En revanche, l'intervalle de confiance est ici beaucoup plus fin. Cela peut laisser penser que cette méthode nous a donné un estimateur plus précis du paramètre de forme.

Voici un tableau récapitulatif des méthodes utilisées :

Méthode	Loi fittée	Paramètre de forme	Seuil	N_u	Distance
Block Maxima	GEV	$\hat{\gamma} = 0{,}538$			
Mean Excess Plot	GPD	$\hat{\gamma} = 0{,}382$	$q_{97{,}50\%}$	214	$1{,}3410 \times 10^{-4}$
Hill Estimator	GPD	$\hat{\gamma} = 0{,}505$	$q_{90{,}15\%}$	840	
QQ Estimator	GPD	$\hat{\gamma} = 0{,}375$	$q_{92{,}48\%}$	641	

N_u correspond au nombre de valeurs de notre distribution supérieures au seuil.

Distance correspond à l'erreur quadratique moyenne entre la fonction de répartition empirique et la la fonction de répartition de la loi ajustée au niveau de la queue de distribution.

Pour conclure sur cette partie, la difficulté des méthodes liés au « Peak Over Threshold » réside dans le fait de devoir choisir arbitrairement des seuils sur un graphique. Parfois, ces seuils ne sont pas d'une évidence limpide. Cependant, chaque paramètre de forme estimé par les différentes méthodes est bien supérieur à 0. Nous pouvons donc être globalement confiant sur le fait que nous nous trouvons en présence d'une queue lourde.

IV – Recherche d'adéquation de loi sur l'ensemble de la distribution : GPD_G_GPD

A) Idée générale

Les parties précédentes nous ont permis d'avoir une meilleure idée de la distribution de la queue droite de notre distribution des pertes. Cependant on aimerait déterminer la loi de l'ensemble de notre distribution et non pas seulement celle des queues. Cette loi pourra par exemple nous servir dans la partie suivante afin de trouver la distribution bivariée adéquate.

L'idée est de produire un modèle hybride dans lequel on modélise le comportement moyen des données par une loi normale, et les comportements extrêmes par deux GPD différentes (une pour chaque queue). Le comportement moyen étant assuré par le théorème centrale limite, et le comportement extrême par le théorème de Pickands, nous indiquant que la queue de distribution peut être évaluée par une GPD à partir d'un certain seuil (que l'algorithme à venir déterminera).

B) Modèle hybride proposé

On définit la fonction de densité associée à notre modèle de la manière suivante :

$$h(x;\ u_1, u_2, \mu, \sigma, \xi_1, \xi_2) = \begin{cases} \gamma_1 g(u_1 - x; \xi_1, \beta_1) & si\ x \leq u_1 \\ \gamma_2 f(x; \mu; \sigma) & si\ x \in]u_1, u_2] \\ \gamma_3 g(x - u_2; \xi_2, \beta_2) & si\ x > u_2 \end{cases}$$

Avec :

- $g\ la\ densité\ de\ la\ GPD :\ g(x;\ \xi, \beta) = \frac{1}{\beta}\left(1 + \frac{\xi}{\beta}x\right)^{-1-\frac{1}{\xi}} \quad \forall x \geq 0\ \text{ pour } \xi > 0$
- ξ_1, ξ_2 les indices de de queue pour la GPD gauche et droite respectivement. On considère dans cette étude que nous traitons uniquement des queues lourdes (paramètres de queue strictement positif)
- β_1, β_2 paramètres d'échelle pour la GPD gauche et droite respectivement
- u_1, u_2 : seuils à partir desquels la distribution change de comportement

- $f(.;\mu,\sigma)$ *densité de la loi normale de moyenne* μ *et d'écart type* σ
- $\gamma_1, \gamma_2, \gamma_3$ poids associés à chaque densité, déterminés de sorte que la fonction h soit une densité et qu'elle soit continue.

Les hypothèses de continuité et de dérivabilité aux points de jonction nous permettent de réduire le nombre de paramètres à estimer. On peut ainsi explicitement exprimer les paramètres $\beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_3$ à partir des autres :

$$\begin{cases} \beta_1 = -\dfrac{(1+\xi_1)\sigma^2}{u_1-\mu} \;;\; \beta_2 = \dfrac{(1+\xi_2)\sigma^2}{u_2-\mu} \\ \gamma_2 = \left[\beta_1 f(u_1;\mu;\sigma) + \beta_2 f(u_2;\mu;\sigma) + \left(F(u_2,\mu,\sigma) - F(u_1,\mu,\sigma)\right)\right]^{-1} \\ \gamma_1 = \beta_1\gamma_2 f(u_1;\mu;\sigma) \;\; ; \;\; \gamma_3 = \beta_2\gamma_2 f(u_1;\mu;\sigma) \end{cases}$$

Avec F la fonction de répartition de la loi normale

On peut également écrire la fonction de répartition de notre loi hybride :

$$H(x;\, u_1, u_2, \mu, \sigma, \xi_1, \xi_2) = \begin{cases} \gamma_1[1 - G(u_1 - x; \xi_1, \beta_1)] & si\ x \le u_1 \\ \gamma_1 + \gamma_2[F(x;\mu,\sigma) - F(u_1;\mu;\sigma)] & si\ x \in]u_1, u_2] \\ 1 - \gamma_3[1 - G(x - u_2; \xi_2; \beta_2)] & si\ x > u_2 \end{cases}$$

Avec :

$G(x;\xi;\beta) = 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}}$ la fonction de répartition de la GPD (pour $\xi > 0$)

C) Algorithme itératif pour la détermination des paramètres

Les paramètres du modèle sont estimés via un algorithme itératif en 2 étapes :

- ➔ Estimer les paramètres de la gaussienne et des deux indices de queue des GPD en considérant les seuils de l'itération précédente (ce qui implique donc qu'il faut initialiser ces deux seuils).
- ➔ Estimer les deux seuils.

Ces estimations sont faites au sens des moindres carrés. On s'arrête lorsque la distance au sens des moindres carrés nous semble suffisamment petite ou bien quand on a atteint le nombre maximal d'itération qu'on ne veut pas dépasser.

Les problèmes de minimisation de cet algorithme sont effectués à l'aide de la méthode de Levenber Marquardt (exploitable par exemple via le package minpack.lm de R)

Un pseudo-code de l'algorithme est proposé ci-dessous.

1: Determination de la fonction de répartition empirique H_n du n-échantillon $X = (X_i)_{1 \le i \le n}$

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}, \quad \forall x \in \mathbb{R}.$$

2: Fixer les paramètres initiaux $\widetilde{u}_1^{(0)}$, $\widetilde{u}_2^{(0)}$, et ϵ

3: Procédure itérative

$k \leftarrow 1$

tant que $d\left(H(X;\theta^{(k)}) - H_n(X)\right) \ge \epsilon$ et $k < k_{max}$

a. Détermination de $\widetilde{\mu}^{(k)}$, $\widetilde{\sigma}^{(k)}$, $\widetilde{\xi}_1^{(k)}$ et $\widetilde{\xi}_2^{(k)}$ les estimateurs, respectivement, de μ, σ, ξ_1 et ξ_2

$$(\widetilde{\mu}^{(k)}, \widetilde{\sigma}^{(k)}, \widetilde{\xi}_1^{(k)}, \widetilde{\xi}_2^{(k)}) \leftarrow \underset{\substack{(\mu,\sigma)\in \mathbb{R}\times\mathbb{R}_+^* \\ (\xi_1,\xi_2)\in \mathbb{R}^2}}{argmin} \left\| H(X;\theta_1^{(k)}) - H_n(X) \right\|_2^2$$

où $\theta_1^{(k)} = \left[u_1^{(k-1)}, u_2^{(k-1)}, \mu, \sigma, \xi_1, \xi_2\right]$.

b. Détermination de $\widetilde{u}_1^{(k)}$ et $\widetilde{u}_2^{(k)}$ les estimateurs, respectivement, de u_1 et u_2

$$(\widetilde{u}_1^{(k)}, \widetilde{u}_2^{(k)}) \leftarrow \underset{(u_1,u_2)\in \mathbb{R}\times\mathbb{R}_+}{argmin} \left\| H(X;\theta_2^{(k)}) - H_n(X) \right\|_2^2$$

où $\theta_2^{(k)} = \left[u_1, u_2, \widetilde{\mu}^{(k)}, \widetilde{\sigma}^{(k)}, \widetilde{\xi}_1^{(k)}, \widetilde{\xi}_2^{(k)}\right]$

$k \leftarrow k + 1$

fin

retour $\theta^{(k)} = \left[\widetilde{u}_1^{(k)}, \widetilde{u}_2^{(k)}, \widetilde{\mu}^{(k)}, \widetilde{\sigma}^{(k)}, \widetilde{\xi}_1^{(k)}, \widetilde{\xi}_2^{(k)}\right]$

D) Résultats sur la fonction de pertes du S&P500

Les résultats de l'algorithme sur les pertes du S&P500 sont les suivants :

$\mu = -0.07\%$

$\sigma = 0.59\%$

$u_1 = -0.76\%$

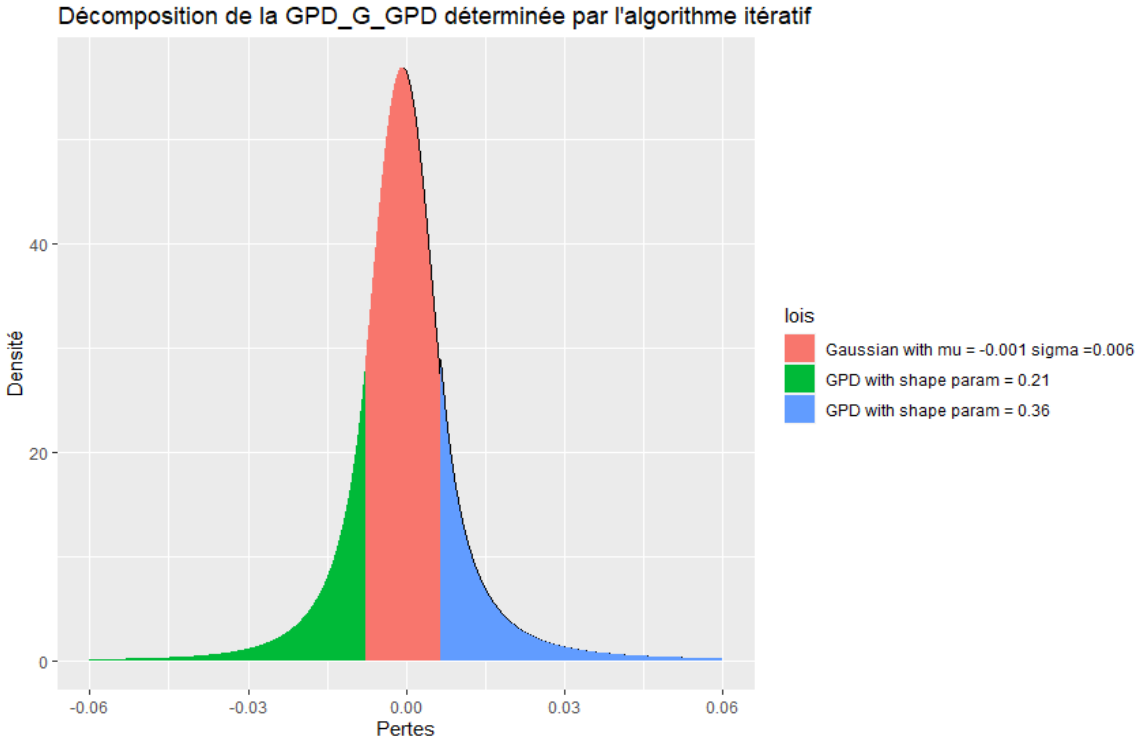
$u_2 = 0.65\%$

$\gamma_1 = 0.21$

$\gamma_2 = 0.36$

Pour plus de lisibilité, on les exprime en pourcentages.

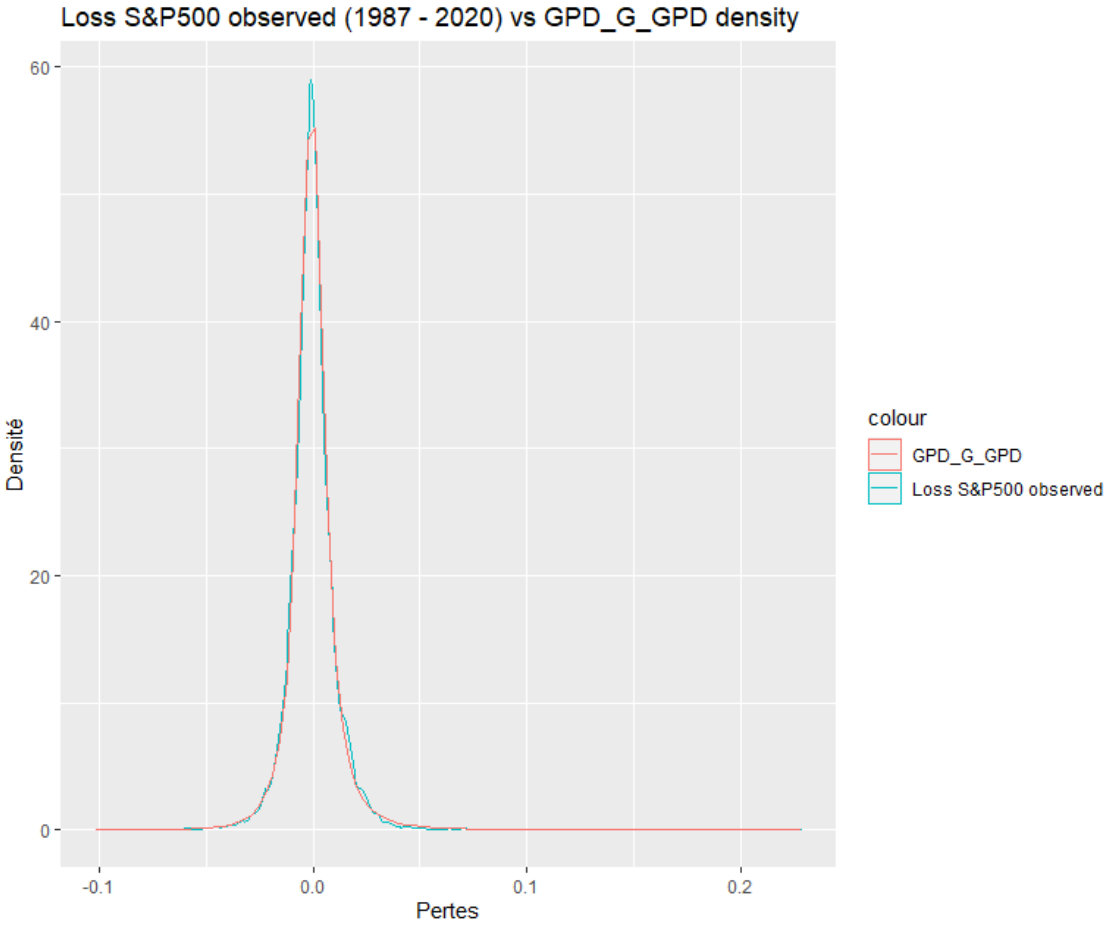
A partir de ces paramètres on peut ainsi en déduire nos autres paramètres et représenter la densité résultante sur un graphe.



Tout d'abord on remarque que la queue des pertes « positives » est plus épaisse que la queue des pertes négatives (i.e des gains), ce qui paraît plutôt cohérent, les log returns les plus extrêmes sont plus souvent des pertes liées à des crises financières.

De plus, le paramètre de forme des pertes est ici de 0.36, ce qui est plutôt cohérent avec le résultat des méthodes précédentes.

Maintenant que l'on a déterminé une fonction de densité théorique pour notre fonction de perte, on peut la comparer avec la densité de nos observations observées depuis 1987.



L'adéquation de notre loi aux observations est plutôt convaincante et semble beaucoup mieux correspondre à notre loi, que la simple fonction gaussienne représentée dans la première description des données.

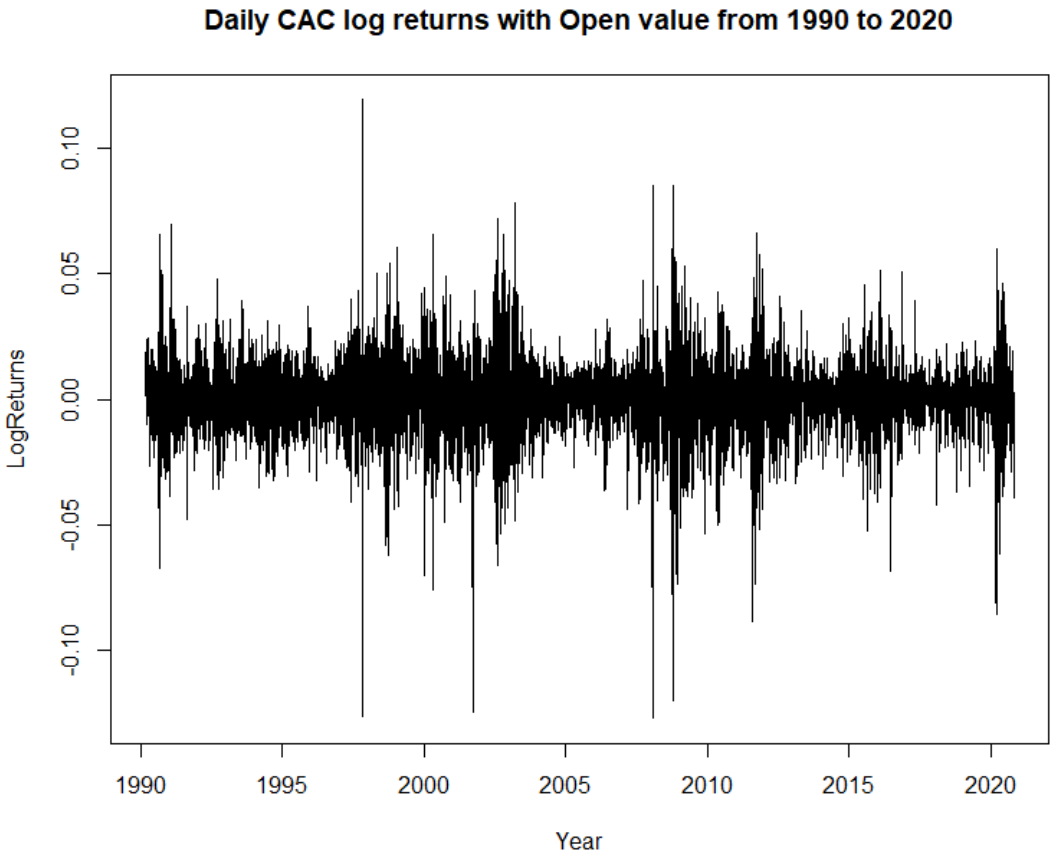
V – Etude bivariée du S&P500 et du CAC40

L'idée de cette partie est d'étudier la dépendance entre l'indice du S&P500 et l'indice français du CAC40. Etudier la corrélation entre ces deux séries peut être utile dans le contexte de l'assurance car il est essentiel dans le cadre d'une perte de savoir si on pourra compenser cette perte sur un autre marché. Ou si au contraire une perte sur le marché américain va aussi aller de pair avec une perte sur le marché français.

A) Description et visualisation des données

a) Description des données du CAC40

Nous ne disposons des données du CAC40 que depuis mars 1990. Nous allons donc faire notre analyse de mars 1990 à octobre 2020, en considérant uniquement les jours qui sont travaillés à la fois aux Etats-Unis et en France (les jours travaillés n'étant pas forcément les mêmes en fonction des pays).



A première vu il semble que la volatilité est plus importante que pour le S&P500. On peut le vérifier à l'aide du tableau des statistiques descriptives ci-après.

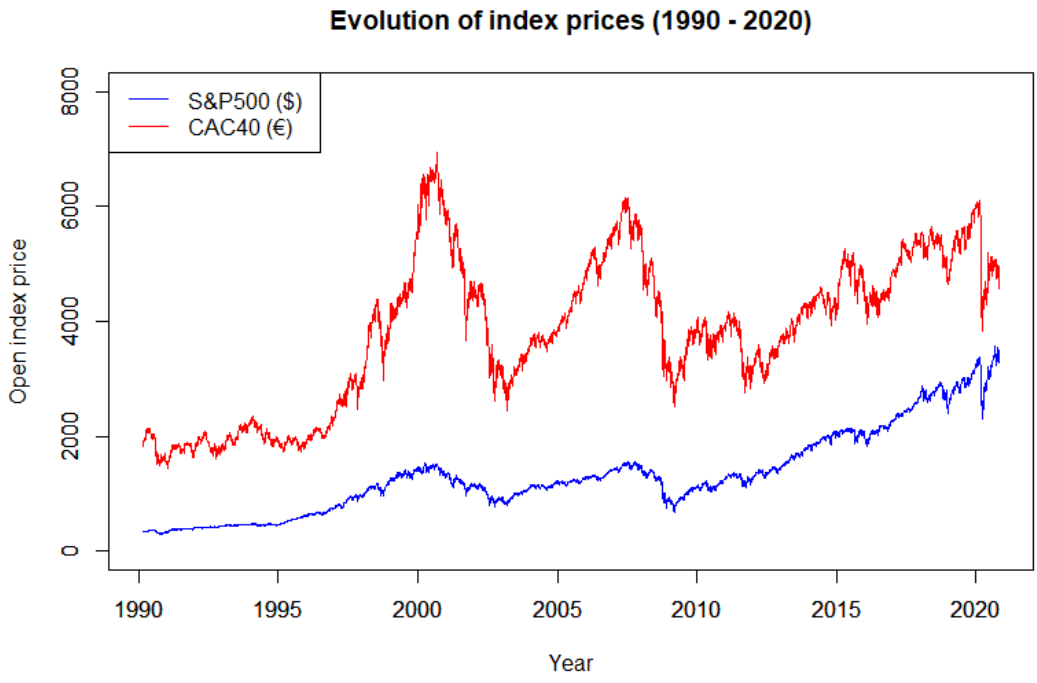
Comparaison des statistiques des log returns pour la période 1990-2020

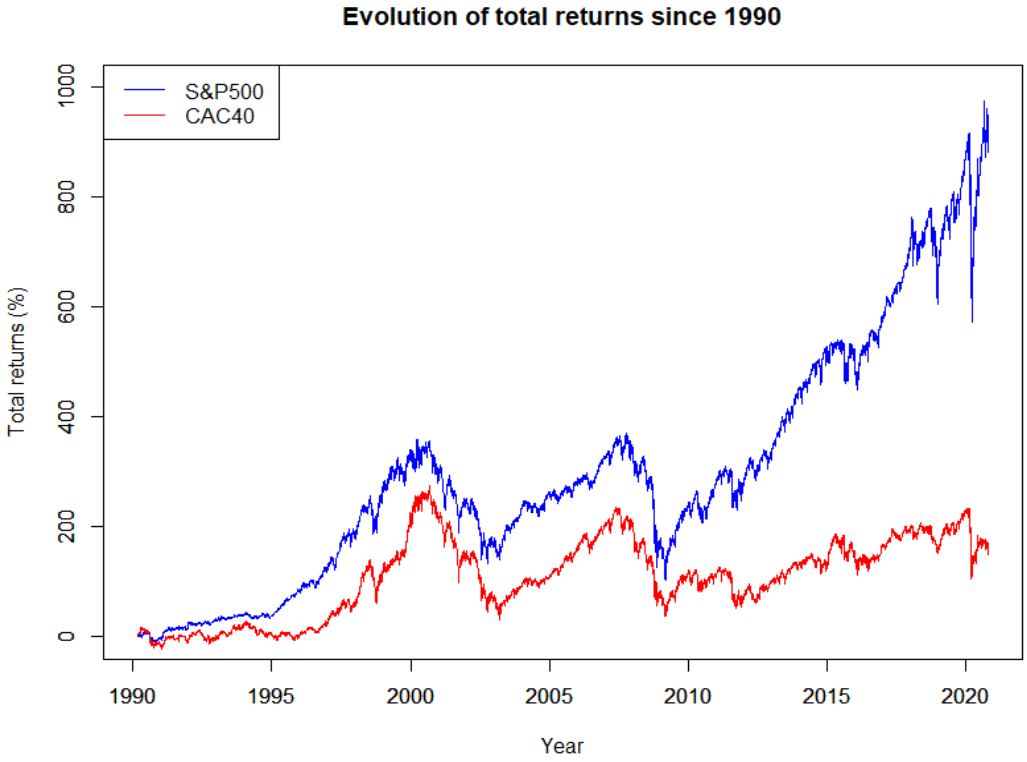
	S&P LogReturns	CAC LogReturns
Moyenne	0.03%	0.012%
Variance	0.012%	0.021%
Minimum	-9.115%	-12.687%
1er quartile	-0.433%	-0.69%
médiane	0.069%	0.067%
3eme quartile	0.553%	0.756%
maximum	10.139%	11.928%
Skewness	-0.339	-0.471
Kurtosis	8.067	6.92

Il est clair que les rendements du CAC40 sont beaucoup plus volatiles que ceux du S&P500. En effet la variance est pratiquement deux fois plus importante et les quartiles sont également beaucoup plus étalés pour le CAC40. Le coefficient d'asymétrie du CAC est également inférieur pour le CAC indiquant une queue de distribution des pertes plus étalée.

b) Description bivariée des données du CAC et du S&P

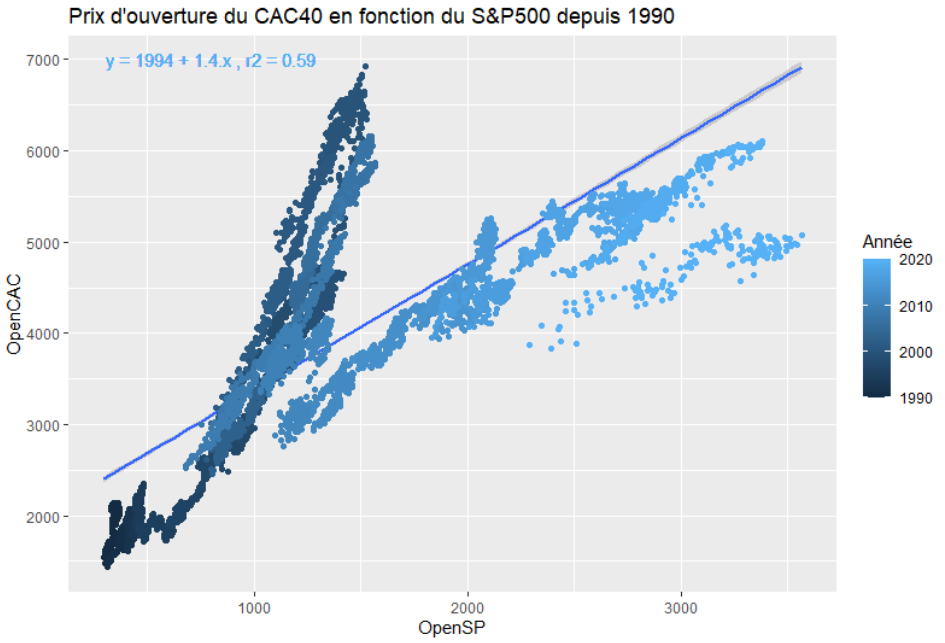
Avant de faire une analyse bivariée des rendements, on peut s'intéresser à l'évolution du prix d'ouverture des indices au cours du temps. La dépendance entre les indices semble ici assez évidente, la crise des années 2000 et 2008 à tout autant touché le S&P que le CAC. On peut cependant remarquer une reprise relativement rapide du S&P après les années 2008 qui a rapidement rattraper son prix d'avant crise, tandis que le CAC peine énormément à rattraper ses niveaux pré-crise.



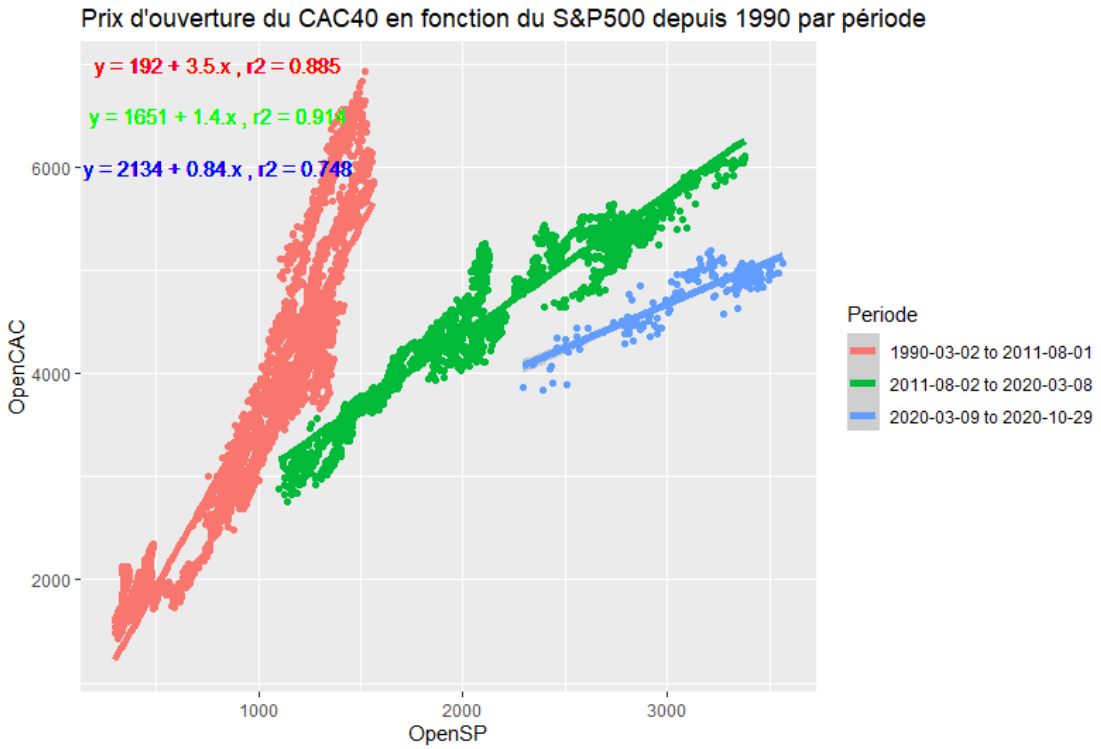


En comparant les rendements totaux depuis 1990, la surperformance du S&P comparé au CAC à partir des années 2010 est assez évidente. Surperformance qui n'empêche cependant pas une certaine dépendance avec de nombreux « mini » krach ces 10 dernières années qui se font ressentir sur les 2 indices.

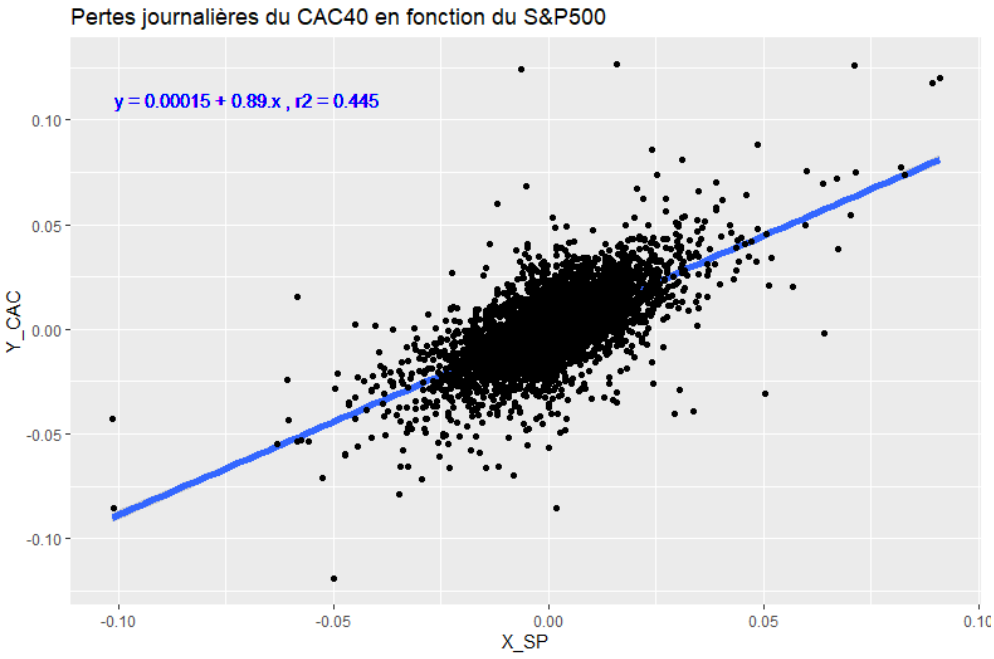
En représentant les valeurs d'ouverture du CAC40 en fonction de celles du S&P500, il est clair que les deux indices sont corrélés. Ob obtient un R2 de 0.59 en effectuant un modèle de régression linéaire sur ces données. Cependant plusieurs périodes bien distinctes semblent sortir du lot avec 3 relations linéaires différentes.



En classifiant les points par période déterminée avec soin au préalable on obtient le graphe suivant :



Ce graphique est intéressant car il permet d'identifier concrètement les dates marquantes à partir duquel le S&P500 à commencer à surperformer le CAC40, à savoir le 2 août 2011 en pleine tempête boursière de l'été 2011, ainsi que le 9 mars 2020 date à laquelle le CAC subit sa pire chute de l'histoire depuis 2008. Si la corrélation linéaire entre les prix d'ouverture est maintenant claire, celle-ci n'implique pas forcément une corrélation entre les pertes, il est donc important de l'analyser également.



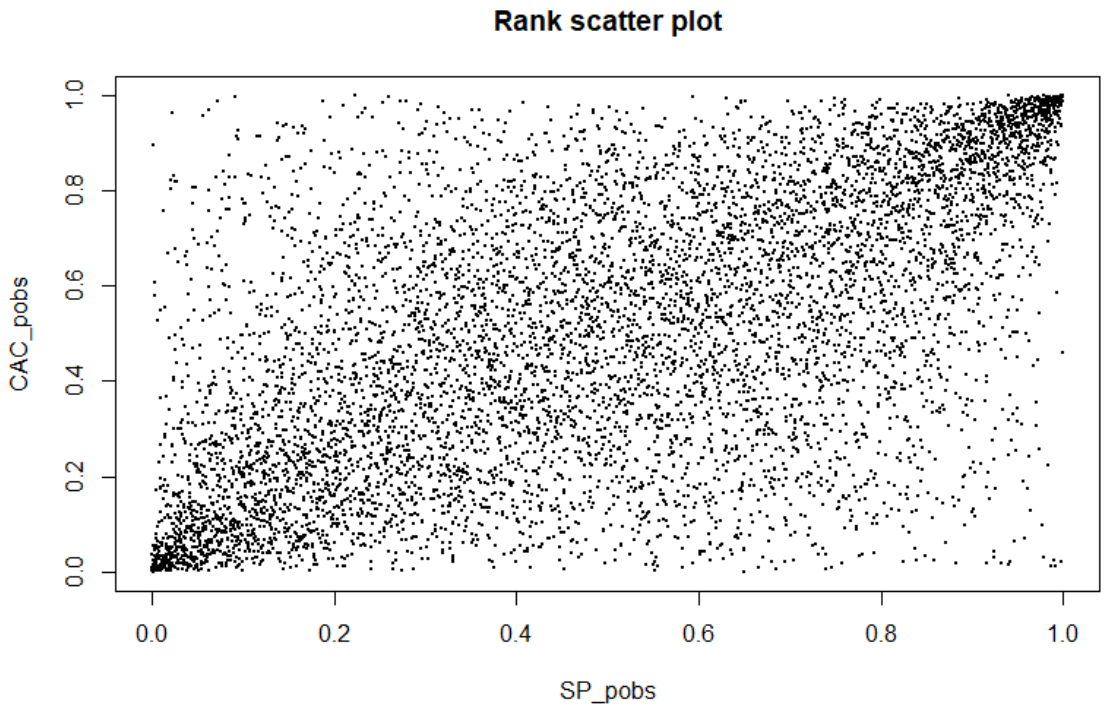
Ce graphe présente les pertes journalières du CAC40 en fonction du S&P500 (on considère à partir de maintenant la fonction $Pertes(x) = -logReturns(x)$ comme dans les parties précédentes). Avec un coefficient de détermination de 44,5% la relation linéaire est toujours là, mais beaucoup moins présente que lorsque l'on considérai uniquement les prix d'ouvertures. Pour mesurer cette dépendance, on peut également regarder les différents coefficients ci-dessous.

```
> cor(X_SP,Y_CAC,method = "pearson")
[1] 0.6670159
> cor(X_SP,Y_CAC,method = "kendall")
[1] 0.4454306
> cor(X_SP,Y_CAC,method = "spearman")
[1] 0.6085729
```

Le coefficient linéaire de Pearson est de 66,7% ce qui indique une corrélation linéaire assez importante.

On peut également s'intéresser aux corrélations de rang permettant d'étudier la corrélation des variables non pas à partir de leur valeur mais à partir de leur rang, captant ainsi des corrélations qui ne seraient pas forcément affines mais qui pourraient être monotone. Ici cela ne semble pas particulièrement le cas, vu que les corrélations sont de 0.44 pour le coefficient de Kendall et de 0.61 pour Spearman.

B) Etude des copules



Ce scatter plot est la représentation graphique du couple $\left(F_{X,n}(X_i), F_{Y,n}(Y_i)\right)$, X et Y étant respectivement les variables aléatoires associées au S&P500 et au CAC40, et n le nombre d'observations. Les fonctions de pseudo-observations $F_{X,n}$ et $F_{Y,n}$ sont définies comme :

$$F_{X,n}(x) = \frac{1}{n+1}\sum_{i=1}^{n} 1(X_i \leq x) \quad et \quad F_{Y,n}(x) = \frac{1}{n+1}\sum_{i=1}^{n} 1(Y_i \leq x)$$

Appliqué à nos observations, ces fonctions ne sont rien d'autres que :

$$F_{X,n}(X_i) = \frac{Rang(X_i)}{n+1}$$

Elles prennent donc leur valeur dans :

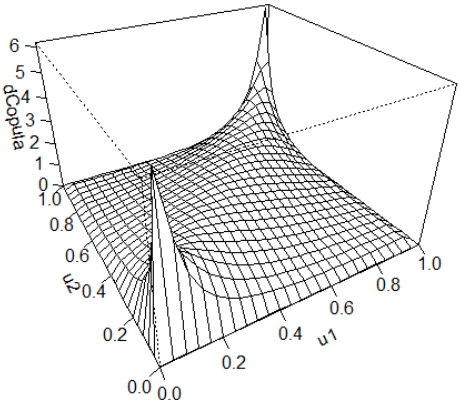
$$\left\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\right\}$$

La fonction de pseudo-observation est en fait très similaire à la fonction de distribution empirique, à la différence près que le facteur n'est pas $1/n$ mais $1/(n+1)$. Cela permet notamment d'éviter les problèmes d'évaluation de la densité aux frontières, le maximum des observations n'étant donc pas égal à 1, mais à $n/(n+1)$.

Concernant les résultats on observe tout d'abord une concentration plus importante des points dans la diagonale principale, ainsi les points ne sont pas uniformément distribués on peut donc en déduire une dépendance au niveau du comportement moyen. Concernant les extrêmes on remarque une concentration plus importante dans cette zone, mais il est difficile de clairement distinguer la répartition des points. Ces premières observations indiquent que nous somme dans le cas d'une copule de Student ou de Gauss. Cependant cette forme en étoile à tendance à nous diriger d'avantages sur une copule de Student.

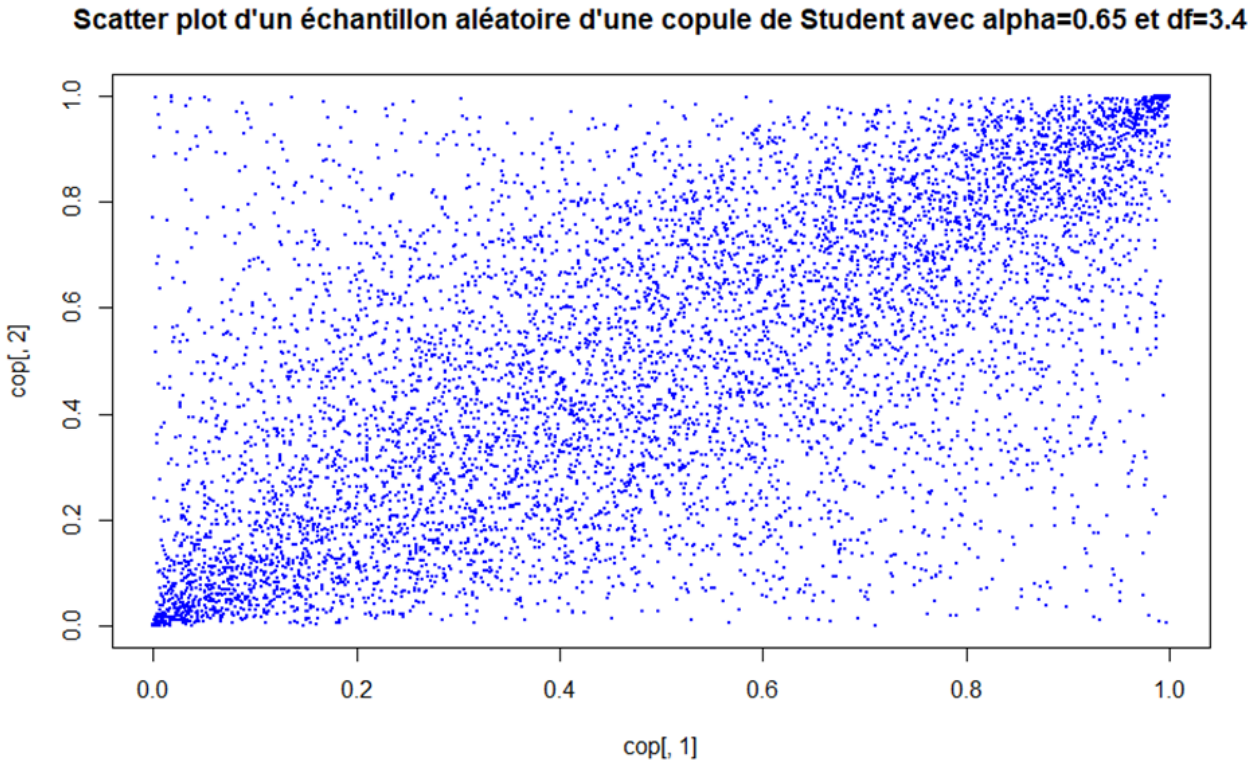
On peut vérifier directement cette hypothèse grâce à la fonction BiCopSelect du package VineCopula. Selon cette fonction la copule la plus cohérente avec nos pertes, est la copule de Student de paramètres $\alpha = 0.65\ et\ \nu = 3.4$

On peut maintenant représenter la densité de cette copule :



A première vue, la densité semble plutôt fidèle à nos observations avec une diagonale principale dominante, ainsi qu'une concentration marquée dans les extrêmes.

Pour pouvoir mieux comparer notre densité théorique avec nos observations, on peut simuler un échantillon aléatoire de même taille que nos données suivant la copule de Student en question :



On obtient un graphe très similaire à celui de nos observations.

C) Lois marginales

Afin de construire une loi jointe bivariée, le théorème de Sklar nous indique que nous avons besoin de deux éléments. D'un côté la structure de dépendance donnée par la copule et de l'autre les lois marginales, c'est-à-dire les lois de chacune des variables que l'on étudie.

- ➔ Concernant la copule nous venons de la déterminer.
- ➔ Concernant les lois marginales nous avons calculé dans la partie IV une distribution pour nos pertes du S&P à partir d'un modèle hybride qui semblait fournir des résultats très convaincants.

Il ne nous reste donc plus qu'à déterminer la loi suivie par les pertes du CAC40. Pour ce faire nous décidons d'appliquer la même méthode que pour le S&P500 en appliquant le modèle hybride.

On obtient les paramètres suivants :

$\mu = -0.09\%$

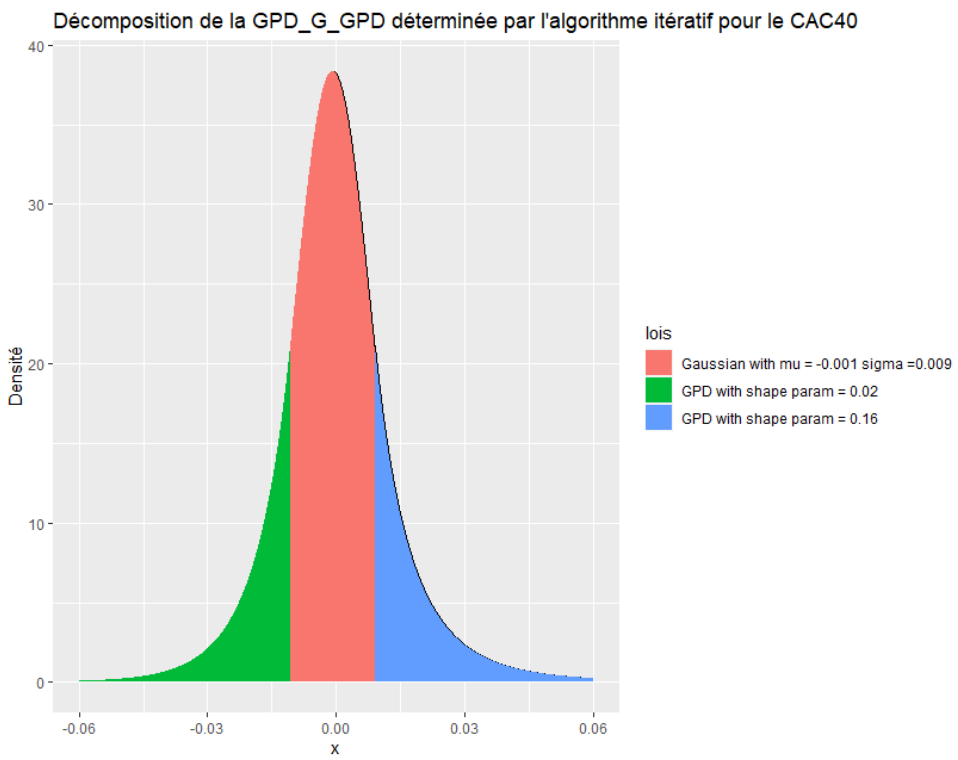
$\sigma = 0.9\%$

$u_1 = -1.06\%$

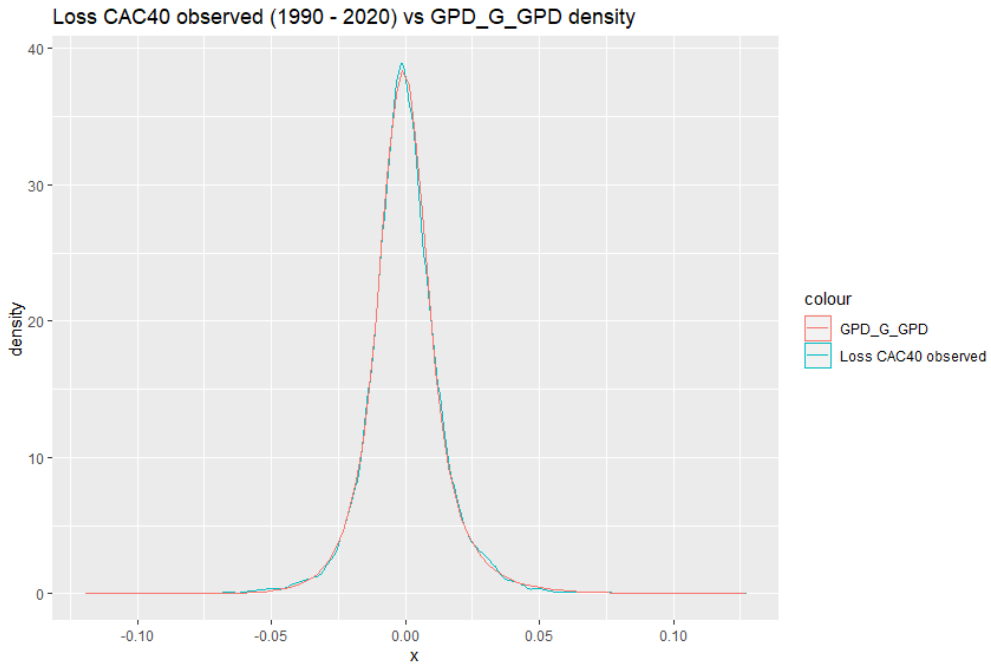
$u_2 = 0.9\%$

$\gamma_1 = 0.02$

$\gamma_2 = 0.16$



Ici nous avons fait l'hypothèse pour appliquer le modèle hybride que les deux queues de distribution des rendements du CAC40 étaient lourdes. Pour la queue des pertes négatives (i.e des gains), il apparait que le paramètre de forme est très proche de 0, indiquant peut-être que notre hypothèse est mauvaise. Pour s'assurer que notre loi est bien en adéquation avec nos observations, on peut superposer les deux :



La loi hybride résultante semble bien correspondre à nos observations.

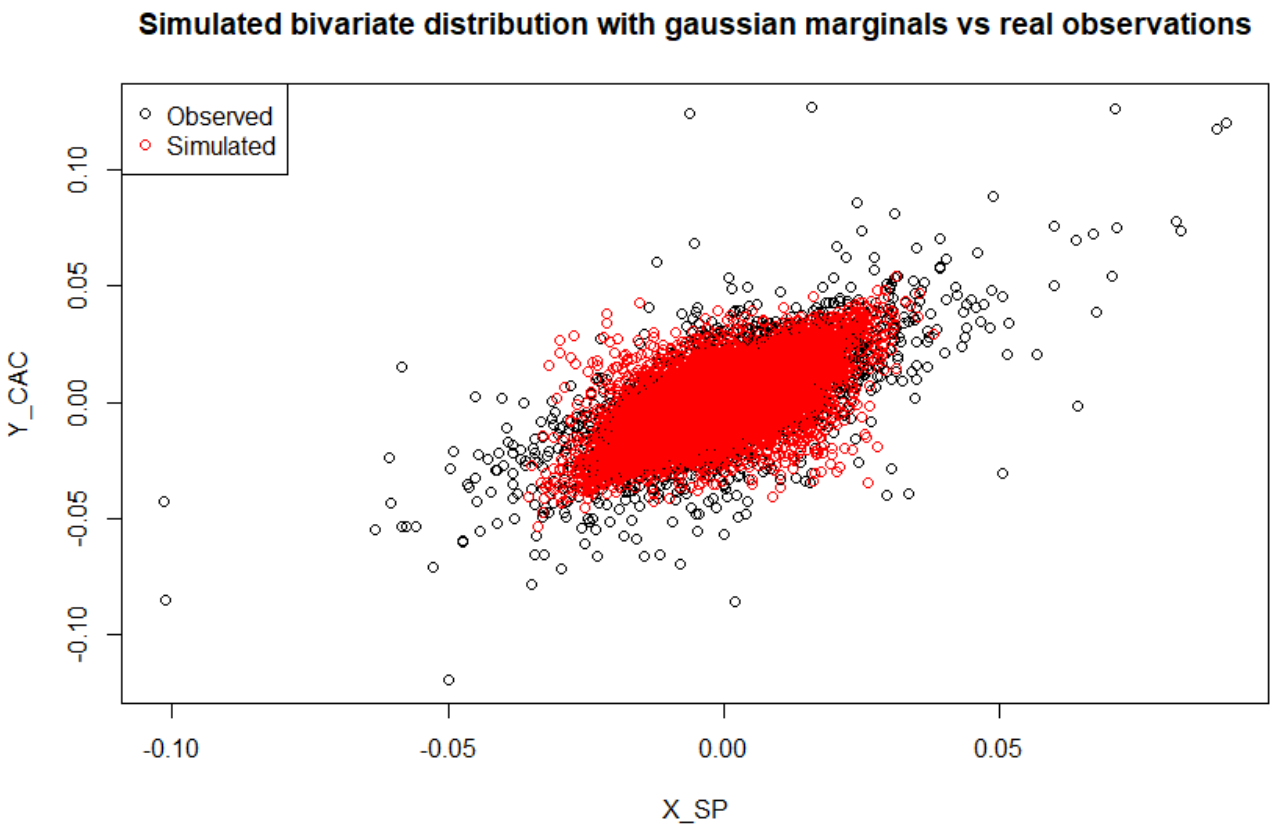
D) Distribution jointe

Maintenant que nous disposons de la copule adéquate et des lois marginales correspondantes à nos deux variables, on peut désormais représenter notre distribution jointe.

a) Distribution jointe avec des marginales gaussiennes

Avant d'utiliser les lois marginales précédemment calculées grâce au modèle hybride GPD_G_GPD on peut s'intéresser à la l'allure de notre distribution jointe si on avait considéré que nos données suivaient des lois gaussiennes comme c'est régulièrement le cas en finance.

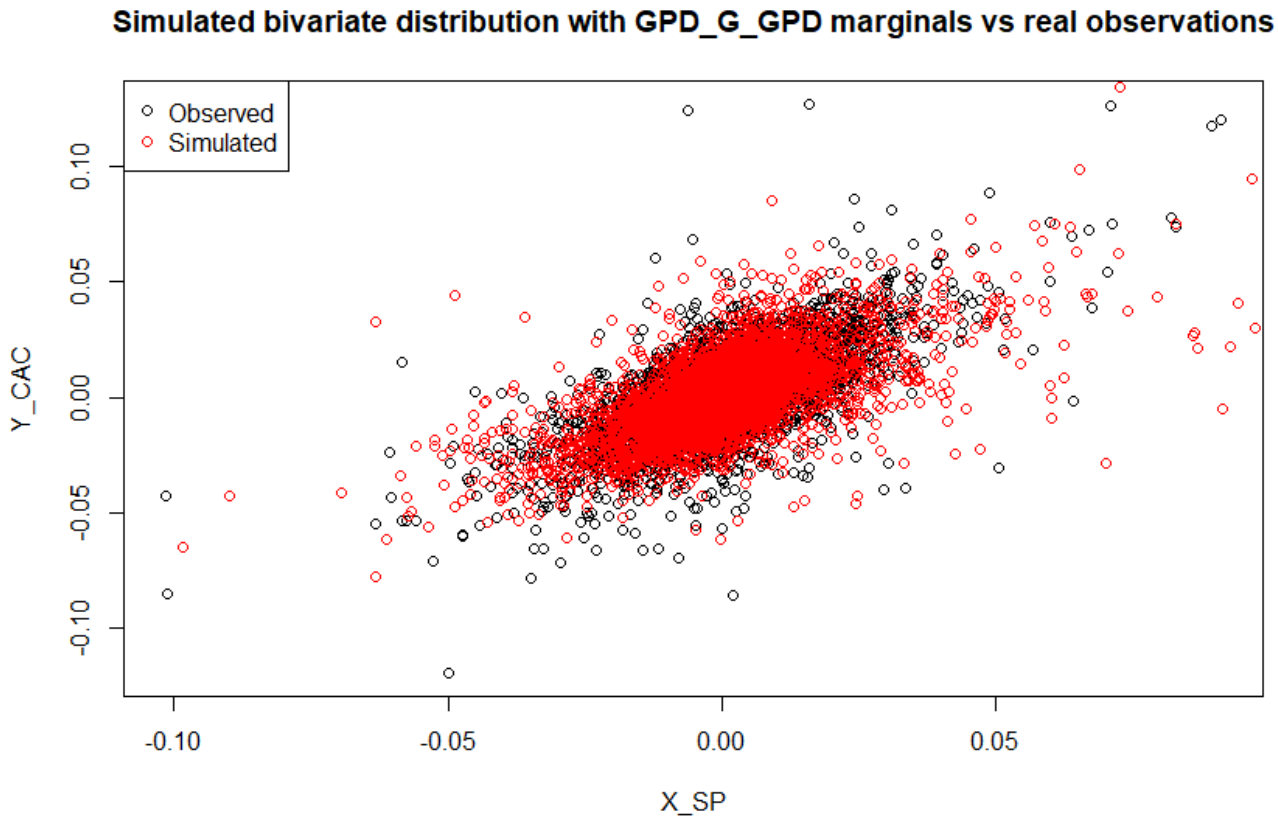
En simulant un échantillon aléatoire de la taille de nos observations qui suit notre loi bivariée on obtient les résultats suivants.



Le corps de nos observations semble correspondre. Cependant les extrêmes sont très mal simulés et ne convient donc pas du tout si l'on veut étudier la dépendance de nos risques dans les cas extrêmes.

b) Distribution jointe avec des marginales GPD_G_GPD

En considérant cette fois-ci que le S&P500 suit la loi hybride calculée dans la partie IV et que le CAC40 suit la loi hybride calculée dans la partie précédente, on obtient les résultats suivants :



Cette fois-ci la modélisation est beaucoup plus convaincante et des extrêmes similaires à ce qu'on a observé sur la période sont simulés par notre loi jointe. On voit ici tout l'intérêt de l'étude des valeurs extrêmes que l'on a mené, qui nous permet ici d'avoir une vision beaucoup plus fidèle de la réalité.

Conclusion

A faire

Références

1) Cours ISUP 2020 théorie des valeurs extrêmes M.Kratz

2) Debbabi, N., El Asmi, S., and Mboup, M. Distribution hybride pour la modélisation de données `a deux queues lourdes : Application sur les données neuronales. In Groupe d'Etudes du Traitement du Signal et des Images,GRETSI (2015).

3) Debbabi N., Kratz M., Mboup M. (2017). A self-calibrating method for heavy tailed data modeling. Application in neuroscience and finance.