Qinghe Gao
2659024
Wenlin Chen
2664925
Group 26

PCA, copula and EVT
Assignment3

2020/05/26

# 1 Introduction

In this report, we did PCA, FA, Copulas and EVT on our portfolio. Firstly, we perform PCA and FA on our portfolio to determine contribution of assets for our portfolio. And we compare the loading composition for two methods. Secondly, we select 5 pairs of assets according to their kendall rank correlation from low to high. Fit 5 kinds of copula (Gaussian, Student-t, Gumbel, Frank and Clayton copula) to each pair to find the best suitable copula and estimate parameters. Thirdly, we perform extreme value analysis on our one asset-Apple stock. We use POT method to get exceed data and fit GPD distribution. Then we calculate VaR and ES and compare them with historical method and student-t method.

# 2 Method and theory

There are total 10 assets in our portfolio and all the historical data is range from 01/04/2011 to 01/04/2020 and has 2235 trading days.

- Stock: Apple(USD), Louis Vuitton(EUR)
- Floating rate bonds: Based on USA 30 years(USD), Based on Germany 10 years(EUR)
- Indices: SPX(USD),AEX(EUR),DOW(USD)
- Commodities: Gold(USD),Silver(USD)
- Other: Bitcoin(USD)

## 2.1 PCA and FA

In this part, we introduce PCA and FA

Principal Component Analysis(PCA) is a statistical method to convert a set of variables that may have correlation into a set of linearly uncorrelated variables, which can reduce the dimension of data to decrease cost of time. For the steps of PCA first we normalize the data. Normalization of data is of vital significance. Because PCA calculates the weight based on the standard deviation of variables. Thus, we need to normalize the variables into same standard deviation. Then, we do PCA on variables and the number of new components is the same as number of assets. Furthermore, we analyze the eigenvalues and loadings of each component.

Factor analysis(FA) is quite different from the PCA. Because PCA aggregates factors but FA produce the new factors. And in process of FA it needs rotation. The purpose of rotation is to redistribute the variance ratio explained by each factor by changing the position of the coordinate axis so that its load factor is closer to 1 or 0, which can better explain and name the variables. And here we use the oblique rotations, which permits the factors to be correlated with one another. And it produces solutions with a simpler structure.

And we do both PCA and FA on our portfolio and compare loading of factors.

## 2.2 Copulas

Copula is a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform on the interval [0, 1]. If both the copula and marginal probability are known, then the multivariate joint distributed could be arrived. Here, Three categories and five kind of copulas are discussed. They are Elliptical copulas (Gaussian, Student-t copula), Archimedean copulas (Gumble, Frank and Clayton copula).

For 2-D joint distribution, the only parameter is $\theta$. And to select the best model, AIC is minimized.

$$AIC = 2p - 2L(\theta) \tag{1}$$

Where $p$ is the number of dimension and $L(\theta)$ is the function of log-likelihood.

### 2.3 EVT analysis

Extretme value analysis provides statistical method to evaluate and describe extreme event. Two methods can be used to evaluate extreme value: Peaks over Threshold method and Block Maxima method. In this report, we focus on Peaks over Threshold method(POT).

POT determines a certain threshold $u$ and collects all the data which excess $u$. And the excess distribution:

$$F_u(x) = P(X - u \leq x | X > u) = \frac{F(x+u) - F(x)}{1 - F(u)} \tag{2}$$

And according to extreme value theory equation 2's natural approximation is Generalized Pareto Distribution(GPD). The GPD equation:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \xi \neq 0 \\ 1 - e^{-x/\beta} & \xi = 0 \end{cases} \tag{3}$$

And there are two parameter $\xi, \beta$. $\xi$ parameter determines the shape of distribution and when $\xi > 0$ the tail of distribution is heavy. $\beta$ is scale.

In this report, first we choose Apple asset, determine the threshold $u$ and use the exceed data to estimate the parameters of GPD. And the we compare the VaR and ES with historical simulation method and Student-t distribution.

# 3 Result and discussion

### 3.1 PCA and FA

This part we do PCA and FA on our portfolio. First, figure 1 shows the results before and after normalization. We can see that before normalization scale of bitcoin is quite large, which will influence the result of PCA. Thus, after normalization for all assets standard deviation and scale of all assets are the same.

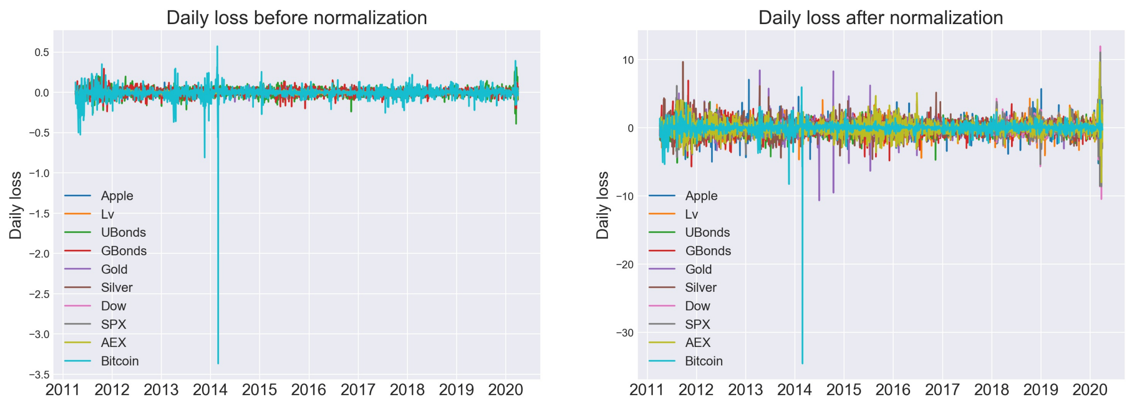



Figure 1: Daily loss of each asset before and after normalization

Figure 2a shows variance contribution of 10 components. We can see the first component have largest variance contribution: 3.95 and we see that first four components' variances are large than 1.Thus, according to Kaiser's Rule only the first four components can be retained. And from 2b we see 78.2% variance is made up by first four components. In short, first four components are retained.

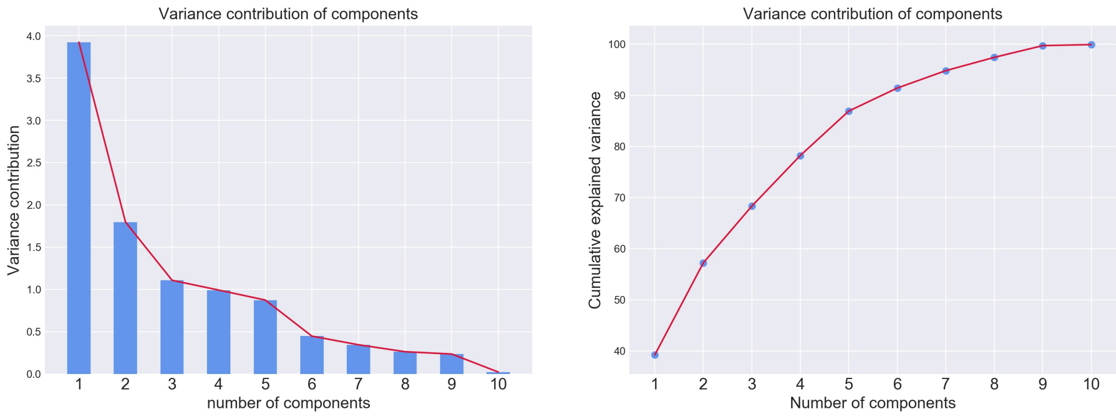



Figure 2: (a)Left plot is variance contribution of PCA each components.(b)Right plot is PCA cumulative explained variance.

Next we analyze the loading of PCA first four components. The first component explains 39.5% variance and we can see that the stocks, bonds and indices have large loading in the same direction for component 1. The possible explanation is component 1 is considered as market $\beta$, which quantifies the degree to which changes in average assets are related to changes in the entire market. Component 2 is basically a "commodities" component and we can see two commodities both have large loading. Additionally two bonds have small percentage in Component 2 from table 1. Component 3 is difficult to interpret. We think this component represents two different market place. Because it is clear to see that most assets from EU have positive loading and else from USA have negative loading. And different signal of loading means EU and USA market are negative correlated. As 5 shows bitcoin hardly has correlation with other assets. Thus, component 4 is only belong to Bitcoin and we can see the absolute value of loading even approaches to 1.

Table 1: PCA loading of each factor

| Component | Apple | Lv | UBonds | GBonds | Gold | Silver | Dow | SPX | AEX | Bitcoin |
|---|---|---|---|---|---|---|---|---|---|---|
| Component1 | 0.335 | 0.364 | 0.326 | 0.248 | -0.007 | 0.075 | 0.455 | 0.46 | 0.404 | 0.021 |
| Component2 | 0.07 | 0.002 | -0.215 | -0.231 | 0.677 | 0.657 | 0.053 | 0.053 | 0.024 | -0.001 |
| Component3 | -0.37 | 0.008 | 0.413 | 0.645 | 0.179 | 0.25 | -0.221 | -0.224 | 0.047 | -0.283 |
| Component4 | 0.161 | -0.0 | -0.089 | -0.184 | -0.068 | -0.051 | 0.06 | 0.069 | -0.036 | -0.957 |

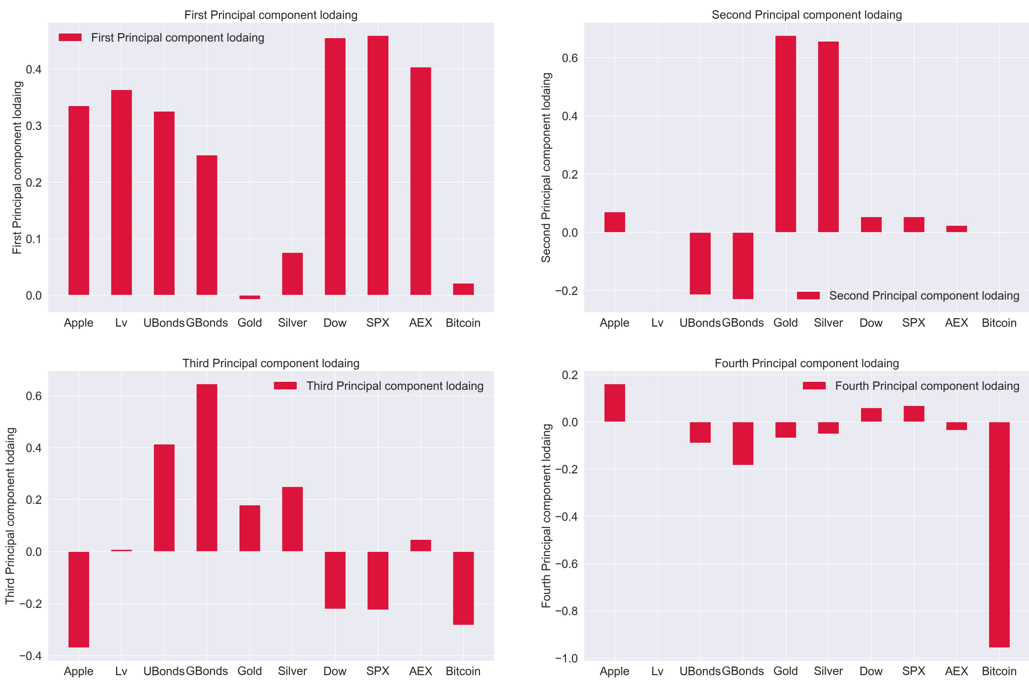



Figure 3: Loading of PCA first six components

Furthermore, we did FA(oblique rotations) to further compare loading composition of PCA. FA yields nine new factors and we choose first four interpretive factors which show as figure 4. Factors' structure from FA is much simpler than PCA. We can see from table 2 that each new factors have own distinguish large loading from each variables. For example, factor 1 has comparatively large loading on Apple stocks, SPX and DOW, which tell us that this factor can be explained by USA stock market. Factor 2 is obviously made up by commodities gold and silver. LV and AEX have the largest loading in factor 3, which can be interpreted as EU stock market. Factor 4 is made up of two bonds. Additionally, we can see that First 4 factors' dominated loading are large, which are all larger than 0.6 as table 2 shows. Additionally, we notice that bitcoin asset have extremely small in first four factors and we find only in factor 8 it has large loading. The possible explanation is because bitcoin has small correlation with the rest of assets, which will bring small contribution to the loading.

In short, FA is better at interpreting loading of factor because it can capture inter-correlations between variables. While PCA can use to reduce the dimension of correlated variables.

Table 2: FA loading of each factor

| Factor | Apple | Lv | UBonds | GBonds | Gold | Silver | Dow | SPX | AEX | Bitcoin |
|---|---|---|---|---|---|---|---|---|---|---|
| Factor 1 | 0.773 | -0.002 | 0.047 | -0.013 | 0.021 | -0.022 | 0.921 | 0.918 | 0.058 | -0.001 |
| Factor 2 | 0.001 | -0.001 | 0.001 | -0.001 | 0.869 | 0.853 | 0.002 | -0.004 | 0.002 | -0.000 |
| Factor 3 | -0.009 | 0.881 | -0.011 | 0.013 | -0.014 | 0.015 | -0.009 | 0.057 | 0.756 | 0.001 |
| Factor 4 | 0.002 | -0.001 | 0.662 | 0.800 | 0.002 | -0.002 | 0.006 | -0.005 | 0.009 | -0.000 |

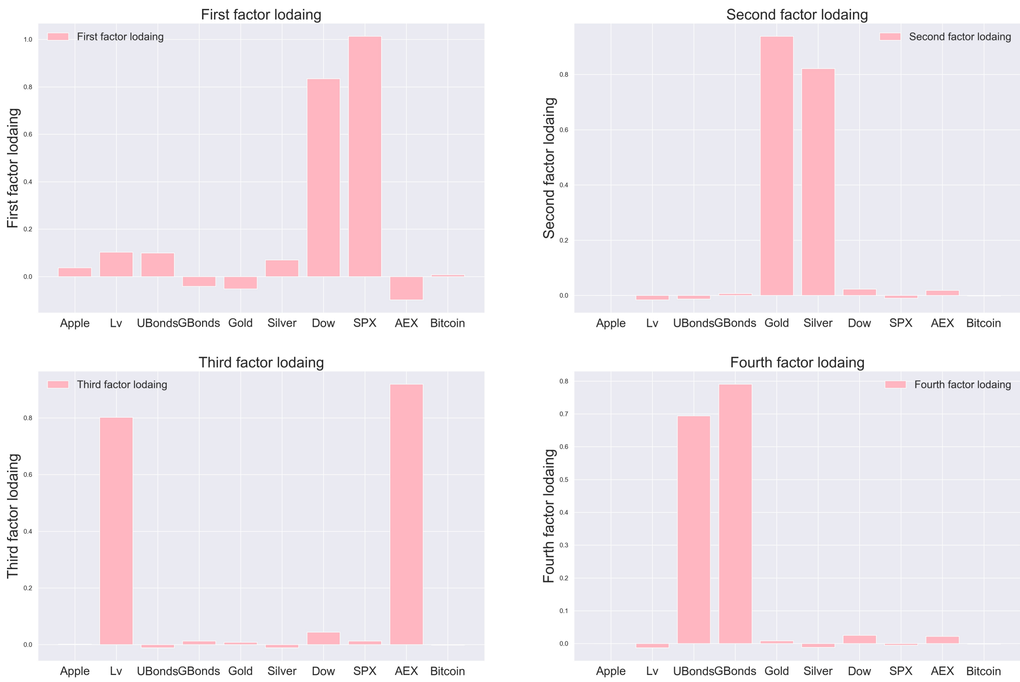



Figure 4: Loading of FA nine components

# 4 Copula

To calculate the value of portfolio, the dependence of each asset is a crucial factor to be considered. In this part, fitting copula to 5 pairs of assets selected from 10 assets, according to the correlation between pairs from low to high. Compared with Pearson's correlation, which assumes the distributions of data are normal, Spearmans rank and Kendall rank correlation are more suitable for data with unknown-distribution. The figure 5 shows the **Kendall rank correlation** between pairs of 10 assets.

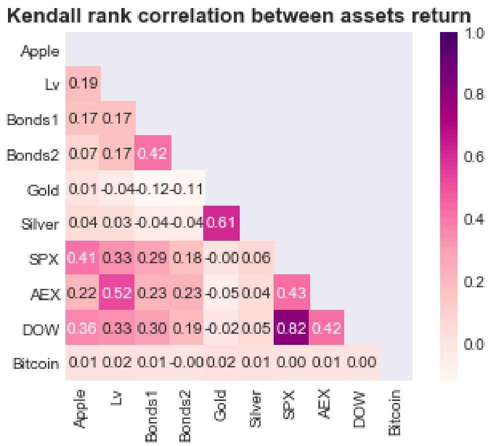



Figure 5: Kendall rank correlation between 10 assets return.

From the shades of pink color, we select 5 pairs of assets with negative, around zero or positive correlation. They are: Gold & bonds **(corr:-0.12)**, Gold & Apple **(corr:0.01)**, Lv & Apple **(corr:0.19)**, Silver & Gold **(corr:0.61)** and DOW & SPX **(corr:0.82)**.

The joint distributions of 5 pairs are plotted in the figure 4. From the upper left to the lower right picture, the correlation increasing from -0.12 to 0.82 $\Rightarrow$ this rule is also shown in these joint distributions, from disordered distribution (elliptical cluster) to centralized distribution (linear cluster).

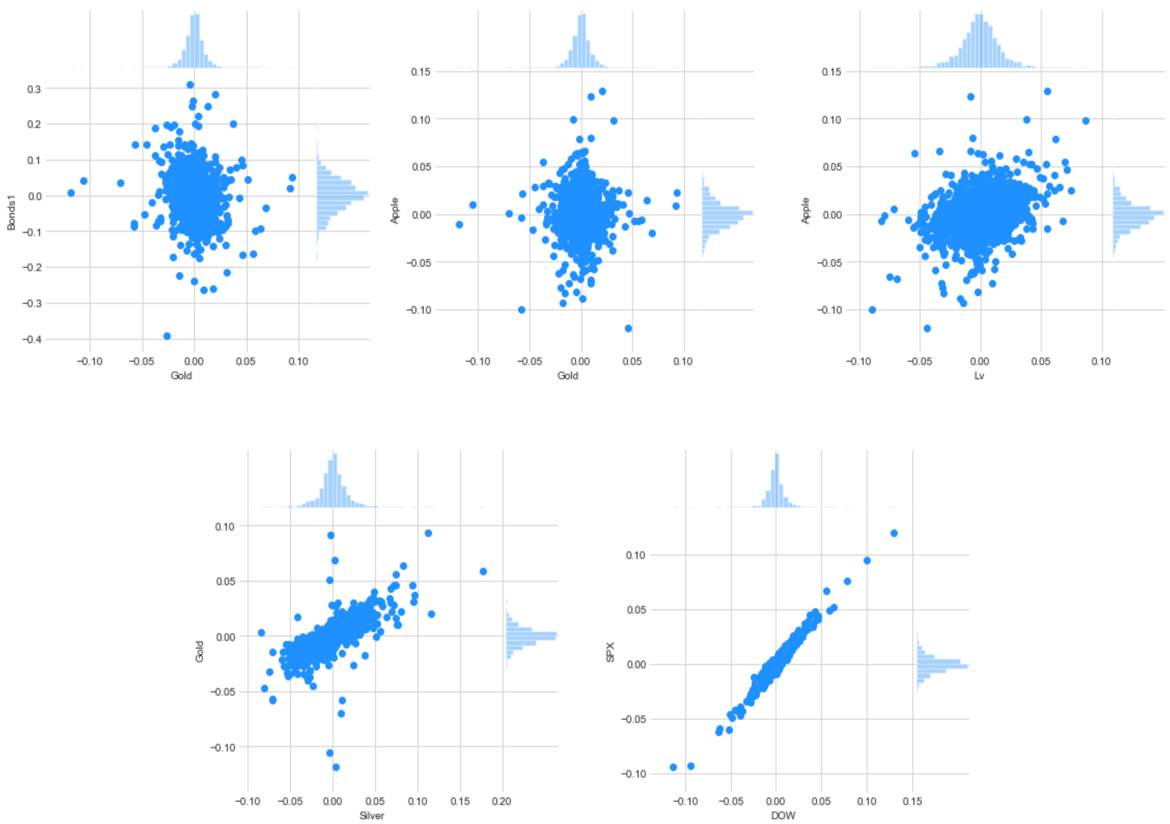



Figure 6: Joint distribution of each pair of assets.

After 25 times copulas fitting, the best copula including parameter for 2-dimension ($\theta$) and AIC, for each pair are shown in the Lower triangular of table 3. And in the upper triangular, the rest copulas methods with Log-likelihood (LLH) are listed.

**In summary**,

1. For positive correlated pair, scatter points of $u$ and $v$ are densly distributed in upper right and lower right tails. For negative correlation pair, like gold and bonds1, $u$ and $v$ are mainly located at upper left and lower right tail.

2. For strongly correlated pair, both the **density** at tail and **log-likelihood** are higher. For weakly correlated, both are lower, which proves that assets are close to 'independent' and $u$ and $v$ are close to uniform distributed.

3. Student-t distribution is more suitable for weakly correlated pairs in the following experiments.

Table 3: Best copula for each pair

| | Gold | Silver | Apple | lv | DOW | SPX | Bonds1 |
|---|---|---|---|---|---|---|---|
| Gold | | LLH Gaussian:1125.96<br>LLH Gumble:1164.80<br>LLH Frank:1131.78<br>LLH Clayton:879.59 | LLH Gaussian:0.11<br>LLH Gumble:1.05<br>LLH Frank:0.08<br>LLH Clayton:0.71 | | | | LLH Gaussian:35.29<br>LLH Gumble:0<br>LLH Clayton:8.89 |
| Silver | Student-t(df=5.07)<br>$\theta = 0.8147$<br>AIC=-1225.99 | | | | | | |
| Apple | Student-t(df=8.63)<br>$\theta = 0.0069$<br>AIC=-8.63 | | | LLH Gaussian:112.72<br>LLH Gumble:126.98<br>LLH Frank:92.24<br>LLH Clayton:83.17 | | | |
| Lv | | | Student-t(df=6.3)<br>$\theta = 0.2999$<br>AIC=-130.37 | | | | |
| DOW | | | | | | LLH Gaussian:2893.48<br>LLH Clayton:2380.68 | |
| SPX | | | | | Gumble<br>$\theta = 5.6698$<br>AIC=-2953.79 | | |
| Bonds1 | Student-t(df=4.54)<br>$\theta = -0.1909$<br>AIC=-75.58 | | | | | | |

The fitting process for each pair is shown below.

### 4.0.1 Fit copulas to pair1 - Gold and Bonds1 (correlation = -0.12)

Gold and Bonds1 have weak **negative** correlation, and their joint distribution in figure 4 are central-located ellipse with scattered points. This shows that they are not highly dependent.

In the figure 4.0.1, for each row, from left to right are 2-D kernel density estimation (kde), 3-D kde and scatter plot of $u$ and $v$. Fit 5 kinds of copula to Pair1, the Student-t copulas with 4.54 degree of freedom, has minimum AIC is the best copula to pair 1.

Table 4: Summary of pair1 fitting

| Copula method | $\theta$ | log-likelihood | AIC |
|---|---|---|---|
| Gaussian | -0.1773 | 35.29 | -31.29 |
| Student-t (df=4.54) | -0.1909 | 79.58 | -75.58 |
| Gumble | 1 | 0 | -4 |
| Clayton | -0.0659 | 8.89 | -4.89 |

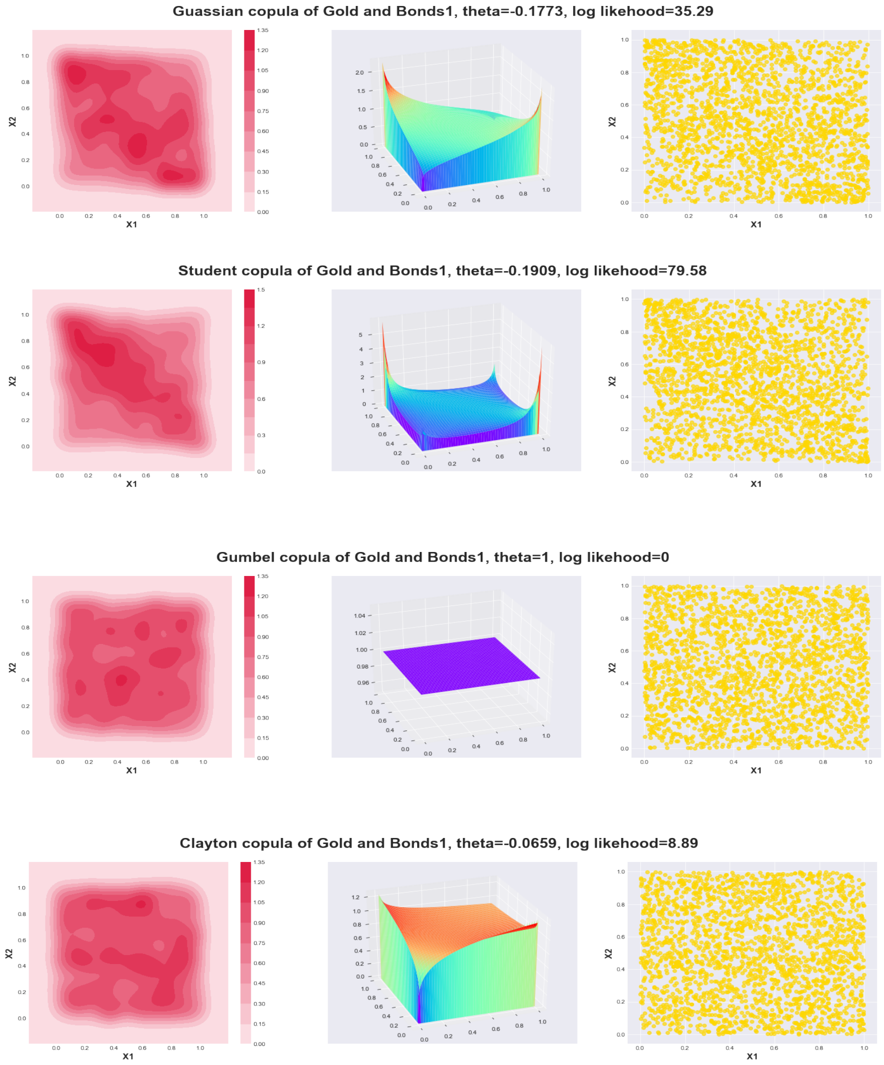



Figure 7: Fit different copulas to Gold and Bonds1 joint distribution.

#### 4.0.2 Fit copulas to pair2 - Gold and Apple (correlation = 0.01)

Gold and Apple have very weak correlation, which is close to zero, and their joint distribution in figure 4 are centrally elliptical cluster. This shows that they seems to be not correlated. Fit 5 kinds of copula to Pair2, the Student-t copulas with 8.63 degree of freedom, has minimum AIC is the best copula to pair 2.

Table 5: Summary of pair2 fitting

| Copula method | $\theta$ | log-likelihood | AIC |
|---|---|---|---|
| Gaussian | 0.0102 | 0.11 | 3.89 |
| Student-t (df=8.63) | 0.0069 | 12.63 | -8.63 |
| Gumble | 1.0149 | 1.05 | 2.95 |
| Frank | 0.0505 | 0.08 | 3.92 |
| Clayton | 0.0246 | 0.71 | 3.29 |

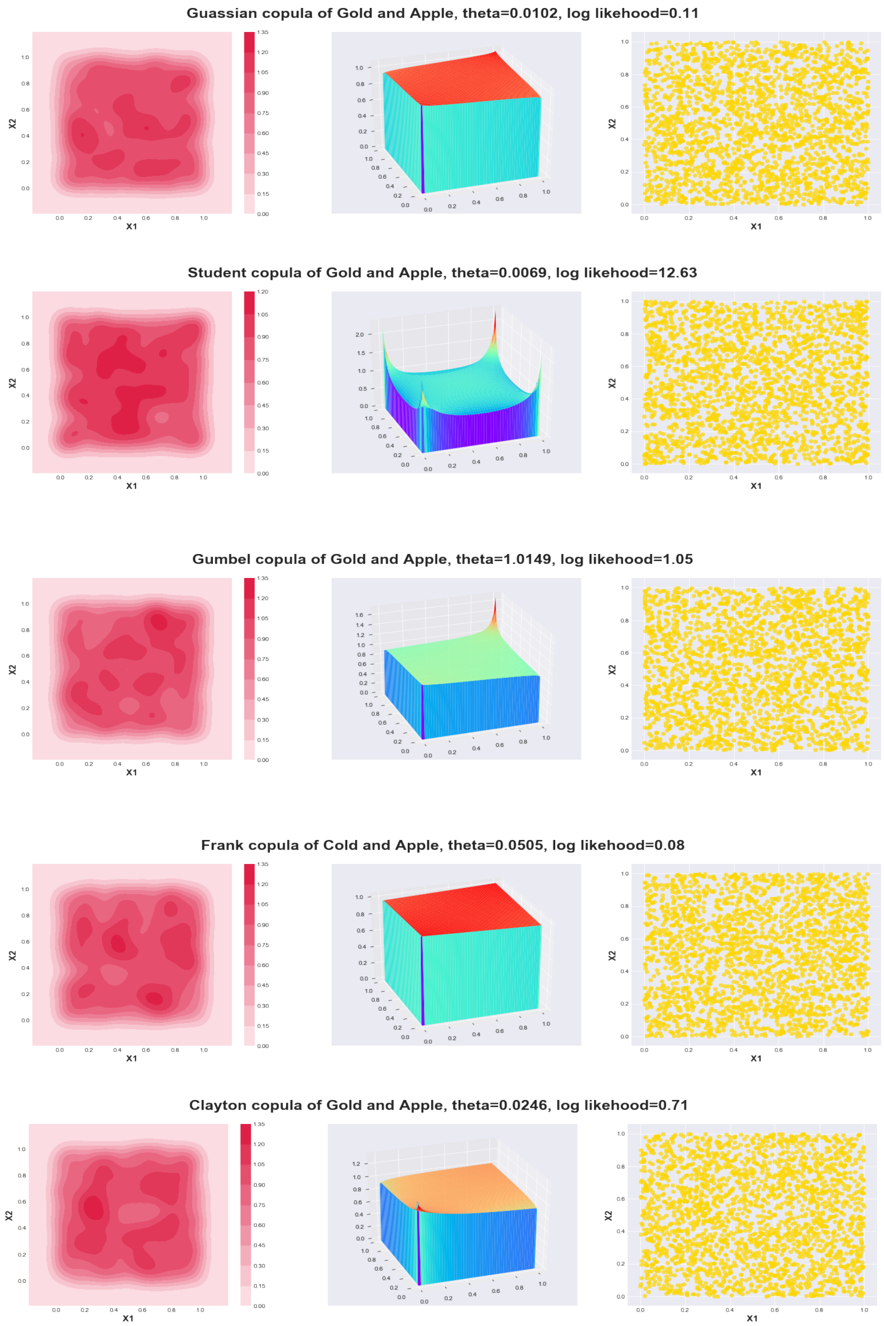



Figure 8: Fit different copulas to Gold and Apple joint distribution.

### 4.0.3 Fit copulas to pair3 - Lv and Apple (correlation = 0.19)

Lv and Apple stocks have weak positive correlation, and their joint distribution in figure 4 are Right leaned ellipse. This shows that they are not highly dependent. Fit 5 kinds of copula to Pair3, the Student-t copulas with 6.03 degree of freedom, has minimum AIC is the best copula to pair 3. It is clear that the Student-t copula has strongest tail dependence.

Table 6: Summary of pair3 fitting

| Copula method | $\theta$ | log-likelihood | AIC |
|---|---|---|---|
| Gaussian | 0.3113 | 112.72 | -108.72 |
| Student-t (df=6.3) | 0.2999 | 134.37 | -130.37 |
| Gumble | 1.2354 | 126.98 | -122.98 |
| Frank | 1.7976 | 92.24 | -88.24 |
| Clayton | 0.3509 | 83.17 | -79.17 |

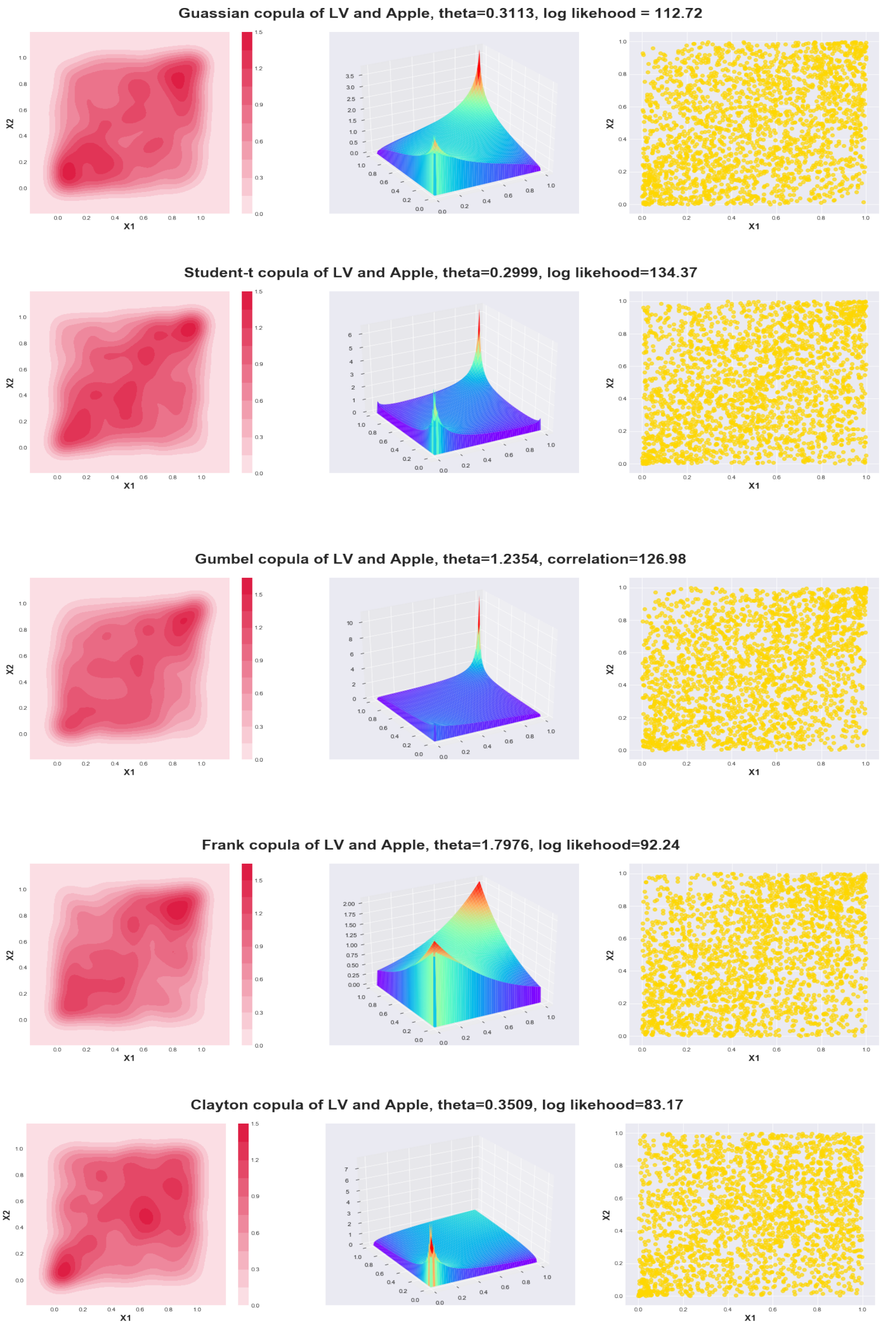



Figure 9: Fit different copulas to Lv and Apple joint distribution.

#### 4.0.4 Fit copulas to pair4 - Silver and Gold (correlation = 0.61)

Silver and Gold have relatively strong positive correlation, and their joint distribution in figure 4 are Right leaned line cluster. This shows that they are highly dependent. Fit 5 kinds of copula to Pair4, the Student-t copulas with 5.07 degree of freedom, has minimum AIC is the best copula to pair 4.

In the figure 4.0.4, it is clear that the Student-t copula has strongest tail dependence with highest density at two tails.

Table 7: Summary of pair4 fitting

| Copula method | $\theta$ | log-likelihood | AIC |
|---|---|---|---|
| Gaussian | 0.7979 | 1125.96 | -1121.96 |
| Student-t (df=5.07) | 0.8147 | 1229.99 | -1225.99 |
| Gumble | 2.4254 | 1164.80 | -1160.80 |
| Frank | 8.2458 | 1131.78 | -1127.78 |
| Clayton | 1.8477 | 879.59 | -875.59 |

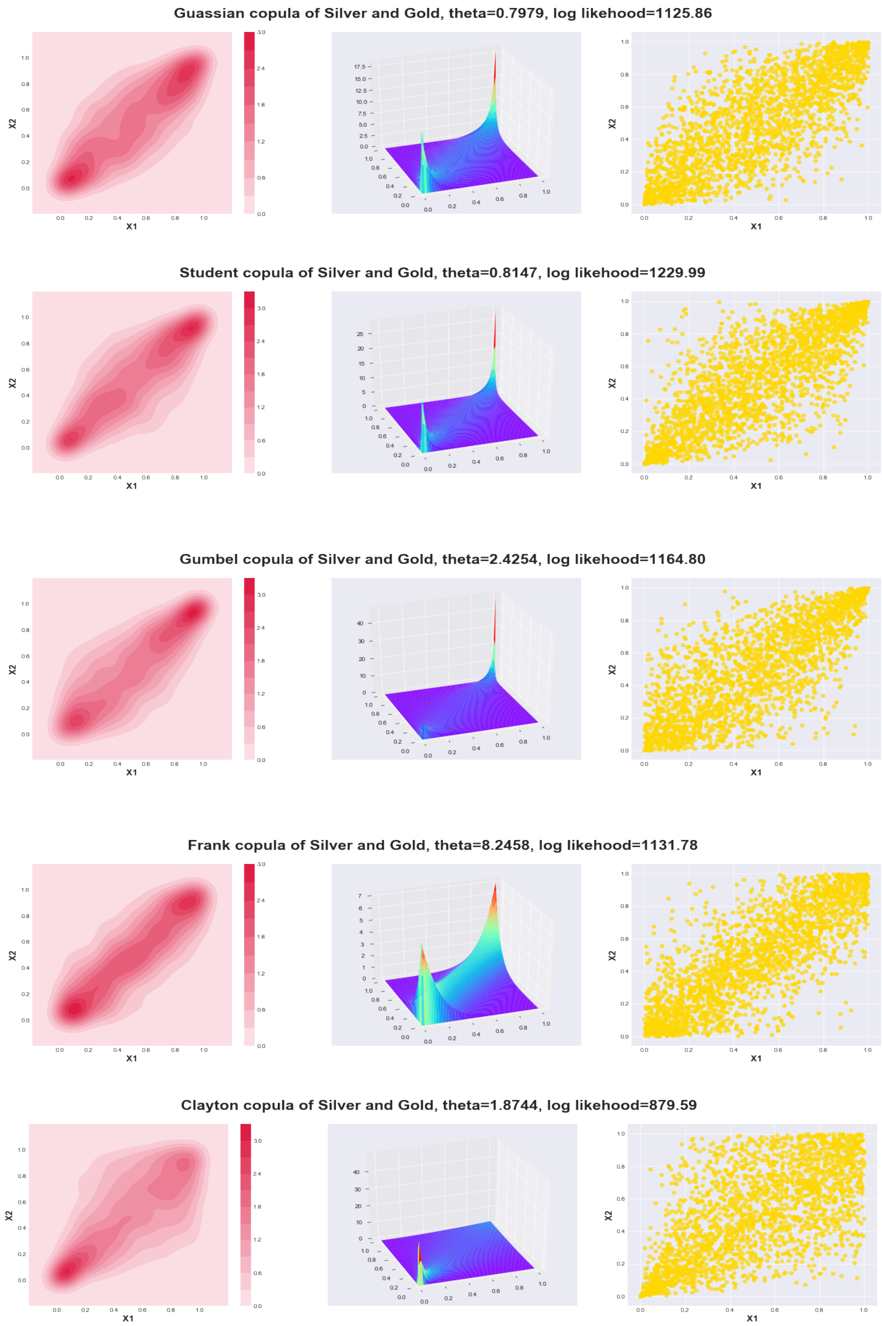



Figure 10: Fit different copulas to Silver and Gold joint distribution.

### 4.0.5 Fit copulas to pair5 - DOW and SPX (correlation = 0.82)

DOW and SPX index have strong positive correlation, and their joint distribution in figure 4 are Right leaned linear square. This shows that they are quite highly dependent. Fit 5 kinds of copula to Pair5, the Gumble copulas with minimum AIC is the best copula to pair 5.

It is clear that the Gumble copula has strongest tail dependence.

Table 8: Summary of pair5 fitting

| Copula method | $\theta$ | log-likelihood | AIC |
|---|---|---|---|
| Gaussian | 0.9619 | 2893.48 | -2889.48 |
| Gumble | 5.6698 | 2957.79 | -2953.79 |
| Clayton | 5.8413 | 2380.68 | -2376.68 |

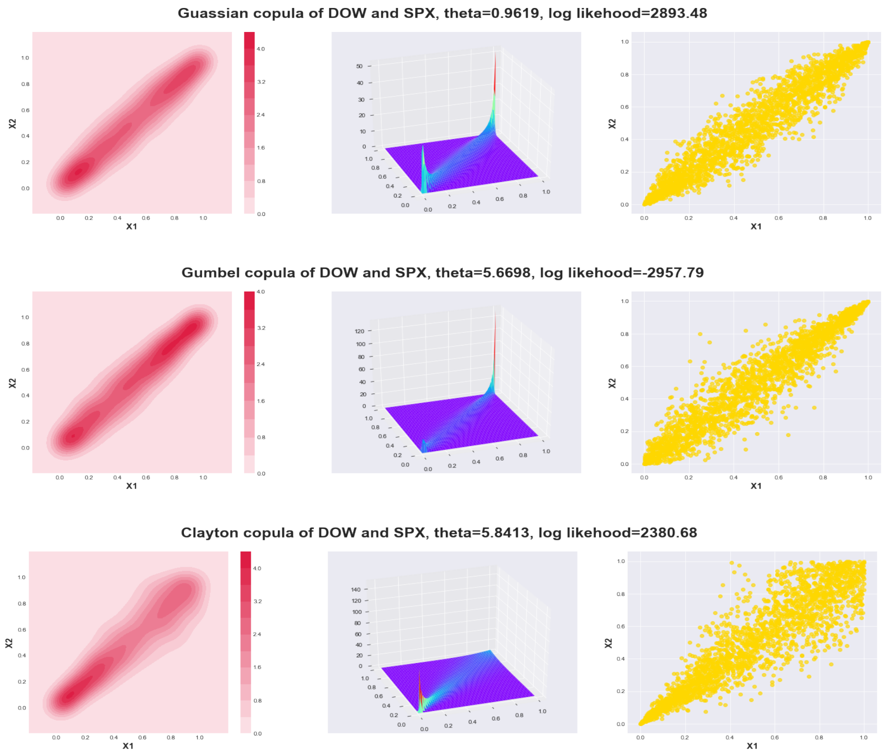



Figure 11: Fit different copulas to DOW and SPX joint distribution.

# 5 EVT

Figure 12 shows QQ plot and hist plot of daily loss of Apple asset. We can see from red line the tail is quite heavy in hist plot. Thus, we use Apple daily loss to investigate the EVT.

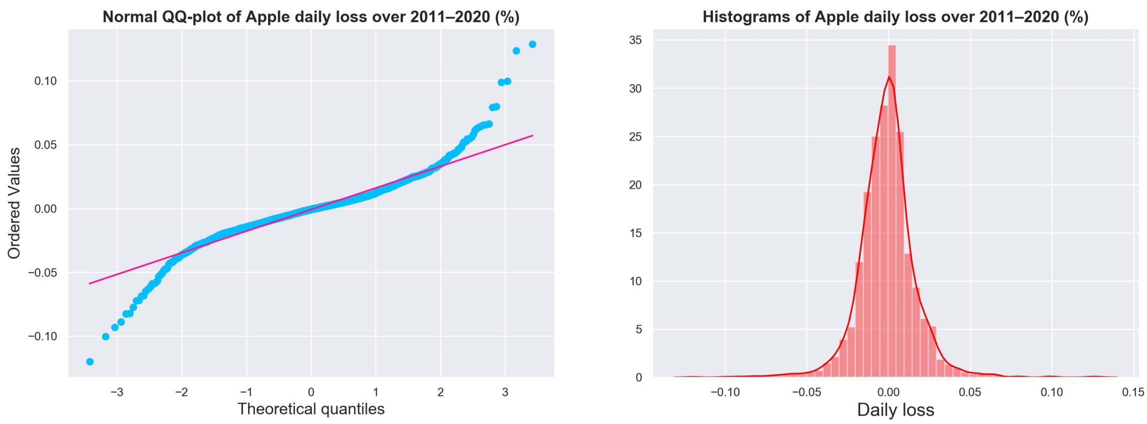
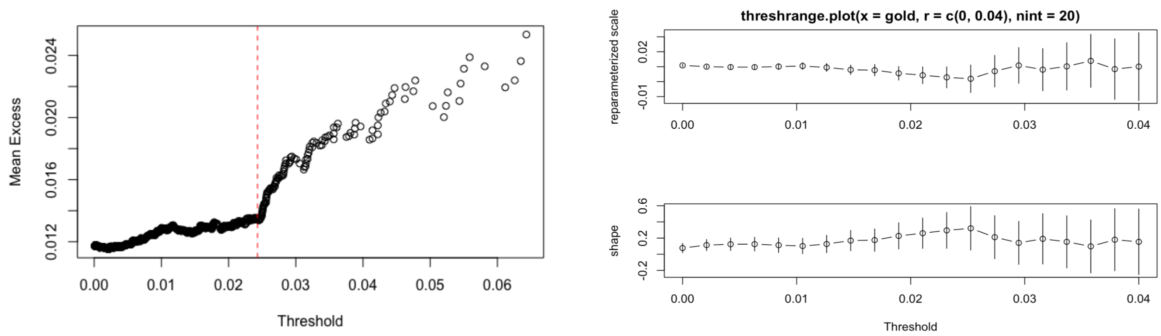



Figure 12: Gold asset

Then, we want to decide the threshold in POT. Figure 13(a) shows mean excess plot with changing threshold and we omit 9 outliner. When threshold is between 0.02 and 0.03 there is a evident turning point. And after that the mean excess is almost linear. Thus we can choose this point as threshold, which shows as red dash line(0.024). Furthermore, figure 13(b) shows scale and shape plots with changing threshold. And it is clear to see that both scale and shape become unstable and variance become large when threshold is bigger than 0.024. Thus, it is reasonable to use 0.024 as threshold.



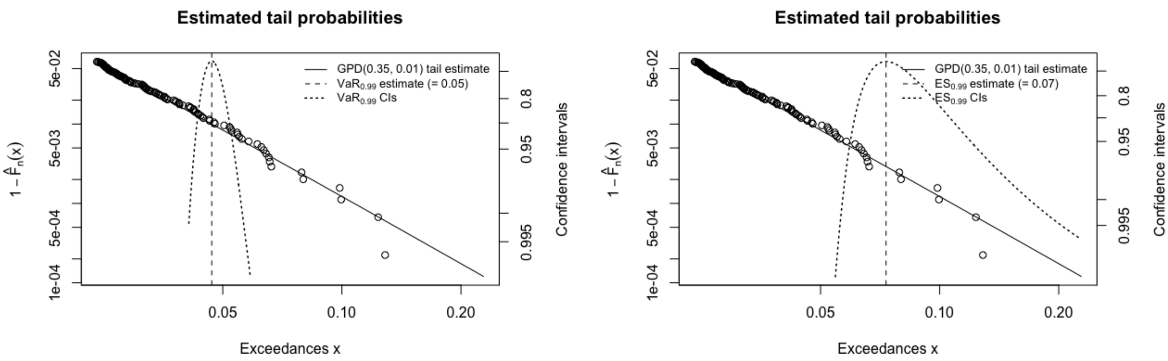

(a) Mean residual plot with different threshold. Red dash line is threshold:0.024

(b) Scale and shape change with different threshold

Figure 13: Distribution of the datasets

Furthermore, we collect the data which exceed the threshold $0.024$ and there are total 43 days. Then we use these data to evaluate the parameter of GPD. And the parameters are $\xi$: 0.351 and $\beta$: 0.009.

Then, we use model to estimate the $VaR_{0.99}$ and $ES_{0.99}$. From 14(a) $VaR_{0.99}$=0.0467 and 95% confidence interval: [0.0422, 0.0531]. $ES_{0.99}$ = 0.0741 and 95% confidence interval: [0.0422, 0.0531]




(a) Estimated tail and estimated $VaR_{0.99}$

(b) Estimated tail and estimated $ES_{0.99}$

Figure 14: Distribution of the datasets

After getting the $VaR_{0.99}$ and $ES_{0.99}$, we compare them with historical simulation(HS) and student t with degree 3. For $VaR_{0.99}$ we see there is no big difference among three methods. GPD is slight bigger than student t and smaller than HS. And $ES_{0.99}$ of GPD is slightly bigger than the other two method. The possible explanation is we use whole nine years data to calculate the $VaR_{0.99}$ and $ES_{0.99}$, which will average the risk. For $VaR_{0.999}$ and $ES_{0.999}$ result from GPD are bigger than the other two.

Table 9: VaR and ES comparison

| Method | $VaR_{0.99}$ | $ES_{0.99}$ | $VaR_{0.999}$ | $ES_{0.999}$ |
|---|---|---|---|---|
| GPD | 0.047 | 0.074 | 0.107 | 0.166 |
| Historical simulation | 0.048 | 0.069 | 0.099 | 0.117 |
| Student-t | 0.045 | 0.070 | 0.103 | 0.155 |

# 6 Conclusion

For PCA and FA, we see that FA is better at interpreting loading of factor. While PCA can use to reduce the dimension of correlated variables. Next, for copula, highly correlated 2-D pair seems to have larger log-likelihood and stronger tail dependence. Elliptical copulas are more suitable for weakly correlated pairs to calculate joint distribution. In the end, for EVT analysis we did not see huge difference of $VaR_{0.99}$ and $ES_{0.99}$ with other two methods. For $VaR_{0.999}$ and $ES_{0.999}$ EVT's results are bigger than other two method.