Assignment 3 PCA, copula and EVT

MeiFang Li(2719570), Yuhao Qian(2684098)

1. Introduction

In this assignment, we create a portfolio consisting of eight stocks and two indices: Shell, Philips, Heineken, TomTom, Aegon, ING, Apple, Microsoft, AEX and S&P500. These assets are equally weighted. We implement Principal Component Analysis(PCA) and Fact Analysis(FA) to reduce the dimension of the portfolio. Then we choose two pairs of asset returns(Apple and Microsoft, Shell and ING) and apply copulas to these paired data. In the last task, we perform Extreme Value Theory Analysis(EVA) to a heavy-tailed asset(Apple) in our portfolio.

2. Theory and Methods

We download 10 year historical data(01/01/2010-01/01/2020) from *Investing.com*[1] through *investpy* python package[2]. Data is synchronized by dropping the missing values.

2.1. Principal Component Analysis

PCA is a data-rotation technique; it reduces the dimensionality of highly correlated data by finding a small number of uncorrelated linear combinations [1]. These linear combinations are principal components(PCs) that explain most of the variance of the original data. Before implementing PCA, we normalize the data with a mean of zero and a standard deviation of one. Then we investigate the explained variance and loadings through *PCA* in *sklearn.decomposition*[3].

2.2. Factor Analysis

FA is another dimension reduction method used to extract maximum common variance from all variables and put them into a common score [2]. It is a linear statistical model investigating a smaller number of unobservable factors through many observable variables. We still use the normalized data in PCA and perform FA with the help of package *factor_analyzer*[4].

2.3. Copula

A copula is a multivariate distribution function with marginal distribution uniform on the interval [0,1]. The copulas can describe the dependence between variables and several copula methods are worthy to be mentioned in practice. For instance, the Gaussian copula is an elliptical copula constructed from a multivariate normal distribution; the t-copula underlies the multivariate student's t distribution; the Clayton, the Gumbel and the Frank copula is derived from the Archimedean class that can take higher dimensions into account.

Regarding fitting copulas to data, we are supposed to perform a two-stage estimation procedure: first estimate the marginal distributions, then form a pseudo-sample of observations from a certain copula function and fit parametric copula by maximum likelihood. However, these can be easily done by using certain functions in *Copulae* Python package like *GaussianCopula, StudentCopula, ClaytonCopula, FrankCopula, GumbelCopula*. Hence, we choose two pairs of asset returns(Apple and Microsoft, Shell and ING) and fit them into five copula models.

2.4. Extreme Value Theory

Extreme Value Theory is used to describe possible distributions of tail events especially the extreme events that have never or barely been observed before. Given a high threshold u, the excess distribution of variables above this threshold could be obtained by: $F_u(x) = P(X-u \leq x | X > u) = \frac{F(x+u)-F(u)}{1-F(u)}$. And EVT provides another natural approximation for this distribution, the Generalized Pareto Distribution(GPD). The GPD is a two-parameter distribution with the following df:

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

[1] https://www.investing.com/
[2] https://investpy.readthedocs.io/
[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
[4] https://pypi.org/project/factor-analyzer/

For $\xi < 0$, this yields Pareto type II distribution; for $\xi = 0$, this yields exponential distribution; for $\xi > 0$, this yields Pareto(reparametrized version) distribution which is heavy tailed. Hence, the excess distribution can be obtained by $\bar{F}(x) = \bar{F}(u)\bar{F}_u(x-u)$ where $\bar{F}(u)$ is estimated empirically by N_u/n and $\bar{F}_u(x-u)$ using Equation 1 and $F_u(x) = G_{\xi,\beta+\xi u}(x)$, then transformed into:

$$\widehat{F}(x) = \frac{N_u}{n}\left(1+\hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-1/\hat{\xi}} \tag{2}$$

where $N_u = \sum_{i=1}^{n} 1_{\{\{X_i>u\}\}}$ is the number of exceedance of threshold u from n samples.

If ξ and β is estimated, then VaR and ES can be computed as follows:

$$\mathrm{VaR}_\alpha = q_\alpha(F) = u + \frac{\beta}{\xi}\left(\left(\frac{1-\alpha}{\bar{F}(u)}\right)^{-\xi} - 1\right) \tag{3}$$

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_x(F)\mathrm{d}x = \frac{\mathrm{VaR}_\alpha}{1-\xi} + \frac{\beta-\xi u}{1-\xi}. \tag{4}$$

According to the theory mentioned above, we perform Extreme Value Theory Analysis(EVA) to a heavy-tailed asset in our portfolio. First, we study the relationship between mean excess and threshold u. Then we choose a rational threshold and use *genpareto.fit* to get estimated ξ and β of GPD. Furthermore, we use these parameters to generate CDF and check whether the exceeds corresponds with this CDF function. Finally, we estimate VaR and ES from EVT results and compare them with those obtained from historical simulation method and Student-t distribution.

3. Results

3.1. Principal Component Analysis

Figure 1 is a bar plot showing the explained variance of different numbers of principles components in PCA. The first two components have the largest explained variance, which is 4.89 and 1.34 respectively. When the number of principal components is greater than 3, the explained variances are less than 1 and form like a straight line. Figure 2 reals the cumulative explained variance ratio. We can see that the ratio increased consistently as the number of principal components increase.

To interpret the PCA result, we need to decide how many principal components to examine. We want to concentrate our initial interpretation only on the largest principal components. According to Kaiser's rule [3], we only choose the principal components with explained variance greater than 1. Hence, we decrease the initial 10 variables to 2 principal components, which together explain 62.34% of the total variance.

The loadings of the first four principle components of PCA are in Figure 3. In the first subplot, all the 10 assets are positively correlated with the first principal component. Since we have both individual stocks and two indices from different countries, we can treat the first component as a general variable like the global market. In the second principal component, Apple, Microsoft and S&P500 have positive loadings while the other assets all have negative loadings. We can interpret the second component as American market and Dutch market, where American assets have a positive direction and Dutch assets have a negative direction. Heineken has the most significant loading -0.82 in PC3; therefore, PC3 might describe Heineken solely. TomTom has large negative loading in PC4, meanwhile Shell has large positive loading, so the fourth component might describe these two assets mostly.

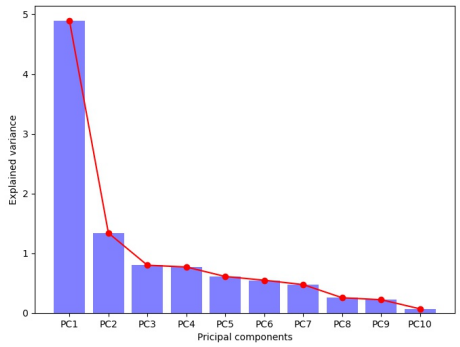

Fig. 1: Explained variance of PCA

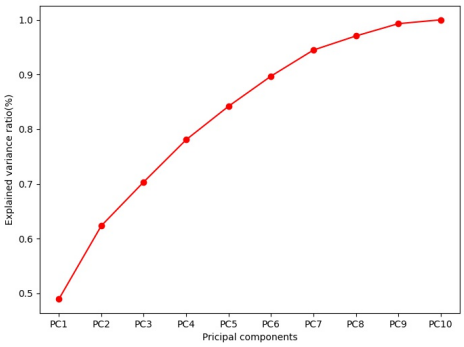

Fig. 2: Cumulative explained variance ratio of PCA

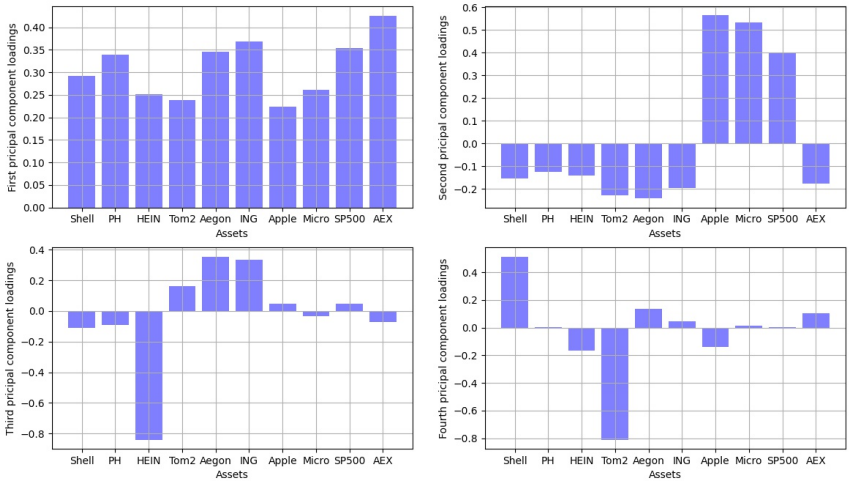

Fig. 3: Loadings of the first four principle components of PCA

3.2. Factor Analysis

We firstly apply Bartlett's test of sphericity to check whether or not the observed variables intercorrelate at all using the observed correlation matrix against the identity matrix [4]. The p-value of Bartlett's test is 0, which means the test is statistically significant. Then we can implement FA with *varimax* rotation.

Figure 4 gives explained variance in FA concerning different numbers of factors. The value of explained variance decreases sharply as the number of factors goes up. Only the first two factors have explained variance greater than 1, while the remaining factors do not have significant variance contribution. Intriguing, we notice that no variation is explained after six factors, given five zeros in the plot. Subsequently, we can get the cumulative variance in Figure 5. Again, we opt to choose the first two factors in this FA model according to Kaiser's rule. Using the chosen two factors, we can explain the 48.78% variation of the initial data. Because the explained variance is zero after six factors, the second part of the curve is horizontal.

Although two factors are contributing explained variance larger than one in Figure 4, we still calculate the loadings for the first four factors in Figure 6. We notice that all the assets have positive loadings, except Aegon has a negative loading in the fourth factor. In the first factor, TomTom has the largest loading 0.632, so we can assume the first factor represents TomTom solely. Shell and Heineken have the largest loadings on factor2, but it is hard to draw a common ground between these two stocks. Hence, we think the second factor mainly represents Shell and Heineken at the same time. The third factor describes Heineken alone since it has only one extensive loading. The fourth factor nearly describes the same thing as factor2 because the distribution of loading is almost identical.

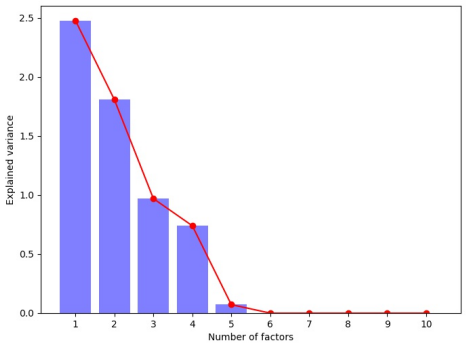

Fig. 4: Explained variance of FA

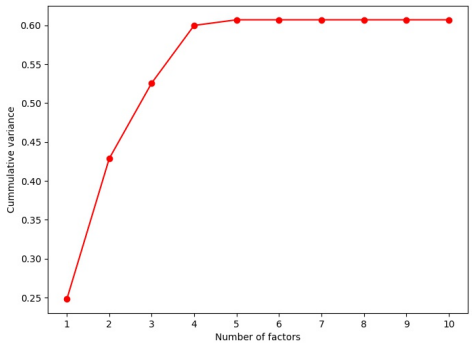

Fig. 5: Cumulative explained variance ratio of FA

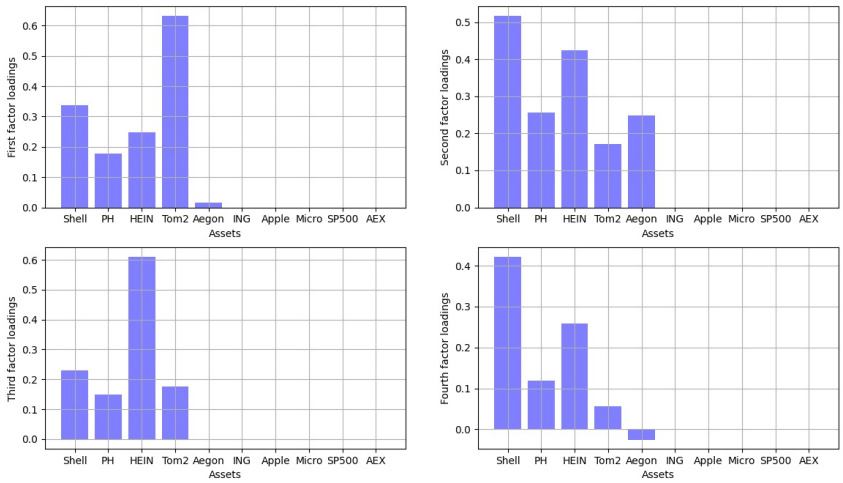

Fig. 6: Loadings of the first four factors of FA

3.3. Copula

We obtain the dependence patterns between Apple and Microsoft shown in Figure 7, 8, 9, 10, 11, 12 by fitting the following five copula methods: Gaussian, Student-t, Clayton, Frank, Gumbel. And we notice that the dependence between empirical data looks most similar to the dependence when using student-t copula because the extreme values cluster on the upper right and bottom left of the figure. Furthermore, the parameters estimated by these methods are displayed in Table 1, where the log-likelihood is also noted. It is easy to find that the largest log-likelihood appears in student-t copula with 368.024 and the estimated $\rho = 0.479$ with 4.042 degrees of freedom. This result corresponds with the speculation we get from the figures so the best copula is student-t copula concerning Apple and Microsoft.

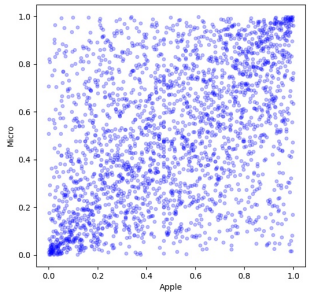

Fig. 7: Empirical-Apple&Microsoft

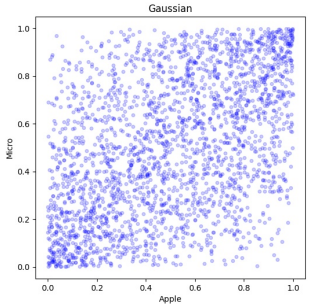

Fig. 8: Gaussian-Apple&Microsoft

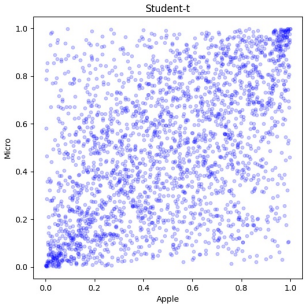

Fig. 9: student t-Apple&Microsoft

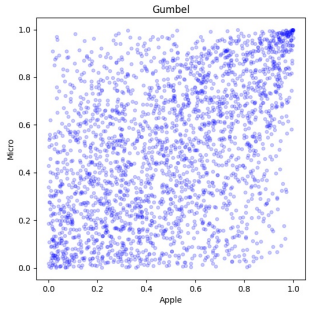

Fig. 10: Gumbel-Apple&Microsoft

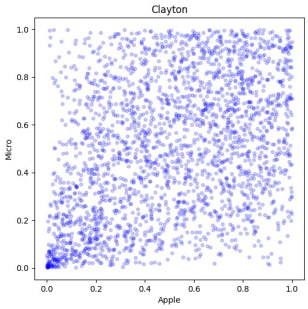

Fig. 11: Clayton-Apple&Microsoft

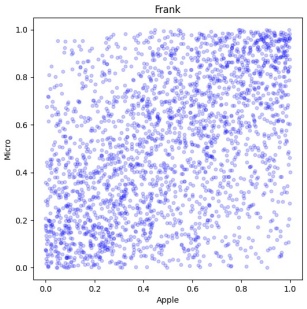

Fig. 12: Frank-Apple&Microsoft

When it comes to the dependence between Shell and ING, we obtain the patterns using the same copula methods, shown in Figure 13, 14, 15, 16, 17, 18. We find out that the dependence between empirical data also looks most similar to the dependence when using student-t copula since the extreme values also cluster on the figure's upper right and bottom left. Moreover, the parameters estimated by these methods are displayed in Table 2. We notice that student-t copula shows the largest log-likelihood with 304.489 and the estimated $\rho = 0.460$ with 8.283 degrees of freedom. This result corresponds with the speculation we get from the figures so the best copula is also a student-t copula concerning Shell and ING.

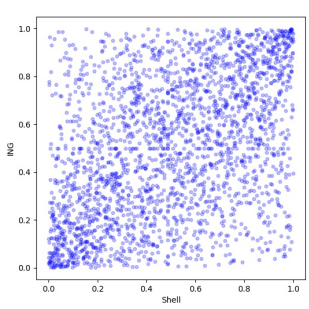

Fig. 13: Empirical-Shell&ING

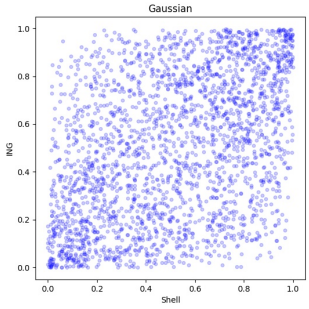

Fig. 14: Gaussian-Shell&ING

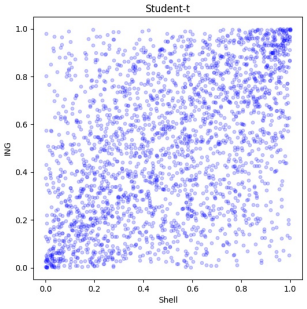

Fig. 15: student t-Shell&ING

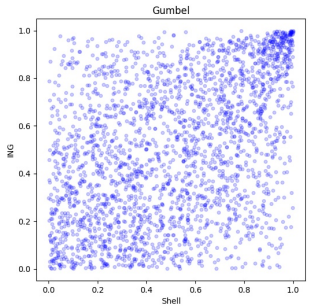

Fig. 16: Gumbel-Shell&ING

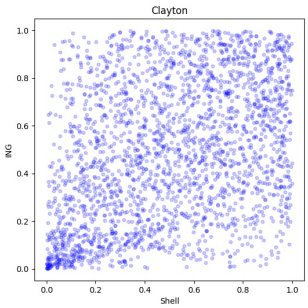

Fig. 17: Clayton-Shell&ING

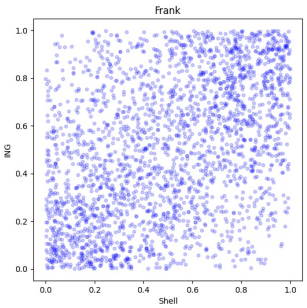

Fig. 18: Frank-Shell&ING

Table 1: The estimated copula parameters for fitting Apple and Microsoft.

Copula	Log-likelihood	Parameter
Gaussian	304.514	ρ=0.467
Student-t	368.024	ρ=0.479, ν=4.042
Clayton	306.582	ρ=0.761
Frank	292.451	ρ=3.193
Gumbel	291.484	ρ=1.420

Table 2: The estimated copula parameters for fitting Shell and ING.

Copula	Log-likelihood	Parameter
Gaussian	284.632	ρ=0.454
Student-t	304.489	ρ=0.460, ν=8.283
Clayton	239.734	ρ=0.629
Frank	277.515	ρ=3.061
Gumbel	263.637	ρ=1.389

3.4. Extreme Value Theory

From the QQ-plots shown in Figure 19, we see that Apple has probably the heaviest tail so we choose it to study EVT. We can see in Figure 20 that the mean excesses steadily grow up as the threshold increases until the threshold reaches a certain value. Thus we choose the threshold of 0.025 to simulate GPD distribution in case there are not enough exceedances. As shown in Figure 22, the exceedances distribute evenly throughout the whole estimated period, and the number of exceedances is neither too large nor too small; so this threshold of 0.025 is suitable for the following analysis. We fit GPD to the exceedances with threshold u=0.025 and get estimated $\xi = 0.4812$ and $\beta = 0.0072$. The ξ parameter indicates that the tail is heavier than that of an exponential distribution. Then Figure 21 can prove that the empirical CDF of exceedances fit relatively perfectly with the CDF of GPD generated by these estimated parameters.

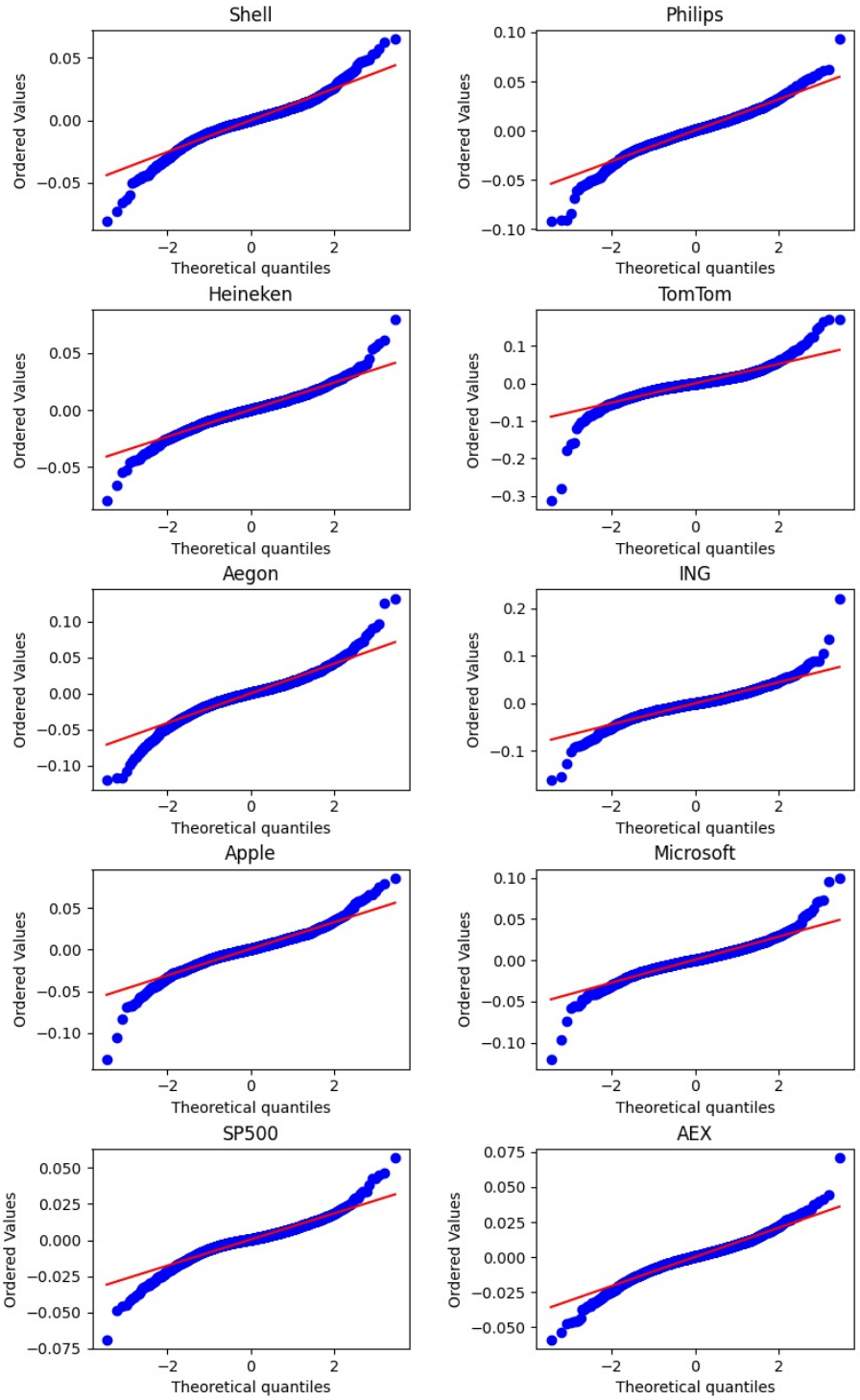

Fig. 19: QQ norm plot of all assets losses

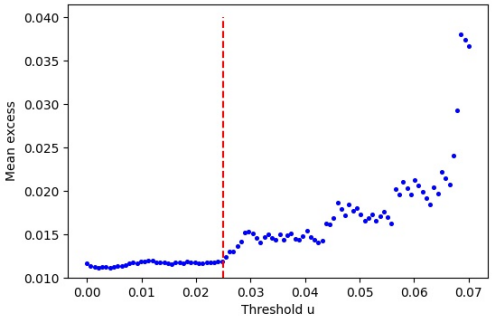

Fig. 20: Mean excess with different thresholds

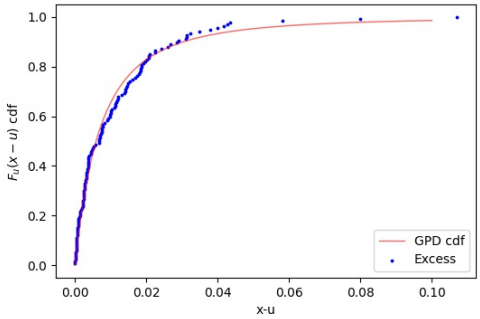

Fig. 21: Cumulative density function of GPD and empirical with threshold u=0.025

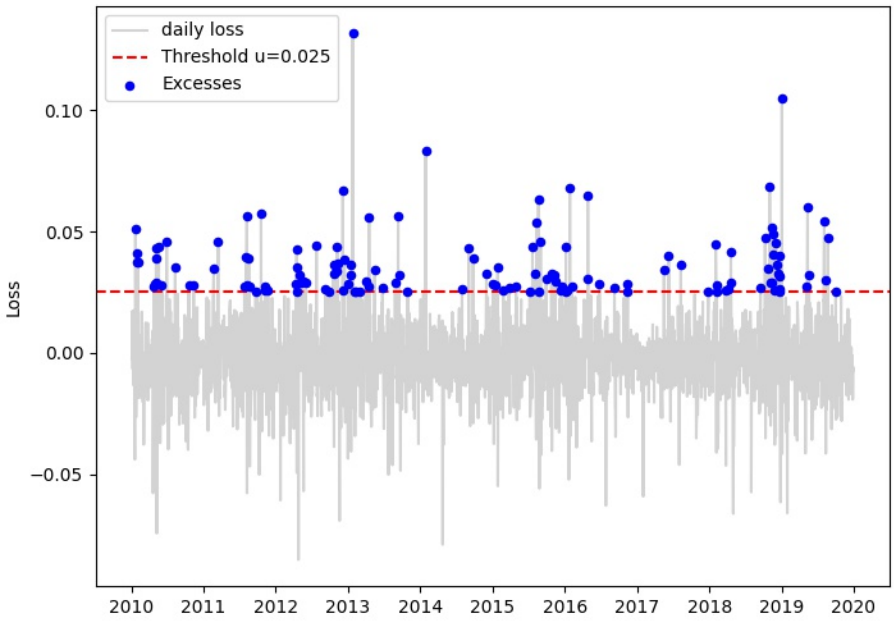

Fig. 22: Exceedances of Apple with threshold u=0.025

Furthermore, the estimated VaRs and ESs from EVT, as well as those obtained from historical simulation method and Student-t distribution are shown in Table 3. We can see that the VaRs of three methods are very close when the confidence level is 0.975 and 0.99; the largest difference is 0.00149 and 0.00061 separately. However, the difference increases when it comes to VaR(0.999). The VaR(0.999) from EVT has the largest value, and the difference between the smallest VaR(0.999) is 0.03618. As for ES, the trend is similar to that in VaR. EVT still has the largest ES under the abovementioned three confidence levels and the biggest difference increase for the higher quantiles.

4. Discussion

Although the numbers of chosen principal components/factors for the PCA and FA are both two, the model interpretation is entirely different. For PCA, the first two principal components mainly represent the global market and American market. However, we think the first factor in FA describes TomTom solely and the second factor mainly describes the linear combination of two unrelated stocks (Shell and Heineken). For us, PCA is easier to

Table 3: The VaR and ES

Method	EVT	Historical simulation	Student-t(df=3)
VaR(0.975)	0.03170	0.03240	0.03091
VaR(0.99)	0.04373	0.04433	0.04372
VaR(0.999)	0.1122	0.07602	0.0972
ES(0.975)	0.05185	0.04698	0.04842
ES(0.99)	0.07505	0.06074	0.06693
ES(0.999)	0.2071	0.1067	0.1462

interpret, while FA is harder to think of a relationship between observed variables and unobserved latent factors. A possible explanation could be that PCA components are fully orthogonal whereas FA does not require factors to be orthogonal.

Different copula methods can be used to describe the dependence between pairs of assets and the best fit is the one with the largest log-likelihood. This can be interpreted by both visual representations and the estimated copula parameters. The student-t copula describes best for both pairs we pick in section 3.3, given the fact that there are two apparent clusters at both sides, indicating two heavy tails.

EVT analysis is strongly related to the chosen threshold. The high threshold could reduce bias in estimating excess function but not too high if there is not enough excess. To estimate the parameters of the excess function, we must choose a suitable threshold and fit GPD to the excess amount over this threshold. The resulting $\xi > 0$ could indicate that the tail is heavier than that of other distributions such as normal and exponential. Due to the heaviness of EVT tails, the VaR and ES with a much higher confidence level would be larger than those of other methods such as historical simulation method and student-t distribution. In contrast, the VaR and ES with a relatively lower confidence level would achieve somewhat similar results among these methods.

5. Conclusion

We apply two dimension reduction techniques(PCA and FA) to a ten-asset portfolio. Although the number of principal components/factors should be two in both cases, we find that the interpretation of the model is quite different. Two pairs(Apple&Microsoft, Shell&ING) are fitted to five different copulas, and it turns out that student-t copulas work best in our task. After applying EVT to a heavy-tailed asset(Apple), we notice that the corresponding VaR and ES are larger than those calculated by Historical simulation and Var-Cov method with a student-t assumption; this difference is more significant for higher quantile like 0.999.

References

1. P. Embrechts, R. Frey, and A. McNeil, "Quantitative risk management.," 2011.
2. DataCamp Community, "(tutorial) principal component analysis (pca) in python," 2019. https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python.
3. A. Rea and W. Rea, "How many components should be retained from a multivariate time series pca?," *arXiv preprint arXiv:1610.03588*, 2016.
4. DataCamp Community, "Introduction to factor analysis in python," 2019. https://www.datacamp.com/community/tutorials/introduction-factor-analysis.