Statistical Analysis of
Cryptocurrency Data

Rik Harng Loong
Student ID: 18061593

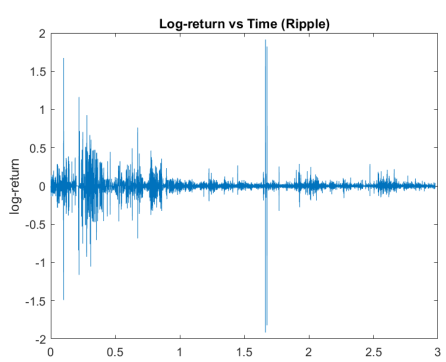
February 12, 2020

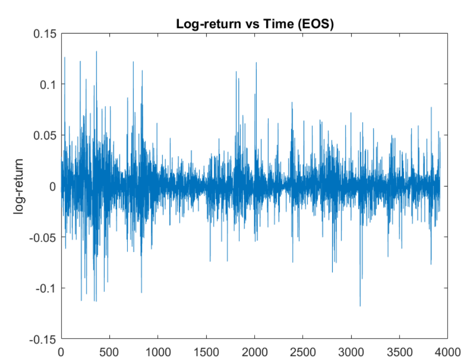
Part I

In this section, we will characterize and compare the data of two main crytocurrencies, namely Ripple and EOS using both parametric and non-parametric approaches. In particular, we will discuss about fitting our data with some probability distributions with the associated parameters, estimating their Value at Risk (VaR) and Expected Shortfall (CVaR) at different time horizons, and finally the statistical significance and robustness of our estimations.

1 Parametric Analysis

We first look at the representation of both Ripple and EOS data by plotting their hourly log-returns. Log-returns are used here as they have a more stable distribution than simply raw returns, and also the distribution of log-returns is roughly normal so it is more convenient to work with [1]. However, we will see shortly that the normality assumption does not hold well in some cases. From the plots below, the log-returns for both assets look very noisy (i.e. a lot of fluctuations across time).





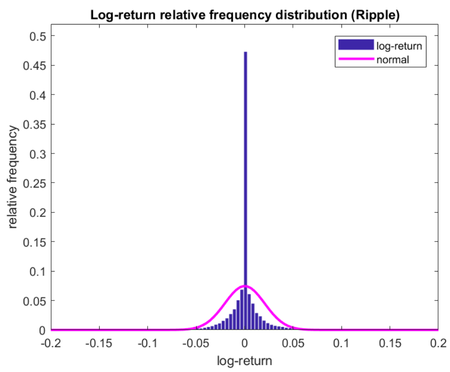
Assuming that the log-returns follow a normal distribution, the Maximum Likelihood estimators (MLE) are given by

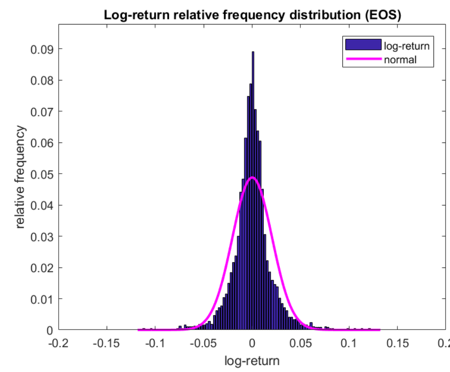
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

	$\hat{\mu}$	$\hat{\sigma}$
Ripple	0.000	0.049
EOS	0.000	0.020

Table 1: Sample mean and std for both assets log-returns

Given these estimators, we try to fit a normal distribution and compare it to the empirical distribution of our data.



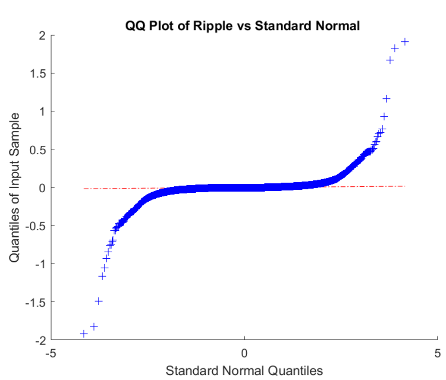


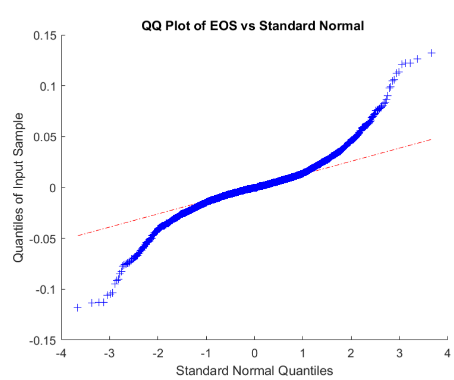
It is clear that in the case of our Ripple log-returns, the normal distribution with MLE parameters poorly represents our sample data, especially at the center of the distribution where some events happen more frequently than we would expect in the normal case. In the case of our EOS, the fit is slightly better although it still suffers a little bit at the center of the distribution. Let us also consider the kurtosis of our data:

	Kurtosis
Ripple	451.949
EOS	9.033

Table 2: Kurtosis of both assets

The kurtosis of both assets exceeds that of the normal distribution which is 3, especially in the case of Ripple the difference of values is huge. Such a result implies a fat tailed distribution, which means the probability of extreme log-returns is higher than one would expect if the distribution was normal. Our analysis shows that the log-returns for both assets seemed not to follow their assumed normal distribution. We can show evidence of this by looking at qqplot which provides a further metric for the accuracy of the normal fit.





One would expect the observed data to fit the dotted line fairly closely if we dealt with an underlying normal distribution. However, it is not the case for both assets, so the plot seems to support our analysis above. More surprisingly, we would expect the EOS data to fit closer to the dotted line than Ripple, but it looked the other way round here. The poor fit in EOS data could be attributed to not having large amount of data and thus do not converge towards a normal distribution by the Central Limit Theorem. In fact, the Ripple data is nearly 8 times the size of that of EOS.

2 Non Parametric Analysis

In this section, we would consider some risk measures such as VaR and CVaR associated with both assets at different time horizons: hourly, daily and monthly. In fact, we have come across one before, that is volatility (by means of standard deviation). However, the main problem with volatility is that it does not take into account the direction of an investment's movement: A stock can be volatile either because of the large price jumps or slumps. Of course we are not so distressed by gains, but more concerned about the odds of losing money. Both the VaR and CVaR are based on this common-sense fact [2].

	$\mathrm{VaR}_{5\%}$	$\mathrm{CVaR}_{5\%}$
Hourly	0.0295	0.0867
Daily	0.1454	0.2597
Weekly	0.2618	0.4566

Table 3: Risk measures for Ripple

	$\mathrm{VaR}_{5\%}$	$\mathrm{CVaR}_{5\%}$
Hourly	0.0303	0.0476
Daily	0.1337	0.1882
Weekly	0.4159	0.6441

Table 4: Risk measures for EOS

Note that we are considering these measures at the 5% level. The q% Value at Risk is the porfolio loss at the (1-q)th-quantile; which means there is a (1-q)% confidence level that the loss will not exceed the VaR value over a certain time-horizon. CVaR is then the expected loss conditional on the loss does exceed the VaR value. The result shows that risk measures value increases with the time interval. It tells us it is more risky to invest long term than short term for both of our assets. Now we wish to test the robustness of out test statistics by bootstrapping.

The idea is that we resample our data n times and we include only 80% of the data to be retained in our bootstrap sample. We then compute the test statistics for each of our sample. In this case, we set n = 1000. We perform bootstrap analysis on our risk measures of daily log-returns since it yields reasonably large values just to be safe.

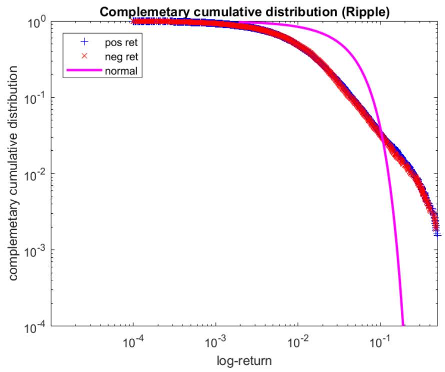
	$\mathrm{VaR}_{5\%}$	$\mathrm{CVaR}_{5\%}$
95% CI	0.135 - 0.160	0.233 - 0.278

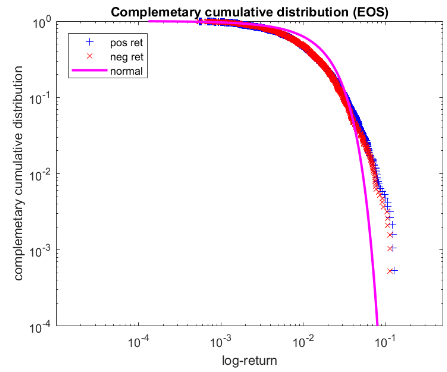
Table 5: Bootstrap confidence interval for Ripple

	$\mathrm{VaR}_{5\%}$	$\mathrm{CVaR}_{5\%}$
95% CI	0.116 - 0.143	0.157 - 0.196

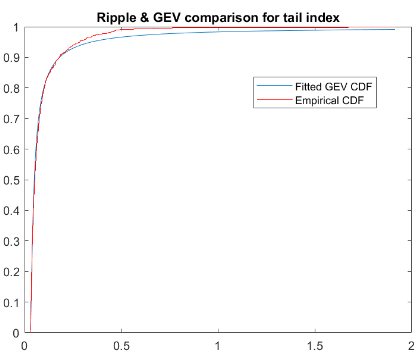
Table 6: Bootstrap confidence interval for EOS

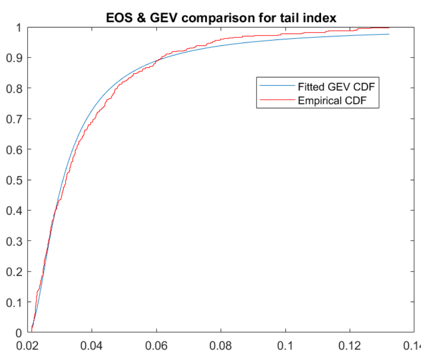
The table above shows the result of a bootstrap two tail p test at 95% confidence level. The result is a good sign since our calculated risk measures fall within the 95% confidence interval. Let us also consider the complementary cumulative distribution (ccd) that maps log-returns to their respective probabilities $P(X > x)$ for our underlying distribution.





The graphs above are plotted log-log. Unsurprisingly, the extreme positive or negative log-returns are more likely to happen than that predicted by a normal distribution for both assets, due to the fatter tail of the unknown distribution. Such extreme event rarely happens, so it is difficult to make an accurate prediction. We can overcome this issue by treating the tail of our distribution as its own separate entity. Assuming that our data is a sequence of i.i.d random variables and has a maximum, we can then model the tail by fitting a Generalized extreme value (GEV) distribution to the extreme values (the top 5% and 10% of log-returns for Ripple and EOS respectively in our case). Note that we consider here the top 10% of EOS data as extreme returns to tolerate with its relatively small dataset.



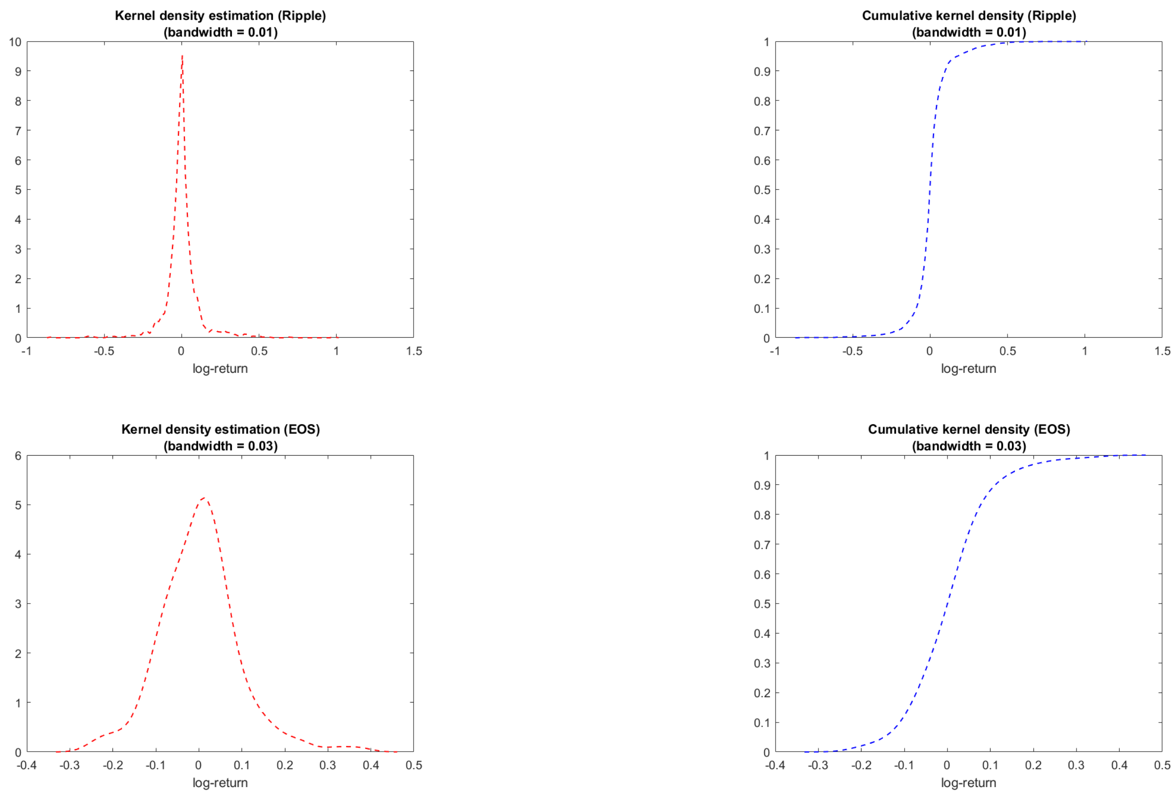


	Ripple	EOS
Exponent α	0.991	1.632

Table 7: α values at top 5% and 10%

$\alpha = 1/\xi$ where ξ is the shape parameter which governs the tail behavior of the GEV distribution. The graphs above show that the extreme log-returns for both assets follow the GEV distribution pretty well. In fact, we perform Kolmogorov-Smirnov test to measure the compatibility of the distribution of our empirical sample and a reference distribution function. We get the resulting p-values: 0.0192 at 1% significance level, 0.3033 at 5% significance level for Ripple and EOS respectively. This indicates we cannot reject the null hypothesis that our data comes from the GEV distribution at the above significance levels, especially for our EOS that has a very high p-value. We could now use these distributions to make predictions on the probability of extreme events occurring by simply referencing the graphs above.

Next, we turn our attention to the Kernel density estimation (KDE) which allows us to approximate our empirical data with a continuous PDF. As usual we apply it to the daily log-returns. Note that the bandwidth is a free parameter h and its choice crucially influences our resulting PDF. Too small bandwidths produce an oscillating PDF, whereas too large bandwidths flatten down the distribution losing information. In general, the bandwidth must decrease with the number of observations [3]. Notice that we use $h = 0.01$ for Ripple (large dataset) and $h = 0.03$ for EOS (small dataset) in order to produce the PDFs that are smooth.



The perks of using KDE over empirical relative frequencies is that no information is lost in discretisation to produce the histogram. These graphs again support our claim earlier that log-returns have fatter tails relative to a normal distribution. We could even use our estimated cumulative distribution graph to estimate daily VaR and CVaR by simply finding the x coordinates of the points corresponding to 0.05 on the y axis (at the 5% level).

Summary

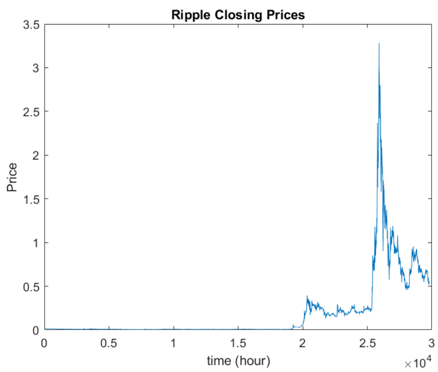
It is found that the log-returns of both Ripple and EOS correspond to a fat-tailed distribution. We have also analyzed the risks associated with both assets at different time horizons. It can be seen that investing in Ripple possesses more risk of "extreme" losses in the tail of the distribution than EOS in general (indicated by higher CVaR), except for the weekly interval. However, we have not taken into account the scaling of our log-returns distribution, and also VaR poses a problem if we considered a portfolio of more than two assets because correlations between assets must also be taken into consideration. This leads us to the the Part II of our report.

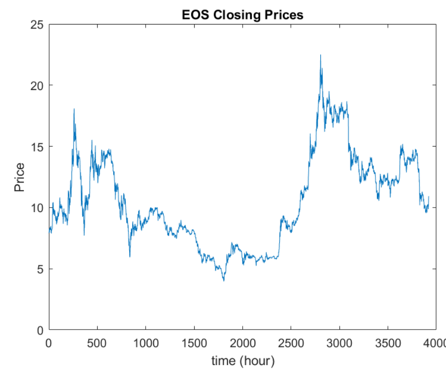
Part II

In this section, we first gain an insight into the scaling laws of time series of both assets Ripple and EOS. We do so by examining their stationarity, autocorrelation, Hurst exponent and tail exponent. Consequently, we use the obtained result to predict certain quantities (e.g volatility, VAR) in a different timescale. Following this, we go on to explore dependencies between different asset returns using linear, non-linear, and information-theoretic measures. Additionally, we also investigate the effect of textual sentiments on asset returns using Granger causality test.

3 Stationarity

First of all, we plot the closing prices of both assets Ripple and EOS for reason discussed later.





Then we examine the stationarity of both closing prices and log-returns on closing prices of the assets. The table below shows our results:

Test	Ripple	EOS
ADF	1	0
KPSS	1	1

Table 8: Test on closing prices

Test	Ripple	EOS
ADF	1	1
KPSS	0	0

Table 9: Test on log-returns

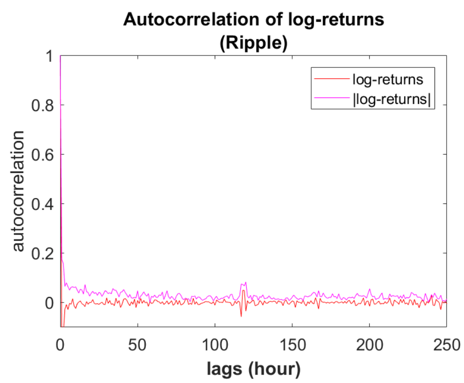
Augmented DickeyFuller (ADF) : The null hypothesis is rejected (i.e. 1) for log-returns of both assets, suggesting that their log-return distribution are stationary. However, the null is only rejected for closing prices of Ripple but accepted in the case of EOS.

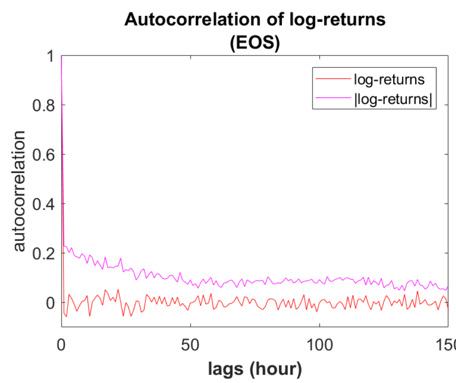
KPSS : In contrast to ADF test, the null hypothesis is accepted (i.e. 0) for log-returns of both assets, in this case accepting that their log-return distributions are stationary. However, the null is rejected for stationarity on the closing prices of both assets.

In summary, it is found that the log-return distributions of both Ripple and EOS are stationary. Indeed, taking logarithm of the returns is equivalent to detrending our time series. Generally, the closing prices of both assets have a unit root (non-stationary) but does not always have a trend. This can be seen from the plot of EOS closing prices above. But note that ADF test suggests stationarity on closing price of Ripple perhaps due to the flat line between (0 - 20000 hours) in the plot above that accounts for most of the stationary properties.

4 Autocorrelations

We can quantify the relationship between the current log-returns values with its past values over short, medium and long term. This allows us to forecast future returns movement of a particular asset, thus be able to make a well-informed decision. The plot below shows our results:

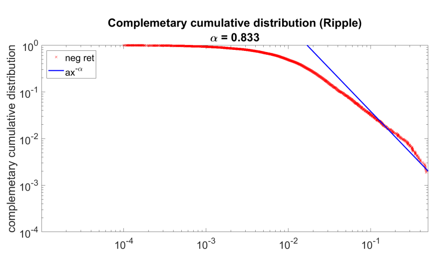


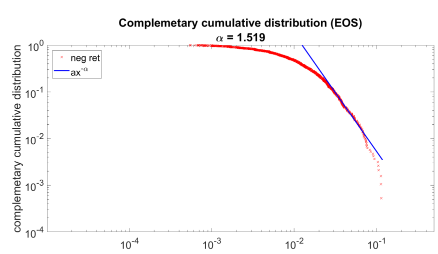


Both graphs show that log-returns trends constantly change and so the variation is mean-reverting (both assets' log-returns tend to move towards their average log-returns over time). However, absolute log-returns trends for both assets are more persistent with EOS lasting longer term than that of Ripple in hourly lags. This implies that a large change in log-returns is more likely to be followed by another large change in log-returns, regardless of whether the trend is positive or negative.

5 Tail Exponent

In this section, we model the complementary cumulative distribution function (ccdf) of negative log-returns (more concerned about loss)' of both assets by fitting a linear function of the form $P(X > x) \sim ax^{-\alpha}$ using Maximum Likelihood method, where α is the gradient of the linear function. The results are summarized below:





	Ripple	EOS
Tail exponent α	0.833	1.519

Table 10: Tail exponent computed from negative log-return

The tail exponent (α) would be useful in the scaling of our log-returns distribution later, where we use it to predict the Value at Risk (VaR) and implied volatility in some future time.

6 Hurst Exponent

We now investigate another quantity called Hurst exponent which is a measure of long-term memory of the time series. Let us consider Generalized Hurst exponent to analyze the scaling properties of our data via kth order moments [4] of the log-returns distribution of both Ripple and EOS.

	H(1)	H(2)	H(3)
Ripple	-0.0351	-0.0279	-0.0542
EOS	-0.0083	-0.0102	-0.0181

Table 11: Generalized Hurst exponent $H(k)$ for kth moment

Clearly $H(k)$ is not constant in our case and therefore we can conclude that the process is multi-scaling (or multifractal) for both assets. In both cases, we have $0 < H(k) < 0.5$ which implies that the time series for log-returns of Ripple and EOS is anti-persistent. This seems to support our previous claim that log-returns are indeed stationary for both assets. Let us also compute the Hurst exponent using closing prices of both assets. We get the following result:

	H(1)	H(2)	H(3)
Ripple	0.4813	0.4305	0.3904
EOS	0.4953	0.4654	0.4340

Table 12: Generalized Hurst exponent $H(k)$ for kth moment

All the values are close to $H(k) = 0.5$ which implies that the closing prices of both assets follow a Brownian motion. This is the case where prediction of future prices using historical data is almost impossible. Similarly, this also supports our previous claim that the closing prices of both assets are non-stationary in general. Since Brownian motion is a Levy process [5], we can compute the tail exponent for closing prices, and then examine its relationship with the Hurst exponent. The results are summarized below:

	Ripple	EOS
Tail exponent α	0.598	2.698

Table 13: Tail exponent computed from negative closing price

7 Modeling, forecasting and validation

We can validate the Hurst exponents above for closing prices using tail exponent α. For a Levy process with tail exponent α, we use the following formula [6]:

$$\alpha \leq 1: \qquad H(k) = 1/k \quad for\ \ k \geq 1 \qquad\qquad \alpha \geq 2: \qquad \begin{aligned} H(k) = H = 1/2 \quad &for\ \ k < \alpha \\ H(k) = \alpha/(2k) \quad &for\ \ k \geq \alpha \end{aligned}$$

Ripple : $H(1) = 1, \quad H(2) = 1/2, \quad H(3) = 1/3$

EOS : $H(1) = H(2) = 1/2, \quad H(3) = 2.698/6 = 0.449$

Indeed the Generalized Hurst exponents $H(k)$ that we computed here for moments $k = 1$, 2 and 3 using tail exponent match closely with the true values on the table above, with EOS a better estimate than that of Ripple. Recall that we mentioned earlier how the tail exponent computed from log-returns could be useful in the scaling of our assets' log-returns distributions. For instance, we can approximate the daily risk measures by using hourly risk measures. Let us do this for VaR and implied volatility σ and then validate our result by comparing it to the empirical values. The daily VaR/volatility is related to the hourly VaR/ volatility by

$$VaR_{day} = 24^{1/\alpha}\ VaR_{hour} \qquad\qquad \sigma_{day} = 24^{1/\alpha}\ \sigma_{hour}$$

where $\alpha = 0.833$ for Ripple and $\alpha = 1.519$ for EOS, the exponents that we computed earlier from negative log-return. One important note is that it only applies for iid processes of stable distribution with $0 < \alpha \leq 2$. When $\alpha > 2$, we must always use 2 in our scaling calculation. The results are shown in the table below:

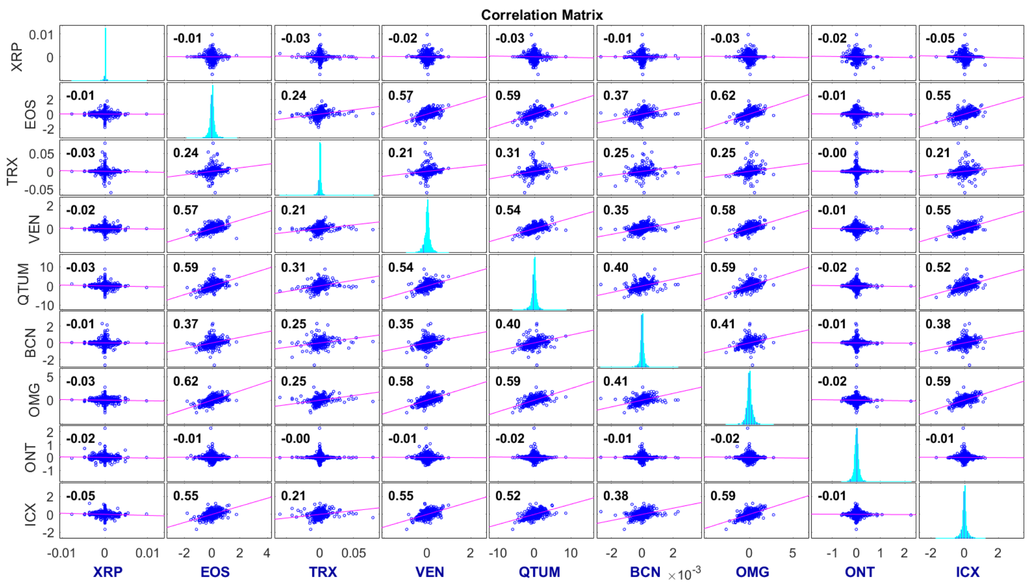
	VaR	σ
Ripple (estimate)	1.3389	2.2239
Ripple (empirical)	0.1454	0.114
EOS (estimate)	0.2455	0.1621
EOS (empirical)	0.1337	0.091

Table 14: Comparison of estimated daily risk measures and empirical values

In both cases, the difference between our estimate and the empirical value is fairly significant. Our estimate for EOS resembles more closely its empirical value than that of Ripple. This probably explains why the power law distribution fit of EOS is slightly better than that of Ripple as can be seen from the graphs above.

8 Linear Dependency

Two variables X_i and X_j are linearly related if they can be expressed in the equation $X_j = a + bX_i$. In this case, the correlation coefficient is equal to either 1 or -1. Any correlation coefficient between -1 and 1 corresponds to the equation $X_j = a + bX_i + \epsilon_{i,j}$, with $\epsilon_{i,j}$ a noise term, where X_i and X_j are only approximately linearly related [7]. Let us consider measure of linear dependence between each pair of asset returns on their closing prices via Pearson correlation coefficient. Note that we only take 2160 observations of each asset return since this is the maximum number of observation for one of our asset ONT. The result is depicted in the plot below:



We consider the 3 highest correlation pairs EOS/OMG, EOS/QTUM and VEN/OMG, then validate the accuracies of their correlation coefficients using permutation test at 95% significance level. The idea is that we generate an artificial reshuffled time series of our asset returns n times (n = 1000 enough for robustness in our case) and then test the significance of each coefficient $c_{i,j}$. Although this method is computationally expensive as we need to consider N(N-1)/2 hypotheses, where N is the number of our assets (N = 9 in our case). Note also that Bonferroni correction is also applied here to control the family-wise error rate. The null hypothesis of permutation test is that there is no correlation between pair of variables [8]. All values from now on are validated through permutation test. The table below shows our results:

	Pearson correlation coefficient	p-value
EOS/OMG	0.61777	1.5528e-227
EOS/QTUM	0.59078	1.986e-203
VEN/OMG	0.57926	8.2367e-194

Table 15: Pearson correlation coefficient and their p -values

The p-values are extremely low (almost close to 0) in all cases, suggesting that the probability of no correlation is close to 0%. Therefore, we can assert that there exists linear dependence between these 3 pairs of assets. We can also examine the cases where the observed dependency can be just the outcome of chance.

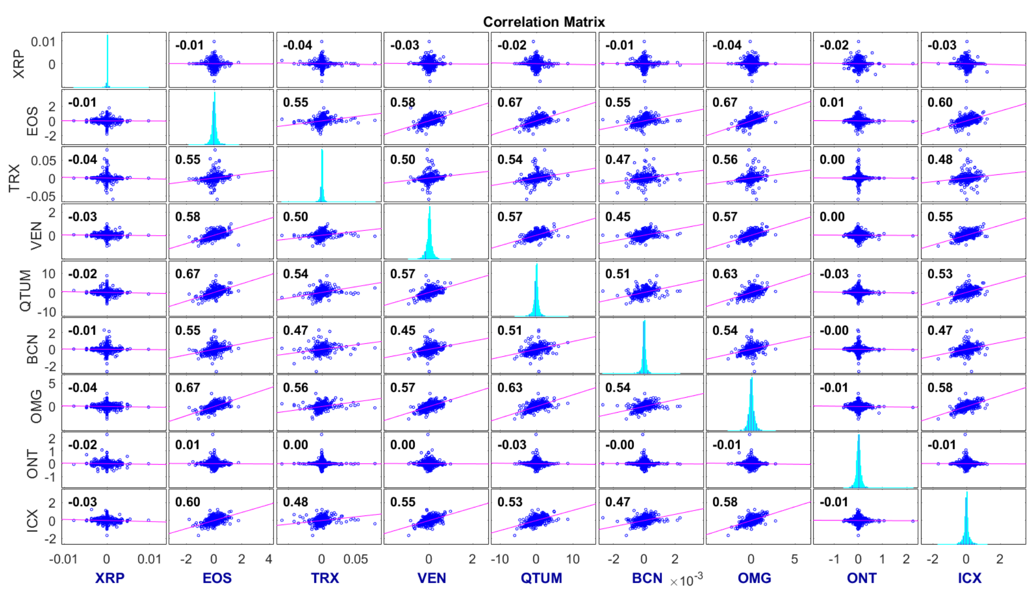
	Pearson correlation coefficient	p-value
XRP/EOS	-0.0053497	0.80376
XRP/VEN	-0.015357	0.47563
XRP/TRX	-0.032799	0.12754

Table 16: Pearson correlation coefficient and their p -values

Our results show that we accept the null hypothesis at the 5% level in all cases that there is no correlation between each pair of assets. In fact by looking at the correlation plot above, XRP (Ripple) attains low correlation coefficients in all pairings. This suggests that XRP returns are very weakly linearly dependent on other assets.

9 Non-Linear Dependency

Recall that when we model the ccdf of negative log-returns of both Ripple and EOS using fat-tailed power law distribution with tail exponent smaller or equal than 2. One problem that arise here is that the standard deviation is not defined with non-normally distributed variables, and so does the correlation coefficient. Therefore, we need to resort to other measure that can relate two random variables in a non-linear way. One such measure is Spearman rho correlation.



Note that the correlation coefficients are in general larger than that in the previous section. One can observe that XRP attains low correlation coefficients with all other assets as before. We are particularly interested in the biggest difference between Spearman rho and Pearson correlation coefficients. The reason is that it will provide an insight into which pair of assets are influenced the most by non-linear dependency. The table below shows 3 pairs of assets with the biggest difference:

	Correlation difference
TRX/EOS	0.31
TRX/VEN	0.29
TRX/ICX	0.27

Table 17: Correlation differences (Spearman - Pearson)

There is a clear trend that the TRX returns are non-linearly dependent on a few other assets. Let us check if our result is statistically significant.

	Spearman rho correlation	p-value
TRX/EOS	0.54916	1.9926e-170
TRX/VEN	0.50197	3.0686e-138
TRX/ICX	0.47749	1.9745e-123

Table 18: Spearman rho correlation and their p -values

Again the extremely low p-values above suggest that the probability of no correlation is nearly 0%. Therefore, we assert that there must exists non-linear dependency between these 3 pairs of assets.

10 Information-Theoretic Measures

The Spearman rho correlation is able to capture non-linear dependencies, but only through arbitrary monotonic function. It would be better if we could quantify dependence from a general perspective (i.e. can capture non-monotonic or a more complicated relationship). One way of doing it is to measure the distance between the joint density function $f(X_i, X_j)$ and the product of the two marginal densities $f(X_i)$ and $f(X_j)$. These two variables are independent if and only if such a distance is zero. Such approach is called mutual information. It is a more general measure of dependency with larger values corresponding to larger dependency. The table below shows our highest results:

	Mutual Information
TRX/BCN	8.0108
TRX/QTUM	7.3491
QTUM/BCN	7.2797

Table 19: Mutual Information values

Clearly both the Pearson and Spearman rho correlations poorly picked up on the 3 pairings above. However, it is found that both the mutual information and Spearman rho correlation yield a consistent result, suggesting that TRX returns appear to have some non-linear relationships with other assets' returns. Similarly, we can identify which pairs of assets are least likely to have dependency on each other.

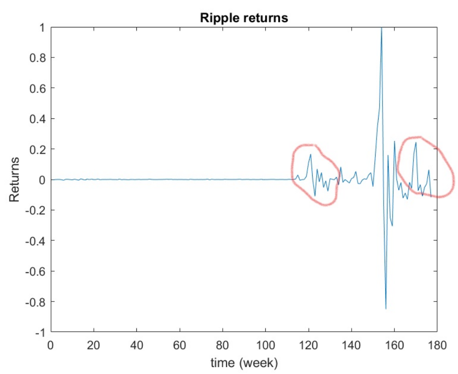
	Mutual Information
XRP/VEN	1.8611
XRP/ONT	1.8946
XRP/EOS	1.9622

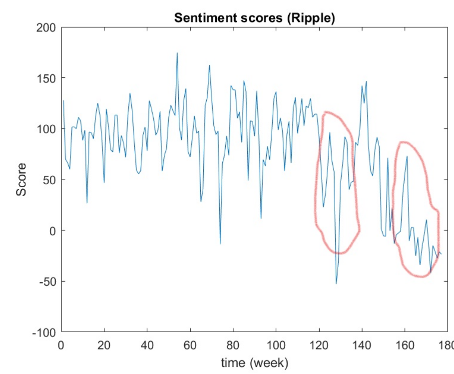
Table 20: Mutual Information values

In this case, it is found that both the mutual information and Pearson correlation yield a consistent result, suggesting that XRP returns indeed are weakly dependent on other assets' returns, whether linear or non-linear.

11 Textual Sentiment on Asset Returns

The dependency measures that we have learned so far does not give any information about causal relations. In particular we are interested in knowing whether events in the past cause the output in the future. We use an approach called Granger causality test to determine whether our time series of sentiment scores of both Ripple and EOS are useful in forecasting time series of their returns. The Matlab function that we use to conduct this test allows us to specify maximum number of lags (in the short or long term) to be considered [9]. Below are the time series returns and sentiment scores of Ripple respectively. Note that we use weekly instead of hourly observations for ease of interpretation on our plots.





Our plots show that the patterns in time series sentiment scores are approximately repeated in time series returns after a certain time lag (2 examples indicated by the red circles). Let us look at the result of Granger causality test at the 5% level:

Max lags	F statistic	critical value
daily	0.1040	3.8418
weekly	4.0982	3.8458

Table 21: Granger causality test (Ripple)

Over a short timescale (daily), we accept the null hypothesis (F statistic < critical value) that the sentiment scores of Ripple do not provide significant information about its returns. However, there exists a causal relationship between longer-term (weekly) sentiment scores and returns. Thus, past values of sentiments can be used for the prediction of future values of returns over a longer timescale.

Similarly, we conduct the same test on EOS data. The table below shows our F statistics and critical values in the short and long term.

Max lags	F statistic	critical value
daily	0.1691	3.8418
weekly	1.0110	3.8438

Table 22: Granger causality test (EOS)

Our result shows that we accept the null in both the short and long term, indicating that past values of sentiments of EOS is not useful at all in predicting the future values of its returns. Indeed if we were to plot the same graphs on both the sentiments and returns of EOS as above, we could not find any patterns between the two time series.

Summary

It is shown that the poor prediction is obtained when we tried to forecast both the VaR and implied volatility using tail exponent by scaling of our log-returns distribution. In fact, this scaling law only applies to processes that are a result of sum of iid random variables. However, many of the real-world processes like our data are clearly not independent of each other. Moreover, both the autocorrelations and Hurst exponents show that asset price returns trend change continuously. No promising result is attained for justification of portfolio formation in the earlier section. We then extend our analysis by considering more assets. It is found that Ripple has a low dependency with all other assets, whether linear or non-linear; whereas EOS is highly dependent on other assets in general. Additionally, it is easier to track Ripple returns than that of EOS using textual sentiments as a result of Granger causuality test. Therefore, it is probably rational to invest in Ripple rather than EOS in order to create a well-diversified portfolio consisting of other assets, so that our total portfolio do not suffer significant financial losses if other assets suffer a serious downturn.

References

[1] Anon, (2019). [online] Available at: https://www.quora.com/Why-are-stock-returns-log-normal.

[2] Macroption.com. (2019). Value At Risk Advantages: Why Use VAR in Risk Management - Macroption. [online] Available at: https://www.macroption.com/value-at-risk-var-advantages/.

[3] Moodle-1819.ucl.ac.uk. (2019). UCL Moodle: Log in to the site. [online] Available at: https://moodle-1819.ucl.ac.uk/pluginfile.php/357987/modresource/content/2/COMP0040 Part1b Probability Further 2019v1.0.pdf.

[4] Mathworks.com. (2019). Generalized Hurst exponent - File Exchange - MATLAB Central. [online] Available at: https://www.mathworks.com/matlabcentral/fileexchange/30076-generalized-hurst-exponent.

[5] Stat.berkeley.edu. (2019). [online] Available at: https://www.stat.berkeley.edu/ pitman/s205s03/levy.pdf.

[6] Moodle-1819.ucl.ac.uk. (2019). UCL Moodle: Log in to the site. [online] Available at: https://moodle-1819.ucl.ac.uk/pluginfile.php/357989/modresource/content/16/COMP0040 Part2 Scaling 2019v1.0.pdf.

[7] Moodle-1819.ucl.ac.uk. (2019). UCL Moodle: Log in to the site. [online] Available at: https://moodle-1819.ucl.ac.uk/pluginfile.php/357991/modresource/content/19/COMP0040 Part3 MultivariateStaistics Dependency 2019v1.0.pdf.

[8] Mathworks.com. (2019). multcomppermcorr - File Exchange - MATLAB Central. [online] Available at: https://www.mathworks.com/matlabcentral/fileexchange/34920-multcomppermcorr.

[9] Mathworks.com. (2019). Granger Causality Test - File Exchange - MATLAB Central. [online] Available at: https://www.mathworks.com/matlabcentral/fileexchange/25467-granger-causality-test.