Fat-tailed Distributions and Extreme Events

by

© Amir Gholizad

A project submitted in partial fulfillment of the requirements
for the course CMSC 6950.

Instructor: Dr. Scott MacLachlan

Department of Mathematics and Statistics

Memorial University

November 2023

St. John's, Newfoundland and Labrador, Canada

Abstract

This project dives into financial data to check out the fat tails in market returns. Using historical info, it calculates returns and makes probability distributions to see how extreme market events happen. The goal is to measure and compare the risk of these extreme events by looking at the fat tails. The study comes up with a number that shows this risk and puts it side by side with a standard Gaussian distribution. The findings highlight the importance of considering fat tails in financial modeling for better risk understanding and management.

Table of contents

Chapter 1

Data Set

This chapter initiates the research journey by outlining the process of obtaining the dataset crucial for subsequent analyses.

The primary tool employed for this purpose is the yfinance Python package, which interfaces with **yahoo!finance**'s API. The choice of yahoo finance is motivated by its extensive coverage and reliability in providing historical data across various financial instruments.

We collected **Bitcoin**'s historical price data in daily, weekly, monthly, and quarterly time frames. The daily data captures finer market nuances, while the broader time frames offer a more comprehensive view of trends.

In order to get the data, we constructed a function and passed in the values

```
ticker = "BTC-USD"
start_date = "2013-01-01"
end_date = "2023-01-01"
```

as our chosen symbol and the interval.

The Bitcoin's price time series in four different time frames are shown in Figure 1.1.

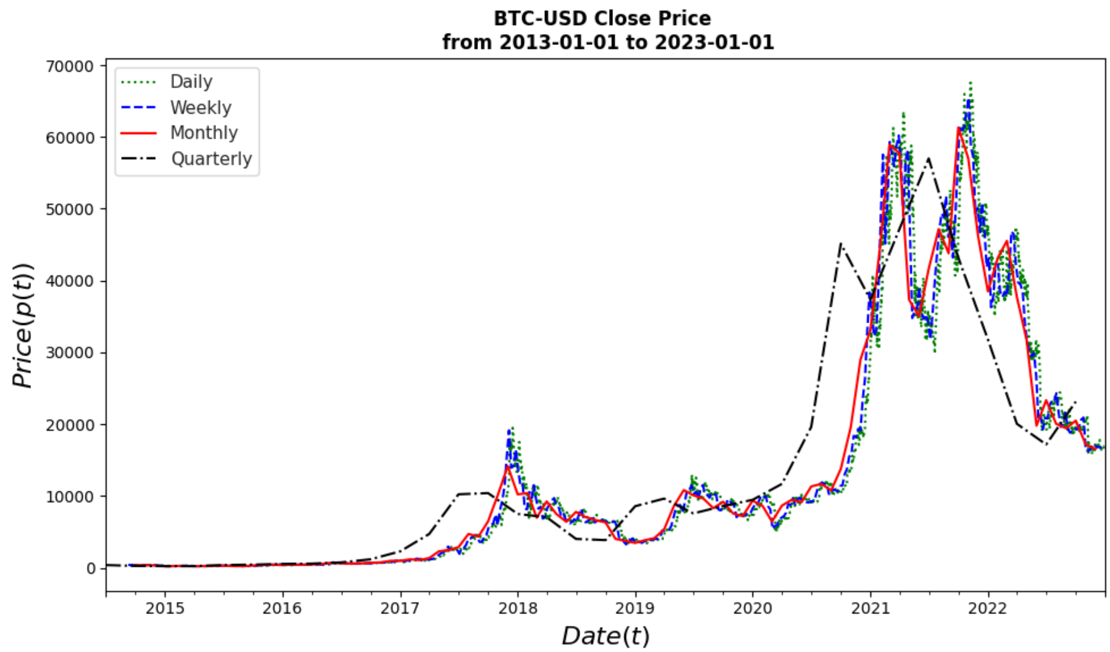


Figure 1.1: Daily, Weekly, Monthly, and Quarterly prices of Bitcoin in USD from Jan. 1st, 2013 to Jan. 1st, 2023

In summary, Chapter 1 sets the stage by introducing the data source and temporal scopes, laying the groundwork for the methodology, subsequent analyses of logarithmic rate of returns and the construction of fat-tailed distributions.

Chapter 2

Methodology

2.0.1 Rate of Returns and Kernel Density Estimation

Let's assume that two prices for a security at times t, $t + \tau$ are $p(t)$ and $p(t + \tau)$ where τ is the time step. We can define the return of these prices as

$$q(t) = ln(p(t + \tau)) - ln(p(t)). \tag{2.1}$$

After finding the returns for the entire historical data, we can observe the rate of occurrence of returns by creating their distribution. Figure 2.1 shows the rate of occurrence vs the return of two subsequent prices as a histogram.

KDE

The primary goal of this project is to apply numerical integration to the tails of distributions, opting for probability density functions (p.d.f) over histograms. However, due to the lack of knowledge about the analytical form of a p.d.f, we employed a technique known as *Kernel Density Estimation (KDE)*[1] to approximate each p.d.f in the project. KDE is a non-parametric method for estimating the density of a random variable. One notable advantage of utilizing KDE is its straightforward form and straightforward implementation.

The kernel density estimator for independent and identical distribution f of each

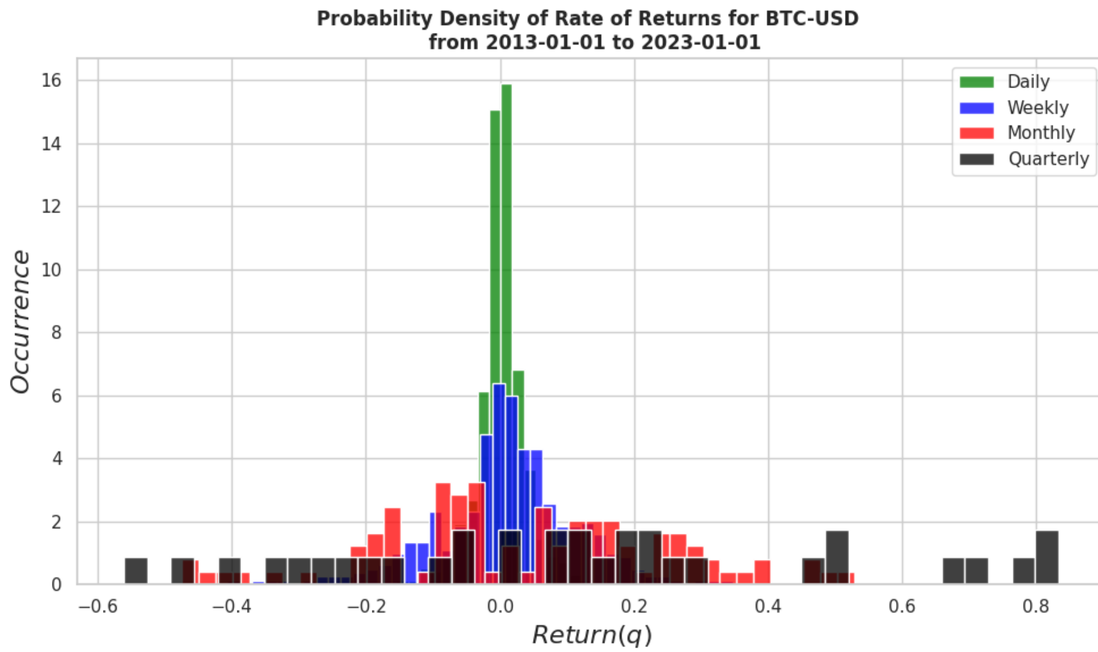


Figure 2.1: Daily, Weekly, Monthly, and Quarterly log-returns of Bitcoin calculated using Eq. 2.1 (Jan. 1st, 2013 - Jan. 1st, 2023)

state q_i in a return time series can be expressed as

$$f_b(q) = \frac{1}{nb}\sum_{i=1}^{n} K(\frac{q - q_i}{b}). \tag{2.2}$$

where K is a non-negative kernel function that integrates to one and has a mean of zero. The parameter b denotes the bandwidth, which is a positive smoothing parameter, and n represents the size of the random variable. Various commonly used kernel functions exist, and we used the Gaussian kernel function

$$K(y) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2}) \tag{2.3}$$

Determining an appropriate bandwidth (b) for each time series holds significant importance. We adhered to the rule of thumb method proposed by **Silverman**[2][3], which provided the optimal bandwidth for our distributions. Our approach involved utilizing the `seaborn.kdeplot()` function from the `seaborn` library in Python to approximate the Daily, Weekly, Monthly, and Quarterly distributions of price returns.

Subsequently, we stored the data from these plots as our refined, regenerated, and improved data set. From this juncture onward, the sole data sets utilized in the project consist of the new q_i and $f_b(q_i)$ grid points derived from the `seaborn.kdeplot()`. The KDE fits over the original histograms are shown in Figure 2.2.

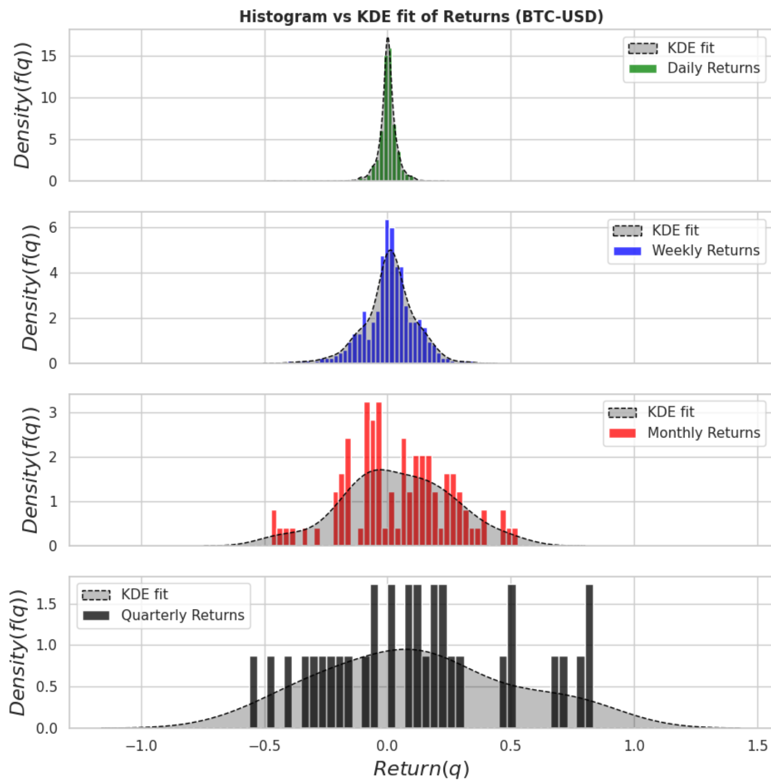


Figure 2.2: KDE fits vs Histograms of log-returns of Bitcoin in Daily, Weekly, Monthly, and Quarterly time frames (Jan. 1st, 2013 - Jan. 1st, 2023)

We can see that the distribution of returns looks like a Gaussian or Normal distribution. Using Gaussian function

$$f(q_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(q_i - \mu)^2}{2\sigma^2}), \tag{2.4}$$

where q_i is the ith price return, μ is the average or the expected value of price returns and σ is the standard deviation of the price returns, we can plot the Gaussian fit for the price returns. Figure 2.3 shows the difference between empirical and Gaussian distributions of price returns for daily, weekly, monthly, and quarterly times cales.

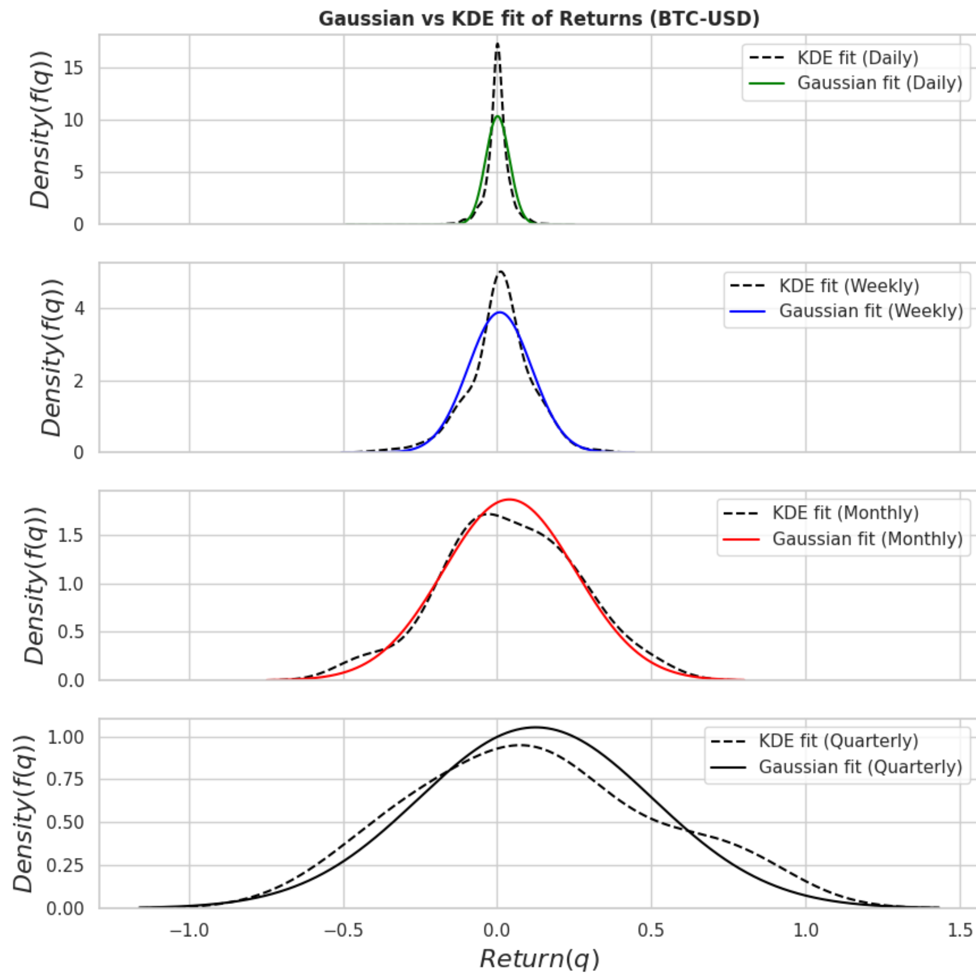
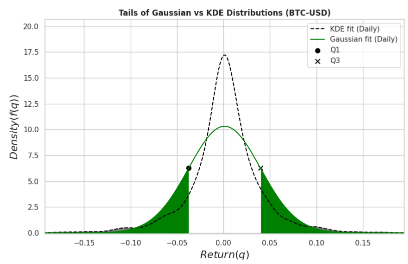
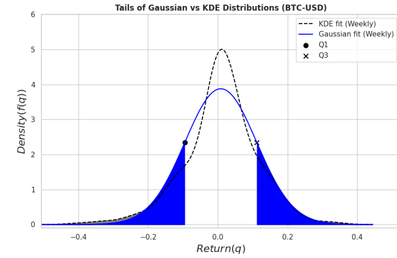


Figure 2.3: KDE vs Gaussian fits of log-returns of Bitcoin in Daily, Weekly, Monthly, and Quarterly time frames (Jan. 1st, 2013 - Jan. 1st, 2023)

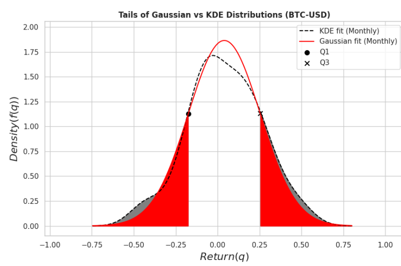
Now that we have a uniform grid points of q and $f(q)$, we can set a criteria for identifying the tails of every distribution. We decided to find the area under the curves in ranges $q \leq Q_1$ and $q \geq Q_3$, where Q_1 an Q_3 are the first and third quantiles. Figure 2.4 shows the area under the curves that we're are going to measure. We used `np.trapz()` to integrate along the q axis using *Trapezoidal rule*.



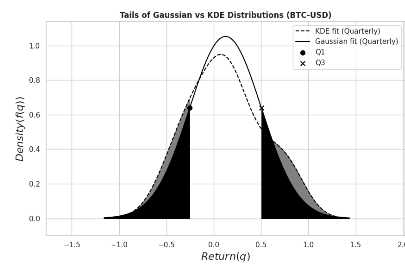
(a)



(b)



(c)



(d)

Figure 2.4: The area under the tails of distributions driven from Gaussian and KDE fits for (a)Daily, (b)Weekly, (c)Monthly, and (d)Quarterly time frames of returns (Bitcoin: [Jan. 1st, 2013 - Jan. 1st, 2023])

Chapter 3

Results

Our aim is to find a feature that indicates the differences of empirical distributions and their normal fits. We calculate

$$Tails \equiv \Delta T = T_{KDE} - T_{Gaussian}$$

where T_{KDE} and $T_{Gaussian}$ are the area under the tails of Gaussian and KDE distributions.

After defining a new feature, it is common to compare it with other statistical features. We compared the trends of Mean, Standard Deviation, Skewness, and Kurtosis with the trend of Tails in 4 time frames. Since we're focusing on the trends only, we used `sklearn.preprocessing.MinMaxScaler().fit()` to normalize the trends between $[0, 1]$. We can also study the correlations between these features. The results are illustrated in Figure 3.1 and Figure 3.2.

	Gaussian/KDE Tails Difference	Mean	Standard Deviation	Skewness	Kurtosis
Gaussian/KDE Tails Difference	1	0.661573	0.546811	-0.556427	-0.575301
Mean	0.661573	1	0.979546	-0.511457	-0.972189
Standard Deviation	0.546811	0.979546	1	-0.341079	-0.93229
Skewness	-0.556427	-0.511457	-0.341079	1	0.645574
Kurtosis	-0.575301	-0.972189	-0.93229	0.645574	1

Figure 3.1: Correlation of four statistical features and the new feature.

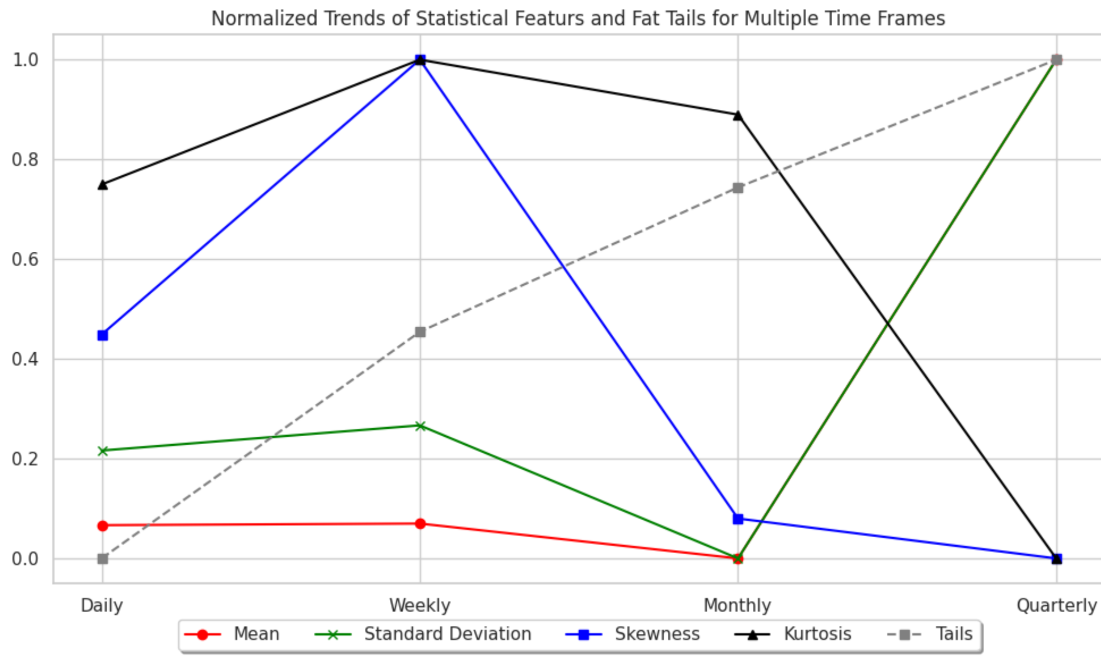


Figure 3.2: The trends of changes of four statistical features and the new feature, Tails by moving to bigger time frames.

In the realm of financial modeling and risk analysis, the study of fat-tailed distributions has gained considerable prominence due to their ability to capture extreme events that may have a significant impact on markets. Observing the behavior of these distributions across varying time frames reveals intriguing patterns. Notably, as the temporal horizon expands, the tails of the distributions exhibit a discernible increase in their magnitude, indicative of a heightened likelihood of extreme events. This phenomenon aligns with the concept of volatility clustering, wherein periods of high volatility tend to cluster together over longer time scales. Concurrently, the parameters of the distributions, such as the *mean* and *standard deviation* undergo a proportional expansion, signifying a broader dispersion of data points. Furthermore, the enlarging area under the tails underscores the escalating probability of extreme outcomes, underscoring the importance of incorporating robust risk management strategies in financial decision-making processes, particularly as the analysis extends to broader temporal horizons. In essence, these findings contribute valuable insights into the dynamic nature of financial markets, offering practitioners a nuanced understanding of risk dynamics across different time frames.

Bibliography

[1] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pages 832–837, 1956.

[2] Simon J Sheather. Density estimation. *Statistical science*, pages 588–597, 2004.

[3] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.