Analysis of high frequency data

BTCUSD(T)

Kosmetsas Tilemahos & Koutkos Christos

A thesis presented for the degree of
Master in Financial Engineering

WorldQuant University
Greece
December 2021

Abstract

High Frequency analysis

BTCUSD

Tilemachos Kosmetsas & Christos Koutkos

Abstract

Dedication

To our beautiful families

Declaration

I will have a winning strategy

Acknowledgements

Thank you all

Contents

1 Introduction **8**
1.1 Problem Statement 8

2 Exploration **10**
2.1 Introduction 10
2.2 Volume 10

3 Sampling **17**
3.1 Introduction 17

List of Figures

1.1 BTCUSD Volume sampled in several timeframes 8
1.2 Volume per trade (tick volume). 9

2.1 Quarterly volume across spot exchanges. 10
2.2 BTCUSD and BTCUSDT spot trading volume (in bitcoin). 11
2.3 BTCUSD and BTCUSDT spot trading volume (in USD(T)). 11
2.4 Mean volume per trade. 12
2.5 Retail trades and BTC price. 13
2.6 Histogram of BTCUSD and BTCUSDT no of retail trades per day . . 13
2.7 Fiat premium. 14
2.8 PCA analysis - 1st principal component and BTCUSD price. 15
2.9 KDE plot of volumes aggregated on 4h timeframe, for all exchanges. . 15
2.10 Eigendecomposition on kendal correlation matrix for positive and negative volume for **BTCUSDT** markets. 16

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

In the past couple of years, a vast inflow of retail and corporate capital has entered the cryptocurrency markets. As time goes by, one may notice a rising interest in these markets, as well as, an almost exponential increase in trading volume. Although the cryptocurrency market has many similarities with the traditional, the authors felt that the differences between them, are enough to differentiate their behavior from the traditional assets and thus investigation and research is deemed mandatory.

The approach the authors will take in this assignment is to analyze existing ideas and implement them on BTC timeseries, but at the same time, explore new approaches and combinations. The difficulty of this project lies with the asynchronous nature of information. The way that information appear in the market must dictate the way they are represented, perceived by the researcher and used by a model. To illustrate this, we will use BTCUSD volume data, from Bitstamp exchange with index ranging from 2020-06-15 to 2020-09-15, aggregated weekly.

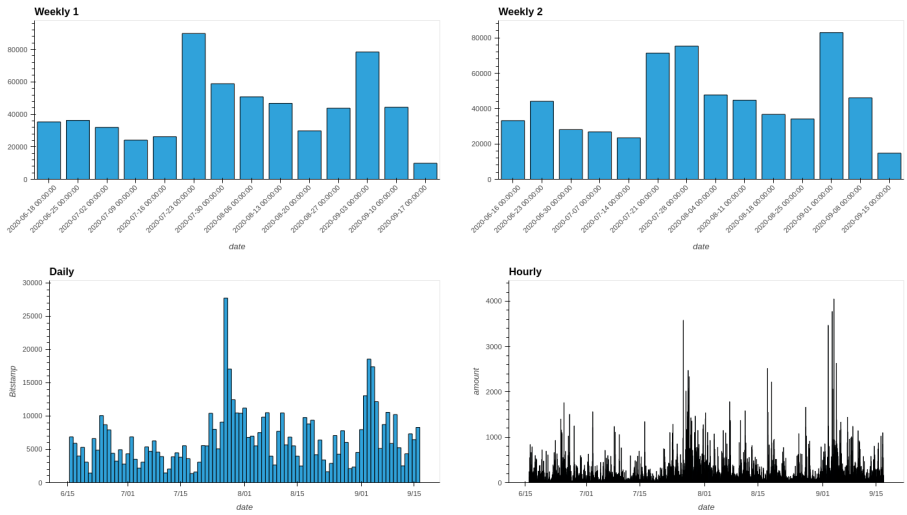

Figure 1.1: BTCUSD Volume sampled in several timeframes

On the *Weekly 1* chart, we observe that the week starting at 2020-07-23, has the biggest spike in volume across these 3 months while the next weeks exhibit declining volume. Another spike at the week starting at 2020-09-03, also takes

place. The *Weekly 2* chart, is drawn on the same data, but before aggregating in weekly timeframe (from daily), the dataset got shifted by 3 days to the left. As a result, the new chart is different from the previous one, as we observe that the 2 week period that begins at 2020-07-21 had significant volume, but the highest spike now occurs at the week that starts 2020-09-08.

By changing the resolution to the daily timeframe, we observe that the volume that was attributed to two weeks in the previous graph, actually took place in 5 days, and the biggest spike in volume occurred in 2020-7-25. Further enhancing the resolution and aggregating to the hourly timeframe, the *Hourly* chart, shows a different story. There is a cluster of volume occuring at 2020-07-25 and persisting for the week to come. More importantly, we observe a second spike around 2020-09-05 that is more pronounced but not as persistent (in terms of lags) as the first one.

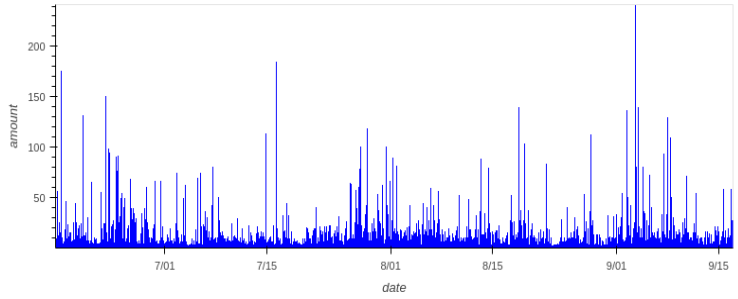

Figure 1.2: Volume per trade (tick volume).

Lastly, the 1.2, is the highest resolution possible and contains all the information we could possible get for volume in Bitstamp during that period. This chart, looks more like a series of impulses (sudden spikes) while some clusters of volume can be seen on the bottom of the graph.

What a researcher and an algorithm might extract from the above data, could be different in each occasion, nevertheless, it is the same data (except for the 3 days shift that illustrate the danger of sampling in large timeframes), containing the same information. The above example used different fixed timeframe intervals but the same applies to sampling based on the side of the trade, or the number of trades.

So, why not always use the highest resolution possible, in order to preserve all the information? This question leads us to the next tradeoff: The lower the resolution, the more information is lost, and the higher it is, the more noisy and less useful the data become.

The above example illustrates the main drive of this project: the necessity for proper sampling in high frequency data. This project, will opt to overcome this problem by sampling asynchronously and dynamically, on the same dataset and across different features.

Chapter 2

Exploration

2.1 Introduction

In this chapter, we will explore the BTCUSD(T) market across 5 major exchanges by following a visual approach on aggregated data. The sampling that is used at this stage is across time, volume and number of trades in a fixed window. Key insights that will be extracted, will serve as the infrastructure of a dynamic way of sampling.

2.2 Volume

Volume is an important aspect of all financial data. Exploring volume across exchanges is a significant task that will provide our analysis with the insights as to how someone should proceed in using trade-to-trade and aggregated volume in several windows, in order to create meaningful signals.

The trade data for BTCUSD begin as early as 2011, with few exchanges offering the opportunity to trade this asset. The first exchange was MtGox. It was launched in 2010 and closed in April 2014 due to fraud, as more than 850,000 BTC were missing Wikipedia, 2021. As time passed by and BTC gained even more traction, the trade volume upscaled significantly and more exchanges, such as Bitstamp, Kraken and Coinbase, appeared. We will consider 2016 - 2017, as the years that BTC became known enough, to attract the first retail and institutional players.

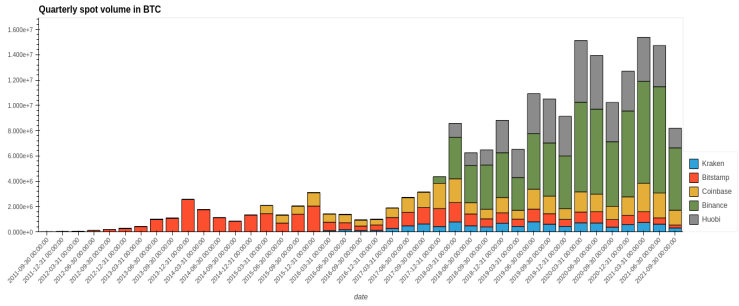

Figure 2.1: Quarterly volume across spot exchanges.

As we can see in 2.1, the overall trading volume begun to rise in early 2017, as more people were attracted to the impressive BTC bull run, up until that point. At this point, we could distinct the BTCUSD from BTCUSDT volume following the

assumption that a retail trader is forced to use fiat currency in order to buy bitcoin for the first time, in some centralized exchange, thus the bitcoin volume on USD, could serve as an indicator of retail activity.

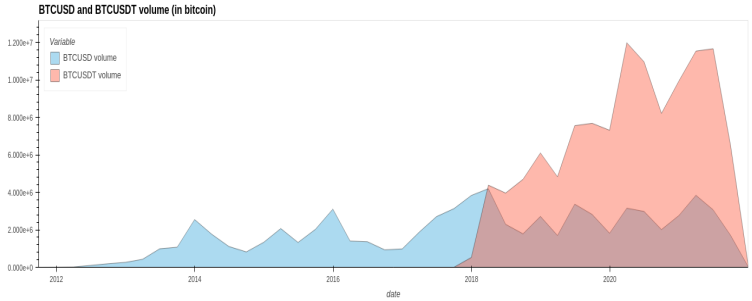

Figure 2.2: BTCUSD and BTCUSDT spot trading volume (in bitcoin).

The first thing to notice in 2.2, is that since the 2017 BTC bullrun, the BTCUSD volume (in bitcoin) is slightly elevated. Furthermore, since the introduction of USDT, the exchanges that offered BTCUSDT trading, easily surpassed those that offered only BTCUSD. The latter is to be expected, since USDT is 'tethered' to the USD (stable coin offering safety from volatility), while being at the same time easily transferable across exchanges in contrast to fiat. On the other hand, the 2.3 shows a steep increase in dollars traded that can be attributed to the increase in bitcoin price.

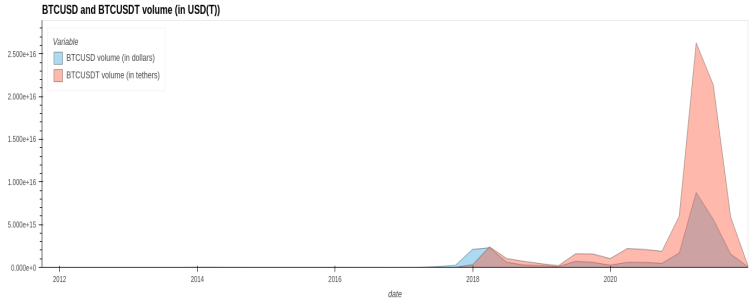

Figure 2.3: BTCUSD and BTCUSDT spot trading volume (in USD(T)).

In the next four graphs 2.4, we can see the mean trading volume in bitcoin and dollars for BTCUSD and BTCUSDT. As we expect, in the upper two graphs, the mean trading volume decreases as bitcoin price increases. In contrast to the above, the bottom graphs, show an increase in mean trading volume, although, this increase, is different for the two markets: the BTCUSD market shows the 'anticipated' behavior that can be explained by the BTC price and the increased interest to this new asset, and the BTCUSDT market, exhibits a smaller increase in mean trading volume (dollars) even though the volume traded in USDT is higher than the volume traded in USD. The latter indicates the existence of many small buy/sell orders in the USDT markets.

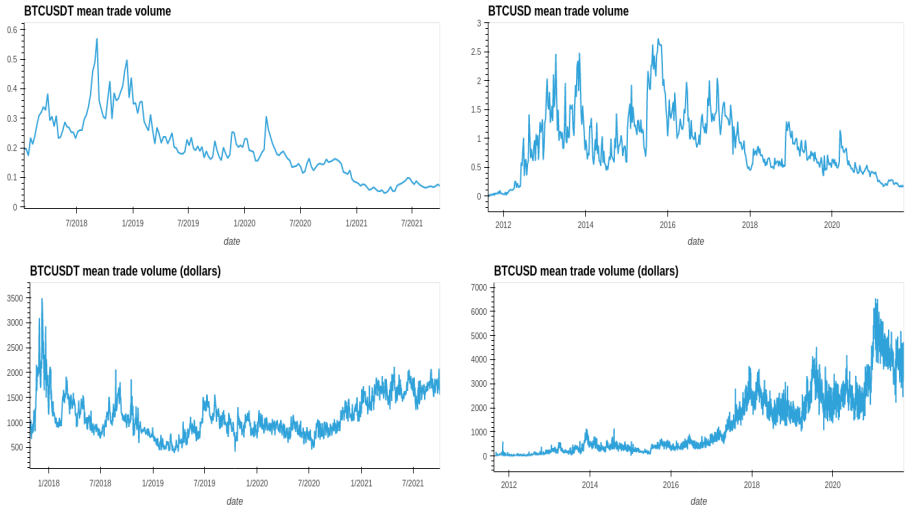

Figure 2.4: Mean volume per trade.

Furthermore, an evolving market such as the crypto market, attracts retail, institutional and high frequency traders. In order to classify a trade as retail or not, two important assumptions must be met:

- Retail traders are trading in integer dollar volumes, and most likely in multiples of 10, and
- Institutional investors will more likely buy and sell in OTC (Over The Counter) markets.

In order to extract the possible retail trades, we chose a mean transaction cost $c = 0.022\%$ per trade, and extracted it from all trades. If a trade was divisible by 10, it was classified as a retail trade. Nevertheless, the fee structure is different across exchanges and even different across traders in the same exchange (volume per month dependent). Therefore, we chose to include an error $e = \$0.15$ as an acceptable distance from the closer mutliple of 10. The trades chosen, should be trades made manually by some trader and not an algorithm (that tends to trade in many decimals). Furthermore, these trades could be made by a professional of a small magnitude and not a retail trader. For briefness purposes, we will refer to these trades, as retail trades, and the traders that initiated them, as retail traders. The above assumptions are flawed in the sense that someone can buy/sell in bitcoin denominated values (0.5 btc or 1 btc), therefore, this metric can capture only a small percentage of retail trades. Nevertheless, based on the data that the authors possess, there is no other way to classify a trade as 'retail trade'.

In the figure 2.5, we can see that the estimated number of retail trades on BTCUSDT, is from 4 to 12 times bigger than the one on BTCUSD. Since a retail trader that wishes to trade for the first time, is forced to use fiat currency, we could assume, that the BTCUSDT trades, were executed from retail traders that were active in previous market cycles as well (2017 bull run and before).

On the top right graph, we can see the ratio of BTCUSDT to BTCUSD trades. We observe that the top is reached during May 2021, when the first large correction of the latest bull markets occured.

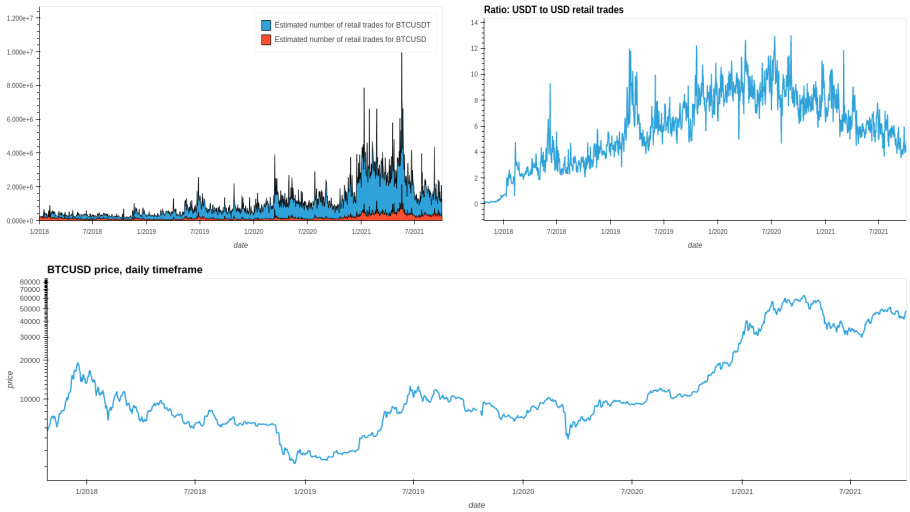

Figure 2.5: Retail trades and BTC price.

The increasing ratio indicates that BTCUSDT trades are relatively more precise in following the bull run (experienced retail traders) while the ratio starts declining, close to market top, indicating the timing when retail activity starts to gain traction in BTCUSD market, where is more likely for a 'first time retail trader' to trade.

On the next histograms, the difference in retail activity between BTCUSD and BTCUSDT becomes even more apparent. In the BTCUSD case, the graph is skewed to the left, with few days distributed to the extremes $> 600,000$. The BTCUSDT markets though, as indicated from standard deviation which is 3 times greater than the one in BTCUSD, show that the retail activity is distributed more evenly. A further search on this, should reveal the events that triggered some of the bellow extreme values.

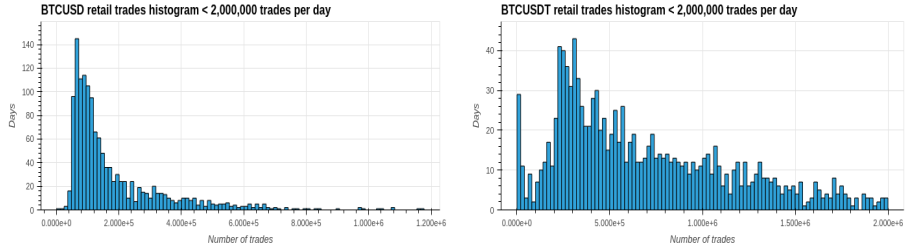

Figure 2.6: Histogram of BTCUSD and BTCUSDT no of retail trades per day

From the summary statistics, we can see that the mean, 25%, 50% and 75% are three to four times greater in BTCUSDT markets, indicative of the preference of retail traders to USDT.

Summary Statistics		
	BTCUSD	BTCUSDT
count	1.372000e+03	1.206000e+03
mean	1.837303e+05	6.846854e+05
std	1.632955e+05	4.669768e+05
min	1.058000e+03	4.790000e+02
25%	8.008925e+04	3.093760e+05
50%	1.186430e+05	5.595545e+05
75%	2.206602e+05	9.985932e+05

The differences between BTCUSD and BTCUSDT markets, extend to the bitcoin price as well. In the next figure 2.7, we can see that there are arbitrage opportunities between BTCUSD and BTCUSDT markets but not among the markets themselves. These opportunities seem to be available in periods of sudden price movements, and could be accredited to the difference in volume between the two markets. Throughout the 2021 bull market, there was a consistent discrepancy in the fiat premium index.

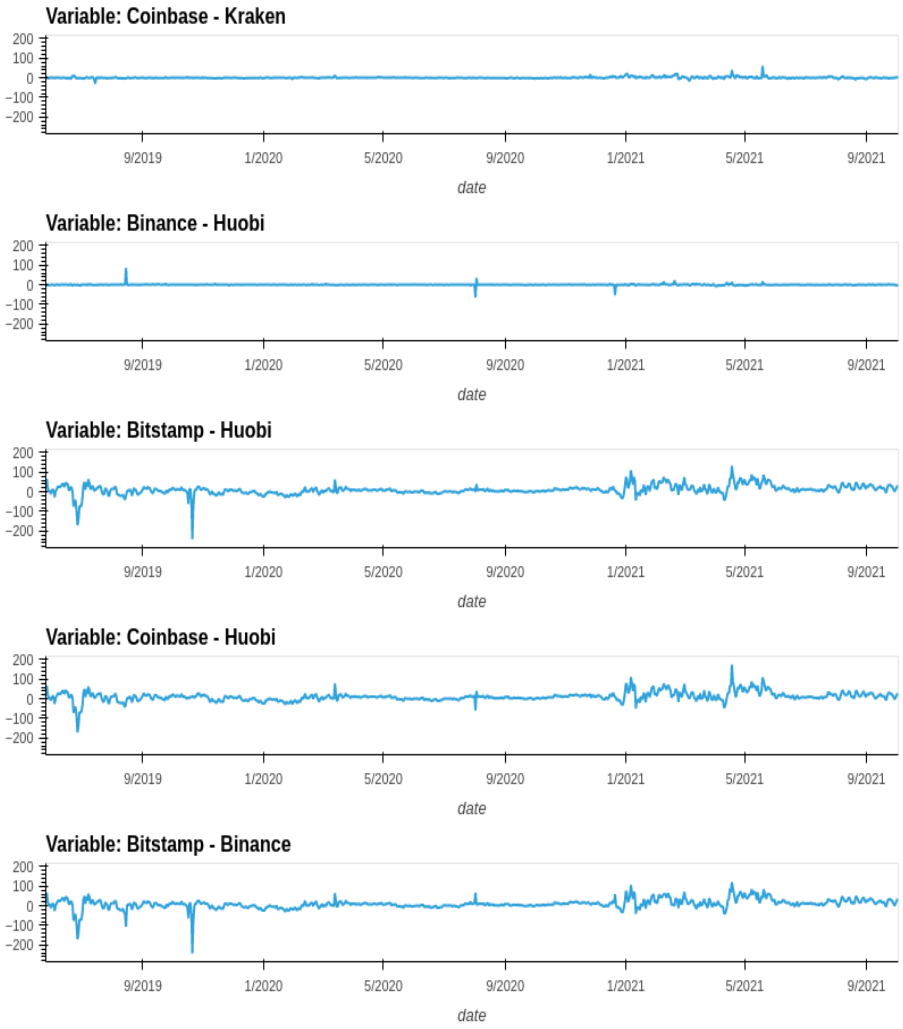

Figure 2.7: Fiat premium.

Such discepancies could be a valuable source of imbalances, that could lead to a more precise sampling. Next, we will explore volume a bit deeper. We will decompose (eigendecomposition) the covariance matrix of volume, of the BTCUSD and BTCUSDT markets. The computation will take place in a rolling fashion under a fixed time interval in order to capture the convergence of volume, between the exchanges in different phases of the market.

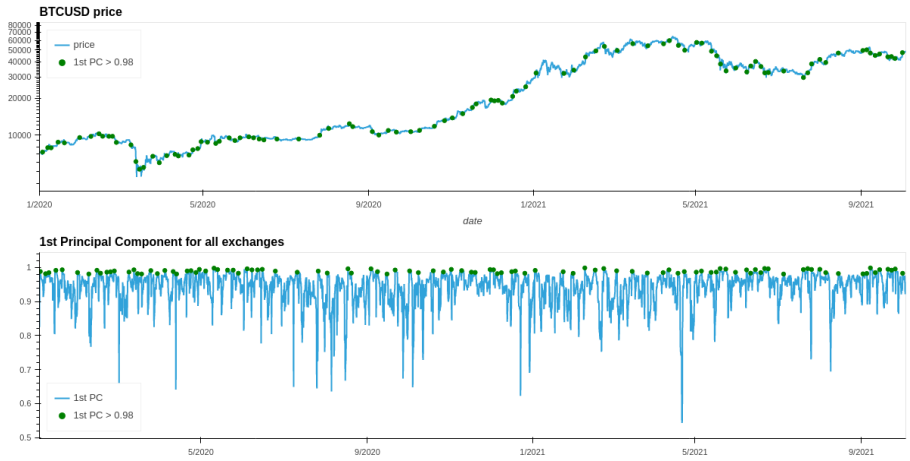

Figure 2.8: PCA analysis - 1st principal component and BTCUSD price.

In the above figure 2.8, we can see that based on the covariance of the volumes across exchanges, the 1st principal component seems to explain almost all variance most of the time. This finding, enhances the idea that information is quickly transfered and volumes generally converge. The same must be tested for metrics other than covariance. An appropriate such metric, is the first principal component computed from the eigendecomposition of the Kendall correlation matrix. Since the volumes are not normally distributed (figure 2.9) , we cannot use neither Pearson or Spearman correlation .

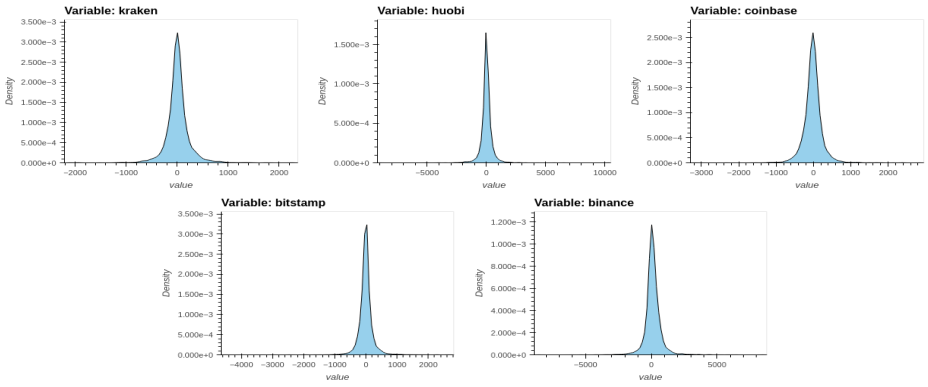

Figure 2.9: KDE plot of volumes aggregated on 4h timeframe, for all exchanges.

Kendall's Tau (τ), is a non parametric test that is used to measure the correlation between two variables. There are three different variations of this test, but mostly the Tau-b (τ_b) is used. The formula is:

$$\tau_b = \frac{2(n_c - n_d)}{\sqrt{n(n-1) - G_x}\sqrt{n(n-1) - G_y}}$$

where:

- n_c is the number of concordant values
- n_d is the number of discordant values

- $G_{x,y} = \sum t_i(t_i - 1)$ where t_i is the number of tied values in the i group of the $\{x, y\}$ variable

For the next figures, we classified the volume V into positive volume and negative volume. The computations involved the sign of the returns $b_t = sign\{p_t - p_{t-1}\}$, where p_t is the price at time t (this computation took place on tick data therefore t is the time measured in number of ticks), multiplied with volume at time $t - 1$: $b_t \cdot V_{t-1}$. This computation created an additional two volumes. The rationale behind this, is that negative volume will be responsible for negative returns and positive volume for positive returns.

Figure 2.10: Eigendecomposition on kendal correlation matrix for positive and negative volume for **BTCUSDT** markets.

In figure 2.10 we can see the convergence of positive and negative volumes among BTCUSDT market. The 1st principal component has consistenly high explained variance ratio > 0.7 which shows that volume between Binance and Huobi, are following the same direction most of the time.

Upon close inspection, it seems that sudden price moves can be associated with higher convergence of volume between the BTCUSDT exchanges. The same seems to be the case, for all exchanges as well (figure 2.8). That leads us to the idea that we could sample when there is convergence in a feature of choice (volume, positive-negative volume, buy/sell volume, number of trades per interval), assuming that in order for such an event to occur, there must be some new information.

Chapter 3

Sampling

3.1 Introduction

In this chapter, we will use the insights from the previous chapter, in order to extract meaningful samples from the dataset. At this point, we will use the BTCUSDT dataset from Binance. For simplicity purposes, before sampling, we will aggregate the data to the The algorithms responsible for sampling, can be used for other datasets as well, after finetuning the parameters associated with the functions that will be created.

Appendix

The appendix with all the code

Bibliography

Wikipedia. (2021). *Mt. gox*. https://en.wikipedia.org/wiki/Mt._Gox. (Cit. on p. 10)