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

List of Figures

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.

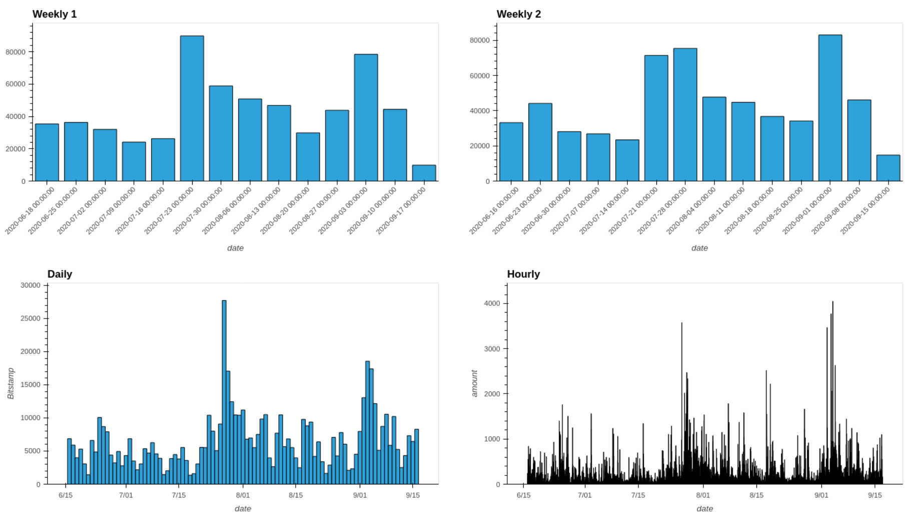

Figure 1.1: BTCUSD Volume sampled in several timeframes

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.

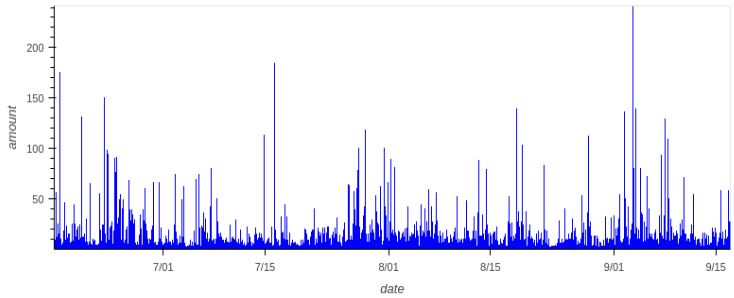

Figure 1.2: Volume per trade (tick volume).

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.

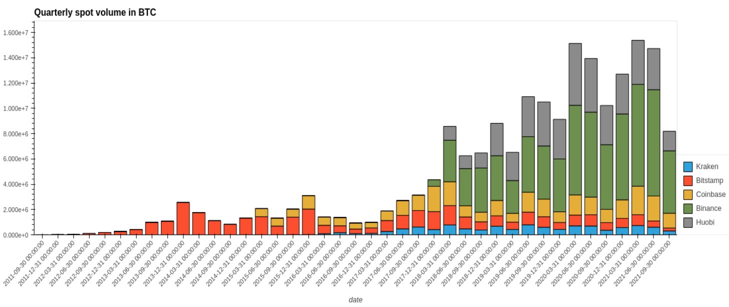

Figure 2.1: Quarterly volume across spot exchanges.

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.

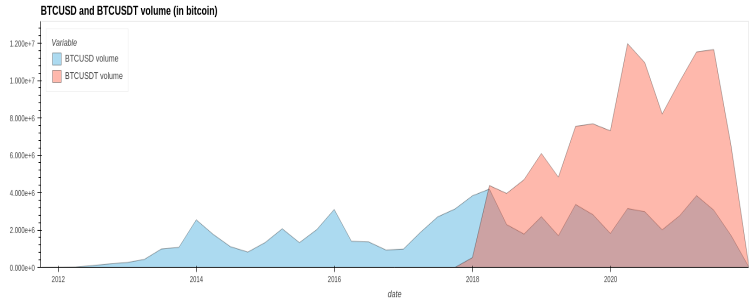

Figure 2.2: BTCUSD and BTCUSDT spot trading volume (in bitcoin).

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.

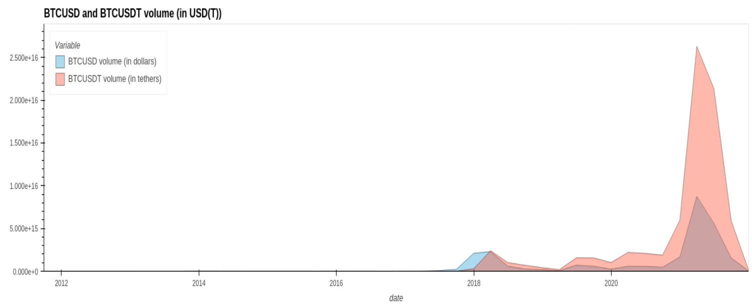

Figure 2.3: BTCUSD and BTCUSDT spot trading volume (in USD(T)).

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.

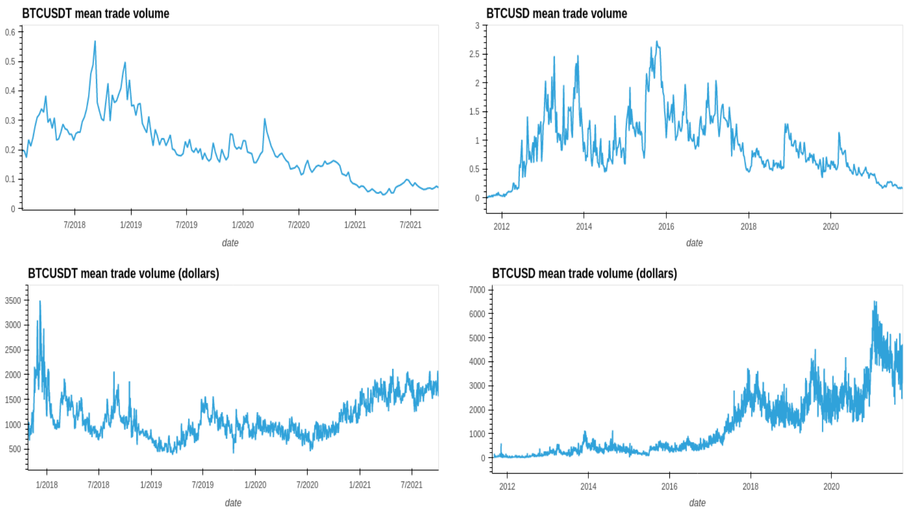

Figure 2.4: Mean volume per trade.

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.

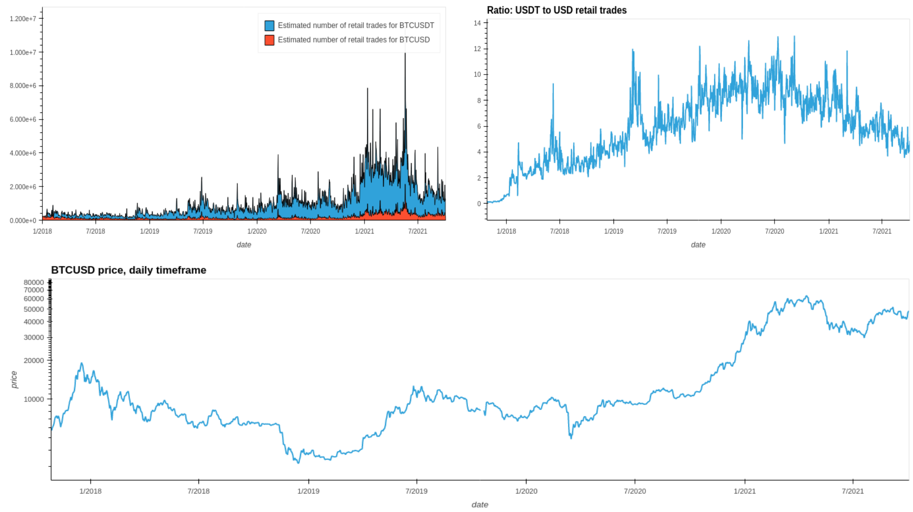

Figure 2.5: Retail trades and BTC price.

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.

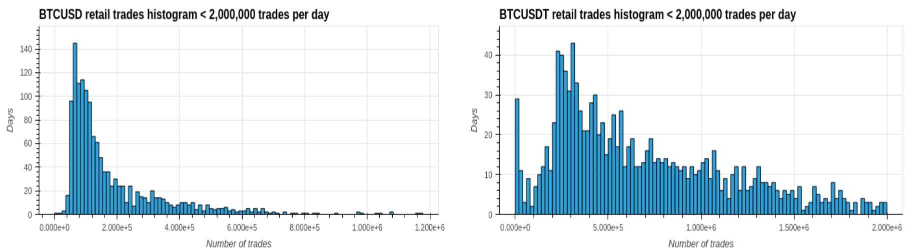

Figure 2.6: Histogram of BTCUSD and BTCUSDT no of retail trades per day

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.

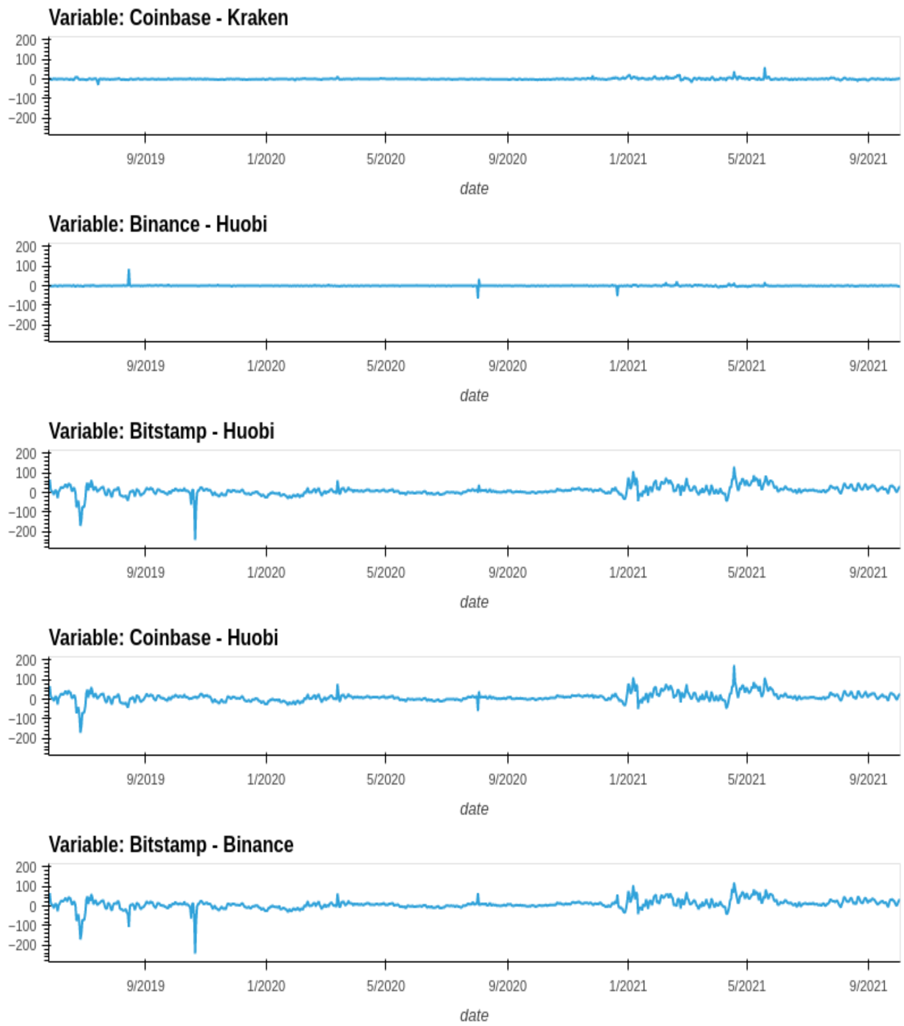

Figure 2.7: Fiat premium.

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.

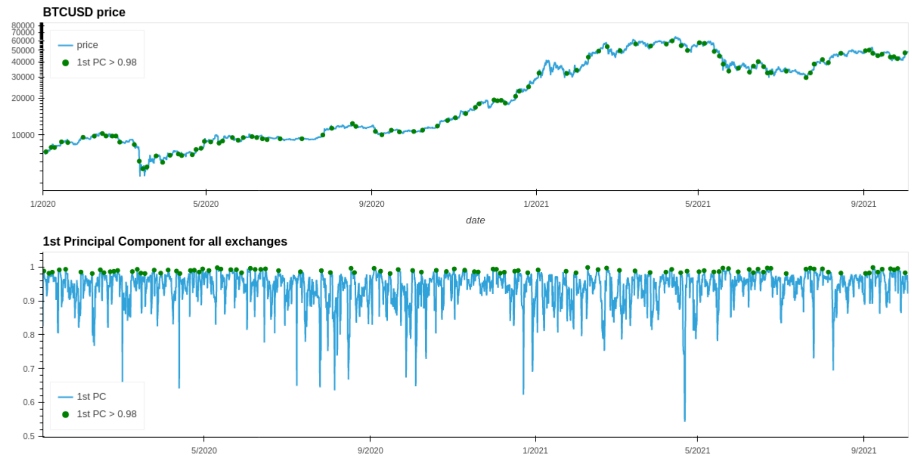

Figure 2.8: PCA analysis - 1st principal component and BTCUSD price.

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 .

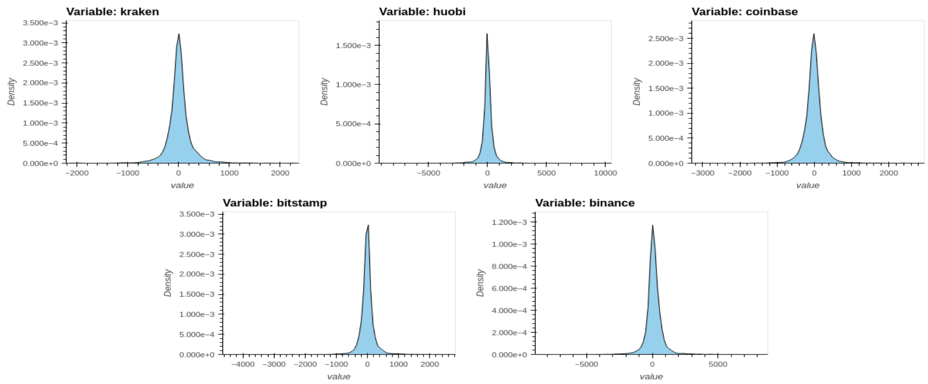

Figure 2.9: KDE plot of volumes aggregated on 4h timeframe, for all exchanges.

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 consistently 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.

Next, we visualize the number of trades that take place in each exchange. In the figure 2.11, we can see the number of trades aggregated in daily timeframe with the upper chard being in logarithmic scale. We observe that the USDT market is processing many more trades than the USD market. By calculating and comparing the mean number of trades per exchange across 2020 and 2021, we find that the mean trades per day of Binance is 5.5 time the mean of Coinbase, 32 times the mean of Kraken and approximately 37 times the mean of Bitstamp. We also observe that the number of trades (aggregated) is presented in waves, with visible spikes around significant price action. Again, these spikes, converge across all exchanges.

Figure 2.11: The number of trades aggregated daily, for each exchange. Note that the upper chart is logarithmic.

In the next figure 2.12, we are separating the buy from the sell trades. First thing to notice is that our data is corrupted, since for approximately 15 days, all trades are classified as buy-side trades. Due to this shortcoming, in any attempt to use the side of the trade, we will have to exclude this portion of the dataset. Furthermore, the Binance, as expected, processes the largest amount of orders, either buy or sell, but in several spikes, Huobi seems to catch or even suprass Binance. Last but not least, it is interesting that the amount of trades in Coinbase's BTCUSD pair, are steadily increasing, catching up those of the BTCUSDT market.

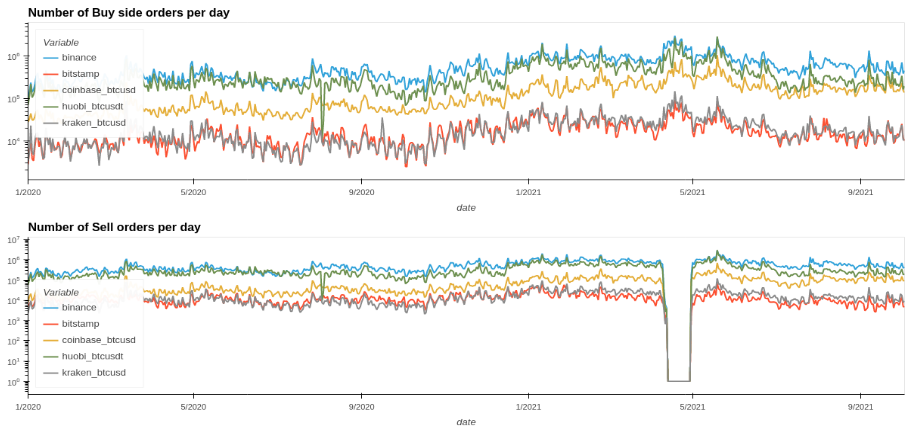

Figure 2.12: The number of buy and sell trades aggregated daily, for each exchange. Both charts are logarithmic.

The next upper graph 2.13, is produced by taking the cumulative sums of buy side and sell side orders across all exchanges, and subtracting one from the other:

$$cumsum\{Buy\_side\_volume - Sell\_side\_volume\}$$

Just before the major bullrun of 2020-2021, the sell volume surpassed by far the buy volume, indicative of the uncertainty of that period (Covid19). Around October 2020, the cumulative sell volume peaked, as shown in the minimum of the graph. From then and on, the buy volume was steadily increasing, which coincides with the price action, at the beginning of the 2020-2021 bull run.

The second graph is produced by taking the cumsum of the difference of buy and sell volume, but for each exchange individually. An interesting finding is that coinbase and bitstamp are processing more buy side volume than sell side, and more buy volume than any other exchange. The exact opposite is true for Binance, where sell volume is the highest. That alligns with our prior findings, in that BTCUSD market is the entrance of the 'first time' bitcoin buyer, who is attracted during the bull run.

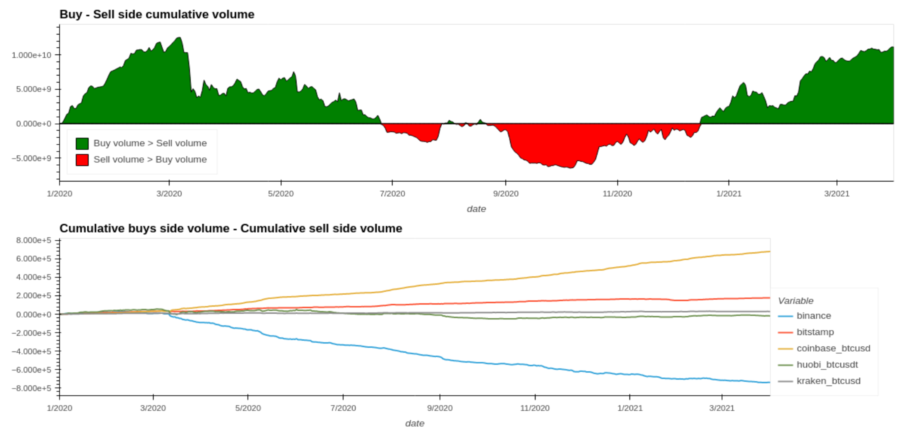

Figure 2.13: The number of buy and sell trades aggregated daily, for each exchange. Both charts are logarithmic.

This disrepancy makes clear that even though each exchange has its own pool of trades, the market dynamics are such, that we need to concern the exchanges not only individually, but as a unified pool of liquidity as well, where one exchange is used mostly from buyers and another on, mostly from sellers. The latter is possible, due to the easy transfering of tokens between exchanges. Especially USDT, XRP, XLM and DOGE amongst others, seem to have enough liquidity, and very low transaction finality time, to be prime candidates for transfering value fast and cheap across exchanges (in contrast to fiat).

Chapter 3

Sampling

3.1 Introduction

In this chapter, we will sample on BTCUSD and BTCUSDT dataset across all features and more specifically volume, side of the trade, speed of the market, and convergence of exchanges in various features. For simplicity reasons, when sampling takes place on all exchanges, the dataset will be aggregated to the 'second' or 'minute' interval (the 'minute' interval sampling will be used mostly for plotting purposes). In other occasions, the sampling will take place directly on raw tick data, nevertheless, in order to create the dataset for all exchanges, some sort of aggregation must be used, most probably in respect to time.

3.1.1 Volume

In this section, we will use volume along with different features of choice such as, positive-negative returns and buy-sell side. We will begin with buy-sell volume, while we illustrate how the sampling could end up in signal creation.

Buy and Sell Volume

The dataset is created by classifying the volume into buy and sell volume, depending on the side of the trade. This classification is taking place directly on tick data. Then, the dataset is aggregated (summation) on the 'second' time interval, where two columns are created: one for the buy volume and one for the sell volume. This procedure is used for all exchanges and as we have shown, all exchanges should be used (see 2.13).

The latter results to a 10 column dataset where 5 columns are created for the buy volume and 5 columns for the sell volume (one column for each exchange). Lastly, the 5 columns of each side, are summed row-wise in order to create one column for buy and one column for sell volume. The last columns represent the volume that took place on all exchanges after being classified as buy and sell.

In order to model the two volumes, we will use a Hawkes process. To provide the reader with a brief explanation of the Hawkes process, we shall begin with a Poisson process, which models the number of occurences at certain time intervals. The key takeaways from Poisson processes are that the expected rate of these occurrences λ is stable (homogeneity) and that the expected rate of occurrences at a future time interval is independent of past occurrences.

In order to study more complex phenomena, a Non-homogeneous Poisson process could be used, where the future events are still independent of past phenomena, but λ is now a function of time. This particular idea fits somewhat with market behavior, in that the expected rate of returns (or as we will later show), the expected rate of volume traded, does seemingly not behave independently of the time interval, but fluctuates locally. There are certain periods of time where the volume that is expected to be traded is higher (moments of behavior co-ordination). For the function $\lambda(t)$ to be called a deterministic function (and also non-homogeneous Poisson process) there are certain axioms that need to be held. This function then is called an "Intensity function".

Lastly, in addition to the above, if the intensity function is not stable, but is affected by the history of the timeseries up to time t, the process is called Hawkes process. A Hawkes process is a self-exciting process, which its past events affect the current value of the process. There also exist other similar approaches to Hawkes, such as convolutional neural networks which are mostly used for image classification that use a weighted average of previous values. The problem with this approach, contrary to Hawkes is that this approach would enforce static dependencies while Hawkes intensity function uses the $N(t)$ counting process as a positive reinforcement that decays exponentially in such a way that past events that are close to time t, affect the value of the function much more than, say older events that are further away from t.

In a volume specific example the formula would be:

$$\lambda(t|H_t) = \lambda_0(0|H_0) + \sum_{i=1}^{N(t)} \mathrm{Vol}(i) \cdot e^{-\delta(t-T_i)}$$

where $\lambda_0(0|H_0)$ is the initial intensity and δ would be a positive dampening coefficient that implies the rate at which the function decays.

Using a Hawkes process with $\delta = 0.2$ and $\lambda_0 = 0$, we modeled the buy and sell volume, and upon the new dataset, we sampled using two parallel moving averages, one slow (larger scope) and one faster (smaller window). The sampling took place, when the fast MA exceeded the slow MA by a threshold. This way, we can have an overview of the buy and sell volume surges. First thing to notice, is that most of the points sampled from the two volumes are different (see 3.1). There is an oversampling in sudden price action and no sampling at all when price goes sideways. Furthermore, the buy volume, is found mostly in local maxima and the sell volume in local minima, which is to be expected.

Figure 3.1: An example of buy (green circle) and sell (red circle) volume. The size of the markers, correspond to the amount volume traded.

At this point, we decided to disregard the mandatory positive counting process and instead use a process that could take negative values. That process would be the difference between positive and negative volumes and it would no longer be a Hawkes process. The first results of our experimentation showed that when the buying side was far greater than the negative selling side, the process would be excited by a positive value and vice versa. Thus, this process could also take negative values, but even if a large negative selling volume appeared, it would take a couple of steps for the function to actually get affected enough and fall below zero (lagging). We consider it natural for sellers to affect buyers and vice versa, and we decided to follow this idea because simple volume self-excitation was not enough.

After using the following custom intensity function:

$$\lambda(t|H_t) = \lambda_0(0|H_0) + \sum_{i=1}^{N(t)} (\text{Vol}_{\text{Buy}}(i) - \text{Vol}_{\text{Sell}}(i)) \cdot e^{-\delta(t-T_i)}$$

After modelling the above difference, we used two moving averages, one slow (larger scope) and one faster (smaller window), as above. We then decided to sample based on two different factors, thus creating 4 states. The rationale behind the 4 states, is that we could sample not only when the volume difference spikes, but also when the spike ends, signaling the end of the price action. The two factors are:

- Faster MA over slow MA (setting a threshold as above)
- Sign of the difference

And the four states:

1. Fast moving average of our process would exceed the slow-moving average, indicating there is incoming positive volume (immediate past) at greater rate than the slow-moving average (larger past time window) and the Buyers volume exceeds the Sellers volume.

2. Fast moving average of our process would exceed the slow-moving average, indicating there is incoming positive volume (immediate past) at greater rate than the slow-moving average (larger past time window) and the sellers volume exceed the Buyers volume. We believe this to be a significant indicator for sampling (specifically a shorting the market indicator) because it shows that although there has been a large Buyers rate recently, Sellers appear to significantly take over the reins (at the very last moment- current moment) just when the market participants think the buying pressure will continue.

3. Fast moving average of our process would fall below the slow-moving average, indicating there is incoming negative volume (immediate past) at greater rate than the slow-moving average (larger past time window) and the Sellers volume exceeds the Buyers volume.

4. Fast moving average of our process would fall below the slow-moving average, indicating there is incoming negative volume (immediate past) at greater rate than the slow-moving average (larger past time window) but the Buyers volume would exceed the Sellers volume. We also believe this to be a significant indicator for sampling (specifically a longing the market indicator) because it

shows that although there has been a large Sellers rate recently, Buyers appear to significantly take over the reins (at the very last moment- current moment) just when the market participants think the selling pressure will continue.

Appendix

The appendix with all the code

Bibliography

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