Sampling on High Frequency BTCUSD data

Koutkos Christos & Kosmetsas Tilemahos
under the guidance of
Professor Greg Cireci

Capstone project for
MSc in Financial Engineering

WorldQuant University
Greece
January 2022

Acknowledgements

First and foremost, we would like to thank our Professor Greg Cireci, who guided us in doing these projects. He provided us with invaluable advice and helped us in difficult periods. His motivation and help contributed tremendously to the successful completion of the project.

Also, we would like to thank our families for their support. Without that support we couldn't have succeeded in completing this project.

Contents

List of Figures

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 [8]. Although the cryptocurrency market has many similarities with the traditional ones, 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 appears 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.

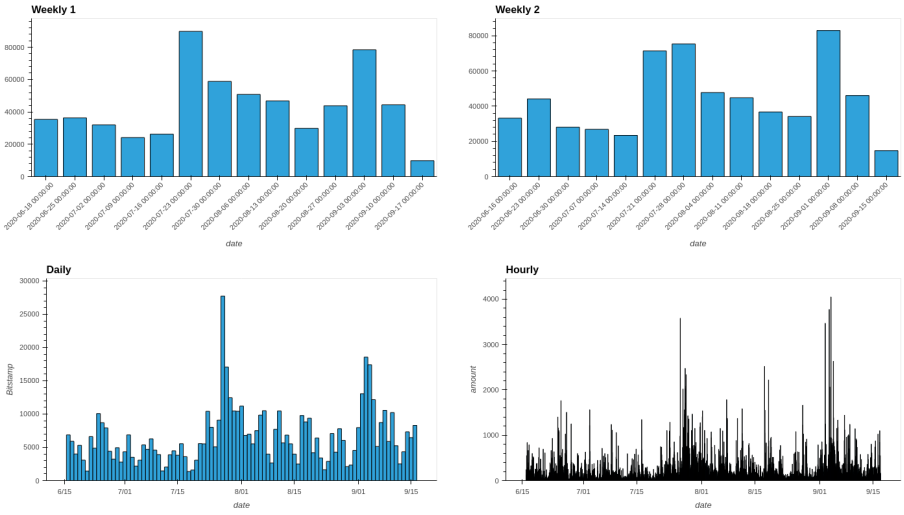

Figure 1.1: BTCUSD Volume sampled in several timeframes

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 ocurring at 2020-07-25 and persisting for the coming week. 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.

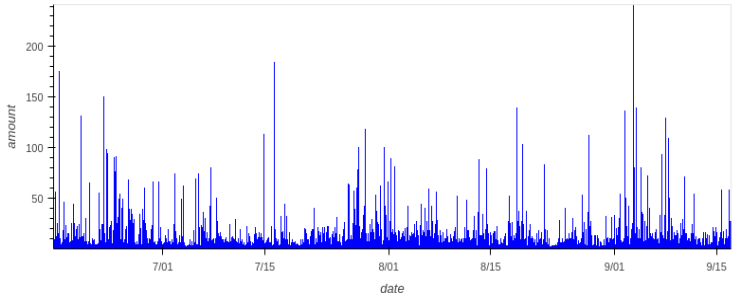

Figure 1.2: Volume per trade (tick volume).

Lastly, graph 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.

1.2 Literature Review

The goal of this literature review is to identify the "avant garde" of researchers in this emerging field, to summarize the up-to-date research results regarding data sampling and pattern recognition and to pinpoint the most cutting-edge results. The authors will then try to place themselves in this large picture and hopefully contribute to technical analysis and data sampling family.

The main driver of this project, is Advances in Financial Machine Learning (De Prado 2018). In his book De Prado gives basic insight on how to sample and prepare data. Specifically, he uses the word "information" in a microstructural sense and proposes the creation of bars, albeit at an entry level, such as tick imbalance bars, volume/fiat bars and tick run bars. These bars could potentially produce signals, that are "triggered" when a certain threshold is exceeded, e.g., a certain amount of volume is being traded, at a certain time, that is beyond the expected level.

In the third edition of his book Analysis of financial time series (Tsay 2010) chapter 12, Gibb's sampling, which is a Markov Chain Monte Carlo method is used. This method enables statistical inference, and has the advantage of "decomposing a high-dimensional estimation" to a problem with lower parameter problem. The insight that can be taken from this, is to approach Bitcoin signal extraction by removing certain correlated features. An in-detail Python application of the MCMC method is illustrated in the Python for Finance (Yves Hilpisch 2015).

An interesting statistical approach on defining and identifying a "Bull or Bear" market can be found on the paper: Defining and Dating Bull and Bear Markets: Two Centuries of evidence. (Gonzalez, Hoang, Powell, Shi 2006). This paper is not cryptocurrency specific but the way it defines these terms is relevant to Bitcoin. The basic tool for identifying the markets being used is the persistence of the time-series above or below one or more moving averages (Turning point methods BB and CC as they are called by the authors).

Another paper published on 2019 called Exogenous Drivers of Bitcoin and Cryptocurrency Volatility – A mixed Sampling Approach to Forecasting (Walther, Klein , Bouri) expands on the mixed sampling method (Garch-Midas) regarding volatility of certain cryptocurrencies during high price movements. This paper in short concludes that exogenous factors are better suited for predicting high volatilities during Bear markets than say the Garch model.

Technical Analysis for Algorithmic Pattern Recognition (Zapranis Tsinaslanidis 2016) is a book expanding on technical analysis and will be exceptionally useful because it provides insight on patters: holding, support and resistance levels recognition. It expands on indicators and tools such as RSI , Bollinger Bands and MA convergence-divergence.

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. Our initial sampling approach will be across time (fixed time window). Key insights that will be extracted, will serve as the infrastructure of a dynamic way of sampling.

2.2 Volume

It is commonly accepted that volume is one of the most important tools, for analyzing timeseries in finance. 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 shut down in April 2014 due to fraud, as more than 850,000 BTC were missing [9]. As time passed by, and BTC gained more traction, the trading volume upscaled significantly and more exchanges, such as Bitstamp, Kraken and Coinbase, appeared. Following 2017, a demographic shift took place: Institutions and retailers, started engaging with the crypto sphere and Bitcoin specifically in growing numbers [1]. This was the period that Bitcoin become "known".

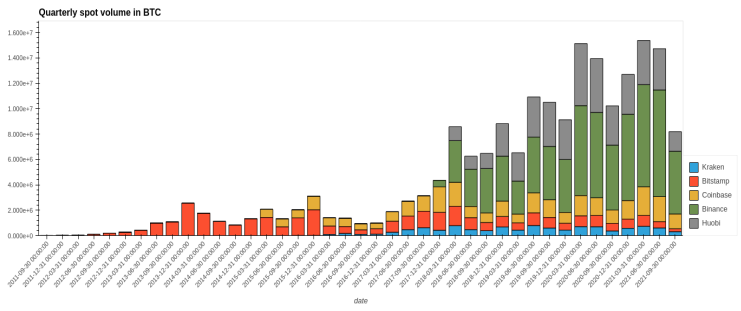

Figure 2.1: Quarterly volume across spot exchanges.

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.

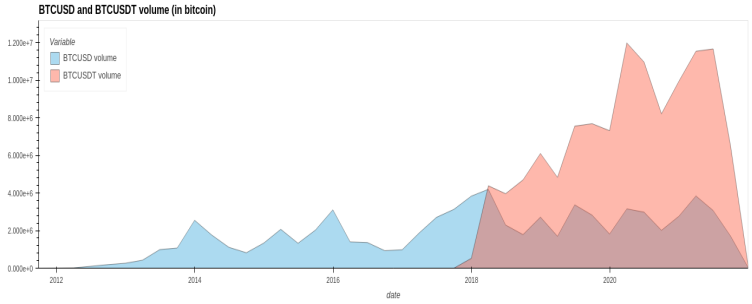

Figure 2.2: BTCUSD and BTCUSDT spot trading volume (in bitcoin).

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.

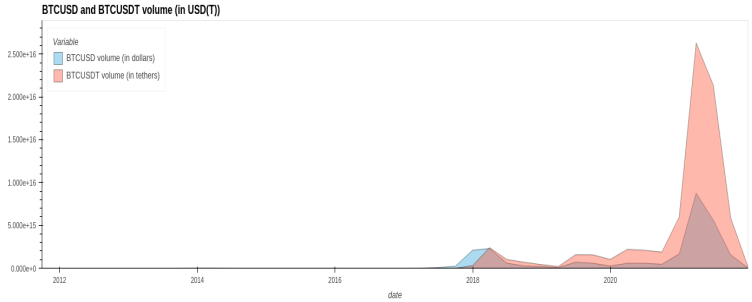

Figure 2.3: BTCUSD and BTCUSDT spot trading volume (in USD(T)).

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.

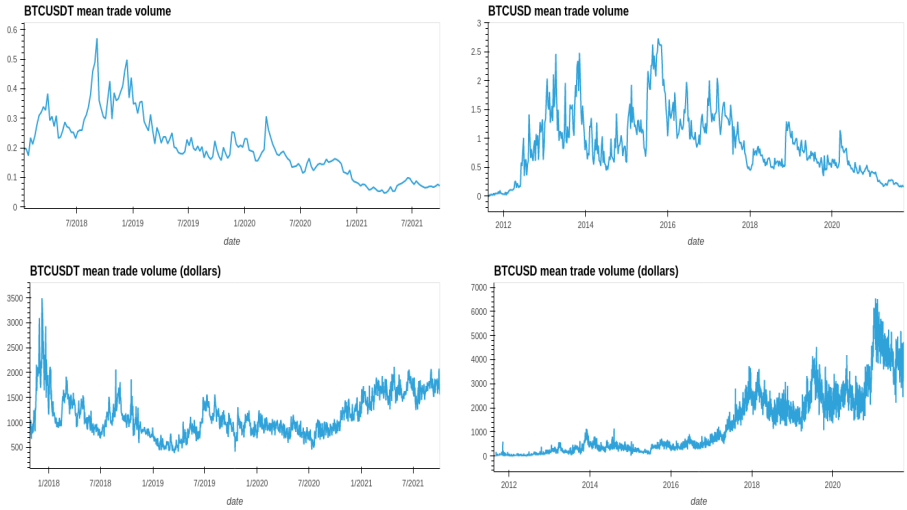

Figure 2.4: Mean volume per trade.

An approach that may disclose the "nature" of the investor based on the trade data at hand, would be to use the following framework of assumptions:

- Retail traders are trading in integer dollar volumes, and
- Institutional investors will more likely buy in OTC (Over The Counter) markets and sell in spot exchanges using sofisticated execution strategies.

In order to extract the possible retail trades, an error of $e = 0.05\$$ was incorporated in creating two boundary conditions:

$$P - \texttt{int}(P) < e \text{ and } , P - \texttt{int}(P) > -e$$

where P is the price in dollars at which a trade was executed. The role of the error e, is to account for floating rounding errors [2].

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.

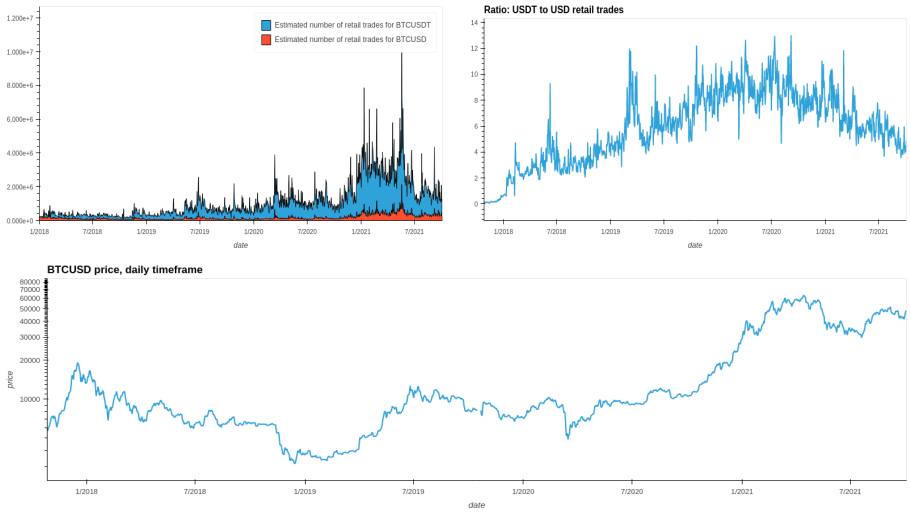

Figure 2.5: Retail trades and BTC price.

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.

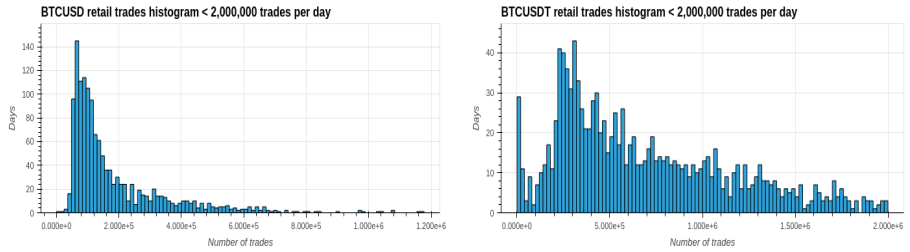

Figure 2.6: Histogram of BTCUSD and BTCUSDT no of retail trades per day

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.

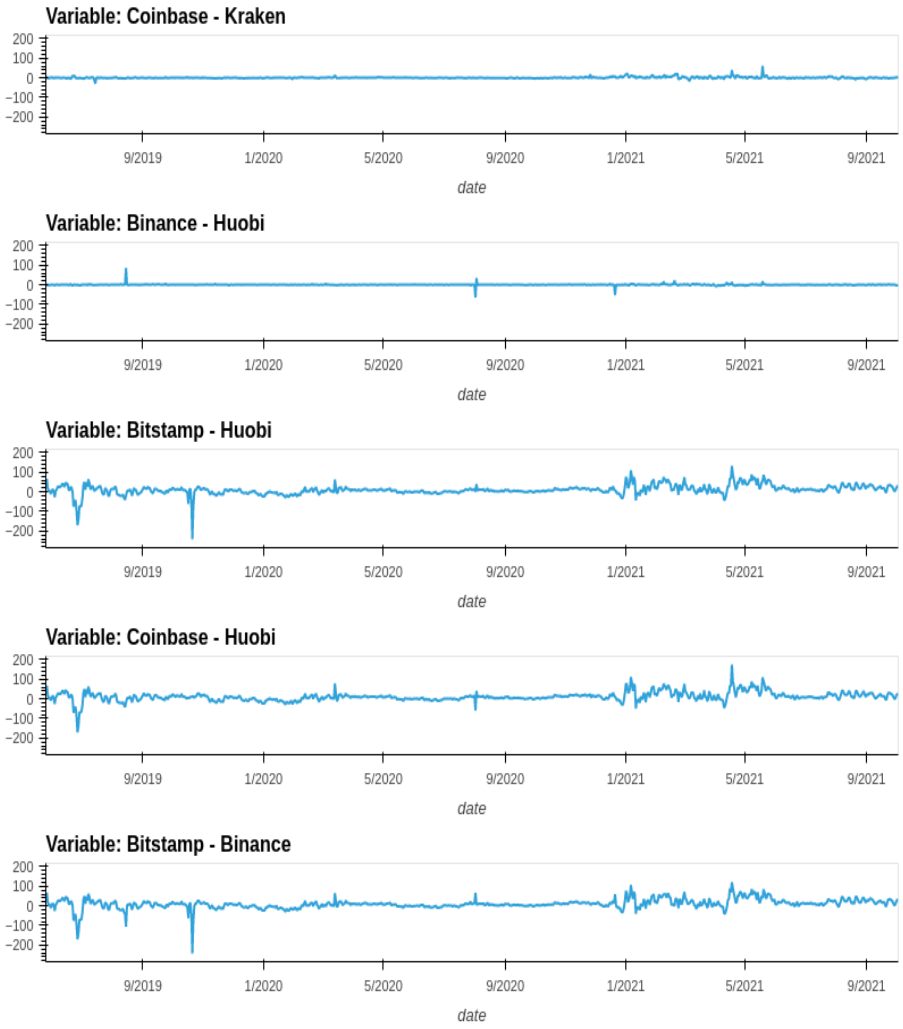

Figure 2.7: Fiat premium.

Such discrepancies could be a valuable source of imbalances, that should be useful for a more precise sampling. Next, we will explore volume, a bit deeper. We will decompose the covariance matrix of volume (eigen decomposition), of the BTCUSD and BTCUSDT pairs. 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 and during different phases of the market.

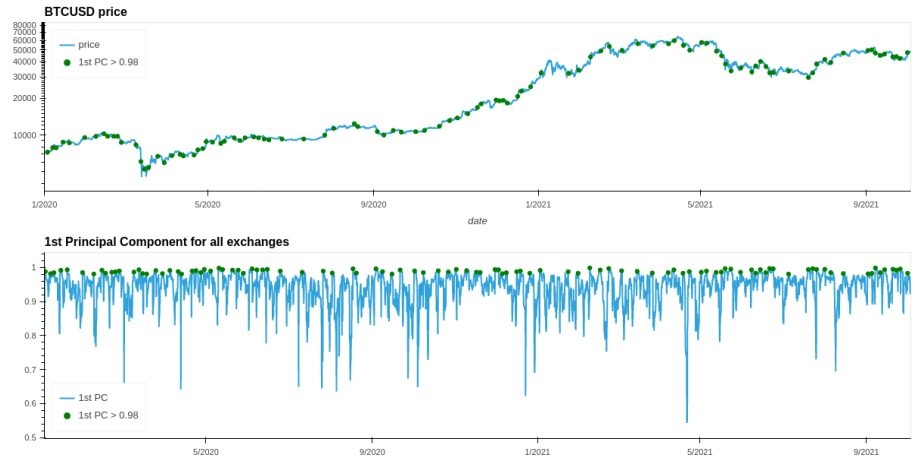

Figure 2.8: PCA analysis - 1st principal component and BTCUSD price.

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 [7].

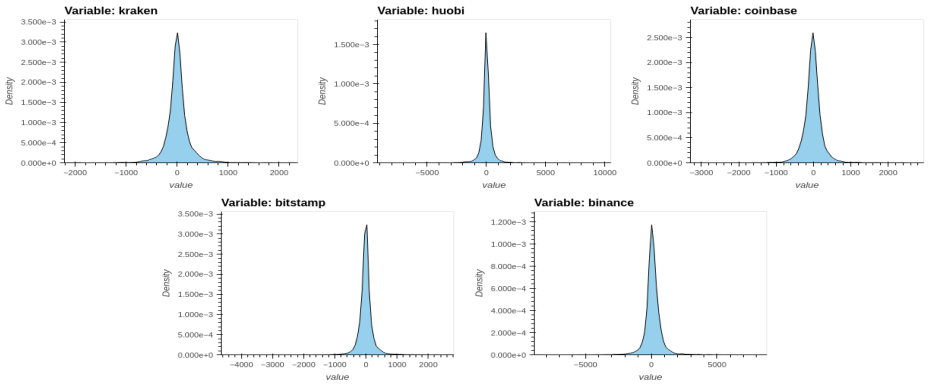

Figure 2.9: KDE plot of volumes aggregated on 4h timeframe, for all exchanges.

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 following graphs, we need to insert, the notion of positive and negative volumes. The notion of a positive volume (and negative accordingly) is the need to differentiate between the trading volume that leads to positive returns and volume in the market that lead to negative returns. 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 $t - 1$: $b_t \cdot V_{t-1}$. Additionally, by classifying volumes, that inherit a positive or negative sign based on the returns, two sets are formed.

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.

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 bid, from the offer trades, depending on who initiated each trade(based on dataset labels buy-sell). First thing to notice is that our data is corrupted, since for approximately 15 days, all trades are classified as trades, initiated by buyers. Due to this shortcoming, in any attempt to use the side of the trade (bid-offer), we will have to exclude this portion of the dataset. Furthermore, the Binance, as expected, processes the largest amount of orders, either buying or selling, 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.

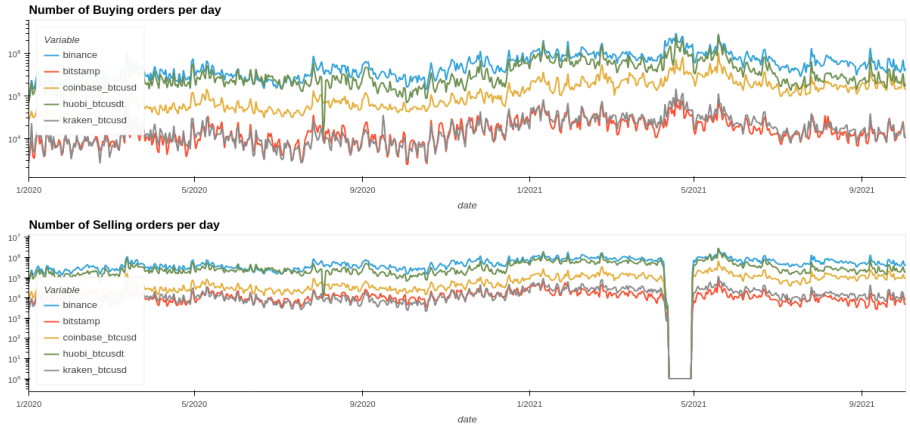

Figure 2.12: The number of buy and sell trades aggregated daily, for each exchange. Both charts are logarithmic.

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:

$$\text{cumsum}\{\text{Vol}_{\text{Buying}} - \text{Vol}_{\text{Selling}}\}$$

Just before the major bull run of 2020-2021, the selling volume, by far surpassed the buying 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.

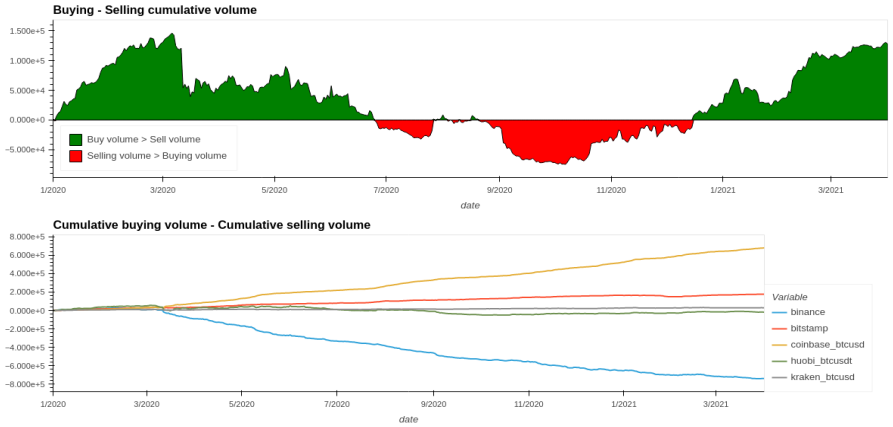

Figure 2.13: The number of buy and sell trades aggregated daily, for each exchange. Both charts are logarithmic.

This discrepancy makes it clear, that even though each exchange has its own pool of trades, the market dynamics are such, that one needs to consider each exchange not only individually, but as a single identity as well with a unified liquidity pool. It appears from data, that it is possible that a single exchange can at times attract buyers and on the other hand, another one can attract sellers. Such a behavior, can be explained by the simplicity of transferring tokens between exchanges (e.g. for arbitrage reasons). Especially USDT, XRP, XLM and DOGE amongst others, seem to have viable liquidity, and very low transaction finality times and fees. Coins such these, are prime candidates for transferring value fast and cheap.

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, mostly with respect to time.

3.2 Volume

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

Buying and Selling Volume

The dataset on which the next sampling is performed, is created by classifying the volume into buying and selling volume, depending on who initiated the trade (bid or ask). 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 buying volume and one for the selling 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 buying volume and 5 columns for the selling 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. The last columns represent the volume that took place on all exchanges after being classified as buying and selling.

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 [4].

In a volume specific example the formula would be:

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

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 buying and selling 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 buying volume, is found mostly in local maxima and the selling volume in local minima, which is to be expected.

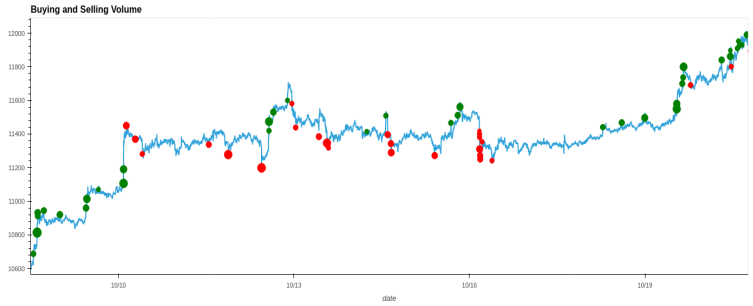

Figure 3.1: An example of buying (green circle) and selling (red circle) volume. The size of the markers, correspond to the amount volume traded.

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.

We used 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 and vice versa (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. Additionaly, it could also indicate the end of an upward price move.

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. Additionaly, it could also indicate the end of an downward price move.

In the figure 3.2, we can see an example of the states. Upon carefull inspection of the local minima, we observe increased sell volume (upper graph, red triangle), with the fast MA above the slow one (lower graph, green circle). In many of these occasions, a reversion occurs right away. The latter is a state 2 example. Using the same notion, we distinguish a state 4 example, by observing that at local maxima, the increased buy volume (upper graph, green triangle) is accompanied with the fast MA being under the slow one (lower graph, red cross). In many occasions, a reversion in the price occurs.

Figure 3.2: An illustration of the four states. The first graph corresponds to the first factor and the second graph to the second factor. In the first graph, the green arrows indicate buy volume and the red ones, sell volume. In the second graph, the green circle corresponds to Fast MA $>$ Slow MA, while the red cross, corresponds to the exact opposite.

The reader should keep in mind, that the sampling took place in the 'second' timeframe, while the price in the figure above, is plotted from the 'minute' timeframe, using the mean price during that minute. Therefore, the real volatility cannot be grasped by this graph, and some points seem to wander off the price plot. In order

to bypass that shortcoming, and furthermore, to illustrate visually how modelling - sampling - parameters look like, a zoom-in figure, with price displayed in seconds will be presented.

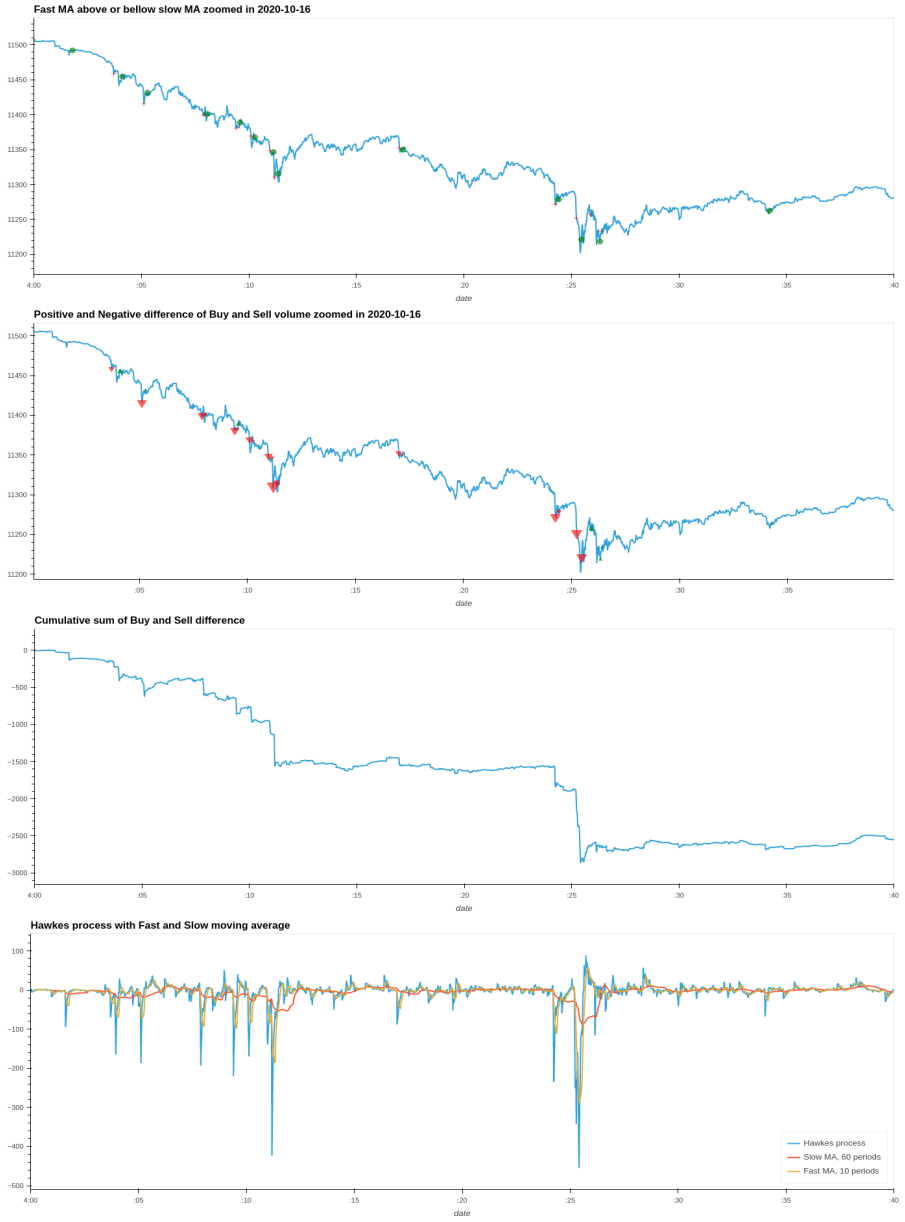

Figure 3.3: Several plots of a 'Hawkes process' on the difference of buy and sell volume along with an example of sampling using moving averages on the above process (in 'seconds' timeframe).

In 2020-10-16 from 04:00 to 04:40, there was a rapid correction in BTCUSD(T) pair. The figure 3.3, presents the 'Hawkes process' on the buy and sell volume difference, and then a sampling effort on this process. At first, the third plot (Cumulative sum of Buy and Sell difference), shows that the sell volume is greater than the buy volume. There are also some points where the sell pressure is higher and periods where buy and sell volume are approximately the same. We observe that in the points where imbalance between buy and sell volume occurs, the 'Hawkes process' spikes and then returns at the mean level (which was set at zero).

The Fast MA (10 periods - orange colour) and Slow MA (60 periods - red colour), cross each other and produce the green circle and red cross of the second plot (Fast MA above or bellow slow MA). The rationale behind this sampling, is that when a falling fast MA crosses a falling slow MA, the price falls (and vice verca), and thus a red arrow of the first plot is accompanied with a red cross of the second plot. Since the 'first point sampled' represents the beginning of one move, the next point that must be sampled, should signify the end of it, where a rising fast MA crosses a falling slow MA. This way, we can sample the beginning (or close to beginning) and the end of the move. Nevertheless, if the price begins to fall gradualy, with no major spikes on the 'Hawkes process', and the fast and slow MA slowly converge, there would be no sampling. Furthermore, this way of sampling is very sensitive to the parameters chosen (fast and slow MA) and as we can see in the second and fourth plot of figure 3.3, there are times that the fast is not fast enough to catch the very fast price moves, but does a better job in sampling correctly the larger moves along with their ending.

Positive and Negative Volume

From a market microstructure perspective, the search of imbalances of signed volumes (positive volume and negative volume), could reveal the presence of information [5]. Based on the work of de Prado, a more precise way of sampling, is to sample when imbalances become apparent in the market. These imbalances could be seen as new information. At this point we need to be precise as to what we will mean from now on, when the word 'imbalance' is used: deviation of a feature, from our expectation that is calculated from the immediate past of the feature (short or long past depending on the parameters chosen). Soon after the imbalance occurs, a new expectation is created that gets updated, as new data arrive.

The 'Hawkes process' of the previous subsection, is used in order to sample the imbalances on the difference between buy and sell volume. A very similar way of sampling will be attempted here as well, by changing the 'buy' and 'sell' with 'positive' and 'negative'. We begin by calculating the sign b_t of returns r_t:

$$b_t = \begin{cases} b_{t-1} \text{ for } r_t = 0 \\ \text{sign}(r_t) \text{ for } r_t \neq 0 \end{cases}$$

The above calculation results to an array with $\{1, -1\}$, depending on the sign of the returns. Lets denote with v_t an array that corresponds to the tick by tick volume, and also keep in mind that those arrays get updated as new data arrive. By multiplying the two arrays row by row (Hadamard product), the array created, has the same values as v_t but signed as positive or negative based on the returns. This array will be used in order to sample any imbalances found to the positive or to the negative side. One way to do the sampling is the proposed by de Prado way, where

$$\theta_T = \max(\sum_{t=t_0}^{T} b_t v_t)$$

represents the imbalance on a non fixed interval of length, T number of ticks. Then the expectation of the next imbalance $\mathbb{E}_0(\theta_T)$ is computed using the expectation of

the next bars length $\mathbb{E}_0(T)$ using the next formula:

$$\mathbb{E}_0(\theta_T) = \mathbb{E}_0(T) \max\{P[b_t = 1]\mathbb{E}_0[v_t|b_t = 1], (1 - P[b_t = 1])\mathbb{E}_0[v_t|b_t = -1]\}$$

The above way of sampling is called *Volume Runs Bars* [5]. Basically, we will sample the imbalances of summed signed volumes. A first thought is to sample when these imbalances cross a fixed threshold. This way, we might be able to sample correctly during 'normal' volume, but at times of high (low) volume we will undersample (oversample).

In order to overcome the fixed threshold problem, we will use a fast exponential moving average (EWMA) and a slow EWMA, which actually serves as a dynamically changing threshold. The sampling will take place, when the fast EWMA crosses the slow EWMA. The EWMA, is a type of moving average, which keeps all past information with decaying weights:

$$\text{EWMA}_t = \alpha \cdot x_t + (1 - \alpha) \cdot \text{EWMA}_{t-1}$$

where x_t is the value of the variable at time t and α controls the importance of the recent value vs the old ones [6]. A slow EWMA has a small *alpha* in order to resist temporary fluctuations, but at the same time, account for them when EWMA of time $T > t$ is calculated. The fast EWMA on the other hand, has a bigger α, chosen in order to prioritize the very recent history. An example of this sampling, can be seen on the figure bellow.

Figure 3.4: An example of volume runs bars, sampled from ticks, for Coinbase and Binance.

In figure 3.4, the green circles represent positive volume, and the red ones, negative volume. Visually the sampling seems to be as expected, since the positive volume is mainly found when the price is rising and the negative, when the price is falling. Furthermore, based on the parameters of the EWMA used in each exchange (Binance and Coinbase), we can see slightly different sampling.

Elaborating on the above effort, we can sample the difference of positive and negative volumes from buy-only and sell-only trades. The rationale behind this decision, is to sample when negative volume, which corresponds to negative returns and subsequently to a falling price (even if very short term), is associated with buy-side trades, and vice versa. Additionaly, for experimenting purposes, we will sample the difference of buy and sell volume from positive and negative only trades, in order to capture the sell (buy) volume on positive (negative) price moves.

For simplicity purposes, we will sample from 'second' timeframe and not directly from tick data. At first we will construct four new columns for each exchange:

1. Positive volume with buying orders
2. Positive volume with selling orders
3. Negative volume with buying orders
4. Negative volume with selling orders

After resampling all exchanges to the 'second' timeframe, the newly created columns, are summed row wise for all exchanges. Using these columns, we construct 4 new ones, by simply subtracting one from the other:

1. Positive volume: Buying volume - Selling volume
2. Negative volume: Buying volume - Selling volume
3. Buying volume: Positive volume - Negative volume
4. Selling volume: Positive volume - Negative volume

The green (red) colour will be used to indicate positive (negative) volume while a triangle (inverted triangle) for buy (sell) volume. In figure 3.5, upper graph, we can see the $\text{Vol}_{Buy} - \text{Vol}_{Sell}$, when $\text{Vol}_{Sell} > \text{Vol}_{Buy}$ under positive volume (returns).

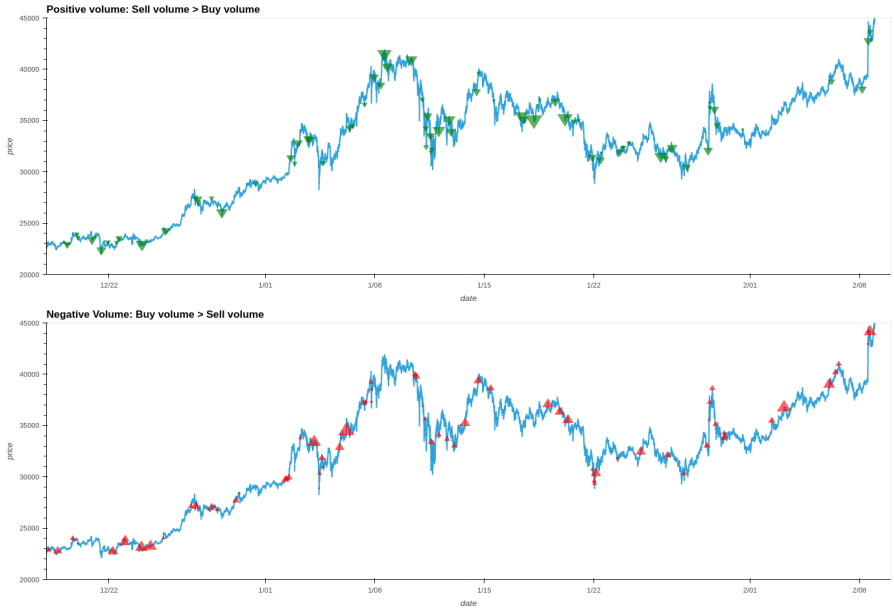

Figure 3.5: Buying - Selling volume imbalance on positive and negative only volume.

What we expect to extract from the above sampling, is the occasions where selling volume spikes during positive returns, signaling thus a short term price reversal. Similarly, in the second graph 3.5, we sample buying volume spikes, during negative returns (or potentialy a falling price). Again, we expect to find a short term reversal, since buying activity should stop or slow down negative returns or a falling price.

Similarly, given the buy volume, we sample the difference $\text{Vol}_{Positive} - \text{Vol}_{Negative}$, when $\text{Vol}_{Negative} > \text{Vol}_{Positive}$, see figure 3.6, upper graph. On the second graph, we can see the excess of positive volume, under sell trades.

This market behavior seems unintuitive. The volume that is computed in this sampling, represents the majority of volume traded in BTCUSD(T) markets. If we assume, that there is no significant volume in another exchange, the rising price with sell trades and falling price with buy trades, should be significant.

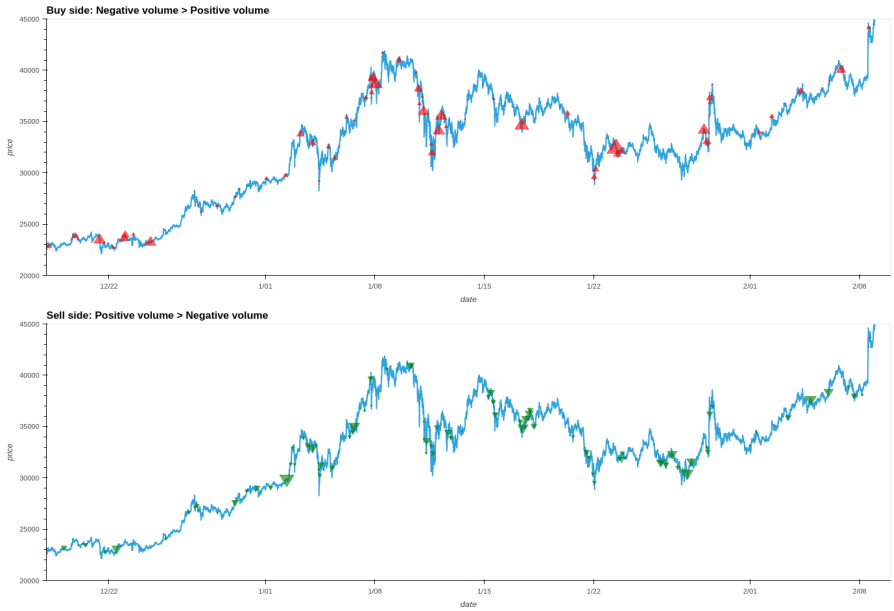

Figure 3.6: Positive - Negative volume imbalance on buy and sell only trades.

3.3 Speed of the market

We define the 'speed of the market' to be the rate of incoming trades. Since the trades arrive in an asynchronous way, the `TimeDelta` between two consecutive trades, vary over time. There will be occasions when the `TimeDelta` contracts and others when it expands. We are interested in those times where it remains contracted for a 'sufficient' number of trades, and as above, the threshold for what is 'sufficient', will dynamically change.

The `DateTime` column of our dataset, is sampled from milliseconds (1 second = 1,000 milliseconds). After been transformed to `Unix` time, we are left with an array of 17 digits integers. The high volume - high transaction rates exchanges, (Binance, Coinbase, Huobi) exhibit many transactions that occur at the same millisecond. That could be an indication of high frequency trading, or iceberg orders.

We transform the `DateTime` array in the following way: at first we compute the first difference in order to calculate the `TimeDelta`. Then we replace the zeroes (trades that occured the same millisecond) with a number z such that $0 < z < 1$.

The resulting array `dt`, contains integers as `TimeDelta` between trades that occured in different time, and a float z, for trades that happened the exact same time. Lastly, we compute `speed` as:

$$\texttt{speed} = \frac{1}{\texttt{dt}}$$

The resulting array `speed` contains floats < 1 everywhere except for the trades that happened the same time, where the values are > 1. At times, when there are few trades or no trades per (milli) second, the `speed` takes values, $< \frac{1}{100000}$. When trades though occur at the same time, the `speed` spikes become pronounced. In order to sample, we will create the cumulative sums of the `TimeDelta`. If transaction rates are steady, there will be no sample. Conversely, if they change rapidly, an imbalance will transpire in the form of a spike, and thus, that point will be sampled.

By slicing the Coinbase dataset, from `2020-12-18 00:00:01.086` to `2021-02-10 23:59:59.700`, we compute the `TimeDelta` between the two dates in milliseconds: $4,751,998,614$. If the arrival of trades were uniformly distributed, we would expect one trade every approximately 295 milliseconds.

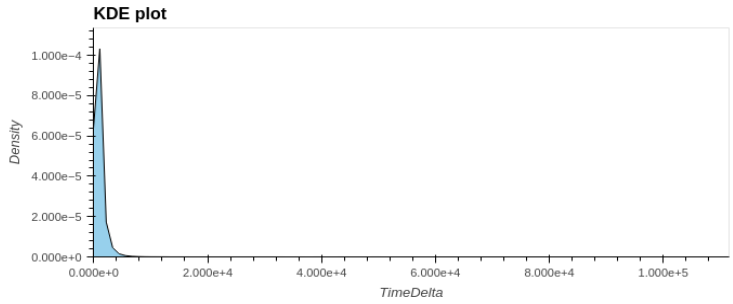

Figure 3.7: KDE plot of timedeltas between consecutive trades, Coinbase.

As we can see in figure 3.7, the `TimeDeltas` are not normally distributed while the proposed distributions to model trade arrival times, are the Exponential and the Weibull [3]. In the above dataset, the 28.57% of the trades, are executed at the same millisecond while the 38.34%, are less than 295.769 milliseconds (0.295 seconds) away. By clustering the trades that were recorded on the same millisecond, we find that the mean amount of volume of these clusters were 0.32 btc, while there were significant clusters with more than 50 btc each.

A characteristic of these clusters, is that the entire volume that was traded whithin the millisecond, was either buying or selling. Furthermore, note, that the 'same millisecond' trades, which account for the 28.57% of all trades, control approximately 44.23% of total volume traded. We conclude that these trades should be sampled carefully, with buying and/or selling volume in mind, positive and/or negative volume, and the effect of these trades in the returns as well.

In the figure bellow 3.8, we showcase an example of sampling with respect to the speed, for the market of Coinbase. The algorithm used to sample the imbalances with regard to speed, is the same as in positive-negative volume sampling. That means, that each imbalance is recorded, and the bar created, consists of all information aggregated between the two samples, while no further information is given on the 'millisecond' trades.

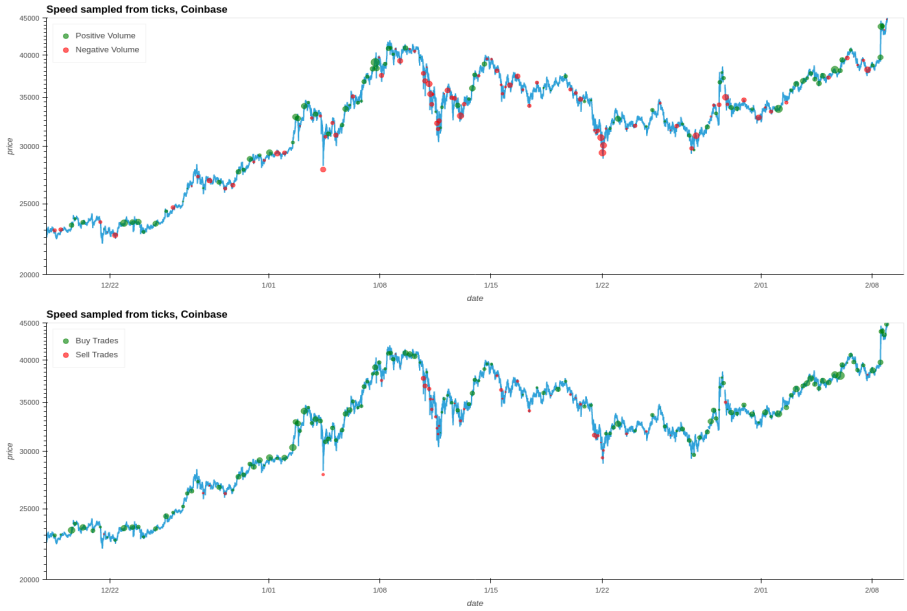

Figure 3.8: Speed sampled from a subset of Coinbase. The area of the circles displayed above, were calculated by aggregating the volume with respect to the sign (positive-negative) or to the side that initiated the trade (buying-selling), between two consecutive samples.

By further enhancing the above sampling, we will extract all milliseconds that 'contain' more than one trade, and sample only on these trades, assuming that they were executed from informed traders. In order to create the figure bellow 3.9, the 'millisecond trades' were extracted and aggregated into one sample each (all trades of the same millisecond, aggregated to one trade), keeping the difference of buy-sell volume, positive-negative volume, number of trades and cumulative returns (`cumprod`). Out of the new dataset, the 2500 samples (out of approximately $2,000,000$) with the most trades, were plotted. The graph bellow, is a zoom-in on the actual plot.

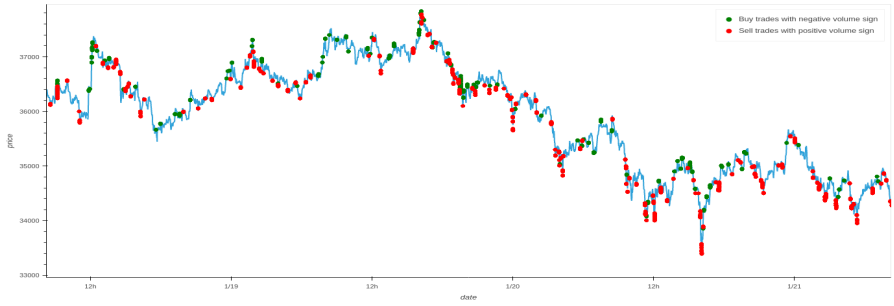

Figure 3.9: Each circle represents one sample that actually contains several trades. The samples shown here, are selected to contain the largest number of trades.

3.4 PCA Sampling

The term *PCA Sampling*, refers to sampling, when all exchanges converge in a specific behavior. That behavior could be a rising volume, increased rate of transactions, increased buying or selling volume etc. In this sampling, it is assumed that all exchanges are of 'equal' importance, even though different sampling for USDT and

USD markets, is presented as well. The rationale behind this idea, is that similar activity on all exchanges, should be indicative of identical information across the market.

The figure bellow 3.10, illustrates a simple PCA sampling effort. The sampling took place on the 'minute' timeframe of positive volume (green circle) and negative volume (red circle). A PCA analysis (with covariance matrix) is applied every 200 minutes in an overlapping manner, and a sample is taken if the explained variance ratio of the 1st principal component, is greater than 0.9. The same could be applied on buying/selling volume, speed of the market, high speed trades (the 'millisecond' iceberg trades from previous section) and generally, on meaningful subsets or transformations of the dataset.

Figure 3.10: An example of PCA sampling on USD, USDT and combined markets on positive/negative volume.

The first thing to notice, is that under the same parameters, the algorithm sampled more frequently on the USDT market. On the other hand, on the USD market, it seems to undersample, since there are fewer times when the 1st PC was greater than 0.9. The combined market, shows a different picture from the USDT (and the USD) market and even after adjusting the parameters in a way to sample approximately the same number of points in both occasions, the sampling was different enough, to justify both sampling efforts in one's arsenal.

We repeat the same thing as above but this time on buying and selling volume (see figure 3.11). The parameters remain the same (window for PCA = 200 minutes,

threshold to sample is 1st PC > 0.9). By contrasting the binance-huobi graph from above with the one from bellow, we observe that the algorithm sampled positive volume with selling activity and vice verca in many cases. That creates an hypothesis to be tested, as to whether this pattern is to be found mostly in local maxima and minima.

Figure 3.11: An example of PCA sampling on USD, USDT and combined markets on buying/selling volume.

3.5 Price-Premium Sampling

TODO

Bibliography

[1] Spencer Bogart. *Bitcoin is (Still) a Demographic Mega-trend: Data Update*. 2020. URL: https://medium.com/blockchain-capital-blog/bitcoin-is-still-a-demographic-mega-trend-data-update-c50df59a6cb3 (cit. on p. 8).

[2] David Goldberg. *What Every Computer Scientist Should Know About Floating-Point Arithmetic*. 1991. URL: https://docs.oracle.com/cd/E19957-01/806-3568/ncg_goldberg.html (cit. on p. 10).

[3] Robert F. Engle Jeffrey R. Russell. *Handbook of Financial Econometrics*. USA: Elsevier, 2010 (cit. on p. 26).

[4] Katarzyna Obral. "Simulation, Estimation and Applications of Hawkes Processes." Master's thesis. University of Minnesota, 2016 (cit. on p. 18).

[5] Marcos Lopez de Prado. *Advances in Financial Machine Learning*. USA, New Jersey: Wiley, 2018 (cit. on pp. 22, 23).

[6] Achilleas D. Zapranis Prodromos E. Tsinaslanidis. *Technical Analysis for Algorithmic Pattern Recognition*. Switzerland: Springer, 2016 (cit. on p. 23).

[7] David Sarmento. *Chapter 22: Correlation Types and When to Use Them*. 2020. URL: https://ademos.people.uic.edu/Chapter22.html (cit. on p. 13).

[8] Statista. *Overall cryptocurrency market capitalization per week from July 2010 to January 2022*. 2022. URL: https://www.statista.com/statistics/730876/cryptocurrency-maket-value/ (cit. on p. 5).

[9] Wikipedia. *Mt. Gox*. 2021. URL: https://en.wikipedia.org/wiki/Mt._Gox (cit. on p. 8).