Stylized Facts of the Bitcoin Market and A Sentiment-Base Model for the Bitcoin

Prepared by Huiyi Wang

Department of Statistics and Actuarial Science

University of Waterloo

December 11, 2017

Abstract

In the recent years, cryptocurrencies become well-known to the public and Bitcoin is the most widespread among all the cryptocurrencies. Bitcoin has become more and more popular, not as a method of payments, but also as an investment. In this paper, we will firstly examine some of the stylized facts and statistical properties of the Bitcoin market and then focus on a sentiment-based pricing model for the Bitcoin. At last, we will discuss some further thoughts about the sentiment-based model. The paper is based on "Some stylized facts of the Bitcoin market" by Naiouf et al. (2017), denoted as "Some stylized facts", and "A Sentiment-Based Model for the Bitcoin: Estimation and Option Pricing" by Cretatola et al. (2017), denoted as "A Sentiment-Based Model".

1 Introduction

The price for Bitcoin has increased more than 1000% since the beginning of 2017[1]. And the market cap is nearly 200B by December 5th, 2017 [2]. As the most widely accepted cryptocurrency, Bitcoin is initially introduced as a method of payments, which could be used as a substitute for fiat currencies such as USD and CAD. While offering high volatilities and high returns, many people have doubt for bitcoins being a kind of currency. As the Chicago Derivatives Exchange, known as CME group, announced to introduce derivatives on Bitcoin, more and more people are considering Bitcoin as a special equity with speculative features [3]. Following "The stylized facts", we will firstly compare Bitcoin with products from other asset classes and examine some important statistical properties in order to determine whether it is appropriate to use the FX model. Long memory will be tested using Hurst exponent to determine if the Bitcoin market is efficient and free of arbitrage. Moreover, return distributions of different asset classes will be discussed and compared. Secondly, we will move to "A Sentiment-Based Model". Three major questions will be raised and explored in this section: 1. Why we need a sentiment-based model? 2. What are the model assumptions and how do we test them? 3. How do we find proxies to the sentiment index? Furthermore, numerical examples will be given on Bitcoin prices from 2015-2017 Q1. Quasi maximum likelihood method (QML) will be employed in parameter estimation and results will be compared with the authors'. Lastly, there will be suggestion on improvements and further thoughts about this topic.

2 Stylized Facts about the Bitcoin

In this section, we will mainly answer two questions. From modeling perspective, should we treat Bitcoin as a currency or an equity? Is the Bitcoin market arbitrage-free? Both the authors for "Stylized Facts" and "A Sentiment-Based Model" answered the first question, and they agreed that Bitcoin should not be modeled as a currency, instead it should be treated as a stock-like asset. We will further work on some statistical properties to justify that statement in Section 2.1. The second question is discussed and answered with details in "Stylized Facts". We will state the idea and replicate the results in Section 2.2.

2.1 Currency or Equity

Cryptocurrencies are first known as substitutes to fiat currencies, with the decentralized feature and belongs to no political entities. As a currency, one can expect Bitcoin to exhibit three major properties:

1) can be served as a medium of exchange; 2) can be used as a unit of account and 3) can be used to store value. Unfortunately, we do not see the above three properties on Bitcoin. Bitcoin is barely used to purchase any non-virtual merchandise or measure the value of any products other than cryptocurrencies. Moreover, due to the large volatility Bitcoin has, it cannot be used to store value. Therefore, from the definition of a standard currency, we can conclude that Bitcoin should not be treated as a standard currency. On the other hand, we can consider Bitcoin as an equity investment since it is invested from a speculative perspective in anticipation of income from dividends and capital gains. Similar to stocks issuing cash dividends or stock dividends, Bitcoin has hard fork which generates another kind of cryptocurrency similar to Bitcoin and could be exchanged on the market with certain value. Bitcoin Cash and Bitcoin Gold are examples of such dividends from Bitcoin's hard fork.

To answer the question whether Bitcoin should be classified as a FX product or an equity product, we firstly compare the return and volatility with the other assets. Secondly, we compare the log return distribution of them. Datasets from 2012-10-01 to 2017-10-20 are used in this section.

To compare return and volatility of Bitcoin with other assets, we have taken some samples from the FX and stocks. We calculated their returns as well as volatilities and plotted them in a risk rewarding diagram.

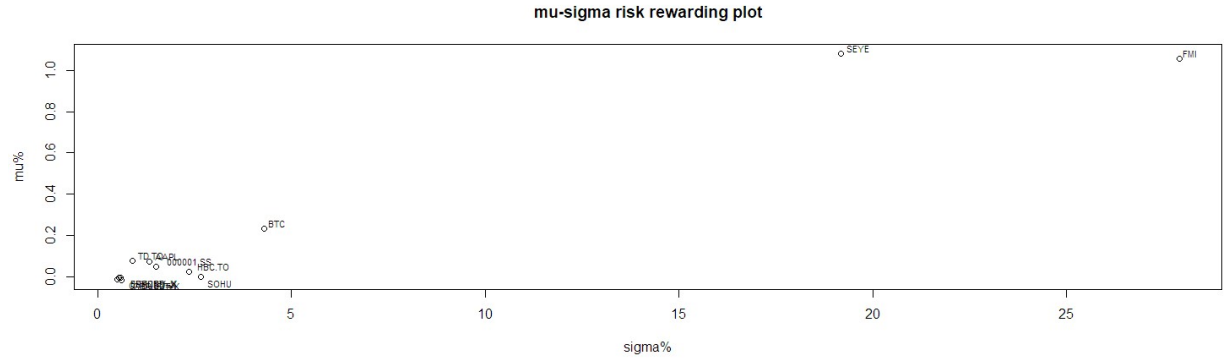

Figure 1: Risk Rewarding Plot on Bitcoin and other assets.

Table 1: Risk Rewarding Table on Bitcoin and other assets

Ticker	01.SS	TD.TO	HBC.TO	SOHU	FMI	SEYE	AAPL	CADUSD	GBPUSD	EURUSD	JPYCNY	BTC
Mu%	0.05	0.08	0.03	0	1.06	1.08	0.07	-0.01	-0.02	0	0	0.23
Sigma%	1.5	0.9	2.37	2.68	27.91	19.18	1.34	0.5	0.62	0.56	0.59	4.29

From Figure 1, we can easily see that Bitcoin has a much higher return and volatility than those standard currencies, which can be found as a cluster at the lower left corner. Though Bitcoin also exhibits higher returns and volatility than most of the stocks, we can still find stocks on the top right corner that are even more volatile than Bitcoin. Thus, it would be more appropriate to consider Bitcoin as a stock with high return and high volatility. Furthermore, we can also plot the historical daily log-return distribution, and compare the return distribution of Bitcoin with that of the other assets. Figure 2 shows the kernel density of the daily log returns with fitted normal distribution overlaid. As we can see from the plot, stocks tend to exhibit fatter tails than FX products. And Bitcoin has an even thicker tail than some of the stocks. From this point of view, we may also classify Bitcoin as a stock-like asset.

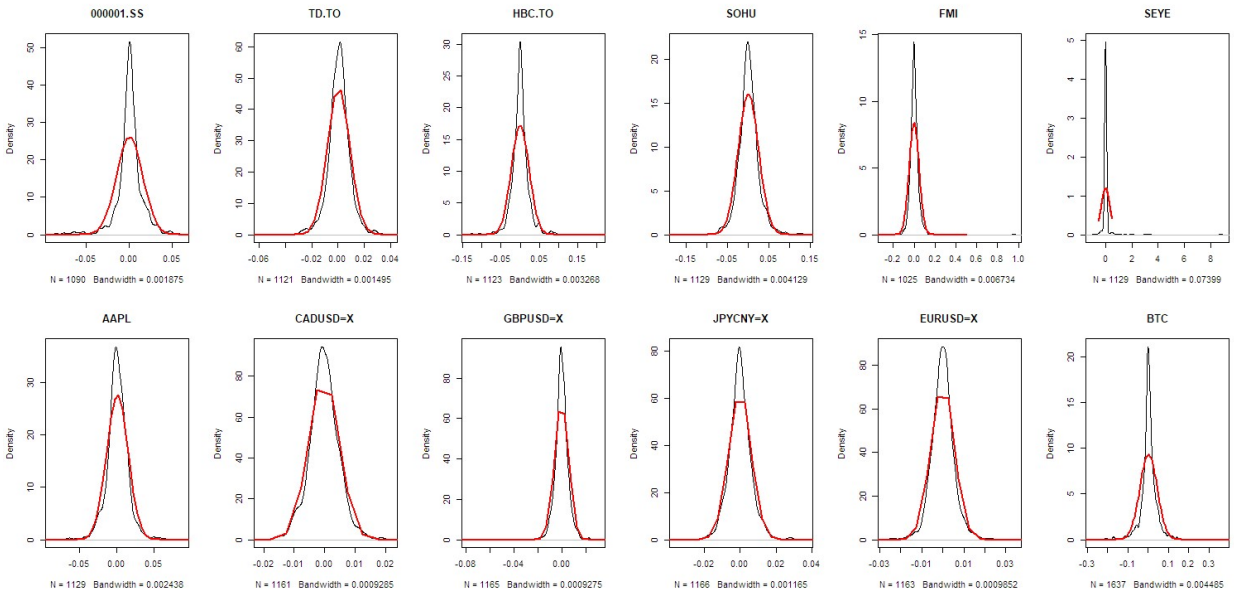

Figure 2: Daily log return distribution with fitted normal distribution overlaid.

To sum up, from the point of stylized facts and statistical properties, Bitcoin is better classified as a stock-like asset with high return and high volatility.

2.2 The Efficient Market Hypothesis (EMH) for the Bitcoin Market

According to Fama.E [4], informational efficiency can be divided into three types: 1) Weak Efficiency, if prices reflect the information contained in the past series of prices; 2) Semi-strong Efficiency, if prices reflect all public information; 3) Strong Efficiency, if prices reflect all public and private information. If a market is informationally efficient, there should not be any arbitrage opportunities. On the other hand,

if there is long memory in the Bitcoin return series, one would be able to find arbitrage opportunities and prove that the market is not informationally efficient. The aim of the long memory test is to see if the Bitcoin market is informationally efficient, since the Efficient Market Hypothesis is used in many models. In order to test the long memory effect of Bitcoin, we employ the Hurst exponent here. The idea of Hurst exponent is similar to autocorrelations of the time series, and the rate at which these decrease as the lag between pairs of values increases. The Hurst exponent H is defined in terms of the asymptotic behaviour of the rescaled range as a function of the time span of a time series as follows [5]:

$$E\left[\frac{R(n)}{SD(n)}\right] = Cn^H \quad as \quad n \to \infty$$

Detailed calculation for Hurst exponent can be found in Appendix A. For the Hurst exponent H we obtained, we have the following rules:

$0 < H < 0.5$: Anti-persistent time series

$H = 0.5$: Random Walk

$0.5 < H < 1$: indicates persistent behavior; the larger the H value the stronger the trend

By taking 500 and 90 data points sliding window respectively in the Hurst Exponent calculation, we obtain the following average results:

Table 2: Hurst Exponent for Bitcoin daily return series

Time range	2012-10 to 2013-12	2014-01 to 2016-12	2012-10 to 2017-10
Hurst Expo (500)	0.62351	0.50528	0.84571
Hurst Expo (90)	0.5591	0.45819	0.56285

The authors also computed the Hurst Exponent for GBP and EUR daily return series and find they are roughly within the interval $\mathcal{H} = (0.45, 0.55)$. If we also take the interval $\mathcal{H} = (0.45, 0.55)$ as the standard, from Table 2, we can easily see that for the three sub time series we take, the full-length series and the sub series before 2014 exhibit long memory property. But for the subsample from 2014 to 2016, there is no evidence showing the long memory property. Consequently, the Efficient Market Hypothesis should be used with extra caution since whether the Bitcoin market is informationally efficient or not depends on the time range taken. But overall, the full-length time series shows some evidence against that hypothesis. This conclusion is partially consistent with the authors'. For daily return time series, the authors did find some evidence against the Efficient Market Hypothesis. Nevertheless,

interestingly, if we take intraday data instead of daily data for Bitcoin prices, the conclusion would be slightly different. The author took 5h, 6h, 7h, 8h, 9h, 10h, 11h and 12h price data from the market and conducted the long memory test for each of them. They found that in all cases, there are significant persistent or procyclical behaviour until 2014. After 2014, the time series seem to stabilize around a value in $\mathcal{H} = (0.45, 0.55)$, introducing some sign of an informational efficient market. However, the authors did not find the reason for such change in the dynamic. One possible reason could be that the informational efficiency is related to volume of transactions or the liquidity level of the market. As there are more and more participants in the market after 2014, informational efficiency could increase.

Another fact that may challenge the Efficient Market Hypothesis is the existence of arbitrage opportunities between different exchanges. As we know if we assume the market is informationally efficient, there should not be any arbitrages. And the price should be unique or within a small range considering the transaction cost. Unlike the standard stock market, Bitcoin has many exchange platforms in different countries and it can be in exchange with different fiat currencies or cryptocurrencies. Bitcoin prices would generally be higher in countries with lower liquidity level or countries experiencing high inflation.

Arbitrage can be realized in the following way:

Setup: Assume we have two Bitcoin Exchanges A and B, the corresponding fiat currencies in the two exchanges are Currency A and Currency B. The two fiat currencies could be the same or different. Assume we have fiat currencies and Bitcoins in both accounts.

Flow: We assume Bitcoin price is lower in Exchange A.

1. Buy Bitcoins in Exchange A with fiat currency A;
2. At the same time sell the same quantity of Bitcoins in Exchange B for fiat currency B;
3. Send Bitcoins bought in Exchange A to Exchange B;
4. At the same time, withdrawal fiat currency B from Exchange B and deposit to account in Exchange A.

There are times that the price difference can be as large as 20% between large exchanges, especially when intraday volatility is large. Examples of price differences can be found in Appendix B. Some of them are for Ethereum, the others are for Bitcoins.

To sum up, the Bitcoin market is neither informationally efficient nor arbitrage-free.

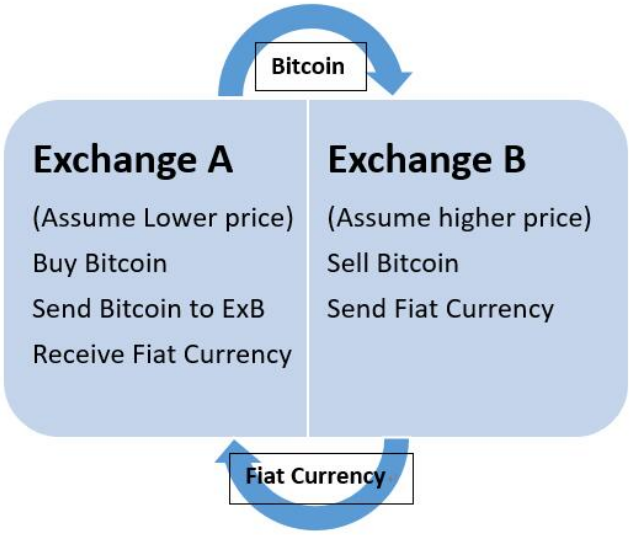

Figure 3: Arbitrage flow between Bitcoin Exchanges.

3 A Sentiment-Based Model

In this section, we will mainly answer two questions. Why should we use a sentiment-based model? Is there a closed form solution to the SDE's? Firstly, we will work on the major factors to Bitcoin prices, and try to incorporate them in one model in Section 3.1, and that will lead us to the answer of why we need a sentiment-based model. Secondly, we will state the model assumptions together with the model and find a closed form solution to the SDE's in Section 3.2. Some statistical properties of the model will also be examined in Section 3.2.

3.1 Rationale for a Sentiment-Based Model

In the recent years, there are some papers discussing the main drivers of the Bitcoin price. Many authors claimed that the high volatility in Bitcoin prices may depend on sentiment and popularity about the Bitcoin market itself [7,8,9]. In other words, unlike standard stocks, the risk-free rate or other general macroeconomic variables would not be major factors to Bitcoin prices. What really matters is people's enthusiasm about the Bitcoin market or even their enthusiasm about the whole cryptocurrency market. One major problem we have to face first is that we cannot observe the popularity or enthusiasm directly. Therefore, we have to find out a way to quantify the popularity or sentiment. Fundamentally, Bitcoin prices are affected by trading policies and major technical adjustments such as hard folks. And those affect people's enthusiasm about the Bitcoin market. The general idea here is that upon policy changes

or technical adjustments, people search to learn those news, and they modify their portfolio positions through trading corresponding to the policy changes afterwards. In this way, the sentiment could be captured and approximated by searching volume or trading volume, which we can observe from the market. In conclusion, we introduce a sentiment index in order to reflect and quantify people's enthusiasm about the Bitcoin market, which is determined as the major price driver for the Bitcoin market.

3.2 The Model and its Solution

We consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ together with a filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ which meets the usual conditions of right-continuity and completeness. Given that probability space, we consider a market in which people buy and sell Bitcoins and we denote the price of the Bitcoin by $S = \{S_t, t \geq 0\}$. We assume that the Bitcoin price dynamics is described by the following equation:

$$dS_t = \mu_S P_{t-\tau} S_t dt + \sigma_S \sqrt{P_{t-\tau}} S_t dW_t, \quad S_0 = s_0 \in \mathbb{R}_+, \tag{1}$$

where $\mu_S \in \mathbb{R}_+ \backslash \{0\}$, $\sigma_S \in \mathbb{R}_+$, $\tau \in \mathbb{R}_+$ is the lag parameter; $W = \{W_t, t \geq 0\}$ is a standard Brownian motion and $P = \{P_t, t \geq 0\}$ is a stochastic factor, representing the sentiment index in the Bitcoin market, satisfying

$$dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, P_t = \phi(t), \quad t \in [-L, 0], \tag{2}$$

where $\mu_P \in \mathbb{R}_+ \backslash \{0\}$, $\sigma_P \in \mathbb{R}_+$, $L \in \mathrm{R}_+$,$\mathrm{Z} = \{\mathrm{Z}_t, t \geq 0\}$ is a standard Brownian motion, which is independent of W, and $\phi : [-L, 0] \to [0, \infty)$ is a continuous deterministic initial function. Here we let ϕ be non-negative, indicating the sentiment index being non-negative. In "A Sentiment-Based Model", the authors consider both drift related terms μ_S and μ_P as constants, so as the volatility related terms σ_S and σ_P. Furthermore, we also consider τ, which is the lag parameter as a non-negative constant. Note that in Equation (2), we also consider the past information, i.e information before $t = 0$. We assume that the sentiment index P_t affects the Bitcoin price S_t up to a preceding time $t - \tau$. We also assume that $\tau < L$ and the sentiment index P_t is observed in the period before time 0. We know the solution to Equation (2) and the closed form solution tells us that P_t has a log-Normal distribution for positive t. And if we look at Equation (1), we will find that the instantaneous variance of the Bitcoin price process S_t is proportional to the delayed process $\sqrt{P_{t-\tau}}$. The idea behind this is that, as the sentiment index increases, which means either people's fear or enthusiasm about the Bitcoin market increases and that will lead to increasing trading activities. Yet, we do not know how or in which direction the price

is changing towards.

We then define the filtration used on the Bitcoin market. We assume that the filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ describing the information on the Bitcoin market, is of the form

$$\mathcal{F}_t = \mathcal{F}_t^W \vee \mathcal{F}_t^Z, \quad t \geq 0,$$

where $\mathcal{F}_t^W$ and $\mathcal{F}_t^Z$ denote the σ-algebras generated by W_t and Z_t respectively up to time $t \geq 0$. Note that $\mathcal{F}_t^Z = \mathcal{F}_t^P$, for each $t \geq 0$, with $\mathcal{F}_t^P$ being the σ-algebras generated by P_t up to time $t \geq 0$ Since at any time t the Bitcoin price dynamics is affected by the sentiment index only up to time $t - \tau$, to describe the trader's information on the Bitcoin market, we consider the filtration $\widetilde{\mathcal{F}} = \{\mathcal{F}_t, t \geq 0\}$ defined by

$$\widetilde{\mathcal{F}}_t = \mathcal{F}_t^W \vee \mathcal{F}_{t-\tau}^P, \quad t \geq 0.$$

We also consider that all filtrations satisfy the usual conditions of completeness and right continuity[10]. Now, we introduce the *integrated information process* $X^\tau = \{X_t^\tau, t \geq 0\}$ associated to the sentiment index P, defined as follows:

$$X_t^\tau := \begin{cases} \int_0^t P_{u-\tau} du = \int_{-\tau}^0 \phi(u)\, du + \int_0^{t-\tau} P_u du = X_\tau^\tau + \int_0^{t-\tau} P_u du, & 0 \leq \tau \leq t, \\ \int_{-\tau}^{t-\tau} \phi(u)\, du, & 0 \leq t \leq \tau. \end{cases}$$

Note that, for $t \in [0, \tau]$, we have $X_t^\tau = \int_{-\tau}^{t-\tau} \phi(u)\, du$ which is deterministic. In addition, for a finite time horizon $T > 0$, let us define the corresponding variation over the interval $[t, T]$, for $t \leq T$, as $X_{t,T}^\tau := X_T^\tau - X_t^\tau$. Obviously, $X_{T,T}^\tau = 0$; moreover, for $t < T$,

$$X_t^\tau := \begin{cases} \int_{t-\tau}^{T-\tau} P_u du & \texttt{if} \;\; 0 \leq \tau \leq t < T, \\ \int_{t-\tau}^0 \phi(u)\, du + \int_0^{T-\tau} P_u du & \texttt{if} \;\; 0 \leq t \leq \tau < T, \\ \int_{t-\tau}^{t-\tau} \phi(u)\, du & \texttt{if} \;\; 0 \leq t < T \leq \tau. \end{cases}$$

Again, note that for $T \leq \tau$, we get $X_{t,T}^\tau := \int_{t-\tau}^{t-\tau} \phi(u)\, du$ which is deterministic.

Now let us go back and put Equation (1) and (2) together, we will have the bivariate stochastic delayed differential equations

$$\begin{cases} dS_t = \mu_S P_{t-\tau} S_t dt + \sigma_S \sqrt{P_{t-\tau}} S_t dW_t, \quad S_0 = s_0 \in \mathbb{R}_+ \\ dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, \quad P_t = \phi(t), t \in [-L, 0] \end{cases} \tag{3}$$

and it has a continuous unique solution $(S,P) = \{(P_t, S_t), t \geq 0\}$ given by

$$S_t = s_0 e^{\left(\mu_S - \frac{\sigma_S^2}{2}\right)\int_0^t P_{u-\tau}du + \sigma_S \int_0^t \sqrt{P_{u-\tau}}dW_u}, \quad t \geq 0, \tag{4}$$

$$P_t = \phi\left(0\right) e^{\left(\mu_P - \frac{\sigma_P^2}{2}\right)t + \sigma_P Z_t}, \quad t \geq 0. \tag{5}$$

If we apply $Itô's$ lemma to $log(S_t)$, we will get

$$d\log\left(S_t\right) = \left(\mu_S - \frac{\sigma_S^2}{2}\right) P_{t-\tau}dt + \sigma_S\sqrt{P_{t-\tau}}dW_t, \tag{6}$$

$$\log\left(\frac{S_t}{S_0}\right) = \left(\mu_S - \frac{\sigma_S^2}{2}\right)\int_0^t P_{u-\tau}du + \sigma_S\int_0^t \sqrt{P_{u-\tau}}dW_u, \quad t \geq 0. \tag{7}$$

Based on Equation (6) and (7), we have the following conclusions:

(i) For every $t \geq 0$, the conditional distribution of S_t, given the integrated information X_t^τ, is log-Normal with mean $log(S_0) + (\mu_s - \frac{\sigma_s^2}{2})X_t^\tau$ and variance $\sigma_S^2 X_t^\tau$.

(ii) For every $t \in [0,t]$, the random variable $\log(S_t)$ has mean $\log(S_0) + (\mu_s - \frac{\sigma_s^2}{2})X_t^\tau$ and variance $\sigma_S^2 X_t^\tau$ respectively given by

$$\mathbb{E}\left[\log\left(S_t\right)\right] = \log\left(s_0\right) + \left(\mu_s - \frac{\sigma_s^2}{2}\right)\mathbb{E}\left[X_t^\tau\right]$$

$$\mathbb{V}ar\left[\log\left(S_t\right)\right] = \left(\mu_s - \frac{\sigma_s^2}{2}\right)\mathbb{V}ar\left[X_t^\tau\right] + \sigma_s^2\mathbb{E}\left[X_t^\tau\right]$$

where $E[X_t]$ and $Var[X_t]$ can be calculated explicitly from the definition of X_t^τ, and are given as

$$\mathbb{E}\left[X_t^\tau\right] = X_\tau^\tau + \frac{\phi(0)}{\sigma_P}(e^{\mu_P(t-\tau)} - 1)$$

$$\begin{aligned}\mathbb{V}ar\left[X_t^\tau\right] = &\frac{2\phi^2(0)}{(\mu_P + {\sigma_P}^2)(2\mu_P + {\sigma_P}^2)}(e^{(2\mu_P + {\sigma_P}^2)(t-\tau)} - 1) \\ &- \frac{2\phi^2(0)}{\mu_P(\mu_P + {\sigma_P}^2)}(e^{\mu_P(t-\tau)} - 1) - \left[\frac{\phi(0)}{\sigma_P}(e^{\mu_P(t-\tau)} - 1)\right]^2\end{aligned}$$

Next, we will show the existence of a risk-neutral probability measure. Let us fix a time $T > 0$ and assume the existence of a risk-free asset, denoted as $\mathrm{B} = \{B_t, t \in [0,T]\}$ given by

$$B_t = e^{\int_0^t r(s)ds}, t \in [0,T], \tag{8}$$

where $r : [0, T] \to \mathbb{R}$ is a bounded, deterministic function representing the instantaneous risk-free rate.

We denote $\widetilde{S}_t = \{\widetilde{S}_t, t \in [0, T]\}$ as discounted Bitcoin price process and it is defined as $\widetilde{S}_t := \frac{S_t}{B_t}$ for $t \in [0, T]$. Then we have the conclusion the pair $(\widetilde{S}_t, P_t)$ satisfies the following system of stochastic delayed differential equations:

$$\begin{cases} d\widetilde{S}_t = \sigma_S \sqrt{P_{t-\tau}} \widetilde{S}_t d\widehat{W}_t, & S_0 = s_0 \in \mathbb{R}_+ \\ dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, & P_t = \phi(t), t \in [-L, 0], \end{cases} \tag{9}$$

Solving for $\widetilde{S}_t$, we get

$$\widetilde{S}_t = s_0 e^{\sigma_S \int_0^t \sqrt{P_{u-\tau}} d\widehat{W}_u - \frac{\sigma_S^2}{2} \int_0^t P_{u-\tau} d_u}, \quad t \in [0, T] \tag{10}$$

where

$$\begin{aligned} \widehat{W}_t :=& W_t + \int_0^t \frac{\mu_S P_{s-\tau} - r(s)}{\sigma_S \sqrt{P_{s-\tau}}} ds, \quad t \in [0, T], \\ \widehat{Z}_t :=& Z_t, \quad t \in [0, T]. \end{aligned} \tag{11}$$

Therefore, we will have the Bitcoin price under the martingale or risk neutral measure given by

$$\begin{cases} dS_t = r(t)dt + \sigma_S \sqrt{P_{t-\tau}} S_t d\widehat{W}_t, & S_0 = s_0 \in \mathbb{R}_+, \\ dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, & P_t = \phi(t), t \in [-L, 0], \end{cases} \tag{12}$$

where r(t) is the risk-free rate.

Thus, we showed the existence of a risk-neutral probability measure.

4 Implementation

In this section, we will work on the implementation of the model. The implementation consists of two parts. Firstly, we have to answer how do we find proxies to the sentiment index. Two proxies to the sentiment index, the Google search volume and volume of transactions in the Bitcoin Market, will be discussed in details in Section 4.1. And we will test the log-Normality of the sentiment index we find. Secondly, there are five parameters we have in the bivariate stochastic process. We will fit in the real data and discuss the ways to estimate those parameters and compare our results with the authors'.

4.1 Proxies for the Sentiment Index

The most important part of the Bitcoin price dynamics is the sentiment index. The authors suggested several proxies that can be used, including the Google search volume, volume of Bitcoin trading transactions and Wiki requests. The Google search volume and Wiki requests would be considered closer to a sentiment index while the volume of transactions is more conventional as an indirect indicator. The idea of Wiki requests is similar to Google search volume and since the full Wiki requests data has relatively less historical data, we will not discuss it here.

Note that if we are looking for proxies, we would prefer daily time series than weekly time series. The reason is that, if we are fitting real data to the series, consider the lag parameter τ, by using weekly data, the least positive integer we can choose is 1. In this sense, we assume that the Bitcoin price would be one week behind the sentiment index. And the filtration we use would only include information up to a week ago. According to the sample correlation we have between the Bitcoin price series and the delayed proxies for the sentiment index series, we could easily see that the correlation is decreasing as a function of τ. By using weekly data, we would lose much more information compared with using daily data. But for some reason, the authors did not get the daily Google search series, instead, they get the weekly Google search series. Here, we will test both daily and weekly Google search series.

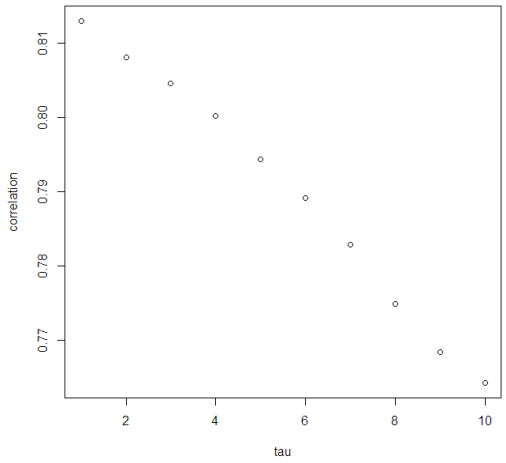

Figure 4: Correlation between Bitcoin Price and the Sentiment Index as a function of τ.

Note that Google search volume is measured in 1 unit. An index from 0 to 100 will be given each day. A score of 100 will be given to the day with the largest search volume in the period requested. And the other scores will be given based on search volume proportional to the largest search volume.

The Google search index can be defined as:

$$P_t^{Google} = \frac{\text{Search volume at time t}}{\text{Max Search volume}} * 100$$

Similarly, we can unitize volume of transactions as:

$$P_t^{Vol} = \frac{\text{Volume of Transactions at time t}}{\text{Max Volume of Transactions}} * 100$$

Now we recall the solution to the sentiment index dynamics stated in Equation (6), and we assumed that P_t would follow a log-Normal distribution. Thus, the first thing we will do is to test the log-Normality of the Google search volume series and the series of volume of transactions. We can either test the observed P_t series by fitting a log-Normal distribution or we can test the log return of P_t and fit in a normal distribution. Here we will conduct the one-sample Kolmogorov-Smirnov (KS) Test for the observed P_t. We find that all full-range series have a p-value far less than 0.05, which indicates a failure to fit in a log-Normal distribution. Thus, we decided to take a sub-sample from 2015-01-01 to 2017-03-31 instead. The p-values of our KS test are given in Table 3.

Table 3: p-value of the one-sample KS test

Source	Volume of Transactions	Daily Google Search	Weekly Google Search
ME	0.2836159	8.49953e-05	0.1313683
Author	0.2152	NA	0.1012

We find that the daily Google search series fail to fit in a log-Normal distribution while Volume of Transactions and Weekly Google Search could be fitted using a log-Normal distribution. This is not good for the Goole search volume since if we use weekly series, we will lose a lot of information. Whereas if we use daily series, it does not fit the log-Normal distribution.

4.2 Parameter Estimation

In this section, we will fit in the real data, which can be considered as a discrete sample. And we will suggest a possible closed form approximation for the joint probability density of the discrete sample. Furthermore, we will give introduce Profile Likelihood to estimate the lag parameter τ.

4.2.1 Quasi Maximum Likelihood (QML)

Maximum likelihood estimation is one of the most widely used estimation method. The idea here is that we use the probability density function to obtain a approximated likelihood function and pick the values that maximize the likelihood function given the realization of the Bitcoin price and the sentiment index.

We firstly define a series A_i^τ with $A_i^\tau := X^\tau_{(i-1)\Delta,i\Delta}$, where $X^\tau_{t,T}$ is the variation of the integrated information process introduced in Section 3.2, and Δ is the time between two discrete sample observations. Since τ is a fixed number, we will just denote as A_i.

Then we can write the unconditional joint probability distribution of $(\mathbf{R}, \mathbf{A})$ by applying Bayes's rule and we will get

$$f_{(\mathbf{R},\mathbf{A})}(\mathbf{r},\mathbf{a}) = f_{A_1}(a_1)\prod_{i=2}^{n} f_{(A_i|A_{i-1})}(a_i)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_S^2 a_i}}e^{-\frac{1}{2}\frac{\left(r_i-\left(\mu_S-\frac{\sigma_S^2}{2}\right)a_i\right)^2}{\sigma_S^2 a_i}} \tag{13}$$

where $\mathbf{R} = (R_1, R_2, ..., R_n)$ is the random variable of discretely observed log-return series of the Bitcoin, conditionally on $\mathbf{A}$ and is jointly normal with covariance matrix $\Sigma = {\sigma_S}^2 Diag(A_1, A_2, ..., A_n)$. And $\mathbf{A} = (A_1, A_2, ..., A_n)$ is a multivariate normal with covariance matrix $Diag(A_1, A_2, ..., A_n)$.

Let $\phi(t) > 0$, for each $t \in [-L, 0]$ in Equation (2) and $\tau < \Delta$. Then, in the market model outlined in Section 3.2, we have:

(i) the distribution of $A_1 - X_\tau^\tau$ is approximated by a log-normal with mean α_1 and ν_1^2 given by

$$\alpha_1 = \log\phi(0) + 2\log\frac{e^{u_p(\Delta-\tau)}-1}{\mu_p} - \frac{1}{2}\log\left(\frac{2}{\mu_p+\sigma_p}\left[\frac{e^{\left(2\mu_p+\sigma_p^2\right)(\Delta-\tau)}-1}{2\mu_p+\sigma_p^2} - \frac{e^{u_p(\Delta-\tau)}-1}{\mu_p}\right]\right)$$

$$\nu_1^2 = \log\left(\frac{2}{\mu_p+\sigma_p}\left[\frac{e^{\left(2\mu_p+\sigma_p^2\right)(\Delta-\tau)}-1}{2\mu_p+\sigma_p^2} - \frac{e^{u_p(\Delta-\tau)}-1}{\mu_p}\right]\right) - 2\log\left(\frac{e^{u_p(\Delta-\tau)}-1}{\mu_p}\right)$$

(ii) the distribution of A_i given A_{i-1} (shortly $Ai|A_{i-1}$), for $i = 1, \cdots, n$, is approximated by a log-normal with means α_i and variances ν_i^2 i given by

$$\alpha_1 = \log(A_{i-1}) + \left(\mu_p - \frac{\sigma_p^2}{2}\right)\Delta, \quad for\ i = 1, \cdots, n,$$

$$\nu_1^2 = \sigma_p^2\Delta, \quad for\ i = 1, \cdots, n.$$

The detailed proof can be found in the Appendix of "A Sentiment-Based Model", we will not demonstrate it here.

Furthermore, given the realized sample $(\bar{\mathbf{r}}, \bar{\mathbf{a}})$, the log-likelihood function $\log\mathcal{L}_{\mathbf{R},\mathbf{A}}(\mu_P, \mu_S, \sigma_P, \sigma_S) : \mathbb{R}^2 \times \mathbb{R}_+^2 \to \mathbb{R}$ is given as

$$\begin{aligned}\log \mathcal{L}_{\mathbf{R},\mathbf{A}}\left(\mu_P, \mu_S, \sigma_P, \sigma_S\right) = &\sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma_S^2 a_i}}\right) - \frac{1}{2}\frac{\left(r_i - \left(\mu_S - \frac{\sigma_S^2}{2}\right)a_i\right)^2}{\sigma_S^2 a_i}\right] \\ &+ \sum_{i=1}^{n}\left[\log\left(\frac{1}{a_i\nu_i\sqrt{2\pi}}\right) - \frac{\left(\log\left(a_i\right) - \alpha_i\right)^2}{2\nu_i^2}\right]\end{aligned} \tag{14}$$

where $(\mathbf{R}, \mathbf{A})$ are random variables and $(\mathbf{r}, \mathbf{a})$ are the corresponding realizations.

Maximum likelihood estimates for the model can be obtained by maximizing the log-likelihood approximation in Equation (14) as

$$(\hat{\mu}_P, \hat{\mu}_S, \hat{\sigma}_P, \hat{\sigma}_S) = \arg\max_{\substack{\mu_P\mu_S \\ \sigma_P\sigma_S}} \log \mathcal{L}_{\mathbf{R},\mathbf{A}}\left(\mu_P, \mu_S, \sigma_P, \sigma_S\right)$$

The reason it is not exactly Maximum Likelihood(ML) but Quasi-Maximum Likelihood (QML) is that the likelihood function we obtain is not an exact one, instead it is approximated as mentioned before. In this case the methodology is referred to as quasi-maximum likelihood, and under suitable conditions, quasi-maximum likelihood estimates are asymptotically equivalent to the maximum likelihood estimates[11,12].

4.2.2 The Profile Likelihood Approach

We will briefly describe the Profile Likelihood approach the authors applied[13,14]. The basic idea is that if there are two sub-vectors in our parameter vector θ, we may split them into two parts. In our case, we have $\theta = (\gamma, \lambda)$, where λ is the parameter of interest, which we also call the nuisance parameter, $\lambda = (\mu_P, \mu_S, \sigma_P, \sigma_S)$. And we have $\gamma = \tau$, the lag parameter.

To estimate (λ, γ) jointly, we should maximize at the likelihood

$$\arg\max_{\gamma,\lambda} \log \mathcal{L}\left(\gamma, \lambda\right)$$

When the above is difficult to achieve, but the likelihood of the nuisance parameter is available, we can apply a two step procedure by maximizing all γ in its parametric space, i.e. we take γ from a possible set $\{1, 2, 3...10\}$, and for each γ, we compute the maximum likelihood, $\log \mathcal{L}\left(\lambda\right)$ given γ_i. At last, we pick the maximum likelihood and get the corresponding γ_i as $\hat{\gamma}$.

The authors also suggest to apply the method of moments to estimate (μ_P, σ_P) separately. And this can be achieved by computing the sample mean and sample variance of the sentiment realizations. If we then plug the estimated values in the likelihood in order to estimate (μ_S, σ_S), these two estimates remain unchanged. In fact the likelihood may be maximized separately with respect to (μ_S, σ_S) and (μ_P, σ_P) since each of the two addend in the likelihood expression depends on just one of this pairs.

4.2.3 Results

In this section we will compare our results with the authers'. Note that, the data source we use is different from the authors'. We are using daily Google search Volume while the authors are using weekly Google search Volume. We used method of moments and applied profile likelihood approach to estimate τ separately.

Table 4: Parameter fit with Daily Google Search Volume

Source	μ_P	σ_P	μ_S	σ_S	τ
ME	0.3895	2.2121	0.0125	0.0725	1 (i.e. 1 day)

Table 5: Parameter fit with Weekly Google Search Volume

Source	μ_P	σ_P	μ_S	σ_S	τ
Author	0.9573	1.0818	0.0181	0.0867	1 (i.e. 1 Week)

From Table 4 and Table 5, we find that the daily Google search volume series seems more volatile than the weekly data. And if we use these estimated $(\hat{\mu_P}, \hat{\sigma_P})$ to simulate a path for P_t by using T = 2 years. We will be getting a proportion of sentiment index numbers P_T larger than even 200. Some details will be further discussed in Section 5.4.

5 Critics and Further Thoughts

Introducing a sentiment index is a brilliant idea in pricing the Bitcoin, since it successfully captured the major price driver and quantified it. Nevertheless, there are also some problems the model is facing, and in this section, we will discuss some of the problems. Firstly, we will challenge the arbitrage-free assumption in Section 5.1. Secondly, we will discuss if it is appropriate to assume a constant volatility for P_t in Section 5.2. Thirdly, as we find that the instantaneous drift term of S_t is always positive, we

will discuss the applicability in Section 5.3. Last but not lease, we will try to simulate the sentiment index and state the difficulties we encountered.

5.1 The Arbitrage-free Assumption

In Section 2.2, we tested the long memory property and find that the Bitcoin market is not always arbitrage-free. According to the results from the long memory test, we should not use any models with an arbitrage-free assumption. But in Section 4.2, following the dynamics of the bivariate model, we find that there exists a risk-neutral probability measure, indicating the Bitcoin market is arbitrage-free. Everything else is built on the arbitrage-free assumption. The authors also discussed the pricing for Bitcoin Options by using the Black-Scholes formula in "A Sentiment-Based Model" , which also requires the market to be arbitrage-free. The violation of the model assumption would be the first problem we have to face. One suggestion to that would be we may try some long memory processes instead. Fractional process would be an example.

5.2 The Assumption of Constant σ_P

In Equation (2), the dyanmics of P_t, we assumed that the instantaneous volatility σ_P is a constant. If we take a rolling window and calculated a sample variance series from the Google search series. We will be getting a volatility series as shown in Figure 5.

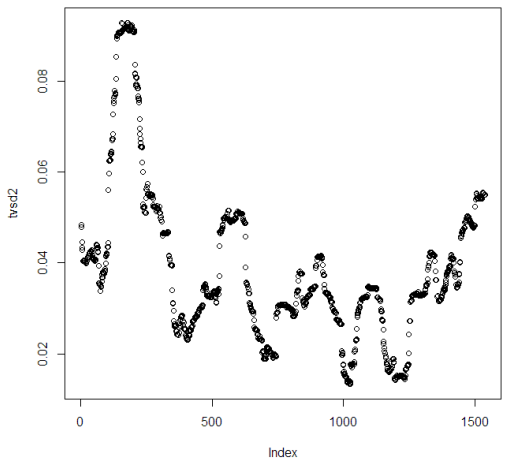

Figure 5: Time varying sample standard deviation as a function of t.

We can easily see that the volatility is not constant and there is a strong pattern in the figure.

Since the volatility varies a lot at different time points, it would be inappropriate to assume a constant volatility throughout all time. Instead, a time-varying volatility process $\sigma(t, P_t)$ would be suggested.

5.3 The Instantaneous Drift Term

In the dynamics of the Bitcoin price stated in Equation (1), we observe that the instantaneous drift term is equal to $\mu_S * P_{t-\tau}$. And we also assume that $P_{t-\tau}$ is always positive. With this being said, we actually assumed that the instantaneous drift term of S_t is always positive or always negative. Moreover, as $P_{t-\tau}$ gets larger, the instantaneous drift term here would be larger and proportional to $P_{t-\tau}$. At the same time, let us look at the instantaneous volatility term, it is increasing at the speed of $\sqrt{P_{t-\tau}}$, which is reasonable.

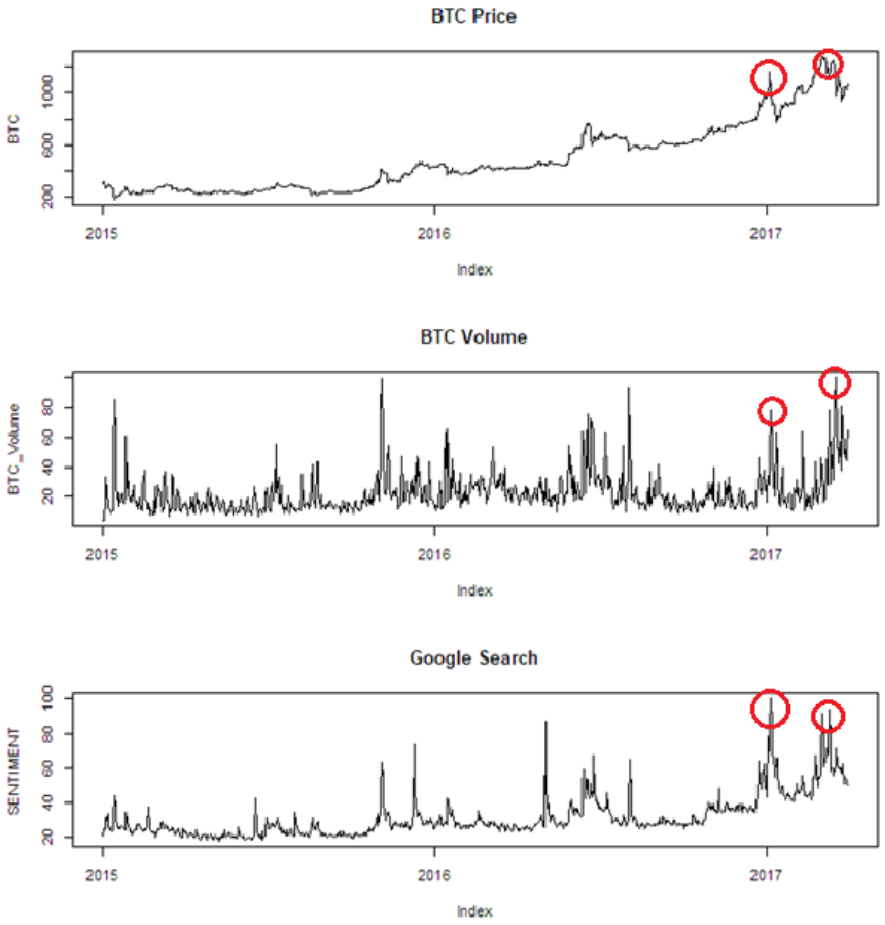

Figure 6: Bitcoin Price and Proxies for the Sentiment Index

As we can see from Figure 6. There are two apparent spikes in the first quarter of 2017. And it

appeared in 2017-01-04 to 2017-01-06 and 2017-03-03 to 2017-03-10, the same for the Google Search volume series. If we look closer at the local behaviour around 2017-01-04, we find that after that day, there were two large drops consecutively in the Bitcoin price, while Google Search volumes remain high. Here we can see when sentiment index is at a very high level, price of Bitcoin may drop dramatically. If we go back to the fundamentals of the sentiment index, we know it measures both people's fear and enthusiasm. However, when people are having fears towards the Bitcoin market, such as the case when China was shutting down all the Crypto Exchanges, we should expect the price to drop significantly. Since we always assume the instantaneous drift term to be positive, this cannot be reflected appropriately.

One improvement we can make here is to allow the sentiment index to go negative. Technically speaking, it may require some natural language processing to get proper proxies. It collects text strings and analyze the feeling behind it. And in this way, the sentiment index will have both magnitude and direction. Our bivariate model will become

$$\begin{cases} dS_t = \mu_S P_{t-\tau} S_t dt + \sigma_S \sqrt{|P_{t-\tau}|} S_t dW_t, \quad S_0 = s_0 \in \mathbb{R}_+ \\ dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, \quad P_t = \phi(t), t \in [-L, 0] \end{cases} \tag{15}$$

5.4 Simulation of the Sentiment Index

As stated in Equation (2), the dynamics of the sentiment index follows

$$dP_t = \mu_P P_t dt + \sigma_P P_t dZ_t, P_t = \phi(t), \quad t \in [-L, 0].$$

We have also stated in Section 4.1, we will have the sentiment index stay non-negative. Now let us consider about pricing an option using the Monte Carlo method. Assume we start at time 0, with $\phi(0) = k$, where k is a relatively small number. Chances are there might be negative numbers we generated for $P(t)$, $t \in [0, T]$. Similarly, if we are using $(\hat{\mu_P}, \hat{\sigma_P}, \hat{\mu_S}, \hat{\sigma_S}, \hat{\tau})$ estimated from our proxies in the simulation, we do expect that $P(t) \leq 100$. But again, we may generate numbers that are larger than 100. Same idea as in Section 5.3, we suggest to allow the sentiment index to go negative. And we will have the dynamics as stated in Equation (15).

6 References

[1] Bitcoin Price, Retrieved October 31, 2017, from https://charts.bitcoin.com/chart/price

[2] Bitcoin Trading Volume . Retrieved October 31, 2017, from https://charts.bitcoin.com/chart/market-cap

[3] Bitcoin Futures Expected to Begin Trading on Two Major U.S. Exchanges This Month. (2017, December 6). *Mondaq Business Briefing*. Retrieved December 10, 2017, from http://www.highbeam.com/doc/1G1-517633172.html?refid=easy_hf

[4] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2, Papers and Proceedings of the Twenty-Eighth Annual Meeting of the American Finance Association New York, N.Y. December, 28-30, 1969):pp. 383-417.

[5] Feder, Jens (1988). Fractals. New York: Plenum Press. ISBN 0-306-42851-2.

[6] Annis, A. A.; Lloyd, E. H. (1976-01-01). The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika*. 63 (1): 111116. doi:10.1093/biomet/63.1.111. ISSN 0006-3444.

[7] Young Bin Kim, Sang Hyeok Lee, Shin Jin Kang, Myung Jin Choi, Jung Lee, and Chang Hun Kim. Virtual world currency value fluctuation prediction system based on user sentiment analysis. PLoS ONE, 10(8):e0132944, 2015.

[8] Ladislav Kristoufek. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3, 2013.

[9] Ladislav Kristoufek. What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis. PLoS ONE, 10(4):E0123923, 2015.

[10] Philip E. Protter. Stochastic differential equations. In Stochastic Integration and Differential Equations, pages 249361. Springer, 2005.

[11] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, pages 125, 1982.

[12] Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681700, 1984.

[13] Anthony Christopher Davison. Statistical models, volume 11. Cambridge University Press, 2003.

[14] Yudi Pawitan. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, 2001.

[15] Bariviera, A., Naiouf, M., Hasperu, W., Basgall, M. J. (2017). Some stylized facts of the Bitcoin

market. *Physica A*, Volume 484(October), 82-90. Retrieved November 10, 2017, from https://doi.org/10.1016/j.physa

[16] Cretarola, A., Fig-Talamanca, G., Patacca, M. (2017). A Sentiment-Based Model for the Bitcoin: Theory, Estimation and Option Pricing. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=30420

[17] H.E. Hurst (1951) Long-term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers 116, 770-808.

[18] R. Weron (2002) Estimating long range dependence: finite sample properties and confidence intervals, *Physica A* 312, 285-299.

7 Appendix

7.1 Appendix A

Hurst Exponent can be calculated in the following way:

Step 1. Calculate the mean of $P(t)$ as m;

Step 2. Create a mean-adjusted series $Y(t) = P(t) - m$;

Step 3. Calculate the cumulative deviate series as sum of $Y(t)$ as S;

Step 4. Compute the range of S as R;

Step 5. Compute the standard deviation of $P(t)$ as SD;

Step 6. Calculate the rescaled range $R(n)/SD(n)$ and average over all the partial time series of length n.

Step 7. Solve for H as $E\left[\frac{R(n)}{SD(n)}\right] = Cn^H$

7.2 Appendix B

Table 6: Arbitrage opportunities in the Etherum and Bitcoin Market

Date	Price CAD	Convert CNY	Price CNY	Diff CNY	Difference
10-Sep	377.02	1949.1934	1600	349.1934	17.91%
	377.02	1949.1934	1633	316.1934	16.22%
	365.02	1887.1534	1640	247.1534	13.10%
	365.11	1887.6187	1650	237.6187	12.59%
	365.04	1887.2568	1650.9434	236.3134	12.52%
11-Sep	380.01	1964.6517	1700.6803	263.9714	13.44%
	380	1964.6	1700.6803	263.9197	13.43%
	368.06	1902.8702	1759.9437	142.9265	7.51%
	375.6	1941.852	1630.7893	311.0627	16.02%
	370.02	1913.0034	1777.5521	135.4513	7.08%
	371	1918.07	1780.2645	137.8055	7.18%
12-Sep	381.26	1971.1142	1807.8512	163.263	8.28%
	386.86	2000.0662	1772.1519	227.9143	11.40%
	380.21	1965.6857	1813.4715	152.2142	7.74%
	385	1990.45	1838.2353	152.2147	7.65%
14-Sep	311	1607.87	1394	213.87	13.30%
	4569.821	23625.972	20600	3025.972	12.81%
	4423	22866.91	19399	3467.91	15.17%
	296	1530.32	1358	172.32	11.26%
15-Sep	322.31	1666.3427	1160	506.3427	30.39%
	327.63	1693.8471	1333.2682	360.5789	21.29%
16-Sep	322.61	1667.8937	1415.9691	251.9246	15.10%
	322.18	1665.6706	1388.5195	277.1512	16.64%
17-Sep	323.11	1670.4787	1400	270.4787	16.19%
	324.18	1676.0106	1376	300.0106	17.90%
	334.87	1731.2779	1404	327.2779	18.90%
27-Sep	374.3	1946.36	1666	280.36	14.40%