National Research University – Higher School of Economics

International College of Economics and Finance

Bachelor Thesis

Price Discovery on the Bitcoin Market

Student: Ruslan Doga

Supervisor: Alexei Boulatov

Contents

1 Introduction

Bitcoin is a decentralized peer-to-peer cryptocurrency protocol first outlined in a paper by Nakamoto (2008) [2]. Over the past nine years it has grown from an experiment into its own market, with capitalization reaching $40 billion. This market is a fully electronic market. In particular, there is no central bank, and a number of trading platforms constitute the only intermediaries. Transactions are verified by a network of nodes that check the accuracy of the latest transaction against the blockchain, a distributed ledger of total transactions. The transaction is subsequently added to the ledger and information is redistributed to other nodes.

Trading Bitcoin is similar to trading stocks on a limit order book market. In particular, the trading platforms do not act as market makers – it's the traders themselves who post market or limit orders which make up an order book. Market orders are immediately executed against standing offers in the book. Limit orders enter the book unless they can be fully executed immediately. Trading is continuous only, there are no auctions or volatility interruptions. Also, the markets never close which allows for trading during 24/7.

The market is fully transparent, that is traders are provided with all the information about the complete state of the order book, as well as the total available volume and the associated price levels. It is also possible to access the entire transaction history for almost each exchange through their APIs.

One of the features of Bitcoin is that it is traded on many exchanges. Here are some of them, sorted by the trade volume in the past 24 hours.

#	Name	Volume (past 24h)
1	BitMEX	$142,027,000
2	Bitfinex	$73,628,500
3	GDAX	$70,368,000
4	Poloniex	$56,366,600
5	Bitstamp	$47,293,700
6	Gemini	$29,442,900
7	Kraken	$24,773,000
8	BTC-E	$20,538,000
9	xBTCe	$11,509,900
...	...	...
15	CEX.IO	$2,893,710
...	...	...
24	HitBTC	$270,756

Figure 1: Bitcoin exchanges sorted by trade volume (for BTCUSD pair only) for the past 24 hours. Source: coinmarketcap.com

The first (and so far, the only) attempt to investigate this multiexchange environment and particularly to study the price discovery was made by Brandvold, Molnar, Vagstad, Valstad (2015) [1]. However, since this topic keeps surfacing the Bitcoin community and knowing which exchange reacts most quickly to new information thus reflecting value of

Bitcoin most precisely is obviously important, in this paper I will attempt to repeat their investigation. Even though their paper is pretty recent, quite a few things have changed in the Bitcoin market, most notably, Mt.Gox, one of the exchanges which was identified by the authors as a leader and which contributed the most to the price discovery process, has filed for bankruptcy, and BTC-E, another leader and also a significant conttributor to the price discovery process has since gave up it's position to other exchanges. My results are supportive of Brandvold et al. (2015) claims in the sense that there still are exchanges who can be thought of as leaders and who contribute to the price discovery process more than others. However the difference among the top exchanges is not as apparent as in Brandvold et al. (2015), with three, instead of two, exchanges having significant information share.

The remainder of this paper is organized as follows. Section 2 (Date) describes the data used. Section 3 describes the price discovery model used, while Section 4 describes the implementation of this model to Bitcoin data. The results are presented in Section 5. Finally, Section 7 draws a conclusion.

For the source code both for computations and the trade history API scrapers, please visit github, for the computations in particular, visit this jupyter notebook.

2 Data

For their study of the Bitcoin market, Brandvold et al. (2015) picked seven Bitcoin exchanges, five big, Bitfinex, Bitstamp, BTC-e, BTC China and Mt.Goxm and two small, Bitcurex and Canadian Virtual Exchange (Virtex). The authors claim that these exchanges had a significant market share at the time, 90%. It helped them make an assumption, that these exchanges represented the whole Bitcoin market. For my purposes, I've collected the trade history data also from seven exchanges, GDAX, Kraken, BitMEX, Bitfinex, Cex, Gemini, and HitBTC (I've later removed HitBTC from the sample because it contained too many missing observations). It can be seen in Figure 1, that these are the exchanges currently occupying a big market share, albeit not 90%. I've tried to also collect the data from Poloniex, Bitstamp, and BTC-E, but failed. Bitstamp and BTC-E don't provide public trade history APIs, and with Poloniex I had technical problems. However, I don't expect that to bias the results, since the market share of the exchanges for which I did collect the data is close to 80%, so that the assumption that the exchanges in the sample might represent the whole Bitcoin market is not too far-fetched.

The time interval for the trades in Brandvold et al. (2015) was set to the period from April 1, 2013 to February 25, 2014. In this paper, the interval is set to the period from September 1, 2016 to Juny 1, 2017 in order to contain the trades from all seven exchanges. Since in my case the data was collected directly from the exchanges' APIs, it is expected to contain all the trade history data without any omissions (Brandvold et al. (2015) downloded the data from an aggregator, www.bitcoincharts.com).

For each exchange the collected data contains (at least) a timestamp (in unixtime) of each trade, its price and volume.

To create a time series of trades for each exchange I sample the trades on five minute intervals by taking the price of the last trade in this interval. If there are no trades happening on any exchange during this time interval, it is treated as a missing observation.

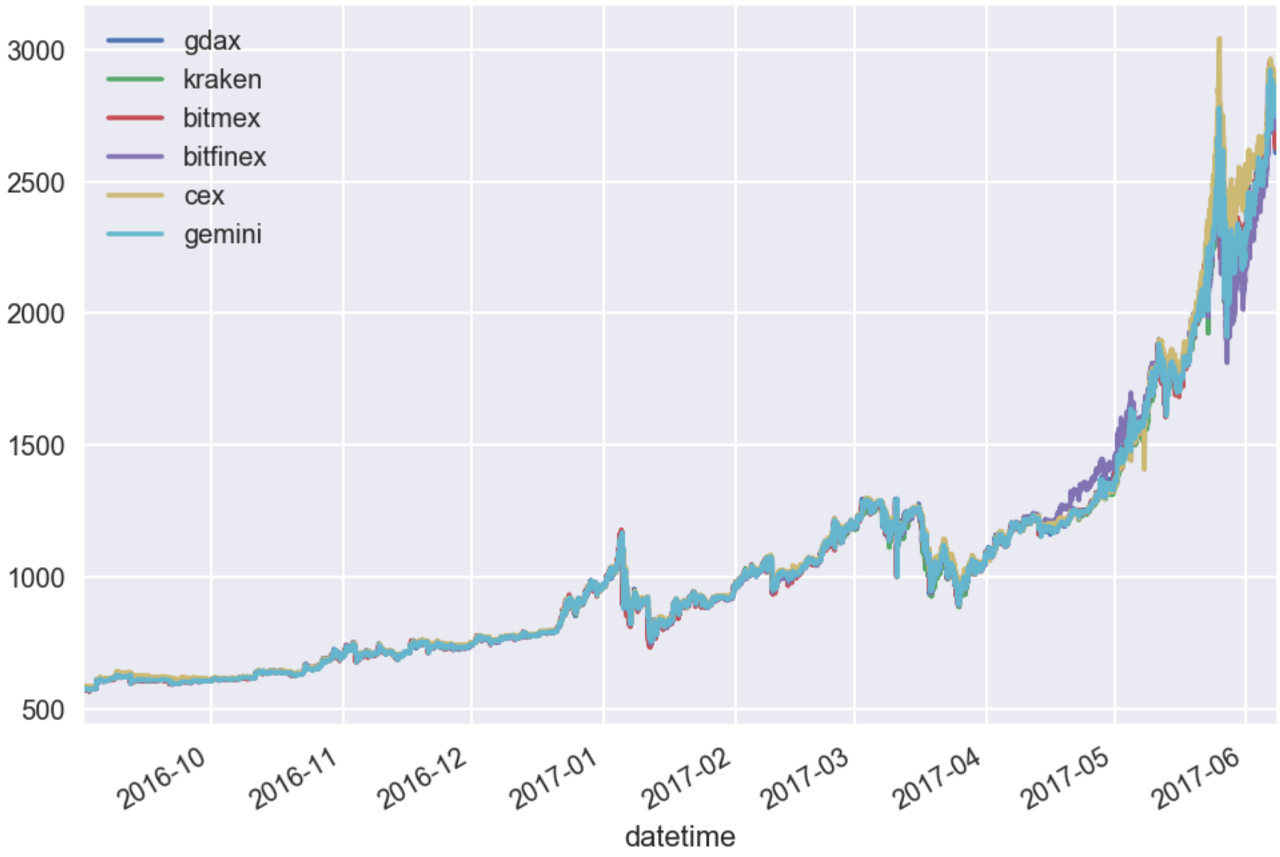

Figure 2: USD Price for a Bitcoin for the sampling period

To get some initial ideas about the 'leaders' and the 'followers' in this market, cross correlations for positive lags and for negative lags can be used. Let's define the 'market' as all the other exchanges. Then if the exchange correlates more with lagged 'market' returns than with concurrent 'market' returns, then the exchange is a follower. If the opposite is true, exchange is leading the 'market'.

However, Figure 3 below shows that no exchange is neither aleader nor a follower. However, all exchanges in this sample can be thought of as informative exchange since they show a symmetrical correlation with the market. This is due to the fact that if they are mostly moving together with the market, some of the price discovery is bound to happen there.

This quick check already shows a result different from Brandvold et al. (2015). In their sample they noted that Bitfinex's highest correlation was not with the concurrent market; Virtex was lagging behind, having almost the same correlation with past as with concurrent market movements; BTC China lagged behind as well; BTC-E and MtGox had higher correlation with future market moves than with past market moves.

No such observations can be made in my case. It seems like the Bitcoin market is become more equal, with noone being neither a clear leader nor a follower. However, it might also be that the time interval (five minutes) I chose is too big.

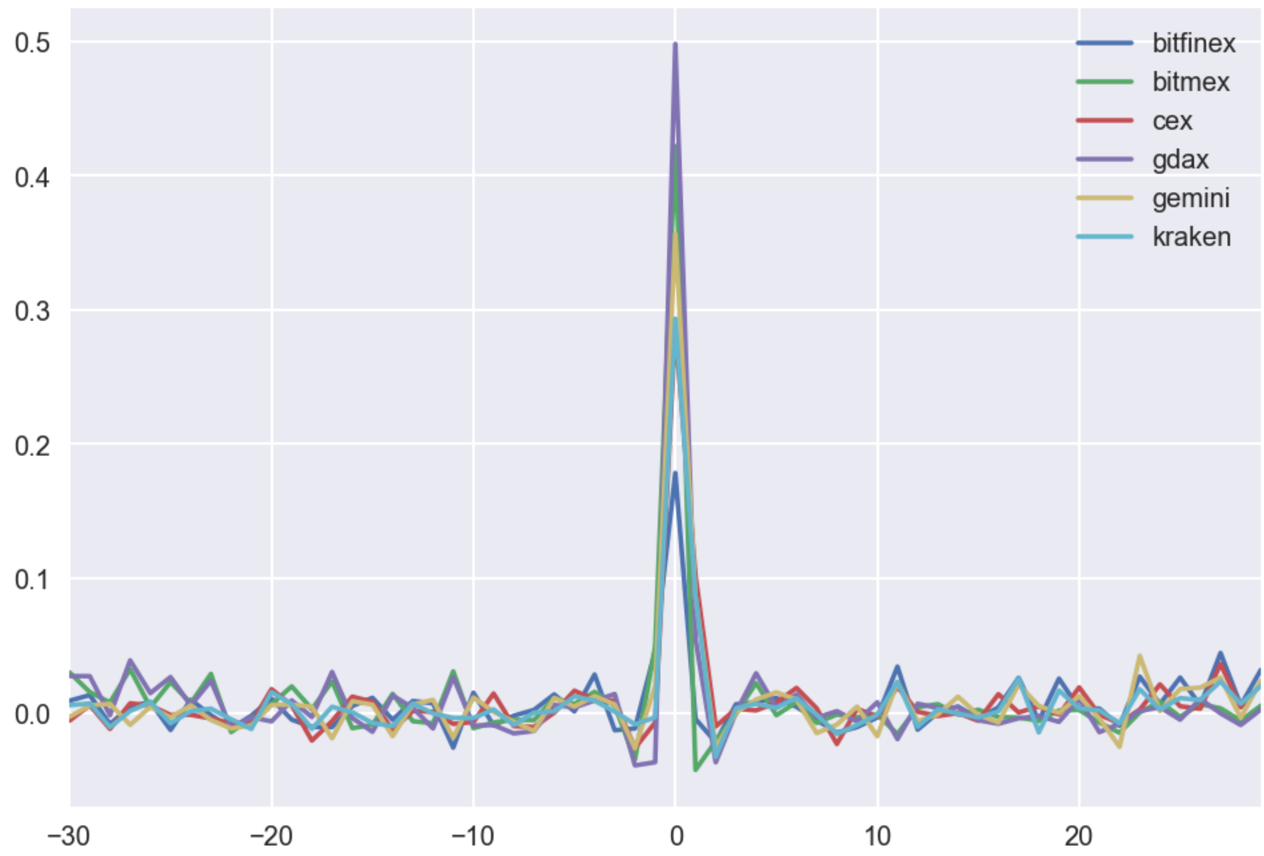

Figure 3: correlation with the market, 30 periods into the future, and 30 periods into the past

3 Model

The results from previous section are not enough to determine the relative contribution of each exchange to the price discovery process.

The price discovery literature uses primarily two methodologies, the information share method by Hasbrouck (1995) [4] and the permanent-transitory decomposition by Gonzalo and Granger (1995) [6]. In this paper I use the unobserved components price discovery model of de Jong et al. (2001) [5] as do Brandvold et al. (2015). According to de Jong et al. (2001) the advantage of this method is that the information share calculated this way is uniquely defined, unlike information share of Hasbrouck (1995), but still takes into account the variance of innovations, unlike Gonzalo and Granger (1995).

In this section I describe the multivariate time series model for price discovery proposed by de Jong et al. (2001), that allows for an assessment of the individual exchenge's information.

The main idea of the model is that the prices of all exchanges are derived from one common unobserved efficient price, such that

$$P_t = P_t^* \cdot U_t \tag{1}$$

where P_t is an n-vector of observed prices at time t, P_t^* – a vector of efficient prices, and

U_t – a vector of exchange-specific idiosyncratic components, that can be thought to reflect noise. The ith elements of P and U refer to exchange i.

Letting $p = \ln P$, $p^* = \ln P^*$, and $u = \ln U$, we get

$$p_t = \iota p_t^* + u_t \tag{2}$$

where $\iota = (1, ..., 1)$ is an n-vector of ones.

Further we proceed to develop a multivariate time series model for this price vector. p^* is assumed to be a random walk with serially uncorrelated increments. The noise term u is assumed to be transitory and takes account of all temporary deviations of the observed prices from the efficient price. Since the prices from all exchanges share the same random walk component, the series are cointegrated, which is intuitive, since the prices on each exchange are expected to revert to the same efficient price in the long run.

Let r_t be the change in the efficient price over the interval $(t-1, t)$, that is

$$r_t = p_t^* - p_{t-1}^* \tag{3}$$

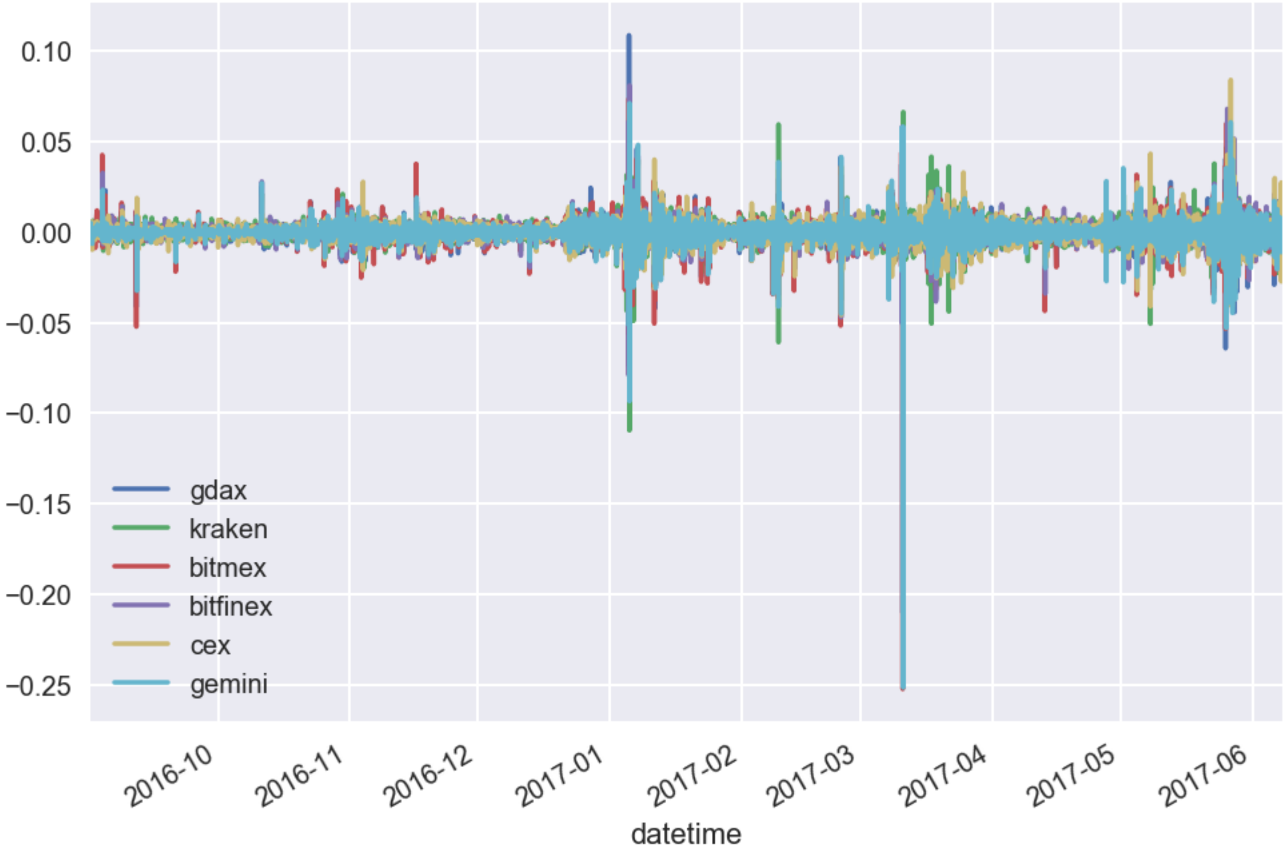

Figure 4: The log returns (observed prices) $p_t - p_{t-1}$

Assume that the unconditional serial covariances of r_t and u_t are stable in the given sampling interval. Further assume that

$$\begin{aligned}
\mathrm{E}[r_t^2] &= \sigma^2 \\
\mathrm{E}[r_t u_t] &= \psi \\
\mathrm{E}[r_t u_{t+l}] &= \gamma_l,\ l \geq 1 \\
\mathrm{E}[r_t u_{t-k}] &= 0,\ k > 0 \\
\mathrm{E}[u_t u_t'] &= \Omega \\
\mathrm{E}[u_t t_{t-k}'] &= 0,\ k \neq 0
\end{aligned} \tag{4}$$

where ψ and γ_l are $(n \times 1)$ and Ω is an $(n \times n)$ matrix.

These assumtions state that the fundamental news r_t is serially uncorrelated. The fundamental news r_t and the idiosyncratic component u_t are uncorrelated at all lags, but may be correlated at leads of r_t to future u_t. The idiosyncratic component u_t is serially uncorrelated. The model therefore has a random walk plus noise unobserved components structure. There is only contemporaneous correlation between the news and the noise.

For the serial covariance properties of the model we consider the vector of price changes Y_t with elements $y_{it} = p_{it} - p_{i,t-1}$, such that

$$Y_t = p_t - p_{t-1} = \iota r_t + u_t - u_{t-1} \tag{5}$$

Given the assumptions (4), the serial covariances of Y_t are

$$\begin{aligned}
\mathrm{E}[Y_t Y_t'] &= \sigma^2 \iota\iota' + \iota\psi' + \psi\iota' + 2\Omega \\
\mathrm{E}[Y_t Y_{t-1}'] &= -\psi\iota' - +\gamma\iota' \\
\mathrm{E}[Y_t Y_{t-2}'] &= -\gamma\iota'
\end{aligned} \tag{6}$$

The most interesting parameter in the model is ψ, the covariance between the fundamental price change and the idiosyncratic shocks. This parameter determines what can be learned from prices from individual exchanges. Consider the conditional covariance between observable and fundamental price changes for an arbitrarily selected exchange i

$$\mathrm{Cov}(p_{it} - p_{it-1}, r_t) = \sigma^2 + \psi_i,\ (i = 1, \ldots, n) \tag{7}$$

Since inference on the fundamental news component r_t given observed price changes is largely driven by this covariance, a higher ψ_i indicates a stronger signal that the price update on exchange i sends, this making it's prices more informative. However, on average, the information generated in each period should equal σ^2, the variance of r_t.

Since the n covariances between the price updates and the news is determined by $n+1$ parameters, an identifying restriction is needed. A natural one is found by considering the average covariance between the price update of an arbitrarily selected exchange and the news

$$\sum_{i=1}^{n} \pi_i \mathrm{Cov}(p_{it} - p_{it-1}, r_t) = \pi'\left(\sigma^2 \iota + \psi\right) = \sigma^2 + \pi'\psi,\ (i = 1, \ldots, n) \tag{8}$$

where π is a vector of weights adding to one. In order to give the natural interpretation of σ^2 as the unconditional covariance of a price update and the news, the restriction $\pi'\psi = 0$ has been imposed by the authors. This restriction is sufficient to identify the parameters σ^2 and ψ. This assumption also leads to a natural definition of the information share of an exchange if π_i is the fraction of trading activity that happened on this exchange, defined as volume and trade count, both given equal weights.

Consider the covariance between the fundamental price revision and the price change on exchange i in (7). Multiplying this covariance with the probability of the trade happening on exchange i gives a natural measure of how much information is generated by the price updates of exchange i. Dividing this by the variance of r_t, the total information generated in the market, the information share of exchange i is

$$\beta_i = \frac{(\sigma^2 + \psi_i)\pi_i}{\sigma^2} = \pi_i \left(1 + \frac{\psi_i}{\sigma^2}\right) \tag{9}$$

Given the restriction $\pi'\psi = 0$, the information shares should add up to 1. From this definition is follows, that, ceteris paribus, an exchange with more trading activity (volume + trade count) has a larger information share. And also that an exchange with a high contemporaneous covariance between its idiosyncratic term and the fundamental news (a high value of ψ_i) has a high information share.

4 Implementation

This section describes how the model from the previous section is implemented using the trade history data. The analysis proceeds in two steps. First, the auto and cross (with the "market") serial covariances from the fixed interval time series are estimated. That is, only the "own" serial covariances $\mathrm{E}[y_{it}y_{it-k}]$ and the covariances with the market $\mathrm{E}[y_{jt}y_{jt_k}]$, where j is the market corresponding to exchange i, both for $k \in [0,1,2]$, are estimated.

These covariances and the restrictions the previous section allow for identification of the parameters $\omega_i^e, \omega_{ij}, \psi_i, \psi_j, \gamma_i, \gamma_j$. σ^2 is estimated as the variance of r_t and interpreted as the total information generated in the whole market.

The resulting equations are

$$\begin{cases} \mathrm{E}[y_{it}^2] & = \sigma^2 + 2\psi_i + 2\omega_i^e \\ \mathrm{E}[y_{it}y_{it-1}] & = \gamma_{1i} - \psi_i - \omega_i^e \\ \mathrm{E}[y_{it}y_{it-2}] & = -\gamma_{1i} \\ \mathrm{E}[y_{jt}y_{it}] & = \sigma^2 + \psi_i + \psi_j + 2\omega_{ij} \\ \mathrm{E}[y_{jt}y_{it-1}] & = -\psi_j - \omega_{ij} + \gamma_{1j} \\ \mathrm{E}[y_{jt}y_{it-2}] & = -\gamma_{1j} \end{cases} \tag{10}$$

However, even though there were "enough" equations to identify these parameters, the for them matrix is degenerate. So a linear optimization technique has to be used to actually identify the parameters.

One point that has not been covered yet, is how π_i, which is is a measure exchange specific activity relative to the total market, is chosen. Since Brandvold et al. (2015) suggest that there is no significant difference in the choice of the weights given to trade count and volume during the computation of π_i, the weights are chosen to be equal.

5 Result

	π	ψ_i	info share
Bitfinex	0.271	$-1.965 \cdot 10^{-3}$	0.259
BitMEX	0.310	$-2.106 \cdot 10^{-3}$	0.296
Cex	0.036	$-1.106 \cdot 10^{-3}$	0.035
GDAX	0.250	$-1.863 \cdot 10^{-3}$	0.240
Gemini	0.066	$-1.729 \cdot 10^{-3}$	0.063
Kraken	0.067	$-1.813 \cdot 10^{-3}$	0.064

Figure 5: information share for each exchange

The result somewhat differs from Brandvold et al. (2015). They had two exchanges with positive ψ_i, BTC-E and MtGox, and the rest – negative. In my case, all exchanges have negative ψ_i. Even though some are more negative than others, it still cannot indicate that the information that comes from some exchanges is more informative than the information that comes from others.

However, mostly due to their market share, the information share for Bitfinex, BitMEX, and GDAX seems to be significantly larger than that of Cex, Gemini, and Kraken. However, unlike BTC-E and MtGox in Brandvold et al. (2015), no exchange has higher information share than activity share (due to negative ψ_i).

It means that, right now (or at least in my sample) the most information is generated/incorporated at Bitfinex, BitMEX, and GDAX.

6 Conclusion

In this paper I have investigated the role of different exchanges in the price discovery process of Bitcoin, following the procedures outlined in Brandvold et al. (2015). The information share of exchange measures the fraction of the price discovery which happens at this particular exchange.

Unsuprisingly, a higher fraction of the price discovery happens at exchanges with more trading activity.

References

[1] Brandvold, Molnar, Vagstad, Valstad (2015) *Price discovery on Bitcoin exchanges*. Journal of International Financial Markets, Institutions and Money, 36, 1835.

[2] Nakamoto (2008) *Bitcoin: A Peer-to-Peer Electronic Cash System*. `https://bitcoin.org/bitcoin.pdf`

[3] Harvey (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge Press.

[4] Hasbrouck (1995) *One security, many markets: determining the contributions to price discovery*. J. Finance 50, 11751199.

[5] de Jong, Mahieu, Schotman, van Leeuwen (2001) *Price Discovery on Foreign Exchange Markets with Differentially Informed Traders*. Tinbergen Institute Discussion Paper Series, No TI 99 032/2.

[6] Gonzalo, Granger (1995) *Estimation of common long-memory components in cointegrated systems*. J. Bus. Econ. Stat. 13 (1), 2735.