CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Pairs Trading in Cryptocurrency Markets

Bachelor thesis

Author: **Miroslav Fil**

Study program: **Economics and Finance**

Supervisor: **doc. PhDr. Ladislav Krištoufek Ph.D.**

Year of defense: **2019**

Bibliographic note

FIL, Miroslav. *Pairs trading in cryptocurrencies.* Prague 2019, 57 pages. Bachelor's thesis (Bc.), Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor - doc. PhDr. Ladislav Krištoufek Ph.D.

Abstract

Pairs trading is a trading strategy which tries to exploit mean-reversion among prices of certain securities. It is market-neutral and self-financing, and has been shown to produce high excess returns in historical backtests.

We employ the most common distance and cointegration approaches on cryptocurrency data from an exchange called Binance spanning the year 2018. The strategy is mostly unprofitable under transaction costs, but certain combinations of hyperparameters can perform well. Overall, the distance method performs far better, being able to achieve 3% monthly profit even in our baseline real-life conditions while the cointegration method always achieves only a slight loss. We also found that increasing the sampling frequency of the data from daily to hourly brings mixed results.

Moreover, since we have to reuse estimates of real-life considerations from equity markets, it is unclear if our results are truly representative of the cryptocurrency market. The strategy is found to be very sensitive to execution difficulties and transaction costs, making their determination crucially important. It is somewhat easy to get returns in excess of 5% monthly under ideal conditions, but whether this could be achieved in real trading conditions is still unclear.

Keywords

pairs trading, cointegration, statistical arbitrage, intra-day trading, cryptocurrencies, distance method

Abstrakt

Párové obchodování je investiční strategie využívající dlouhodobého ekvilibria v hodnotách cenných papírů. Navíc je tržně neutrální s nulovou čistou investicí, a zároveň historicky vykazuje velké zisky.

S pomocí kointegrační a vzdálenostní metody analyzujeme data z burzy jménem Binance za rok 2018. Ukazuje se, že strategie je díky transakčním poplatkům převážně neprofitabilní, ale jisté kombinace parametrů dosahují dobrých výsledků. Vzdálenostní metoda je obecně ziskovější a dosahuje až 3% měsíčního zisku i v naší základní simulaci reálných podmínek, zatímco kointegrační metoda je vždy mírně ztrátová. Navíc se ukazuje, že obchodování s hodinovými místo denními daty má smíšený efekt.

Naše výsledky mají do jisté míry omezenou výpovědní hodnotu, jelikož jsme spoustu fenoménů odhadovali podle ekvivalentních konceptů z amerických akciových trhů. Přitom jsme ukázali, že úspěch strategie je velice citlivý vůči transakčním poplatkům a potížím při exekuci obchodů, což znamená, že jejich přesné určení je kritické. Pokud bychom tyto faktory zanedbali, dosáhnout ziskovosti i nad 5% měsíčně by bylo lehké. Nelze však říci, že by naše výsledky byly robustní vůči podmínkám v reálném obchodování.

Klíčová slova

párové obchodování, kointegrace, statistická arbitráž, vnitrodenní obchodování, kryptoměny, vzdálenostní metoda

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 29, 2019

Signature

Acknowledgment

I would like to express my deepest gratitude to doc. PhDr. Ladislav Krištoufek Ph.D. for his advice and guidance during the writing of this thesis. I would also like to thank my family and friends, namely Tereza Tížková, Daniel Štancl and Marie Viktorinová, for their support.

Bachelor's thesis proposal

Author	Miroslav Fil
Supervisor	Ladislav Krištoufek
Proposed topic	Pairs trading in cryptocurrencies

Research question and motivation:

This thesis aims to examine the efficiency of pairs trading, a trading strategy mostly examined in traditional equity markets, in newly formed cryptocurrency markets. We research various implementations of the pairs trading strategy, all based around the idea of mean-reversion to long-term equilibria, and how they compare to benchmark strategy, both in and out of sample. The thesis should explain whether pairs trading is a viable strategy in cryptocurrency markets, using methods centered around statistical arbitrage and developed for traditional equity markets.

Contribution:

The thesis should assess the difference between trading in cryptocurrency markets and well established markets such as the NYSE. It should also examine the differences between approaches to modelling the pairs trading signals, as well as discuss the formation of suitable pairs in the rather unorthodox cryptocurrency market.

Methodology:

The pairs trading strategy is centered around long-term mean-reversion found in suitable pairs, which generates trading signals upon short-term divergences from the assumed equilibria. We conduct appropriate backtesting for pairs trading based on approaches such as the distance and cointegration methods. We collect data from selected highly liquid cryptocurrency exchanges and analyze it using statistical software using our implementations of the selected trading strategies.

Outline:

1. Introduction

2. Pairs trading

3. Data analysis

4. Results

5. Conclusion

References:

E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, “Pairs Trading: Performance of a Relative-Value Arbitrage Rule,” 2006.

R. J. Elliott, J. Van Der Hoek *, and W. P. Malcolm, “Pairs trading,” Quant. Financ., vol. 5, no. 3, pp. 271–276, Jun. 2005.

H. Rad, R. K. Y. Low, and R. Faff, “The profitability of pairs trading strategies: distance, cointegration and copula methods,” Quant. Financ., vol. 16, no. 10, pp. 1541–1558, Oct. 2016.

D. Bowen, M. C. Hutchinson, and N. O’Sullivan, “High Frequency Equity Pairs Trading: Transaction Costs, Speed of Execution and Patterns in Returns.” 01-Mar-2010.

P. S. Lintilhac and A. Tourin, “Model-based pairs trading in the bitcoin markets,” Quant. Financ., vol. 17, no. 5, pp. 703–716, May 2017.

C.-F. Huang, C.-J. Hsu, C.-C. Chen, B. R. Chang, and C.-A. Li, “An Intelligent Model for Pairs Trading Using Genetic Algorithms,” Comput. Intell. Neurosci., vol. 2015, pp. 1–10, Aug. 2015.

Contents

Acronyms

AIC Akaike information criterion.

API Application programming interface.

AR(p) Autoregressive process of order p.

BIC Bayesian information criterion.

BTC Bitcoin.

CAGR Compund annual growth rate.

CLT Central limit theorem.

DF Dickey-Fuller test.

KDE Kernel density estimator.

OLS Ordinary least squares.

Q-Q plot Quantile-quantile plot.

SE Standard error.

SSD Sum of squared deviations.

US United States.

VaR Value at Risk.

List of Figures

List of Tables

Introduction

Pairs trading is a mean reversion strategy that is widely believed to originate from Morgan Stanley in the 1980s, pioneered by Gerry Bamberger and Nunzio Tartaglia (Bookstaber, 2007).

The strategy aims to find the so-called pairs of securities which are assumed to have a common long-term relationship in some of their characteristics. Taking price as the quantity of interest, short-term deviations from the purported relationship where the securities become mispriced compared to what is expected in the long-run would lead a pairs trading strategy to short the overvalued and long the undervalued security. This should later lead to a realization of profit once the pair starts converging towards its long-term equilibrium. Pairs trading is thus a market-neutral and mean-reverting strategy.

The strategy proceeds in two stages. First, in the pair formation period, it is necessary to identify suitable pairs. One thus has to find a metric that chooses just the pairs that tend to move together with a stable long-run relationship. Second, a measure of spread from the long-run equilibrium is defined and a trading strategy is defined based on the values attained by the spread. Typically, the spread is constructed so that it should, in theory, oscillate around zero. Any extreme values attained are then a signal for opening a position, as it should be reverting back towards zero shortly after. Either stage of the strategy supports many different approaches that can be freely combined.

The most frequent approaches to pairs trading are based on the distance, cointegration or stochastic spread methods, although other methodologies have been studied. The most renowned analysis using the distance method was conducted by Gatev et al. (1999), whereas the cointegration method was introduced by Vidyamurthy (2004) based on the work of Engle and Granger (1987). Those two approaches are the most represented in academic literature, but even for those, evidence outside of the US is scarce.

This thesis will adapt those common approaches for use on the cryptoccurency market, which has, due to its novelty, so far not been studied anywhere near the extent of the US stock market. In short, cryptocurrencies are a form of electronic cash that have emerged in 2008. They promise several advantages over traditional currencies, including anonymity, speed of transactions and low fees. However, their adoption, while improving, is still suboptimal. They can be understood as a new asset class and can in fact be traded like many traditional securities for which pairs trading has already been studied.

The main contribution of this thesis is in investigating applicability of standard pairs trading approaches on highly nonstandard cryptocurrency data, with benchmarks set by the already existing papers on pairs trading, such as those by Gatev et al. (2006). The most popular methods are compared between each other and their overall viability on the cryptocurrency market is investigated.

The thesis also contributes by extending the intra-day pairs trading literature using high-frequency data, and it compares various trading strategy evaluation criteria, which have, again, been mostly studied with respect to traditional securities.

However, it is important to note that the purpose of this thesis is not necessarily to determine the best method for pairs trading on cryptocurrencies, especially when also considering selection of hyperparameters for each method. While optimizing them could improve the performance significantly, it is enough for us to simply demonstrate that pairs trading is a promising strategy for cryptocurrency trading.

This bachelor thesis is structured as follows. In Section 1, we provide a comprehensive literature review, covering both standard results in the field as well as recent development. The next Section treats the source of our data, as well as the preprocessing applied. In Section 3, a theoretical overview of both distance and cointegration methods is given. Then, Section 4 discusses the backtested trading performance of our strategy. Finally, the last section summarizes our findings and proposes further avenues for exploration.

1 Literature review

A certain mean-reverting behavior of stocks has been examined earlier than pairs trading. Fama and French (1988) found long-term negative autocorrelation in stock returns, able to predict 25-40% of 3 to 5-year return variance, which suggests either a market inefficiency or time-varying expected returns by rational investors. Those results are supported by the work of Poterba and Summers (1988), who found evidence for transitory components in stock prices that support mean-reverting behavior, suggesting a positive short-term and, again, negative long-term autocorrelation in stock price returns.

The possible autocorrelation and its implications for the efficiency of markets caused significant controversy. Kim, Nelson, et al. (1988) raise objections to statistical methodology used for the previous mean-reversion results and they argue that mean-reversion is a feature of pre-war stock markets, but is absent in the post-war period.

Counterevidence is further refined by Jegadeesh (1991), who supports the existence of mean-reversion even in the post-war period, but additionally discovers that the effects are entirely concentrated in the month of January, proposing seasonality in the pattern of mean-reversion. The special importance of January has seen much discussion in academic literature, for example by Bondt and Thaler (1987), who show how stock portfolios composed of past losers tend to outperform portfolios made of winners, with the excess returns materializing mostly in January. This phenomenon is attributed to market overreaction related to stock price changes.

Furthermore, it is shown that the effect is not a reiteration of the size or risk characteristics of the winning and losing firms, although it is unclear if time-varying discount rates (as proposed by Fama and French (1988)) do not play an important role.

If there indeed were mean-reverting forces on the stock market, a simple contrarian strategy should be able to perform well. Jegadeesh (1990) finds that monthly stock returns exhibit significant negative first-order correlation and a significant positive higher-order correlation. Over an initial period, a regression for stock returns is estimated, the stocks are sorted according to forecasted returns and ten portfolios are assembled from stocks with consecutively ranked expected returns. It is then found that the difference between the two most extreme portfolios (one with all the lowest expected return stocks and the other with all the highest) is 2.49% per month over a period as long as 1934-1987, meaning that stock prices do not really follow a random walk.

It might thus be that pairs trading is simply a case of contrarian investing, performance of which is unexpected as it contradicts the efficient market hypothesis. This question among others is answered in one of the first pairs trading papers with empirical evidence written by Gatev et al. (1999), who investigate the US stock market on data from 1962 to 1997.

They use the minimized normalized distance approach to identify potential pairs and trade them when the spread exceeds two standard normal deviations. Such a strategy turns out to be highly profitable, even after adjustment for transaction costs to correct inaccuracies caused by bid-ask bounce examined by Jegadeesh (1990).

Importantly, a bootstrap test is done, showing that just purely random pairs are very unlikely to generate profit, proving that pairs trading exploits more than just the aforementioned mean-reversion. The profitability of the strategy is hypothesized to be caused by the Law of One Price, as pairs trading might be understood as a relative mispricing of two close substitutes.

In a later revision of the paper (Gatev et al., 2006), the examination period is extended by 5 years and decreasing raw returns (but with consistent risk-adjusted returns) in more recent periods are found. Also, some common objections to the results are ruled out, such as short-sale constraints or unrealized bankruptcy risk. Those are important because the excess returns are in fact asymmetric in origin, being mostly generated by the short portfolio (made of stocks that increased in value relative to their counterparts prior to opening of the pair). Two explanations for the decreasing returns are offered, either they are caused by increased hedge fund competition, or by the existence of a common (but unknown) risk factor that drives pairs trading profits but that also decreased in significance in the later periods.

The decreasing returns have been further documented by Do and Faff (2010), who replicated the methodology of Gatev et al. (2006), again on data from the US stock market extended until 2010, and confirmed the decreasing but still positive returns to pairs trading.

A large-scale study of the US and 34 other stock markets conducted by Jacobs and Weber (2015) shows that the distance method is persistently profitable, but with a significant time-varying component for which two main causes are identified, time-varying arbitrage constraints and investor attention. Of the 35 markets studied, pairs trading was the most profitable in both emerging markets (where significant arbitrage constraints are more likely) and in markets with a large number of eligible pairs (where there is information overload).

Published detailed evidence for markets other outside the US is limited. Perlin (2009) has applied the distance method to the Brazilian stock market and again found that it significantly outperformed a naive buy-and-hold benchmark. Similar conclusions were reached by Broussard and Vaihekoski (2012) who also used the distance method to investigate the Finnish stock market where multiple share classes are common, and pairs formed by those generated particularly high profit.

While the model proposed in Gatev et al. (1999) was nonparametric and based on simple statistical relationships, several other approaches to pair formation and/or generating trading signals have been suggested. Elliott et al. (2005) proposed to model the mean-reversion as an Ornstein-Uhlenbeck process, which became known as the stochastic spread method, and Do, Faff, and Hamza (2006) introduced a variation of the aforementioned referred to as the stochastic differential residual, whereas Vidyamurthy (2004) outlines a cointegration-based approach.

Liew and Wu (2013) note that methods based on correlation or cointegration force linearity on the pair's underlying dependence structure, which fails to appropriately describe the relationship in cases such as tail dependence. They thus used copulas to model trading rules and benchmarked it against the traditional distance/cointegration methods. Their empirical results on a specific pair of stocks considered in the period 2009-2012 show that the copula method achieves the best results.

Various other more complicated methods have been proposed, able to outperform the classical approaches at least in small samples. Most recently, advances in computing have led to applications of methods such as neural networks or genetic algorithms with promising results, for example by Huang et al. (2015) or Huck (2010). Hurst exponents were used by Ramos-Requena et al. (2017) to generate returns superior to the distance and cointegration methods using the Dow Jones index as a benchmark through 2000 to 2015. Bogomolov (2013) proposes a new nonparametric approach based on renko and kagi constructions, which originate from Japanese charting indicators and that trade on the volatility of the spread process, rather than its mean. It is shown to be theoretically profitable for the Ornstein-Uhlenbeck process and it is also backtested on multiple datasets from Australia and the US.

The most standard models were occasionally directly compared in uniform setting (though many studies describing novel approaches also use classical methods as benchmarks). The breadth of both documented approaches and used datasets means that comparability of methods is far from rigorously researched.

Rad et al. (2015) examine the distance, cointegration and copula methods on US stocks from 1962 to 2014. Returns (on both raw and risk-adjusted basis) are found to be very similar among distance and cointegration, with the copula method lagging behind. However, all methods except copulas exhibit decreasing trading opportunities over time. Also, while copulas have comparable performance to the other two methods in converged pairs, its relatively high proportion of diverging pairs severely limit its performance. Whether the copula method could be adjusted to negate its downsides while maintaining its distinguishing characteristics from the other methods is left for further research.

Additionally, Carrasco Blázquez et al. (2018) compare the correlation, cointegration, distance, stochastic and stochastic differential residue methods on stocks belonging to the US financial sector from 2008-2013, with the reasoning that a market-neutral strategy such as pairs trading should be able to perform both in and outside of crises.

Studies applying trading strategies (not just pairs trading in particular) on cryptoccurency markets are very limited in general. Lintilhac and Tourin (2017) have, using stochastic control theory to find the optimal trading rules, constructed 4 profitable pairs trading strategies trading BTC-USD across different exchanges, but the empirical results only include comparison amongst the derived strategies. Nakano et al. (2018) use artificial neural networks for prediction of Bitcoin price direction. Depending on the magnitude of predicted returns, predictions are classified into weakly/strongly upward/downward moving and three strategies are constructed. Sensitivity analysis of various hyperparameters of the neural network reveals that it is fairly robust with regards to model specification and the model generally significantly outperforms a buy-and-hold benchmark even after accounting for transaction costs.

Most relevant research focuses on trading issues which are either more fundamental or specific to the cryptocurrency space. For example, Feng et al. (2018) study informed trading in Bitcoin markets. Already existing metrics of informed trading are shown to be inappropriate for Bitcoin markets due to its nature as an order-driven, highly volatile market with nonstationary trading volume, and no samples of reported informed trading being available. A new metric based on the size of buyer-initiated (seller-initiated) orders is proposed and correlation to large positive (negative) events is observed. Profits of informed traders are also roughly estimated to be in the hundreds of thousands USD per event on the Bitstamp exchange alone.

Balcilar et al. (2017) try to estimate the volume-return relationship that has seen much attention for traditional assets. Since non-linearity and structural breaks are detected, they propose a novel nonparametric causality-in-quantiles test that deals with the aforementioned issues. Results show that volume can predict returns while the market is in ”normal” mode (that is, outside of major bear/bull runs), but fails to predict volatility in either case.

The macroeconomic drivers of Bitcoin prices were discussed by Ciaian et al. (2016). Standard supply and demand factors known from ordinary currency price formation are able to explain Bitcoin prices to a large extent, particularly as it became more established compared to earlier periods. On the contrary, the impact of social awareness (measured by metrics such as views on Wikipedia) was more prominent in the early periods and the effects are particularly pronounced in short-run speculative trading leading to bubbles. Major financial indicators such as the Dow Jones Index or oil price are found to be important only in the short run.

2 Data

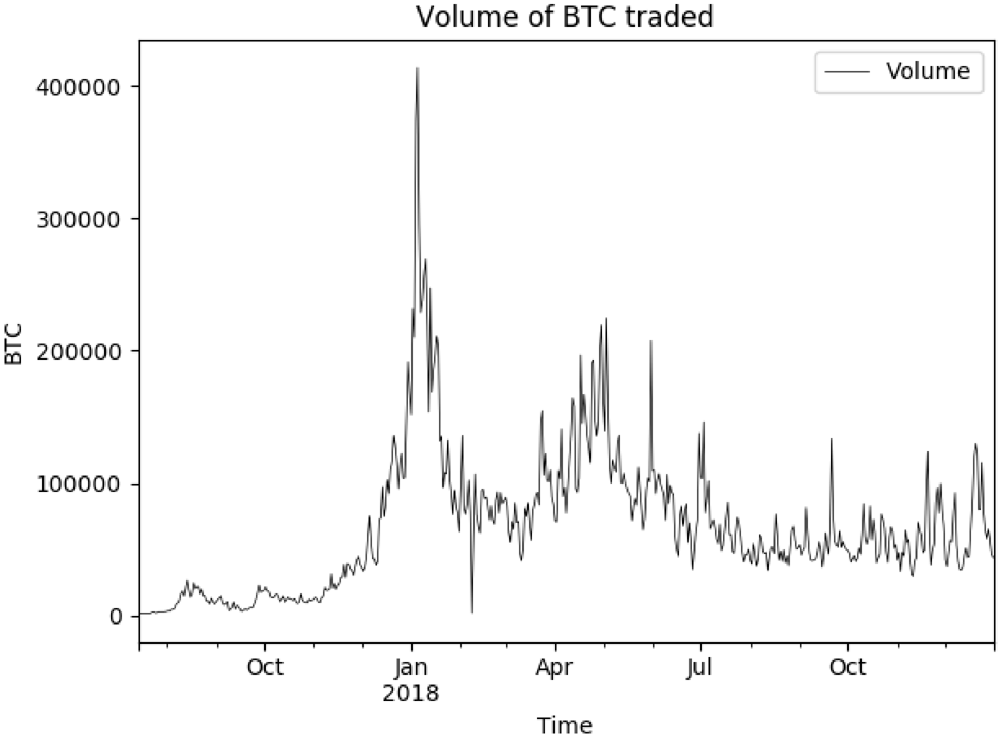
We use historical data from Binance, a global cryptocurrency exchange founded in 2017 that quickly rose to prominence and was one of the biggest exchanges by early 2018. Through its public API, it provides historical trade data at up to 1-minute resolution on all its traded pairs. The rapid increase in BTC trading volume on Binance is displayed in Figure 1.

From the API, we pull data on Open, High, Low, Close and Volume of each traded cryptocurrency for the year 2018. Because Binance does not deal with fiat, all currencies are denominated in Bitcoins with the exception of Ethereum and Bitcoin which also have pairs quoted in terms of Tether (a cryptoccurency designed to maintain 1:1 parity with USD). The full list of traded pairs can be found in the Appendix, Table 9.

We will restrict ourselves to cryptocurrencies that have data spanning the whole year 2018, and we will in fact only use data from 2018/01/01 to 2018/12/31 in our whole analysis.

New coins are added to Binance throughout the whole year, making the

Figure 1: BTC volume traded on Binance



panel data unbalanced. This is solved by simply removing the offenders to avoid distortion in our analysis. Likewise, we also remove the the currencies that got delisted at some point during 2018.

Moreover, our focus is only on the upper 30% percentile in terms of volume. We primarily hope this helps to minimize the magnitude of problems related to liquidity.

Nonetheless, even with the restrictions described above, our dataset is still of significant size. After applying both the preprocessing and volume cutoff, there are 23 cryptocurrencies left as our final set.

While this number is seemingly low compared to, say, the number of traded US equities, we actually have a bigger amount of data than studies using the US stock market since ours is sampled at 1-min frequency compared to the more common daily frequency. However, for the empirical part, only data upsampled to either hourly or daily frequency is used due to computational constraints.

To conclude, we will touch up on why we do not try to trade across multiple exchanges, and instead limit ourselves to just one.

The crypto space is actually well-suited to this, particularly if we avoided fiat altogether, since the strengths of cryptocurrencies, such as transaction speed and low fees, tend to reduce the trading barriers between exchanges, and arbitrage seemingly has fewer obstacles than in traditional securities. The crypto markets are in this sense unusually global, and we do not really get to see anything akin to things like national stock markets.

However, implementation of such a strategy would also require things like capital management across exchanges and treatment of different environments on each exchange, adding another layer of difficulty for implementation.

Makarov and Schoar (2018) study this sort of inter-exchange arbitrage and find that the profits are significant, with price spreads across exchanges much smaller when fiat is not involved. So while such a strategy appears viable, researching it would perhaps be more valuable from the perspective of arbitrage rather than pairs trading, as the the trading strategy itself would remain pretty much the same and trying to use this approach would add little value to this thesis while risking to devalue our results due to likely being forced to gloss over other complicated issues that might arise due to increased complexity.

3 Methodology

A pairs trading strategy is best understood in terms of two stages: pairs formation and trading period. Each stage may be executed independently of each other. The distance and cointegration methods are in fact fairly similar. It will soon be seen that the most common implementations differ only in the pairs formation stage. In the following text, we will try to stress parts of the process that are directly equivalent between each method, as well as point out the few present differences.

The pairs trading procedure can be summarized as follows. First, a distance metric is defined to detect pairs that are suitable for pairs trading. Optionally, the dataset might first be pre-filtered, primarily to help prevent false positives. Once suitable pairs are identified, a measure of spread is defined that is thought to have predictable long-run properties. Positions are then opened and closed based on the spread deviations from the long-run equilibrium.

At most steps of the process just described, some parameters have to be determined. For the most part, we will try to unify the parameters of choice with other literature as much as possible for better comparability. Since our motivation is not to find the best possible hyperparameters for optimal performance, this is not a big issue. Most literature does not provide justification for its choice of parameters either, and the original papers determine them arbitrarily as well.

To better see why the methods have so much in common, consider that the initial motivation behind pairs trading is finding pairs that ”move together” and have a ”long-run equilibrium”. Therefore if everything worked perfectly, both methods should identify the same viable pairs which would be traded the same, irregardless of how they were selected.

We shall now move on to details. The distance method is conceptually simpler compared to the cointegration method, so we will cover it first.

3.1 Distance method

In order to describe the distance method, we follow the methodology from Gatev et al. (2006). First, we decide on the backtesting period, which is split into pairs formation and trading periods. The pairs formation comprises of the following steps:

1. For each time-series, per-period returns are approximated using logarithms:

$$r_{it} = log(\frac{P_{it}}{P_{i,t-1}}). \tag{1}$$

Equivalently, we can say that our returns will be calculated from log prices.

2. The returns over the whole backtesting period (meaning both formation and trading periods) are then normalized by subtracting the sample mean and dividing by sample deviation calculated from the formation period only to avoid look-ahead bias. If we let superscript F denote the formation period and N the normalized variable, the formula is:
$$\hat{r}_{it}^{N} = \frac{r_{it} - \hat{\mu}_i^F}{\hat{\sigma}_i^F}. \tag{2}$$

3. Normalized price series is the normalized price index, set to start at one and to evolve additively using the returns from step 2, yielding the series $R_{\cdot t}$. An exhaustive pair-wise search across those series is then performed to find the pairs that minimize the sum of squared deviations (SSD) between their normalized price indexes and a portfolio is constructed from equally weighted top 20 pairs with the lowest distance measure. The SSD can be written as:
$$SSD_{ij} = \sum_t (R_{it} - R_{jt})^2 \qquad i \neq j. \tag{3}$$

Next, we take the spread between two stocks to be the difference between their cumulatively summed normalized returns, meaning:

$$spread_{ijt} = R_{it} - R_{jt}.$$

The spread then gets normalized. For the normalization step, we must again make sure to only calculate the mean and standard deviation from the formation period.

We open and close the pair position when the normalized spread crosses a predetermined threshold. Gatev et al. (2006) uses two standard historical deviations for opening the pair and closes the pair when the prices next equal, equivalent to the spread crossing zero. We do the same, although it is set arbitrarily. We then go long on the lower priced stock and short on the higher priced stock by equal amounts. At the end of the trading period, the position is liquidated regardless of convergence.

In order to calculate excess returns, we follow Broussard and Vaihekoski (2012), who use a slightly modified, but equivalent version easier to interpret than the one originally proposed by Gatev et al. (2006). Total return of the pair for each period can be obtained by combining returns for the long and short position as

$$r_{ijt} = w_{it} r_{it}^{L} - w_{jt} r_{jt}^{S}, \tag{4}$$

where $w_{\cdot t}$ stands for relative weighting with respect to the initial investment. The weights for subsequent periods can be calculated as

$$w_{it} = w_{i,t-1}(1 + r_{i,t}).$$

Inbetween threshold crossings, this setup corresponds to a buy-and-hold strategy. While it is possible to readjust the portfolio so that the weights are equal in each period, it would cause prohibitive transaction costs and it is typically avoided in literature. Instead, our cashflows are realized only upon threshold crossings or at the end of the trading period.

The opening of a pair is self-financing with zero net investment since we go both long and short an equal amount, typically 1$, which can better be seen as 1 unit of capital. For the purpose of calculating returns, it is assumed that we deployed 1$ of capital for the long position and the short position is bought on margin. Conveniently, having initial unit weights allows us to interpret the changes in weights as percentage profit.

Importantly, such a naive structure is only approximately market-neutral. While the capital deployed on both securities is equal, we do not know their respective market correlation betas that would determine truly market-neutral weighing of capital to be deployed. However, accounting for this is an uncommon practice in related literature and for the sake of comparison, we will follow the same schema applied in other studies. Moreover, even if we did setup the initial weights to achieve market-neutrality, it would be gone after the first period unless we adjusted the portfolio every period, which we already explained is not feasible.

On the closing of pairs, positive cashflows are realized and at the of the trading period when all positions are necessarily liquidated, we get either positive or negative cashflow depending on whether the pair has converged or stayed divergent. It is worth mentioning that even the convergent pairs could have negative profit due to transaction costs. This might happen if the standard deviation of the spread is low so that the gain from pair convergence is lower than transaction costs incurred.

There is one more special issue we need to take care of. Since pairs trading is a contrarian strategy, it might be subject to bid-ask bounce (documented for example by Jegadeesh (1990)). Our strategy sells stocks that have done well and buys those that did not, and since we only observe prices of orders filled at either a bid or ask quote, we need to account for the tendency that the winner's price is more likely to be an ask quote and the loser's price a bid quote. Otherwise, we might be implicitly buying losers at bid quotes and vice versa, biasing our returns.

Following Gatev et al. (2006), we rectify this by implementing a one-period waiting rule for the opening/closing of pairs in order to circumvent execution difficulties. The resulting change in profits can be taken as an approximation of transaction costs arising from the bid-ask spread.

3.2 Cointegration method

We will first clarify some key terminology in time series theory following canonical definitions. We need to introduce stationarity, in particular with respect to integration and cointegration. We also describe the Dickey-Fuller test used for deciding the stationarity of a time series.

Following Wooldridge (2008), a stochastic process $\{x_t : t = 1, 2, ..\}$ is *stationary* if for every collection of time indices $1 \leq t_1 \leq t_2 \leq ... \leq t_n$, the joint distribution of $(x_{t_1}, x_{t_2}, .., x_{t_n})$ is the same as of $(x_{t_1+h}, x_{t_2+h}, .., x_{t_n+h})$ for every integer $h \geq 1$. We also recognize *covariance stationarity*, for which it is necessary that the expected value and variance are constant (and finite) across time and the covariance $Cov(x_t, x_{t+h})$ depends only on h. Covariance stationarity is largely sufficient for our desired properties of estimation. A process that is not stationary is called *non-stationary*.

Using OLS on non-stationary time-series is dangerous and known to produce spurious regressions (Granger and Newbold, 1974). Such regressions are devoid of econometric meaning, even though they might have good values of fit such as R^2 and coefficient significance. However, various estimators can be designed that deal with non-stationarity and allow meaningful inference, for example the Fixed effects or First difference estimators. Those include transformations of a single non-stationary series to make it stationary.

A special feature of the generating process called *cointegration*, which exploits a certain similarity of the process to another one, can reduce the problem to estimation of stationary series and is introduced next.

In order to develop this concept, we build on the work of Engle and Granger (1987) to establish a methodology for testing cointegration between two non-stationary time-series, which will allow us to construct a new stationary time-series from the original two.

We will say a process is *integrated of order d* if the d-th difference is covariance stationary. Then Engle and Granger (1987) define cointegration as follows:

> The components of the vector x_t are said to be co-integrated of order d, b, denoted $x_t \sim CI(d, b)$, if all components of x_t are $I(d)$ and there exists a vector α ($\neq 0$) so that $z_t = \alpha^T x_t \sim I(d - b), b > 0$. The vector α is called the co-integrating vector.

We will focus our attention only on the $CI(1, 1)$ with a two dimensional vector x_t case from now on. We can thus interpret cointegration as the requirement that there exists a a linear combination such that $\alpha_1 x_{1t} + \alpha_2 x_{2t}$ is a stationary time series, and in that case, x_{1t} and x_{2t} are said to be cointegrated. We will explicitly standardize the linear combinations so that one of the coefficients is always equal to 1.

We notice that the definition above requires that each of the time series is I(1) by itself. Such series are said to have have a *unit root*. Unit-root processes do not exhibit mean-reversion after realization of the error term, which means that the error term at time t continues to affect all subsequent periods. Standard econometric methods only apply to time series without unit roots and it is necessary to remove them prior to estimation. A common approach to test for the presence of unit roots is the Dickey-Fuller test (for detailed treatment, see Dickey and Fuller (1979)).

As a side note, we remark that unit roots are just one of the possible sources of possible non-stationarity, which all tend to cause problems with standard econometric procedures. The possibilities are endless, but the other most commonly recognized forms are trend-stationarity, for which the series becomes stationary after detrending, and non-stationarity in variance. The individual forms can occur independently of each other.

The DF test is therefore not exactly a test for non-stationarity, but rather just for one of its forms. It is somewhat common not to draw this distinction, however.

Let us now look at specifics of the pairs trading procedure itself. The usual reference for the cointegration method is Vidyamurthy (2004), which we also follow. The Engle-Granger two step test (originally developed by Engle and Granger (1987)) is used to detect cointegrated pairs.

The test for two price series x_t and y_t can be summarized as follows:

1. **Establish individual non-stationarity** For two time series to be cointegrated, both of them must be I(1). We can assess this with the Augmented Dickey-Fuller test, which detects unit roots. If both series are I(1), we proceed to step 2.

2. **Establish cointegration** Run the OLS regression

$$y_t = \mu + \beta x_t + u_t, \tag{5}$$

and save the residuals $\hat{u}_t$. If the estimated residuals are stationary (which can be tested using methods identical to those applied in step 1), the two series are cointegrated. Care should be taken to choose the proper form of ADF test in both steps of the test.

We remark that all hypothesis testing is conducted at 5% significance level unless explicitly stated otherwise. That said, let us take a closer look at the two steps.

In place of the general x_t and y_t, we will be using log prices per suggestion of Vidyamurthy (2004), who proclaims that the assumption that logarithms of stock prices follow a random walk is a standard one. Indeed, log prices were used in the distance method as well.

Importantly, the distribution of the test statistic from the ADF test applied throughout the Engle-Granger test will not have the usual Dickey-Fuller distribution (which is already nonstandard) due to the necessity of estimating the cointegration coefficient β. This becomes an issue because y_t and x_t are generated from two different processes, whereas an ordinary Dickey-Fuller test uses only one series from a single process (see Engle and Granger (1987) for a more rigorous discussion).

Since $y_t - \hat{\beta}x_t = \hat{\mu} + \hat{u}_t$, if step 2 shows stationarity of the estimated residuals, we get that the time series $y_t - \hat{\beta}x_t$ is stationary (at least at our specified significance level). This is the cointegration method's *spread*, calculated as

$$spread_{xyt} = y_t - \hat{\beta}x_t.$$

As in the distance method, the spread is then normalized using data from the formation period.

If we did not use log prices during estimation in (5), the long-run equilibrium μ would not be invariant under percentage returns. If μ was non-zero, an α percentage increase in both securities X and Y would also increase the long-run mean by α percent. This however does not happen if log prices are used since then the estimated μ is essentially an approximate percentage difference between the two series, and taking logarithms thus makes the series more suitable for our estimation, as our spread will be invariant under percentage changes.

Proceeding to the trading period, we trade on oscillations of the spread around its equilibrium. The standard deviation threshold is most frequently set to 2, as that is what Gatev et al. (1999) used in the original paper on the distance method.

It is also important for the estimate of β in (5) to be positive, so that their prices “move together”. Otherwise, the pair could not be traded using a contrarian strategy as we would have to go either long or short on both pairs, whereas our initial motivation was to find a profitable trading strategy regardless of market conditions (that is, a *market neutral* strategy).

Vidyamurthy (2004) suggests calculation of profit in period t, which can be rearranged into change in spread, as follows below:

$$[log(p_t^L) - log(p_{t-1}^L)] - \beta[log(p_t^S) - log(p_{t-1}^S)] = spread_t - spread_{t-1}. \quad (6)$$

Note that this is explicitly the special case when we have 1 unit of the long pair and β units of the shorted pair. In the general case, the summands need to be weighted according to their current weights in period t. The calculated returns can then be summed.

Note that this formula is in no way inherent to pairs trading. It is nothing more than just the usual portfolio return calculation, and it should not be surprising that it is the same for both distance and cointegration methods.

By the design of our cointegration model, the initial weights assigned to the pairs are of the form $1 : \beta$. For the purpose of calculating returns, we will still assume 1$ of deployed capital, with the rest of necessary capital being borrowed.

Some issues, such as the bid ask spread, were discussed along with the distance method and of course apply here as well. There are a couple more problems concerning the cointegration method specifically that warrant more thorough treatment.

The first issue to be considered is the multiple comparisons problem since we are bound to conduct a high number of tests. Just by test construction, we are expected to falsely reject the true null hypothesis of no cointegration in a certain amount of tests depending on our desired significance level (i.e. we get *Type 1 errors*).

To illustrate the issue, consider that the Engle-Granger test is conducted pairwise. In Section 2, we saw that our final dataset consists of 23 stocks. In general, the testing procedure will be conducted number of times equal to

$$\frac{n(n-1)}{2}.$$

In our case with $n = 23$, this amounts to 253 procedures. At the 5% significance level, we thus expect to get approximately 13 pairs that are spuriously cointegrated.

There is one more issue related to this that was not yet mentioned. The Engle-Granger test is not symmetric with respect to the choice of x_t and y_t in the regression equation, and it is possible that we might get different results if we just permute the series.

This further accentuates our multiple comparisons problem. While there exist symmetric cointegration tests, we will keep using the Engle-Granger test as it is the most common choice in pairs trading literature, and treat the asymmetry by simply testing only one ordering out of the possible two. Not doing so would further double the number of tests, and since true cointegration is a symmetric relationship, getting different results based on ordering would just point towards faultiness of the chosen method anyways.

Gatev et al. (2006) tries to deal with false positives by considering stocks within the same sector. Similarly, Vidyamurthy (2004) proposes to choose stocks exposed to common risk factors based on Arbitrage pricing theory. Those methods are however hardly applicable to cryptocurrencies because they are weakly related to fundamental factors, and we cannot use qualitative analysis for pre-selection this way.

The second big issue is proper use of the Dickey-Fuller test. We have already noted its proper utilization is important. If we misuse it, the risk of spurious cointegration is increased which will in turn decrease the returns of our strategy. Therefore in order to properly execute the Engle-Granger test procedure, we need to understand the Dickey-Fuller test well.

The test initially considers OLS estimation of an AR(1) process

$$y_t = \rho y_{t-1} + u_t \tag{7}$$

This process has a unit root if $\rho = 1$, otherwise it is either stationary ($\rho < 1$) or exploding ($\rho > 1$). In order to test for the presence of unit root, we consider $\rho = 1$ as the null hypothesis against the alternative that $\rho < 1$ (stationarity), leading to a one-sided test. It is implicitly assumed that our price series can not be an exploding time series since economic data is unlikely to behave that way.

We can also rewrite the model by subtracting y_{t-1} from both sides as

$$\Delta y_t = (\rho - 1)y_{t-1} + u_t = \delta y_{t-1} + u_t \tag{8}$$

Testing $\rho = 1$ is then equivalent to testing $\delta = 0$. The test statistic is

$$DF = \frac{\hat{\delta}}{SE(\hat{\delta})}.$$

There are three main alternatives of the test, depending on the kind of unit root suspected. The difference lies in the form of (7) since it might also include a drift or a linear trend. Based on that, we distinguish three main forms of the Dickey-Fuller test.

1. Unit root only:

$$\Delta y_t = \delta y_{t-1} + u_t, \tag{9}$$

2. Unit root with drift:

$$\Delta y_t = \mu + \delta y_{t-1} + u_t, \tag{10}$$

3. Unit root with drift and deterministic time trend:

$$\Delta y_t = \mu + at + \delta y_{t-1} + u_t. \tag{11}$$

Care must be taken to choose the proper model, especially regarding interpretation of results since each specification has a different null hypothesis. For example, the second model considers random walk with drift as the null hypothesis process.

Improperly selected form distorts size and power of the test and prior knowledge (perhaps coupled with visual inspection) should be used to pick the proper form. Non-discretionary methods are beyond the scope of this text (see for example Enders (2014)).

Both steps of the Engle-Granger cointegration test make use of the DF test, but each case is different. Let us first discuss its application in Step 1 for determining individual non-stationarity.

When deriving (8), notice that we what we did was subtract y_{t-1} from both sides in (7). This results in a first difference on the left-hand side, but it is not first differencing.

This is important because it means that to determine the presence of drift and trend in the DF test formulation, it suffices to decide whether the original generating process of the series included them.

However, we are still in a dire situation, as the number of our cryptocurrency time series is high enough to make visual inspection unfeasible.

Furthermore, if we picked just one series, made a decision and applied it to all, we would be likely to misclassify the other series. Indeed, for each form of the DF test, it is not hard to imagine that it would be possible to find a series in our data that would appear to exhibit the desired properties.

Instead, we will rely on heuristics to decide the form. We are going to be using the DF test with a constant included for the following reasons. It is widely believed that stock prices have changing means, meaning at least one of drift or trend has to be included.

First of all, notice that adding drift in (8) in fact gives us a linear trend, since the left-hand side is differenced. Adding linear trend in (8) would actually give us a quadratic trend in the original series, since for there to be a linear trend in the difference, there has to be a quadratic one in levels.

Second, since we will be doing the testing with log prices, drift seems appropriate, since a linear trend (for which we need to include drift) in log prices means exponential trend in actual prices. If our assumptions are based on a stylized fact like "markets grow on average x% every year", which implies exponential growth, including just drift appears to be the correct choice.

Table 1: # of unit roots per test specification

	neither	μ	$\mu + t$	$\mu + t + t^2$
# of unit roots (n=23)	23	21	20	20

Luckily, it turns out that even if we choose the wrong form, it is not that big of a problem. If we try all aforementioned forms of the test, the results are not very different.

As can be seen in Table 1, in the worst case, we might inappropriately exclude about 10% of our data. In terms of sheer magnitude, the rest of our preprocessing has removed a much greater chunk of our original dataset. It remains however unclear if the pairs removed are not just the ones who would end up spuriously cointegrated.

For Step 2, the situation is a bit different. Importantly, there is a potential fallacy to be made. Since we desire to test OLS residuals coming from (5), we might reason that such residuals have zero mean by construction and the inclusion of neither drift nor trend is appropriate.

Whether the residuals actually have a zero mean or not depends on the inclusion of intercept/trend in (5). We should include the intercept/trend in either the cointegrating regression or the unit root test, but not both and the critical values with respect to inclusion of constant/trend should be chosen based on their presence in either the cointegrating regression or the DF test (Harris (1995)).

Another issue to keep track of is that the DF test statistic is not t-distributed and each type of the test has its own set of critical values. In fact, the distribution of the test statistic has to be computed using Monte Carlo simulation. The critical values depend on both the model selected and the sample size. For most practical use cases, it is necessary to interpolate critical values from a table, for example from MacKinnon (2010) or Fuller (1995).

Up until now, we assumed that the residuals u_t are homoskedastic and serially independent. We know from the Gauss-Markov linear model assumptions that if those conditions are not fulfilled, hypothesis testing is affected.

It turns out that heteroskedasticity does not tend to be a big problem. Kim and Schmidt (1993) find that in finite samples with GARCH errors, the DF tests tend to over-reject, but the problem is not very serious. Pantula (1988) demonstrates that under suitable conditions, heteroskedasticity makes no difference asymptotically. Nonetheless, a number of special tests was constructed that outperform the default Dickey-Fuller test at least in certain situations (Cavaliere and Taylor (2009), Ling et al. (2003)).

Serial correlation in residuals is conceptually simpler to deal with. It is particularly relevant because autocorrelation in stock returns is well documented (for example by Jegadeesh (1990)). An extension called the Augmented Dickey-Fuller test has been developed which includes a number of lags of the dependent variable in (8) to deal with autocorrelation in residuals, so that the resulting model is

$$\Delta y_t = \delta y_{t-1} + \sum_{i=1}^{n} \gamma_i \Delta y_{t-i} + u_t \tag{12}$$

Correctly choosing the lag structure is key to avoid further decreasing the already low power of the test, as well as avoid size distortions.

The amount of lags can be determined either heuristically or by optimizing metrics such as the Akaike Information Criterion, Bayesian Information Criterion or Hannan–Quinn information Criterion, which try to penalize the structure of model parameters with regards to its performance. Another frequently used rule of thumb is to add lags until the residuals appear to be white noise.

For our purposes, we will use AIC. Not only it is one of the most frequently used criteria, but in comparison with BIC, it has the added advantage of penalizing included lags far more weakly than BIC where the penalty scales with number of observations. The difference between other metrics is not so clear-cut.

Since we have a huge dataset (especially when using hourly data), BIC might thus be prone to underfitting. Because losing observations due to having more lags is not a big problem given the size of our dataset, we decide to rather include more than fewer lags as autocorrelation is likely to be a more serious issue.

However, neither approach is universally applicable and another non-standard information criterion with more desirable properties was developed by Ng and Perron (2001). While the asymptotic distribution of the DF test statistic is the same regardless the amount of lags, finite sample properties are significantly affected by the lag length as shown by Cheung and Lai (1995) along with tabulated values.

The choice of lag length is therefore both important and without an optimal solution. At the same time, it is clear that proper execution of the test is most crucial as it directly influences our chosen pairs and improper testing will make it more likely for us to trade spuriously cointegrated pairs.

The last issue we cover is that it might occur that our time series have some missing values (and it in fact does happen). In that case, unit root tests can not be applied, at least directly, since they rely on some sort of serial dependence between time series values and we risk getting misleading results.

Ryan and Giles (1998) have compared linear interpolation, shifting the series back to "close" the gap and filling the missing values with the last known value. Replacement with the last observed values before the gap was found to have the most desirable properties in a wide variety of situations, so we will proceed analogously in case we encounter missing data for any reason.

The Dickey-Fuller test is well studied in literature, being a common unit-root test of choice in many applications. That said, there are several alternatives for testing unit roots with different strong points, such as the Phillips-Perron, Kwiatkowski–Phillips–Schmidt–Shin or ADF-GLS test. However, their treatment is beyond the scope of this text and the previously discussed Dickey-Fuller test is the one most widely used in economic literature, which is why we will focus on it (and it is also the test of choice originally utilised by Engle and Granger (1987)). Of course, a better chosen testing procedure could potentially improve performance of the strategy.

3.3 Measure of returns

As mentioned before, the calculation of returns will be based on the assumption that we have used \$1 of capital to trade each pair. Since short-selling is essentially margin trading, it must be assumed that going long \$1 will be enough to borrow for the short position, meaning our required capital to open the position is just the \$1 everytime.

This is directly related to the amount of leverage we are allowed to use. The schema used by Gatev et al. (2006), which was also described above, relies on 2x leverage ratio (meaning that for every 1\$ of deployed capital, we are able to go 1\$ long and 1\$ short).

This is particularly important for the cointegration method, where the pairs ratio is not pre-determined and the necessary leverage ratio varies. However, we will assume that it is always possible to trade all pairs in their respective cointegration ratios with just 1\$ of deployed capital and subsequently use it in profit calculations. However, not all literature uses the same leverage. For an empirical study using 5x leverage, see Liu et al. (2017).

While we have already described how to calculate individual pair returns, we have yet to propose a measure of return for portfolios. Gatev et al. (2006) used two different measures - return on committed capital and on employed capital.

The idea behind committed capital is that some capital has to be set aside for pairs that are nominated for trading, but do not actually trade and thus generate no returns, whereas employed capital only considers pairs that were in fact traded.

We will assume that our usage of capital is sufficiently flexible and only calculate the employed capital portfolio return, which can be calculated as the average return over the desired period of n individual pairs that were traded:

$$R = \frac{\sum_{i=1}^{n} r_i}{n}. \tag{13}$$

This actually still underestimates the profit figures since a pair position is seldom open the whole period, but our portfolio return metric treats all profit equally regardless of over how long a period it was realized.

Apart from raw returns for each pair, risk-adjusted measures are also reported. Earlier, we have discussed the leverage used by our trading strategy. Clearly, leverage amplifies not only returns, but also volatility. While this is going amplify the magnitude of raw returns, it will cancel out in risk-adjusted measures of return.

First, we calculate the Sharpe ratio, by far the most popular metric for this purpose. It can be computed as

$$Sharpe = \frac{R_p - r_f}{\sigma_p}, \tag{14}$$

where R_p is our portfolio return, r_f is the risk free rate and σ_p is the standard deviation of portfolio returns. As our risk-free rate, we will use 2% p.a. As of June 2019, this is very close to the 10-year Treasury yield [1].

It is common practice to annualize daily or monthly Sharpe ratios using a formula such as

$$Sharpe_{\text{annual}} = Sharpe_{\text{in 1 period}} * \sqrt{(\text{\# of periods in a year})}. \tag{15}$$

This holds exactly only under restrictive assumptions (Lo (2003)), but we will use it in all cases. An important thing to keep track of is that a traditional market trading year has only 252 trading days, but crypto markets never close, giving us 365 trading days. Upsampling from any sort of Sharpe ratio is done analogically.

It is not immediately clear whether the Sharpe ratio is a good quality ranking. It has been shown to be inadequate under certain assumptions,

[1] The yields are published by the United States Department of the Treasury and can be found at https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield.

the consequence of which was development of other, more complex criteria (Farinelli et al., 2008). Some of its problems include not properly accounting for non-normal return distributions and considering both downward and upward volatility equally.

It is also important to consider to what extent the Sharpe ratio makes sense, particularly if we are looking to use it for comparison. It is known that high-frequency trading strategies have high Sharpe ratios (Baron et al., 2012), and double digit values are not uncommon.

Of course, in the context of low-frequency equity strategy returns, such figures are absurdly high. To intuitively see why high-frequency trading might not have a meaningful Sharpe ratio, let us consider an example. If we imagined a strategy that essentially does just pure arbitrage across two assets (in our setting, it might be the same cryptocurrency on two different exchanges), its Sharpe ratio would be infinite by design, as there is no risk involved.

The risk arbitrageurs face in practice is related to execution, since they can not actually sell and buy in an infinitely small amount of time. Here, a finite Sharpe ratio is essentially a deficiency of the trading procedure. In contrast, a long-run equity strategy would not have such high ambitions.

It follows that we do not get much value out of trying to compare our calculated Sharpe ratios to commonly cited thresholds of viability. In fact, it is best to be careful with applying the same evaluation metrics across different types of trading strategies, as was just seen. We will prefer to make use of the Sharpe ratio for ranking of possible scenarios rather than to focus on its absolute magnitude. We will soon see that even changing the trading frequency markedly changes the risk-adjusted results.

Some of the Sharpe ratio's fundamental deficiencies are supposed to be corrected by more modern formulas. Hence we also compute metrics based on lower partial moments or drawdown, namely the Sortino and Calmar ratios, as well as Excess return on Value at Risk. However, those and other measures of return were found to produce mostly equivalent rankings for all sorts of traditional assets by Eling (2008), though it is unknown if the same holds here.

Indeed, if it was true that the rankings are pretty much the same irrespective of the chosen metric, there would be no need to worry about having chosen the appropriate one since they would all be the same.

Let us shortly introduce the other, somewhat non-standard performance ratios.

1. **Calmar ratio** tries to measure risk by maximum drawdown. The ratio, when first devised by Young (1991), was meant to be calculated over the last 36 months. Our time series is not that long and we have to use the whole dataset. Using the entire history during computation actually gives the so called MAR ratio. Of course, given the length of our dataset, this distinction is moot.

   The Calmar ratio can be computed as

   $$Calmar = \frac{CAGR}{MaxDrawdown}, \tag{16}$$

   where *CAGR* is the compound annual growth rate and *MaxDrawdown* is the biggest peak-to-trough decline.

2. **Sortino ratio** aims to only penalize downside risk rather than both downside and upside volatility like the Sharpe ratio. First proposed by Sortino and Meer (1991), it can be calculated as

   $$Sortino = \frac{R_p - r_f}{\sigma_d}, \tag{17}$$

   which is the same as the Sharpe ratio except for the denominator, which is the standard deviation of negative returns only.

3. **Value at Risk** tries to estimate how much an investment might lose at a given confidence level in a set period. Its definition is not constructive, leaving the specifics of calculation to practicioners.

   Indeed, there are several common approaches to calculating Value at Risk. Some of the most popular are the variance-covariance method, which forces an assumption of normality on returns and proceeds analytically, and the historical method, which just examines the empirical distribution function.

The historical method has the advantage of being non-parametric, which is why we will use it. Assuming normality would be too restrictive, and we will further examine to what extent would the normal approximation be appropriate in Section 4.3.

Anyways, the historical method calculation goes along the lines of

$$VaR = position * \hat{Q}(c), \tag{18}$$

where *position* is our investment, Q is the empirical quantile function of returns and c is the confidence level (for example, 5%).

Then the Excess returns on VaR can be calculated easily:

$$Excess\ return\ on\ VaR = \frac{R_p - r_f}{VaR}, \tag{19}$$

where again R_p stands for realized portfolio return, r_f is the risk-free rate and VaR is as in (18).

To investigate the correlation between risk-adjusted measures, we will turn to Spearman's correlation coefficient, also called Spearman's ρ. This coefficient is analogous to the usual Pearson correlation coefficient. In fact, it is the same up to the exception that it uses ranks rather than true values of the data to calculate correlation.

This makes it effective for detecting not just relations that are linear, but more generally monotonic. It thus detects a much wider class of functional relationships between variables.

Hypothesis testing for ρ can be carried out with Fisher's z-transformation. It can be shown that

$$F(\rho) = arctanh(\rho)$$

has an approximately normal distribution with mean $F(\rho)$ and standard error $\frac{1}{\sqrt{n-3}}$.

From there on, hypothesis testing can be carried out as usual for the normal distribution. In particular, we can calculate z-score in order to obtain approximate confidence intervals. Those can then be converted back to confidence intervals for ρ by the inverse Fisher transformation. The result might lack accuracy for small samples with significantly non-normal distributions, however, for our needs, we will only make use of the theory outlined above.

3.4 Sensitivity analysis

The implementations of pairs trading based on and including Gatev et al. (2006) mostly do not attempt to find optimal hyperparameters for their strategies. We will however try a small search among variations of the previously discussed methods, which might be particularly important due to the unusual nature of our data.

Furthermore, Bowen et al. (2010) have documented high sensitivity of pairs trading profits to transaction costs and execution windows. Huck and Afawubo (2015) explicitly compare distance and cointegration methods on US stocks and discover that while the methodology first applied by Gatev et al. (2006) yields negative returns, other hyperparameter choices can, particularly when using cointegration, yield great profits. It thus appears that the conclusion of our analysis is contingent on being able to find the right design of our trading strategy.

That said, we stress that the point of this thesis is not to find the optimal hyperparameters for our implemented trading strategies, but rather to establish whether such a hyperparameter search is a reasonable thing to do. A proper optimization procedure would be far too demanding on computer resources and it would contribute little, if at all, to the value of the thesis. Since most research does not try to optimize the parameters, it follows that even the robustness with respect to parameter mis-specification is not well known.

Keeping this in mind, we will now describe how we are actually going to approach this issue in detail.

It is clear that virtually all parameters of the trading strategy can be varied in effort to obtain the optimal result. Additionally, there is no intrinsic reason for one set of hyperparameters to work not only across asset classes (for example, across US equity and cryptocurrencies), but also within each class (with different cryptocurrencies requiring different choices).

A basic list one might consider in trying to design an optimal trading strategy could be:

- **Spread threshold for pair opening**

  While the most common threshold for opening a position is at +-2 standard deviations of the spread, there is no reason to believe it might have optimal properties, and is perhaps most valuable for convenience. Indeed, according to Gatev et al. (2006), the value was chosen based on discussion with practitioners.

- **Length of training/testing period**

  In particular, a reasonable hypothesis to test might be that more immature markets work better with shorter training/testing windows because the trading strategy gets outdated quicker, perhaps since we expect more structural breaks, and thus more frequent retraining is needed. However, there is a delicate balance to strike between fitting to noise with short formation windows and inability to adapt with longer windows.

  Liu et al. (2017) calibrate its intra-day pairs trading model with a formation period ranging from 30 to 100 days, but with just a single day of trading period, which is highly successful. However, doing so might turn out to be excessively computationally intensive, provided we use a high number of tradable assets and/or high sampling frequency.

- **More complicated trading strategy**

  For example, it might be advantageous to implement a stop-loss on overly divergent pairs, both in terms of too high absolute values of the spread or perhaps too long a holding time before convergence. This might help treat the problem of spuriously cointegrated pairs, and allows us to cut losses quicker. Similarly, another possible adjustment would be not to automatically close pairs at the end of the trading period, perhaps conditioned on how close to the end was the position opened.

- **Various sampling frequency of prices**

  Structure of the market can get obscured by aggregation into bigger chunks. We could also be interested in seasonality with respect to best trading times of the day/month. Since crypto markets are global and never close, we could perhaps try to find interesting patterns based on the time of day, as different regions of the world are likely to be more active during their local daytime.

- **Further qualitative pre-selection of traded pairs**

  While cryptocurrencies lack the same fundamental properties as traditional securities, there is room for improvement in that regard. For example, it should be possible to design a system of sectors (akin to traditional industry sectors) to group up the cryptocurrencies and focus on finding suitable pairs within sectors. To illustrate the idea, it could prove reasonable to put together all cryptocurrencies that primarily focus on anonymity of transactions.

However, we will not concern ourselves with most of those and only discuss a few in the empirical part. The options are limitless, and the list above is hardly exhaustive.

Our first concern will be frequency of sampling. In fact, we have no information regarding intra-day behavior in cryptocurrency markets available and it remains to be seen how it affects our strategies. We will also alter the standard deviation trading trigger, as well as experiment with implementing stop-loss on overly divergent pairs. The modified trading rules must also be observed in tandem with things such as transaction costs.

3.5 Transaction costs and other practical considerations

Most literature only roughly proxies transaction costs, even though they are obviously vital to estimating viability of trading strategies. Their detailed treatment is rather hard and modelling them properly would be far outside the scope of this thesis. Instead, we will base our analysis of the costs magnitude on the work of Do and Faff (2012).

The three key aspects of transaction costs are commissions, market impact and short selling costs. Commissions are easy to find out since cryptocurrency exchanges have explicitly listed maker/taker fees which also typically go down as a function of trading volume. For example, Binance has a baseline of 10/10bps for maker/taker that goes down to 2/4bps at the highest trading volume level.

Quantifying market impact is much harder and optimal real-world execution of large trades is a complicated issue. Do and Faff (2012) estimated the market impact costs to be 20bps for US equities over a period centered around the year 2000. We will use the same estimate, but we recognize that it is likely inaccurate for our purposes. Nevertheless, it still provides a frame of reference and gives us an estimate that has some grounding in reality.

In total, this gives us a one-way estimate of transaction costs of at most 30bps, which gets applied twice in case the pair closes at some point (including at the end of the trading period).

We have yet to mention short selling costs. Those are particularly interesting because most cryptocurrency exchanges do not actually support short selling and if they do, it is usually limited to select few currencies. The data we work with comes from Binance which does not support margin trading whatsoever as of June 2019.

D'Avolio (2002) estimated US equities in 2000-2001 to have short loan fees mostly below 1% per annum. However, Kraken, one of the leading European exchanges dealing with fiat, charges a 1bp rollover fee every 4 hours. Other exchanges with margin support have comparable rates which are all enormous compared to costs in traditional securities, making serious use of short-selling difficult.

That said, what we just described is just scratching the surface. In the real world, less liquid securities would typically also have higher shorting costs, which is a problem we will ignore entirely. Similarly, it might be reasonable to model the transactions costs dynamically, perhaps making them dependent on the trading volume of each pair or some proxy of liquidity, as well as allowing to it to change over the course of time. Also, margin trading comes with the danger of getting margin called in case our investment loses too much value, which might prove to be dangerous as cryptocurrencies are assumed to be very volatile assets, and pairs trading implicitly relies on leverage.

3.6 Introduction to cryptocurrencies

We will now try to not only briefly introduce the concept of cryptocurrencies and related phenomena, but also outline some of the main challenges a real-world application of pairs trading on such assets has to face that are not present in traditional markets. Our exposition, particularly in the area of technical details, is rather brief, but a more comprehensive introduction can be found for example in Narayanan et al. (2016). To begin, we will talk about Bitcoin, the first ever and still the most influential cryptocurrency.

Bitcoin is a decentralized cryptocurrency protocol first described in an influential whitepaper by Nakamoto (2008). Its first implementation went online in 2009. Roughly speaking, it aims to be a form of electronic cash that is free of regulatory oversight (unlike traditional currencies issued by central banks) by relying on decentralized consensus instead. Transactions are verified through cryptographical means which utilize computational power of volunteers (called *miners*). In exchange for their processing power, miners get Bitcoin as a reward for maintaining the network. The total amount of Bitcoins is predetermined and mining becomes more difficult over time.

Bitcoin price is affected by multiple factors. Several of the most frequently claimed price drivers were examined by Kristoufek (2015), who found that that fundamental factors such as usage in trade or money supply are important, as well as interest in the form of search engine queries.

The usability of Bitcoin is limited due to multiple issues. There is a great deal of uncertainty regarding its legal status and it is far from universally accepted as a payment method, making the possibility of converting it to fiat currencies crucial. There are numerous other cryptocurrencies, such as Ethereum, Ripple or Litecoin, which have often popped up as responses to certain technological limitations of the Bitcoin protocol, for example transaction speed or anonymity. Other cryptocurrencies are typically called *altcoins*. In the same vein, any cryptocurrency in general is also commonly referred to as a *coin*.

There are specialized cryptocurrency exchanges which typically support trading between Bitcoin and altcoins, but not necessarily fiat since dealing with real-world money involves regulatory restrictions such as Know-Your-Customer or Anti-Money-Laundering laws. Exchanges greatly differ in qualities like number of tradable currencies, ease of deposits/withdrawals and overall trustworthiness. Importantly, short-selling is a scarcely supported feature frequently limited to only few cryptocurrencies.

Exchange quality was investigated by Moore and Christin (2013), who examined 40 Bitcoin exchanges established prior to 2013, 18 of which have closed, sometimes even wiping client funds, and he also showed that exchanges with more volume are less likely to close.

We have just described what amounts to barriers in free trade in the crypto space. This causes price differences among individual exchanges. Arbitrage is often limited by the technological and legal aspects discussed above, making complete price correction difficult.

Indeed, Makarov and Schoar (2018) find that in the period from Dec 2017 to Feb 2018, arbitrage profits were over 1$ billion. At the same time, arbitrage opportunities are primarily found across rather than within regions and since spreads in crypto-to-crypto trading are far smaller than in transactions involving fiat, it suggests that currency controls are one of the main drivers of this discrepancy.

Interestingly, a lot of Bitcoin issues preventing mainstream usage, such as unreliable regulatory framework or security difficulties, are caused by its decentralized nature. Furthermore, highly influential groups like the largest miners or developers of the core protocol hold great power in directing Bitcoin's development. For example, Eyal and Sirer (2018) shows how mining is not secure against colliding groups, who do not even need to command the conventionally assumed lower bound of 1/2 of total hashrate to make it feasible for them to take control over the network.

As a consequence of the aforementioned problems, the price of Bitcoin has seen great volatility over the course of its existence. To give an example, Bitcoin in Jan 2017 was valued at roughly \$1000 and it rose to around \$20 000 in December 2017, only to fall back to \$6000 in February 2018. The complete price history of BTC/USDT on Binance can be found in the Appendix, Figure 5, which confirms that the dollar value varies wildly. The total cryptocurrency market size evolved similarly, as Bitcoin tends to make up around 50% of the total market cap, and reached a peak of \$800M in Jan 2018.[2]

Technical problems are also much more common compared to traditional exchanges (Chohan, 2018). Those and other significant political, legal or technical risks are things we can not account for. This in turn is likely to lead to overestimated risk-adjusted measures of return.

To sum up, while we can try to apply traditional trading methods on new assets such as cryptocurrencies, the actual execution of short selling tends to be literally impossible and there is a much higher degree of risk compared to traditional markets.

[2] The market-wide statistics come from a free and widely used site https://coinmarketcap.com/ which aggregates trading statistics from most existing exchanges. No effort was made to verify these numbers.

4 Empirical results

As we touched upon earlier, two base scenarios will be considered, which we will refer to as Scenarios 1 and 2, that are designed to mimic the methodology applied by other studies.

The main difference between our two scenarios is the sampling frequency. Scenario 1 uses daily data whereas Scenario 2 uses hourly data. Intra-day pairs trading is only rarely explored in other literature. However, a recent study by Liu et al. (2017) applied pairs trading to oil companies stocks at 5-minute intervals and achieved a remarkable 188% annualized return, albeit with 5x leverage compared to our 2x. We thus hope to confirm that trading on finer timescales is a promising avenue to explore.

One complication we have to face is that our dataset is actually pretty small in the temporal dimension. Our dataset spans just the year 2018 and since Binance was founded in late 2017, there is simply not much more data to get.

This is aggravating because Gatev et al. (2006) uses a 12-month formation period followed by a 6-month trading period. That is obviously unreasonable for our data, meaning an adjustment is needed.

For daily data, we will instead use a 4-month formation followed by a 2-month trading period. Consecutive backtests will have the starting date shifted by a month, meaning there will be multiple portfolios ”running at once”.

For hourly data, we will further shorten the periods to 20-day formation followed by 10-day trading, with consecutive backtests starting 10 days apart.

We aim to hit a balance between the length of formation and training periods while also achieving certain robustness of results with respect to timing of the backtest. Fundamentally, it is like conducting cross-validation to limit the possibility of overfitting.

Of course, when using hourly data, this is somewhat easier to achieve. The daily dataset has 365 observations per cryptocurrency, while the hourly dataset has 8760.

Other important parameters were already discussed earlier and we will let them be common to both scenarios. In summary, transaction costs were estimated to 30bps. Execution lag is 1 period. Positions are opened when the normalized spread crosses 2 standard deviations.

A summary of both scenarions can be found in Table 2.

Table 2: Description of basic scenarios

Frequency	Formation	Trading	Jump	Tx cost	Exec. lag	Threshold
Daily	4 months	2 months	1 month	30bps	1 day	2 stds
Hourly	20 days	10 days	10 days	30bps	1 hour	2 stds

Furthermore, even though we report the annualized Sharpe ratio, it is good to keep in mind that as per our discussion earlier, its magnitude by itself is mostly meaningless, albeit we will still comment on it. We will soon see for ourselves that changing the frequency of trading introduces a disparity that prevents comparison.

4.1 Base scenarios

The results for our two base scenarios can be found in Table 3, which contains values averaged over all the backtest periods. From now on, we will refer to each scenario by the capitalized frequency and method, e.g. Daily Distance. Unless specified otherwise, all statistics are calculated over the trading period.

In some cases, we will recalculate a statistic to make it the same frequency across all scenarios for comparability, for example the number of trades. That said, it is time to focus on the results themselves.

The Distance method appears to be doing better in our sample overall. In the Daily case, the monthly profit is negative at -0.01%, and a pair trades on average only 0.469x per month, the lowest of all scenarios. Furthermore, only 22.3% of trades are round-trip. The average length of a position is 31.7

days (out of total 60 trading days). The performance is fairly poor risk-wise, with an annualized Sharpe ratio of 0.17 (meaning the strategy is excessively risky for the profit it generates) and maximum drawdown of 24.6%.

However, the Hourly Distance method is doing by far the best out of the four backtests. The profit is very high at 2.87%, and the Sharpe ratio is 2.6 while the number of trades is higher than before at 3.29. Also, 27.9% of all trades are now round-trip. Each position is open for about 5.5 out of 10 trading days. Interestingly, the percentage of winning trades is 0.7% lower than in the Daily case despite the heightened profit.

Table 3: Results of base scenarios

	Daily		Hourly	
	Distance	Cointegration	Distance	Cointegration
Monthly profit	-0.01%	-0.07%	2.87%	-1.09%
Annualized Sharpe	0.17	-0.78	2.6	-2.8
Monthly number of trades	0.469	0.526	3.29	4.38
Round-trip trades	22.32%	21.90%	27.90%	36.37%
Length of position (days)	31.7	30.3	5.5	5.0
% of winning trades	46.25%	45.35%	45.52%	48.40%
Max. drawdown	24.64%	22.65%	13.33%	16.19%
Nominated pairs	20.0	17.4	20.0	20.0
Traded pairs	81.88%	86.33%	82.00%	92.43%

The Cointegration cases fare somewhat worse in comparison. On Daily scale, the profit is negative at -0.07% monthly. The pairs trade on average 0.526x with 21.9% of round-trip trades, and the percentage of winning trades is the lowest out of all 4 scenarios at just 45.3%. Unsurprisingly, the Sharpe ratio is then also negative at -0.78.

Most interestingly, the behavior of Hourly Cointegration is quite specific. The loss is far greater, giving -1.09% monthly loss, but its other metrics are very positive in the sense that they are what we would hope to achieve.

It trades on average 4.38x per month, over 33% higher than the second highest and achieves an even more unusual 36.4% of round-trip trades. Furthermore, even though it has the highest percentage of winning trades at 48.4%, it is still very unprofitable. It therefore appears that success in those auxiliary metrics does not really translate to good returns.

In conclusion, we can say that the Hourly versions of our trading strategies are at least sometimes performant while also having a higher Sharpe Ratio and smaller drawdowns. The pairs converge much more often, suggesting a lower proportion of spurious cointegrations, but this is not necessarily reflected in the proportions of winning trades, nor in profit. Switching to Hourly data helped the Distance method a lot, but Hourly Cointegration is doing worse, meaning it is so far inconclusive with respect to whether it is a good idea or not. But since both Daily backtests had negative profit, it can still be considered an improvement.

Certain features of our results might be a bit tricky to interpret, so we will now focus on discussing those at length.

We notice that the average length of position is a bit misleading. It is always roughly half of the trading period in question, which is essentially the effect of most pairs only closing automatically at the end of trading. If we assume that pairs are equally likely to open during the whole trading period, it then follows that the average holding period should be the halved trading length.

Typically, only around 25% trades are round-trip. It would not be helpful to calculate the average amount of time a pair is open for just the converged pairs because their convergence times are likely to be on the lower end of the real distribution of those times, and any produced estimate would be biased downwards. We also are not able to tell whether a pair is taking long to converge simply because the process is slow or because it is effectively divergent.

Furthermore, it is not immediately clear why, despite the better auxiliary metrics of the Hourly Cointegration method, the profit lags behind. One possible explanation is that there are tighter and less volatile spreads among the pairs chosen by the Cointegration method. This causes more frequent trading, but also limits the potential upside and results in more transaction costs incurred. Even with convergence, the pairs can then fail to be profitable due to other costs incurred.

On top of that, it also means that while the potential upside is small, the downside is essentially unlimited. A single spuriously cointegrated pair can thus have negative profit great enough in magnitude to offset many converged pairs.

This demonstrates one more interesting aspect of pairs formation we did not consider previously. In the presence of transaction costs, our goal is not to just find pairs that are likely to converge back to long-run equilibrium. In addition, the pair spread also has to be sufficiently volatile for the potential upside to be big enough, in order to exceed costs of trading. Proper pairs formation procedure thus also has to be a function of transaction costs.

The transaction costs we are working with here are one-off and get realized when we open/close position. However, short-selling fees are charged per unit of time. If we were to take those into account, then we would also have to try to minimize the amount of time during which our position is open. It could be said that we have to estimate the expected profit of a pair from the get-go to see if its worth trading, and this condition is dependent on exogenous factors.

It is thus unwise to consider pairs formation rules in isolation, and they should instead be tailored to the expected transaction costs and other real-life considerations.

Lastly, we mention the average nominated and traded pairs. The Distance cases nominate top 20 pairs everytime by design, but Cointegration has a variable amount, although it is quite close to 20 on average anyways.

Cointegration has in general a higher percentage of traded pairs than Distance, which is what we would expect, given that it also has a higher number of monthly trades.

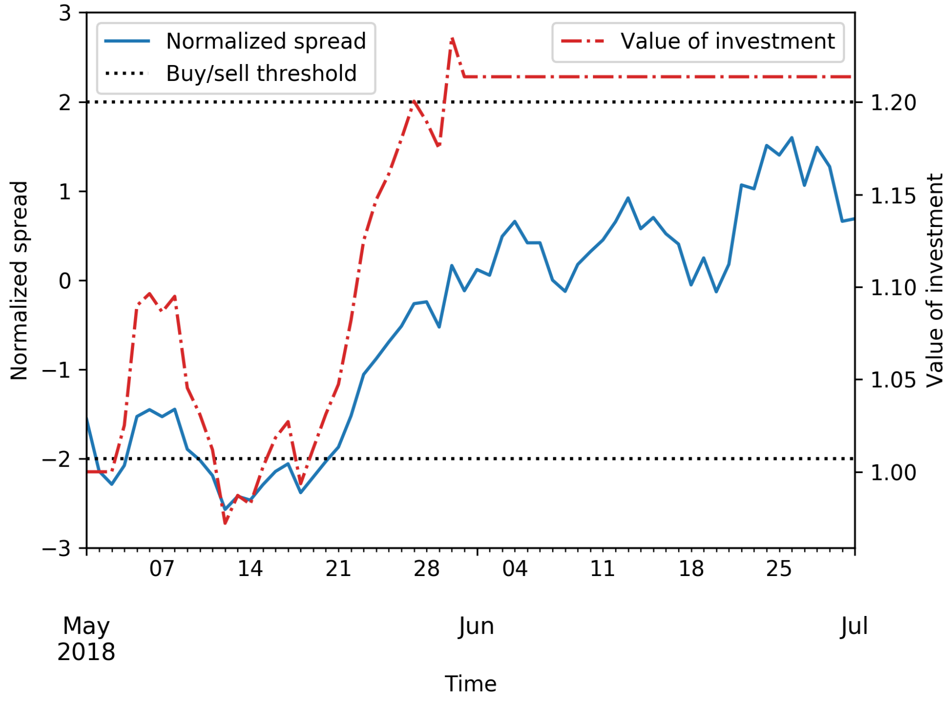
Let us now visualize the behavior of a pair during the trading period.

Specifically, we will consider the Daily Distance backtest with formation from 2018/01/01 to 2018/05/01 and trading period from 2018/05/01 to 2018/07/01. From that data, the pair made of WAVES and XRP has its normalized spread and cumulative profit shown in Figure 2, with the dotted lines marking the 2 standard deviations threshold for opening/closing positions.

The pair goes Long on 2018/05/03 and successfully reverts past zero, closing the position on 2018/05/31, by which it has accumulated 21.3% profit. It then does not trade again during the observed period.

We notice that the shape of the profit line matches the spread line. The exactness of this behavior is due to our simplifying assumptions, namely the non-existence of short-selling loan fees (which would gradually be decreasing our profit while the position is open) and no return on idle capital (which implies the flatness of the profit curve once the position is closed).

Figure 2: Exemplary pair during trading period



Note: Both lines use different y axes. The legend shows to which side the line belongs.

Next, we investigate whether the ranking of pairs based on risk-adjusted measures changes with respect to the chosen measure. If it did, we would have to be careful about choosing the right metric, especially since our conclusions could be adjusted simply by choosing a convenient risk-adjusted measure.

To this end, we use Spearman's correlation coefficient and also calculate an approximate 5% confidence interval, as explained in Section 3.3.

As can be seen in Table 4, the correlation appears to be very strong. It is at most slightly lower than what Eling (2008) reported on the correlation between different US asset classes, where he found almost exclusive above 0.95 correlation coefficient with sample sizes greater than ours across many different classes.

That said, we reach the same conclusions. All the common alternatives to Sharpe ratio tends to produce very similar rankings, particularly when considering the Sortino and Excess returns on VaR metrics. The Calmar ratio appears to be somewhat less correlated with the others, but it is most likely an issue of the length of our dataset, meaning the ratio can not be properly computed.

To illustrate, in order to calculate the Calmar ratio, we had to not only annualize the return since our trading period is nowhere a year long, but also take the maximum drawdown originally meant to be over a year just over the trading period. We have shown earlier how annualizing Sharpe ratios might produce dubious results due to unmet assumptions, and it likely is also the case here, at least partly.

For some pairs, it was not even possible to calculate all the ratios. On some occasions, a pair would have only positive returns (typically when the position was open for only a few periods), meaning there was literally no drawdown. As an example, the Sortino ratio would be undefined in that case.

Many of the other metrics used by Eling (2008) would be similarly undefined. As it is now, we just skip the pairs where not all our used metrics are defined and calculate the rankings without them.

Table 4: Correlations between different risk-adjusted measures

	Distance				Cointegration			
	Sharpe	Sortino	Calmar	VaR	Sharpe	Sortino	Calmar	VaR
Sharpe	1.0	0.98	0.92	0.98	1.0	0.98	0.94	0.99
Sortino	0.98	1.0	0.92	0.99	0.98	1.0	0.95	0.99
Calmar	0.92	0.92	1.0	0.93	0.94	0.95	1.0	0.95
VaR	0.98	0.99	0.93	1.0	0.99	0.99	0.95	1.0

The approximate 5% confidence intervals can be found in the Appendix, Table 12. It further confirms that the correlation is likely to be very strong, and we can conclude that the correlations are above 0.9 at a commonly accepted level of confidence.

To conclude, let us investigate the degree of pairs formation similarity among the two methods.

In our backtest, the distance and cointegration method have average overlap of chosen pairs of about 25%. This contrasts with the results in Huck and Afawubo (2015) who examined the SP 500 with a vaguely similar backtest design, but far smaller overlap (frequently below 1% with a maximum of 13%).

Of course, since we consider a much smaller sample of available cryptocurrencies than the number of US stocks, there are far less possible pairs in total, meaning a higher overlap is expected. Despite that, the reality of those facts is still very worrying by itself. Since both of our methods are looking for pairs with the same characteristics, the inconsistency between them is troubling.

We have already discussed the multiple comparisons problem back in Section 3.2. In that light, one possible explanation for the just observed situation would be that despite our best efforts, the majority of our pairings are essentially spurious.

On top of that, we also recognize that even if a pair is cointegrated in the formation period, it might not be anymore in the trading period simply because the generating process is not stationary (indicating there was something like a structural break inbetween).

Ideally, since there are the same cryptocurrencies traded over the whole dataset, we would like to have pair selection consistent both across methods and across time. Clearly, those conditions are not met.

However, financial data is known to have low signal-to-noise ratios, so having difficulties in finding the good pairs is to be expected. It is then not too surprising that our approximately market/dollar-neutral trading strategy has results such as nigh zero expected return with about 50% profitable trades.

Since most of our ”cointegrated” pairs are likely not actually cointegrated, our trading results look a bit like as if we just went long/short on random cryptocurrencies. Even if our pairs selection procedure is able to do slightly better than random pairings, we still have to overcome hurdles such as transaction costs, making it even harder to make consistent profit.

Table 5: Identical chosen pairs between methods

	daily	hourly
% of identical pairs	24.38%	22.86%

4.2 Other scenarios

We will now look at several what-if scenarios, as discussed in Section 3.4. We hope this will shed some more on light on the performance of our strategy. Most of the metrics apart from profit/Sharpe ratio will stay the same or nearly the same as in the base scenarios. For example, the number of trades does not change due to execution lag unless the trade would be initiated so close to the end of the trading period that the position does not actually get the chance to open in time. We drop the statistics on average and nominated pairs completely since they never change.

First of all, let us see what happens when we have zero execution lag, meaning the position is opened/closed in the same period that the spread crosses our threshold. We remind that the lag was meant to approximate the bid-ask spread.

As can be seen in Table 6, the improvement in performance for every of our baseline strategies is remarkable. Profit increased across the board, and not a single case is now unprofitable. The Sharpe ratios are also greatly improved. The Hourly Distance scores a Sharpe of 6.6 while also being by far the most profitable, having a 6.08% monthly profit.

The Hourly Cointegration also managed to get 51% of winning trades, the first time this metric went above 50%, and while it is over 4% more than for Hourly Distance, the profit again lags far behind at 2.34%.

If the lag indeed approximated the bid-ask spread well, this would imply that the spread is very significant, as our returns are now way better. Taking it as a proxy for liquidity, it would also imply that liquidity is low.

Table 6: Results with no execution lag

	Daily		Hourly	
	Distance	Cointegration	Distance	Cointegration
Monthly profit	0.69%	0.46%	6.08%	2.34%
Monthly profit (committed)	0.55%	0.47%	4.81%	1.15%
Annualized Sharpe	0.85	-0.23	6.6	2.4
Monthly number of trades	0.472	0.526	3.29	4.39
Round-trip trades	22.74%	22.45%	27.82%	36.14%
Length of position (days)	32.5	31.0	5.5	5.0
% of winning trades	45.94%	43.14%	46.95%	51.24%
Max. drawdown	24.34%	22.56%	13.25%	15.97%

We offer one more alternative explanation for the execution lag effect, especially compared to other papers like Gatev et al. (2006), who found the effect to be in the order of going from 1.44% to 0.9% in monthly profit.

While those results are akin to ours in the Daily case, the Hourly scenarios get much larger performance gains. The importance of undelayed execution could, for example, be explained by the volatility of crypto pairs, causing faster reversion speeds, meaning a one-period wait here is a ”longer” waiting time than in US equities. Of course, one might also argue that since stock markets are supposedly more mature and efficient, they should be reversing faster, and the issue is not entirely clear.

Lastly, we point out that we also reported the return on committed capital rather than just on employed capital, as was discussed in Section 3.3. Not doing so would be somewhat unnatural since assuming no execution lag while also not having to have capital ready on standby seems far-fetched. It appears reasonable that if we wish to have no trade delay, we must have capital prepared ahead of time.

This introduces a trade-off between quick and slow execution. But as Table 6 shows, our conclusions remain pretty much unchanged. The profit is somewhat lower when calculating return on committed capital, but still far higher than it was in the base scenarios. Notably, Daily Cointegration even has a higher profit than usual, which can happen due to different trading activity of pairs in each backtest, as going from employed to committed capital is essentially a reweighting of the results.

We take this occasion to discuss one other related and similarly subtle issue that permeates our whole analysis. If we use only the employed capital measure, the capital base relative to which we compute returns will be different for each backtest, since the number of actually traded pairs will vary. Furthermore, while the Distance method always nominates top 20 pairs, the Cointegration method also has variable amount of nominations. Then, once we try to average the returns across all backtests, we weight those equally, even though each of them considers a different amount of capital.

This means we have a problem with capital deployment. Posting a 5% return over the trading period is quite different when we trade three pairs total compared to when we trade twenty five pairs. This could be amended for the Distance method if we used the committed capital measure everytime, but even that would not suffice for the Cointegration method due to non-constant amount of nominated pairs. Consequently, we suggest additional caution while interpreting the monthly profit figures.

Next, we test the importance of transaction costs. Table 13 in the Appendix shows the situation with no transaction costs.

Disregarding transaction costs has an effect quite similar to not having any execution lag. Again, there is a dramatic improvement in performance for every strategy. Every pair now again has positive profit, although the Daily cases are still very moderate in that regard, scoring returns of around 0.3%. Hourly Distance is again dominant in terms of monthly profit at 5.29%, followed by the Hourly Cointegration at 1.77%.

It is also quite interesting that the Daily scenarios were helped far less by ignoring transaction costs than the Hourly ones. This can be explained by the difference in number of trades, since Hourly scenarios trade over 6x more often than the Daily ones.

The results are best seen in contrast with the base scenarios. In the base cases, only the Hourly Distance was profitable and even the returns of this scenario doubled after leaving out transaction costs.

We remind that our transaction costs are 30bps, which is on the lower side of estimates proposed in literature. This shows that pairs trading is very sensitive to transaction costs, and we did not even take into account the short-selling fees.

One thing we have not tried yet is to alter the trigger threshold, and also introduce a stop-loss. Let us first see what effect it has on Daily scenarios in Table 7.

The Distance method is discussed first. There are a couple interesting patterns here in both dimensions of the table. First, we can see that the highest trigger threshold generally wins out in raw profit followed by the smallest threshold. Also, it appears that higher stop-loss produces better results.

However, it is necessary to be careful, as almost all scenarios have negative profit and if the higher threshold scenarios trade less often, they essentially get to commit less mistakes, and having less negative profit is not what we aim for, ultimately. It is noteworthy that the highest threshold is also the only one to ever achieve positive profit.

The Sharpe ratios interpretation is also somewhat tricky. For negative ratios, getting closer to zero can be achieved either by increasing raw returns or by increasing volatility. However, it is quite unlikely that the changes made to the model parametrization here would increase volatility, as they actively limit the variance, and we have already noticed that raw returns tend to increase with respect to both the opening threshold and stop-loss. Therefore, we conclude that the improvement in Sharpe ratios is a positive sign.

Table 7: Various Daily stop-loss and opening triggers

			Distance			Cointegration		
			Threshold			Threshold		
			1	2	3	1	2	3
Stop-loss	2	Monthly profit	-0.50%	None	None	-0.37%	None	None
		Annualized Sharpe	-1.3	None	None	-0.82	None	None
	3	Monthly profit	-1.22%	-1.92%	None	-1.18%	-1.48%	None
		Annualized Sharpe	-1.1	-1.9	None	-1.5	-1.9	None
	4	Monthly profit	-1.12%	-1.74%	-0.28%	-0.23%	-0.69%	0.75%
		Annualized Sharpe	-1.2	-1.6	-0.32	-0.63	-0.98	0.61
	5	Monthly profit	-0.33%	-1.10%	0.04%	-0.32%	-0.86%	-0.41%
		Annualized Sharpe	-0.21	-0.93	0.33	-0.78	-1.3	-0.65
	6	Monthly profit	-0.00%	-0.79%	1.04%	-0.07%	-0.36%	-0.10%
		Annualized Sharpe	0.29	-0.53	1.3	-0.68	-0.9	-0.54

Note: if the threshold was equal or smaller than stop-loss, the results would not make any sense, so we mark them as None

The Cointegration method fares a bit differently, with results slightly less clear-cut. This would suggest that the parametrization of each method needs to proceed differently. This further builds up on the statistics we saw earlier describing the percentage of round-trip trades and others, where we saw that the two methods behave quite differently. That said, we should be careful here as comparing values across such a small amount of options as we have here is not very robust.

Overall however, we notice that the profits of the base case scenarios seen in Table 3 are somewhat hard to beat. Taking a closer look at the results with threshold set to 2 that matches our base scenarios, we see that introducing a stop-loss actually lowers our performance significantly across both methods, and the improvement from increasing the stop-loss trigger can perhaps be interpreted as simply getting it out of the way.

We also notice that the scenarios with threshold set to 2 are most of the time the least performant out of the three possible thresholds, regardless of stop-loss. This might indicate that the baseline strategy formulation we adopted from other literature is highly suboptimal in our case.

While we first assumed that once the normalized spread gets above 2 in absolute magnitude, it is bound to reverse, we now see that if it gets as high as 6, it is still not a valid reason to believe that it has diverged permanently. It thus appears that at the very least, a reasonable stop-loss has to be set very high.

We offer one more view on this matter. If a pair is actually cointegrated and we stop-loss it, it is a mistake. But if it is not cointegrated, then the spread is equally likely to go up and down no matter its value. Stop-loss in this case would be neutral, which means that overall, stop-lossing is at best neutral and on average harmful.

Analogous tables for the Hourly scenarios are in the Appendix, Table 15. This time, the lowest threshold is the best, regardless of stop-loss or the method in question, and we again see a clear uptrend in performance as we increase the stop-loss trigger. The 2 standard deviations threshold is still the worst, as in the Daily table.

In fact, most of our previous comments still apply. In particular, introducing stop-loss appears to do more harm than good. Just going off our results, it appears that the way to go is a threshold of 1 with a high stop-loss trigger. Generally, the Cointegration method has a particularly poor performance, with its best result at -0.79%, while Distance has multiple cases of profit above 2%, and even up to 5%.

Nonetheless, given the wealth of different reactions to parameter adjustments across our scenarios, attempting to optimize through a grid-search as we did delivers unclear results and might be just a case of data snooping.

Still, our analysis showed some promise and a more elaborate optimization scheme might prove to be more helpful. Given the amount of possibilities we outlined during our discussion on sensitivity to parametrization, our results are just scratching the surface and we have more or less shown that the problem does not have an immediate, high-level solution.

4.3 Distribution of returns

Now, let us briefly discuss the distribution of returns. We have already mentioned that normality of returns is important in many respects - be it for metrics like Sharpe Ratio or hypothesis testing. Likewise, parametric statistical procedures often use some assumption of normality which might be amendable by large sample sizes via the central limit theorem, but "large" depends on context.

Normality can be assessed by a number of tests based on many different characteristics, ranging from skewness/kurtosis analysis to comparing empirical distribution functions. Popular options include the Jarque-Bera and Shapiro-Wilks tests, as well as many other alternatives.

However, such tests lack power in small samples and are sensitive to even small deviations from normality in large samples (Ghasemi and Zahediasl, 2012). Of course, CLT assures that at large sample sizes, the non-normality is not severe in most use cases and hypothesis testing can be carried out by asymptotic arguments, which is what we will ultimately rely on here.

First of all, let us see basic descriptive statistics of returns in Table 8. The table lists values computed from all backtesting periods across all pairs.

There are several quantities of interest. We notice the great volatility, seen in multiple metrics. The standard deviation is high, but what is really telling is the kurtosis. Clearly, the returns are very strongly leptokurtic, indicating fat tails compared to the normal distribution and a high number of "outliers".

Indeed, this is further evidenced by the range - the absolute values of both minimum and maximum are very high as well. It is possible for a pair to lose or gain half its invested value within an hour.

However, neither strategy quite achieves a percentage of positive returns above 50%. In spite of this, the mean return tends to be slightly above zero. Likewise, the skewness is negative for Daily data, indicating that the left (negative) tail is longer. However, the situation is reversed for Hourly data.

The t-statistic for testing

$$H_0 : mean = 0$$

is in favor of not rejecting the null hypothesis. No configuration has mean return significantly different from zero at the 5% level. Our sample sizes are quite high in all cases (roughly 5000 for Daily and 100 000 for Hourly), so we need not worry about the underlying data distribution being non-normal too much. But as the volatility of returns is just too high, the t-test does not allow for rejecting any null hypotheses.

Table 8: Descriptive statistics of return distributions

	Daily		Hourly	
	Distance	Cointegration	Distance	Cointegration
Mean	-0.00007	0.00035	0.00005	-0.00003
Std	0.056	0.0537	0.0148	0.0174
Max	0.474	0.395	0.368	0.663
Min	-0.713	-0.673	-0.409	-0.364
Jarque-Bera p-value	0.0	0.0	0.0	0.0
Skewness	-1.91	-1.82	-0.0193	0.789
Kurtosis	27.7	29.5	51.9	68.9
Positive	47.24%	47.26%	48.78%	48.35%
t-stat	-0.086	0.44	1.1	-0.47

The Jarque-Bera test rejects normality at any common significance level with the p-value extremely close to zero. To further emphasize the breach of normality, Figure 4 in the Appendix displays kernel density estimations (using the Gaussian kernel) compared to the normal distribution with location and scale parameters estimated from the data in each case.

The graphical representation confirms our interpretation of higher moments. The return distribution displays a very strong peak near the mean, but has a significant number of extreme outliers compared to the normal distribution that is more spread out. The jaggedness of the KDEs is likely attributable to heavy tails making rule-of-thumb bin width selection difficult (Buch-larsen et al., 2005).

We also include a Q-Q plot in the Appendix, Figure 6 to compare the empirical quantiles to those of a normal distribution. The "S" shape of our data clearly confirms the heavier tails.

Furthermore, there is a noticeable difference in slopes between Hourly and Daily data, both in Figure 4 and Figure 6. This is probably caused by the different sampling frequency since the Daily returns are a sum of the Hourly ones. This causes greater dispersion away from 0, as well as the greater range.

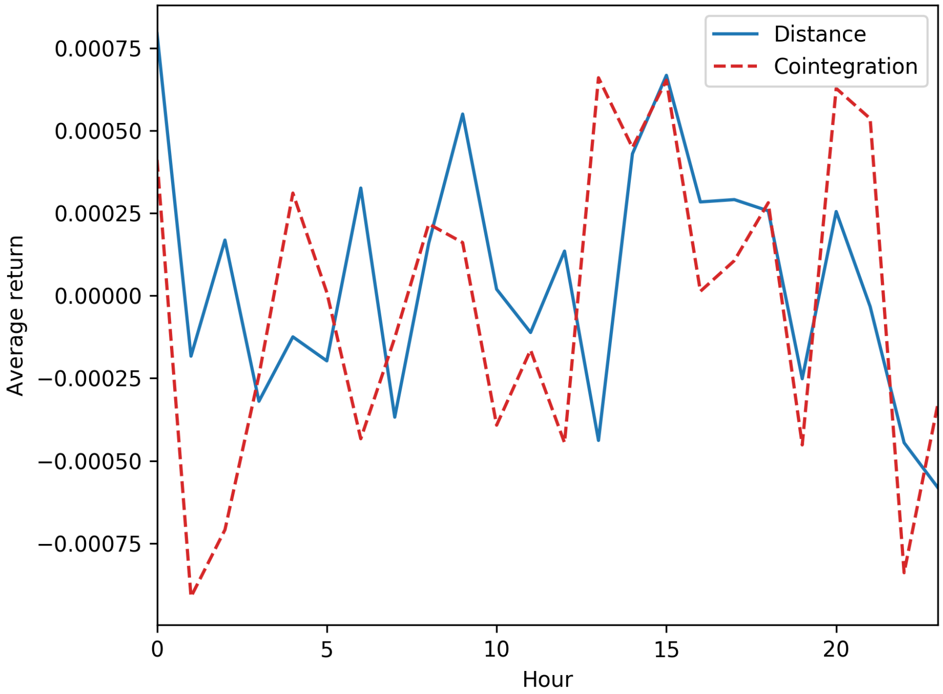
Additionally, for the Hourly base scenarios, let us also group the results by hour. Summary statistics can be found in the Appendix, Table 14. Judging off that, there appears to be a noticeable difference between the mean returns grouped by hour of day (all hours are in UTC).

While the standard deviations of returns are relatively high, we also include a t-statistic for testing against the null hypothesis of zero mean. We are ultimately able to reject that null hypothesis in multiple cases, as our number of observations is also quite high, almost 4000 for every hour. It follows that we would also likely reject the null hypothesis of equality among expected returns per hour, as they tend to be more distant from each other than from zero.

A visualization can be found in Figure 3. Indeed, even by visual inspection, the distribution appears to be somewhat consistent across the two methods, which is mildly supportive of the difference being real and common to both methods. The Distance and Cointegration lines have a Pearson correlation coefficient of 0.42.

Based on that analysis, we might try to hypothesize influence of some exogenous factors. For example, it appears that the afternoon, and particularly 2-4pm, produces significant positive profits. That said, to what extent this difference is meaningful and whether it is actually caused by different geographies being active is far beyond the scope of this thesis.

Figure 3: Average returns per hour of day



We will however note that analogous effects in traditional equities are not at all uncommon. Some prominent examples include the weekend effect (French, 1980) or the January effect (Thaler, 1987). Keeping those auxiliary results in mind, what we found out is all the more suspect and a more detailed investigation is a promising topic left for further research.

Conclusion

The point of this thesis was to investigate the applicability of pairs trading, a technique developed primarily for equity markets, to cryptocurrencies. We took the two most common approaches, the distance and cointegration methods, and backtested them on a year of data from Binance, one of the largest exchanges at the time of writing.

In many ways, our results confirm the more recent literature on the topic. The original strategy first described by Gatev et al. (1999) is not profitable, although certain combinations of hyperparameters can be found that do much better.

In particular, the results are very sensitive to transaction costs and execution windows. Even modest transaction costs hurt our strategies immensely, and so does the usage of a one period execution lag. It also appears that the distance method does much better overall, and is able to achieve good positive profits, whereas the cointegration method only rarely has returns above zero.

However, in ideal conditions, all the strategies perform well. Through further fine-tuning, it might be possible to consistently profit even in spite of transaction costs. The extent to which the execution lag is a faithful representative of the bid-ask spread, and subsequently liquidity, is much harder to answer.

On the other hand, we were able to confirm that all the most commonly used risk-adjusted measure of returns are essentially equivalent, making the choice among them inconsequential.

It also appears to be somewhat advantageous to use more granular time series. Trading based on hourly data tends to outperform the daily case in risk-adjusted metrics, and also in raw returns. However, we have pointed out that Sharpe ratio does not generalize well across different classes of strategies, and it might be more prudent to say that hourly data amplifies the magnitude of returns.

We observe that with daily data, all our strategies have approximately zero profit. But if we use hourly data, the distance method becomes highly profitable, although the cointegration method's loss is slightly magnified. So while this is technically an improvement as we managed to get at least some profitable scenario, it is not ideal.

We also uncovered several deficiencies present in currently available literature that are fit for further research. Importantly, robustness of the results is mostly ignored and risk management is neglected. This extends to not only explicit hyperparameters like spread thresholds or lengths of periods, but also, particularly in the cointegration method, the setup of statistical tests. Even in the presence of several alternatives, the difference amongst them is hardly ever explored.

Another unaccounted for uncertainty stems from applying results coming from traditional markets to cryptocurrencies. Given that our goal was to determine the difference between crypto and standard markets, it is without doubt inappropriately restrictive to base our approximations of transaction costs on research studying US equities. However, literature on the same matters in the crypto space is lacking.

Moreover, we were unable to carry over some concepts such as pairs pre-selection based on sector, since there are no established sectors in cryptocurrencies. However, sensible sector design might be invented, based on either qualitative analysis or unsupervised cluster analysis.

To sum up, while pairs trading appears to be viable, the lack of research on factors exogenous to the trading strategy such as transaction costs or execution difficulties explicitly done on cryptocurrencies means that our results do not quite fulfill our original ambitions to assess the viability of pairs trading in cryptocurrencies.

In ideal conditions, we have shown that our strategies are highly profitable. However, we have also seen how quickly can the tables turn due to seemingly small real-life obstacles, and their precise determination is therefore most vital. That said, doing so is left for further research.

References

Balcilar, Mehmet et al. (2017). "Can volume predict Bitcoin returns and volatility? A quantiles-based approach". In: *Economic Modelling* 64, pp. 74–81. DOI: 10.1016/j.econmod.2017.03.019.

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko (2012). *The Trading Profits of High Frequency Traders.* DOI: 10.2139/ssrn.2106158.

Bogomolov, Timofei (2013). "Pairs trading based on statistical variability of the spread process". In: *Quantitative Finance* 13.9, pp. 1411–1430.

Bondt, Werner F. M. De and Richard H. Thaler (1987). "Further Evidence on Investor Overreaction and Stock Market Seasonality". In: *The Journal of Finance* 42.3, pp. 557–581. DOI: 10.2307/2328371.

Bookstaber, Richard (2007). *A Demon of Our Own Design: Markets, Hedge Funds, and the Perils of Financial Innovation.* Wiley. 288 pp. ISBN: 978-0-471-22727-4.

Bowen, David, Mark C. Hutchinson, and Niall O'Sullivan (2010). *High Frequency Equity Pairs Trading: Transaction Costs, Speed of Execution and Patterns in Returns.* SSRN Scholarly Paper ID 1611623. Rochester, NY: Social Science Research Network.

Broussard, John Paul and Mika Vaihekoski (2012). "Profitability of pairs trading strategy in an illiquid market with multiple share classes". In: *Journal of International Financial Markets, Institutions and Money* 22.5, pp. 1188–1201. DOI: 10.1016/j.intfin.2012.06.002.

Buch-larsen, Tine et al. (2005). "Kernel density estimation for heavy-tailed distributions using the champernowne transformation". In: *Statistics* 39.6, pp. 503–516. DOI: 10.1080/02331880500439782.

Carrasco Blázquez, Mario, Carmen De la Orden De la Cruz, and Camilo Prado Román (2018). "Pairs trading techniques: An empirical contrast". In: *European Research on Management and Business Economics* 24.3, pp. 160–167. DOI: 10.1016/j.iedeen.2018.05.002.

Cavaliere, Giuseppe and A. M. Robert Taylor (2009). "Heteroskedastic Time Series with a Unit Root". In: *Econometric Theory* 25.5, pp. 1228–1276.

Cheung, Yin-Wong and Kon S. Lai (1995). "Lag Order and Critical Values of the Augmented Dickey-Fuller Test". In: *Journal of Business & Economic Statistics* 13.3, pp. 277–280. DOI: 10.2307/1392187.

Chohan, Usman W. (2018). *The Problems of Cryptocurrency Thefts and Exchange Shutdowns*. SSRN Scholarly Paper ID 3131702. Rochester, NY: Social Science Research Network.

Ciaian, Pavel, Miroslava Rajcaniova, and d'Artis Kancs (2016). "The economics of BitCoin price formation". In: *Applied Economics* 48.19, pp. 1799–1815. DOI: 10.1080/00036846.2015.1109038.

D'Avolio, Gene (2002). *The Market for Borrowing Stock*. SSRN Scholarly Paper ID 305479. Rochester, NY: Social Science Research Network.

Dickey, David A. and Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series With a Unit Root". In: *Journal of the American Statistical Association* 74.366, pp. 427–431. DOI: 10.2307/2286348.

Do, Binh and Robert Faff (2010). "Does Simple Pairs Trading Still Work?" In: *Financial Analysts Journal* 66.4, pp. 83–95.

— (2012). "Are Pairs Trading Profits Robust to Trading Costs?" In: *Journal of Financial Research* 35.2, pp. 261–287. DOI: 10.1111/j.1475-6803.2012.01317.x.

Do, Binh, Robert Faff, and Kais Hamza (2006). "A New Approach to Modeling and Estimation for Pairs Trading". In:

Eling, Martin (2008). "Does the Measure Matter in the Mutual Fund Industry?" In: *Financial Analysts Journal* 64, pp. 54–66.

Elliott, Robert J., John Van Der Hoek *, and William P. Malcolm (2005). "Pairs trading". In: *Quantitative Finance* 5.3, pp. 271–276. DOI: 10.1080/14697680500149370.

Enders, Walter (2014). *Applied Econometric Time Series*. 4 edition. Hoboken, NJ: Wiley. 496 pp. ISBN: 978-1-118-80856-6.

Engle, Robert and Clive Granger (1987). "Co-integration and Error Correction: Representation, Estimation, and Testing". In: *Econometrica* 55.2, pp. 251–76.

Eyal, Ittay and Emin Gün Sirer (2018). "Majority is Not Enough: Bitcoin Mining is Vulnerable". In: *Commun. ACM* 61.7, pp. 95–102. DOI: 10.1145/3212998.

Fama, Eugene and Kenneth French (1988). "Permanent and Temporary Components of Stock Prices". In: *Journal of Political Economy* 96.2, pp. 246–73.

Farinelli, Simone et al. (2008). "Beyond Sharpe ratio: Optimal asset allocation using different performance ratios". In: *Journal of Banking & Finance* 32.10, pp. 2057–2063. DOI: 10.1016/j.jbankfin.2007.12.026.

Feng, Wenjun, Yiming Wang, and Zhengjun Zhang (2018). "Informed trading in the Bitcoin market". In: *Finance Research Letters* 26, pp. 63–70. DOI: 10.1016/j.frl.2017.11.009.

French, Kenneth R. (1980). "Stock returns and the weekend effect". In: *Journal of Financial Economics* 8.1, pp. 55–69. DOI: 10.1016/0304-405X(80)90021-5.

Fuller, Wayne A. (1995). *Introduction to Statistical Time Series.* 2 edition. New York: Wiley-Interscience. 728 pp. ISBN: 978-0-471-55239-0.

Gatev, Evan G, William N Goetzmann, and K. Geert Rouwenhorst (1999). *Pairs Trading: Performance of a Relative Value Arbitrage Rule.* Working Paper 7032. National Bureau of Economic Research. DOI: 10.3386/w7032.

Gatev, Evan, William N. Goetzmann, and K. Geert Rouwenhorst (2006). *Pairs Trading: Performance of a Relative-Value Arbitrage Rule.* SSRN Scholarly Paper ID 1095996. Rochester, NY: Social Science Research Network.

Ghasemi, Asghar and Saleh Zahediasl (2012). *Normality Tests for Statistical Analysis: A Guide for Non-Statisticians.* Vol. 10. 486 pp. DOI: 10.5812/ijem.3505.

Granger, Clive and P. Newbold (1974). "Spurious regressions in econometrics". In: *Journal of Econometrics* 2.2, pp. 111–120.

Harris (1995). *Using Cointegration analysis in Econometric modelling.* 01 edition. London: Prentice Hall. ISBN: 978-0-13-355892-0.

Huang, Chien-Feng et al. (2015). "An Intelligent Model for Pairs Trading Using Genetic Algorithms". In: *Intell. Neuroscience* 2015, 16:16–16:16. DOI: 10.1155/2015/939606.

Huck, Nicolas (2010). "Pairs trading and outranking: The multi-step-ahead forecasting case". In: *European Journal of Operational Research* 207.3, pp. 1702–1716. DOI: 10.1016/j.ejor.2010.06.043.

Huck and Komivi Afawubo (2015). *Pairs trading and selection methods: Is cointegration superior?* Vol. 47. DOI: 10.1080/00036846.2014.975417.

Jacobs, Heiko and Martin Weber (2015). "On the determinants of pairs trading profitability". In: *Journal of Financial Markets* 23, pp. 75–97. DOI: 10.1016/j.finmar.2014.12.001.

Jegadeesh, Narasimhan (1990). "Evidence of Predictable Behavior of Security Returns". In: *The Journal of Finance* 45.3, pp. 881–898. DOI: 10.2307/2328797.

— (1991). "Seasonality in Stock Price Mean Reversion: Evidence from the U.S. and the U.K". In: *The Journal of Finance* 46.4, pp. 1427–1444. DOI: 10.2307/2328865.

Kim, Kiwhan and Peter Schmidt (1993). "Unit root tests with conditional heteroskedasticity". In: *Journal of Econometrics* 59.3, pp. 287–300. DOI: 10.1016/0304-4076(93)90027-3.

Kim, Myung Jig, Charles R Nelson, and Richard Startz (1988). *Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence.* Working Paper 2795. National Bureau of Economic Research. DOI: 10.3386/w2795.

Kristoufek, Ladislav (2015). "What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis". In: *PLOS ONE* 10.4, e0123923. DOI: 10.1371/journal.pone.0123923.

Liew, Rong Qi and Yuan Wu (2013). "Pairs trading: A copula approach". In: *Journal of Derivatives & Hedge Funds* 19.1, pp. 12–30. DOI: 10.1057/jdhf.2013.1.

Ling, Shiqing, W. K. Li, and Michael McAleer (2003). "Estimation and Testing for Unit Root Processes with GARCH (1, 1) Errors: Theory and Monte Carlo Evidence". In: *Econometric Reviews* 22.2, pp. 179–202. DOI: 10.1081/ETC-120020462.

Lintilhac, P. S. and A. Tourin (2017). "Model-based pairs trading in the bitcoin markets". In: *Quantitative Finance* 17.5, pp. 703–716. DOI: 10.1080/14697688.2016.1231928.

Liu, Bo, Lo-Bin Chang, and Hélyette Geman (2017). "Intraday pairs trading strategies on high frequency data: the case of oil companies". In: *Quantitative Finance* 17.1, pp. 87–100. DOI: 10.1080/14697688.2016.1184304.

Lo, Andrew (2003). *The Statistics of Sharpe Ratios.* Vol. 58. DOI: 10.2469/faj.v58.n4.2453.

MacKinnon, James G. (2010). *Critical Values for Cointegration Tests*. 1227. Queen's University, Department of Economics.

Makarov, Igor and Antoinette Schoar (2018). *Trading and Arbitrage in Cryptocurrency Markets*. SSRN Scholarly Paper ID 3171204. Rochester, NY: Social Science Research Network.

Moore, Tyler and Nicolas Christin (2013). "Beware the Middleman: Empirical Analysis of Bitcoin-Exchange Risk". In: *Financial Cryptography and Data Security*. Ed. by Ahmad-Reza Sadeghi. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 25–33. ISBN: 978-3-642-39884-1.

Nakamoto, Satoshi (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*.

Nakano, Masafumi, Akihiko Takahashi, and Soichiro Takahashi (2018). "Bitcoin technical trading with artificial neural network". In: *Physica A: Statistical Mechanics and its Applications* 510, pp. 587–609. DOI: 10.1016/j.physa.2018.07.017.

Narayanan, Arvind et al. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Google-Books-ID: LchFDAAAQBAJ. Princeton University Press. 335 pp. ISBN: 978-1-4008-8415-5.

Ng, Serena and Pierre Perron (2001). "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power". In: *Econometrica* 69.6, pp. 1519–1554.

Pantula, Sastry G. (1988). "Estimation of Autoregressive Models with Arch Errors". In: *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)* 50.1, pp. 119–138.

Perlin, Marcelo Scherer (2009). "Evaluation of pairs-trading strategy at the Brazilian financial market". In: *Journal of Derivatives & Hedge Funds* 15.2, pp. 122–136. DOI: 10.1057/jdhf.2009.4.

Poterba, James M. and Lawrence H. Summers (1988). "Mean reversion in stock prices: Evidence and Implications". In: *Journal of Financial Economics* 22.1, pp. 27–59. DOI: 10.1016/0304-405X(88)90021-9.

Rad, Hossein, Rand Kwong Yew Low, and Robert W. Faff (2015). *The Profitability of Pairs Trading Strategies: Distance, Cointegration, and Copula Methods*.

SSRN Scholarly Paper ID 2614233. Rochester, NY: Social Science Research Network.

Ramos-Requena, J. P., J. E. Trinidad-Segovia, and M. A. Sánchez-Granero (2017). "Introducing Hurst exponent in pair trading". In: *Physica A: Statistical Mechanics and its Applications* 488 (C), pp. 39–45.

Ryan, Kevin F. and David E. A. Giles (1998). *Testing for Unit Roots With Missing Observations.* 9802. Department of Economics, University of Victoria.

Sortino, Frank A. and Robert van der Meer (1991). "Downside risk". In: *The Journal of Portfolio Management* 17.4, p. 27. DOI: 10.3905/jpm.1991.409343.

Thaler, Richard H. (1987). "Anomalies: The January Effect". In: *Journal of Economic Perspectives* 1.1, pp. 197–201. DOI: 10.1257/jep.1.1.197.

Vidyamurthy, Ganapathy (2004). *Pairs Trading: Quantitative Methods and Analysis.* 1 edition. Hoboken, N.J: Wiley. 224 pp. ISBN: 978-0-471-46067-1.

Wooldridge, Jeffrey M. (2008). *Introductory Econometrics: A Modern Approach.* Google-Books-ID: 64vt5TDBNLwC. Cengage Learning. 889 pp. ISBN: 978-0-324-58162-1.

Young, Terry W (1991). "Calmar ratio: A smoother tool". In: *Futures* 20.1, p. 40.

Appendix A - Tables

Table 9: Cryptocurrencies on Binance

ADA, ADX, AE, AGI, AION, AMB, APPC, ARDR, ARK, ARN, AST, BAT, BCC, BCD, BCHABC, BCHSV, BCN, BCPT, BLZ, BNB, BNT, BQX, BRD, BTCU, BTG, BTS, BTT, CDT, CHAT, CLOAK, CMT, CND, CVC, DASH, DATA, DCR, DENT, DGD, DLT, DNT, DOCK, EDO, ELF, ENG, ENJ, EOS, ETC, ETH, ETHU, EVX, FET, FUEL, FUN, GAS, GNT, GO, GRS, GTO, GVT, GXS, HC, HOT, HSR, ICN, ICX, INS, IOST, IOTA, IOTX, KEY, KMD, KNC, LEND, LINK, LOOM, LRC, LSK, LTC, LUN, MANA, MCO, MDA, MFT, MITH, MOD, MTH, MTL, NANO, NAS, NAV, NCASH, NEBL, NEO, NPXS, NULS, NXS, OAX, OMG, ONG, ONT, OST, PAX, PHX, PIVX, POA, POE, POLY, POWR, PPT, QKC, QLC, QSP, QTUM, RCN, RDN, REN, REP, REQ, RLC, RPX, RVN, SALT, SC, SKY, SNGLS, SNM, SNT, STEEM, STORJ, STORM, STRAT, SUB, SYS, THETA, TNB, TNT, TRIG, TRX, TUSD, USDC, VEN, VET, VIA, VIB, VIBE, WABI, WAN, WAVES, WINGS, WPR, WTC, XEM, XLM, XMR, XRP, XVG, XZC, YOYO, ZEC, ZEN, ZIL, ZRX

Table 10: Cryptocurrencies left after preprocessing

ADA, ARN, BNB, DASH, ENJ, EOS, ETC, ETH, IOTA, LINK, LTC, MDA, MTL, NEO, OMG, QTUM, TRX, WAVES, WTC, XMR, XRP, XVG, ZRX

Table 11: Correlations between risk-adjusted measures for Hourly data

	Distance				Coint			
	Sharpe	Sortino	Calmar	VaR	Sharpe	Sortino	Calmar	VaR
Sharpe	1.0	0.99	0.9	0.99	1.0	0.99	0.86	0.99
Sortino	0.99	1.0	0.91	0.99	0.99	1.0	0.86	0.99
Calmar	0.9	0.91	1.0	0.9	0.86	0.86	1.0	0.86
VaR	0.99	0.99	0.9	1.0	0.99	0.99	0.86	1.0

Table 12: 5% confidence intervals for Spearman correlations

			Sharpe	Sortino	Calmar	VaR
Daily	Distance	Sharpe	[1.0, 1.0]	[0.98, 0.99]	[0.89, 0.94]	[0.98, 0.99]
		Sortino	[0.98, 0.99]	[1.0, 1.0]	[0.89, 0.94]	[0.98, 0.99]
		Calmar	[0.89, 0.94]	[0.89, 0.94]	[1.0, 1.0]	[0.9, 0.95]
		VaR	[0.98, 0.99]	[0.98, 0.99]	[0.9, 0.95]	[1.0, 1.0]
	Coint	Sharpe	[1.0, 1.0]	[0.97, 0.98]	[0.91, 0.96]	[0.98, 0.99]
		Sortino	[0.97, 0.98]	[1.0, 1.0]	[0.93, 0.96]	[0.98, 0.99]
		Calmar	[0.91, 0.96]	[0.93, 0.96]	[1.0, 1.0]	[0.93, 0.97]
		VaR	[0.98, 0.99]	[0.98, 0.99]	[0.93, 0.97]	[1.0, 1.0]
Hourly	Distance	Sharpe	[1.0, 1.0]	[0.99, 0.99]	[0.88, 0.92]	[0.99, 0.99]
		Sortino	[0.99, 0.99]	[1.0, 1.0]	[0.89, 0.92]	[0.99, 0.99]
		Calmar	[0.88, 0.92]	[0.89, 0.92]	[1.0, 1.0]	[0.89, 0.92]
		VaR	[0.99, 0.99]	[0.99, 0.99]	[0.89, 0.92]	[1.0, 1.0]
	Coint	Sharpe	[1.0, 1.0]	[0.99, 0.99]	[0.84, 0.88]	[0.99, 0.99]
		Sortino	[0.99, 0.99]	[1.0, 1.0]	[0.84, 0.88]	[0.99, 0.99]
		Calmar	[0.84, 0.88]	[0.84, 0.88]	[1.0, 1.0]	[0.83, 0.88]
		VaR	[0.99, 0.99]	[0.99, 0.99]	[0.83, 0.88]	[1.0, 1.0]

Table 13: Results with no transaction costs

	Daily		Hourly	
	Distance	Cointegration	Distance	Cointegration
Monthly profit	0.31%	0.26%	5.29%	1.77%
Annualized Sharpe	0.5	-0.42	6.5	1.5
Monthly number of trades	0.469	0.526	3.29	4.38
Round-trip trades	22.32%	21.90%	27.90%	36.37%
Length of position (days)	31.7	30.3	5.5	5.0
% of winning trades	46.25%	46.66%	47.21%	50.58%
Max. drawdown	24.47%	22.52%	13.21%	16.02%

Table 14: Summary statistics for returns by hour of day (UTC)

	Distance				Cointegration			
	Returns distribution				Returns distribution			
	Mean	Count	Std	T-stat	Mean	Count	Std	T-stat
Hour								
0	0.00079	3886	0.017	3	0.00041	4513	0.024	1.2
1	-0.00018	4154	0.015	-0.81	-0.00091	4758	0.017	-3.8
2	0.00017	3496	0.014	0.72	-0.00071	4102	0.018	-2.6
3	-0.00032	3510	0.014	-1.4	-0.00024	4142	0.017	-0.91
4	-0.00012	3542	0.014	-0.52	0.00031	4162	0.017	1.2
5	-0.00020	3561	0.014	-0.86	0.00001	4183	0.016	0.04
6	0.00033	3593	0.013	1.5	-0.00043	4219	0.015	-1.9
7	-0.00037	3609	0.013	-1.7	-0.00013	4230	0.016	-0.52
8	0.00016	3627	0.014	0.69	0.00022	4238	0.017	0.85
9	0.00055	3646	0.013	2.6	0.00016	4272	0.016	0.68
10	0.00002	3667	0.014	0.086	-0.00039	4300	0.017	-1.5
11	-0.00011	3682	0.013	-0.52	-0.00016	4326	0.017	-0.65
12	0.00014	3701	0.014	0.57	-0.00045	4333	0.018	-1.6
13	-0.00044	3715	0.016	-1.7	0.00066	4367	0.019	2.3
14	0.00043	3730	0.016	1.6	0.00045	4384	0.018	1.6
15	0.00067	3744	0.018	2.3	0.00065	4396	0.02	2.2
16	0.00028	3772	0.016	1.1	0.00001	4428	0.018	0.051
17	0.00029	3792	0.016	1.1	0.00011	4455	0.017	0.41
18	0.00026	3800	0.018	0.9	0.00028	4456	0.015	1.2
19	-0.00025	3808	0.016	-0.96	-0.00045	4471	0.021	-1.5
20	0.00026	3824	0.016	0.99	0.00063	4486	0.017	2.5
21	-0.00003	3840	0.015	-0.14	0.00054	4490	0.016	2.2
22	-0.00045	3862	0.013	-2.1	-0.00084	4512	0.014	-3.9
23	-0.00058	3882	0.013	-2.7	-0.00033	4515	0.016	-1.4

Note: A corresponding Daily table does not exist, as it does not differentiate among hours of day

Table 15: Various Hourly stop-loss and opening triggers

			Distance			Cointegration		
			Threshold			Threshold		
			1	2	3	1	2	3
Stop-loss	2	Monthly profit	-5.99%	None	None	-5.93%	None	None
		Annualized Sharpe	-4.5	None	None	-6.8	None	None
	3	Monthly profit	1.45%	-2.24%	None	-1.53%	-4.02%	None
		Annualized Sharpe	-0.031	-5.7	None	-4.0	-8.6	None
	4	Monthly profit	2.11%	-0.70%	0.04%	-1.42%	-3.29%	-2.93%
		Annualized Sharpe	1.8	-2.6	-0.61	-3.8	-7.4	-4.8
	5	Monthly profit	3.24%	0.45%	1.16%	-0.79%	-2.08%	-1.34%
		Annualized Sharpe	3.6	-0.46	0.71	-3.5	-6.0	-2.1
	6	Monthly profit	4.58%	1.86%	2.32%	-0.79%	-2.45%	-1.83%
		Annualized Sharpe	5.6	1.2	2.3	-3.7	-6.3	-2.2

Appendix B - Figures

Figure 4: Normal and KDE approximations of return distributions

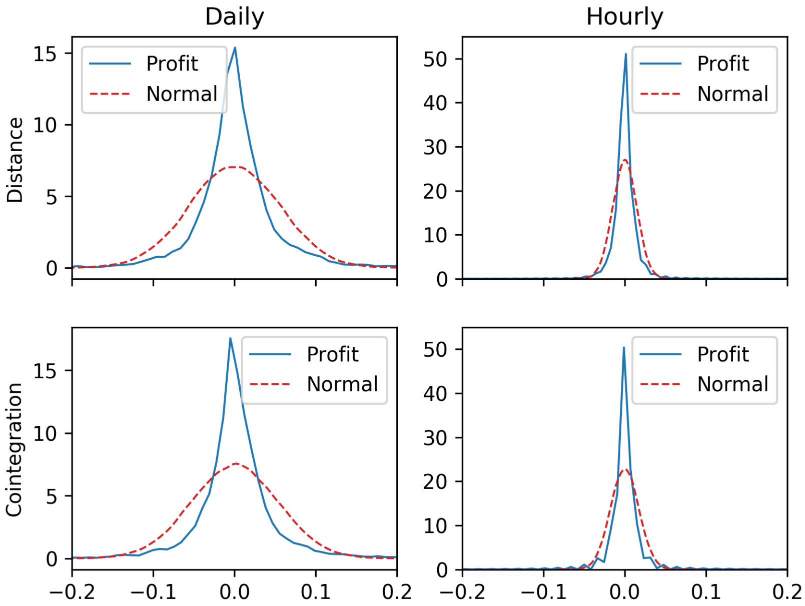


Figure 5: Price history of BTC/USDT on Binance

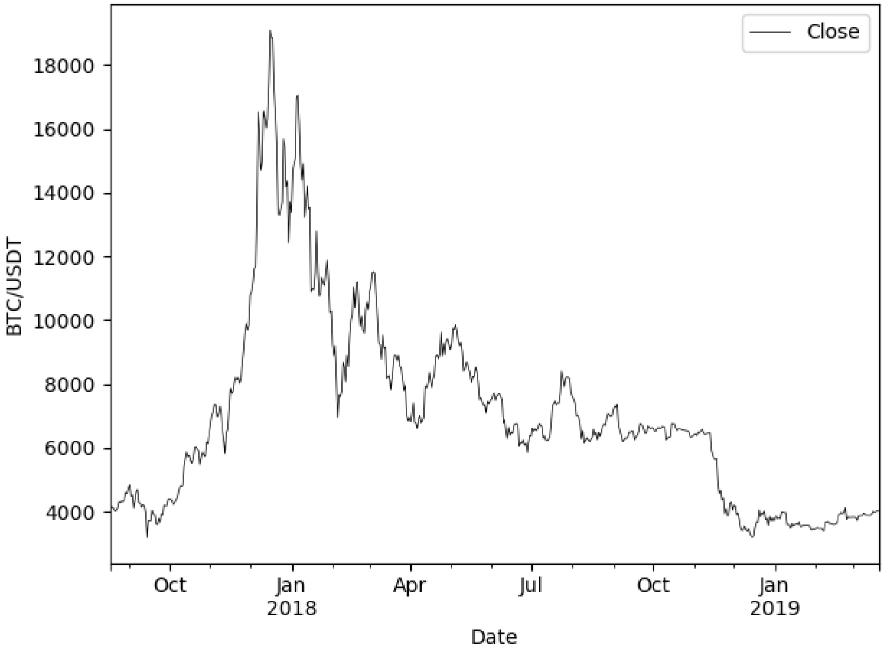
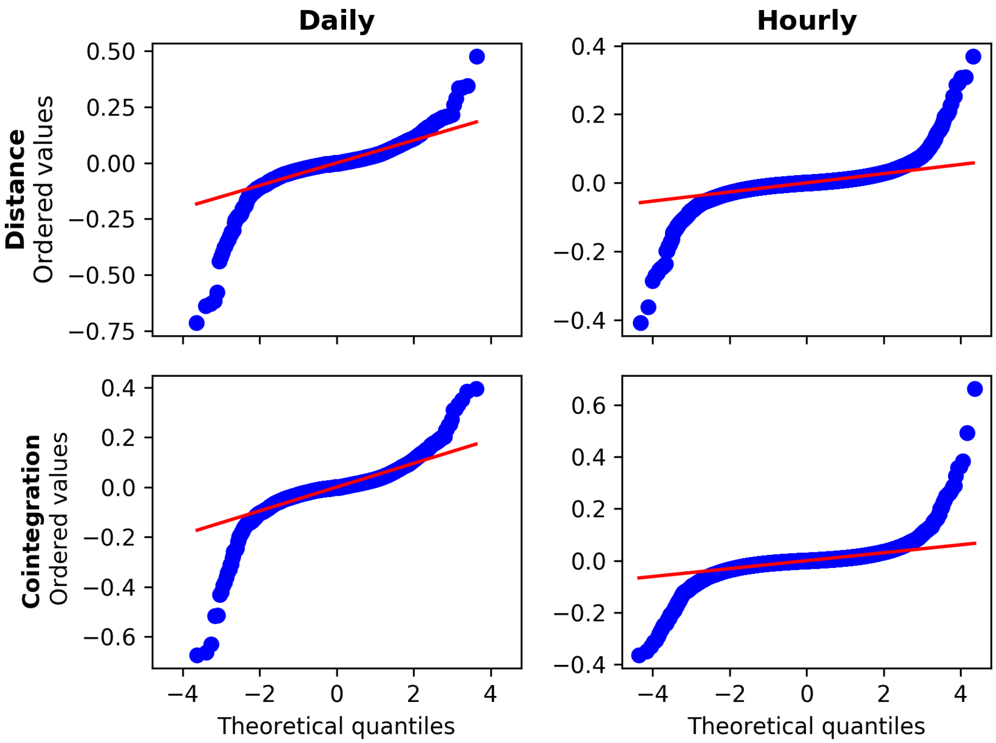


Figure 6: Q-Q plot of return distribution


Daily
Hourly
Distance
Ordered values
Cointegration
Ordered values
Theoretical quantiles
Theoretical quantiles