CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling

Bachelor's thesis

Author: Tomáš Barhoň
Study program: Economics and Finance
Supervisor: prof. PhDr. Ladislav Krištoufek Ph.D.
Year of defense: 2024

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, August 18, 2024

Tomas Barhon

Abstract

The abstract should concisely summarize the contents of a thesis. Since potential readers should be able to make their decision on the personal relevance based on the abstract, the abstract should clearly tell the reader what information he can expect to find in the thesis. The most essential issue is the problem statement and the actual contribution of described work. The authors should always keep in mind that the abstract is the most frequently read part of a thesis. It should contain at least 70 and at most 120 words (200 when you are writing a thesis). Do not cite anyone in the abstract.

JEL Classification	C01, G00, F23, H25, H71, H87
Keywords	Cryptocurrency, Bitcoin, Ethereum, Litecoin, Machine Learning, PCA, Noise Reduction
Title	Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling

Author's e-mail	`tomas.barhon@hotmail.cz`
Supervisor's e-mail	`ladislav.kristoufek@fsv.cuni.cz`

Abstrakt

Nutnou součástí práce je anotace, která shrnuje význam práce a výsledky v ní dosažené. Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). V abstraktu by se nemělo citovat.

Klasifikace JEL	C01, G00, F23, H25, H71, H87
Klíčová slova	KryptomÄ›ny, Bitcoin, Ethereum, Litecoin, StrojovÃ© uÄŤenÃ, PCA, Redukce Ĺˇumu
Název práce	Redukce šumu a extrakce rysů pomocí analýzy hlavních komponent pro modelování cen kryptoměn
E-mail autora	tomas.barhon@hotmail.cz
E-mail vedoucího práce	ladislav.kristoufek@fsv.cuni.cz

Acknowledgments

The author is grateful especially to prof. PhDr. Ladislav Krištoufek Ph.D..

Typeset in FSV LaTeX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Barhoň, Tomáš: *Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling.* Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 43. Advisor: prof. PhDr. Ladislav Krištoufek Ph.D.

Contents

List of Tables

List of Figures

Acronyms

BTC	Bitcoin
ETH	Ethereum
LTC	Litecoin
ML	Machine Learing
DL	Deep Learing
ANN	Artificial Neural Network
SGD	Stochastic Gradient Descent
LR	Linear Regression
SVM	Support Vector Machines
SVR	Support Vector Regression
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
PCA	Principal Component Analysis
SVD	Support Vector Decomposition
ARIMA	Autoregressive Integrated Moving Average
PoW	Proof of Work

Bachelor's Thesis Proposal

Author	Tomáš Barhoň
Supervisor	prof. PhDr. Ladislav Krištoufek Ph.D.
Proposed topic	Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling

Motivation Crypto assets have always been exceptionally volatile compared to traditional assets such as stocks or gold. The historical window is relatively short, thus modeling their price or volatility proposes quite a difficult challenge. It is generally believed that noise in any data decreases the precision of predictions. This effect might be reduced, which will improve the performance of traditional models that are used for cryptocurrency price modeling.

The main motivation for researching this topic is that there is still an ongoing discussion about the role of different features in crypto pricing dynamics. (Kukacka; Kristoufek 2023) have shown that a lot of the pricing dynamic emerges from complex interactions between fundamental and speculative components. They also show the different correlations between all of the explanatory variables which have a direct connection to principal component analysis. It is crucial to study the real impact of those variables in different models as many of them might turn out to be obsolete.

There is currently little use of this dimensionality reduction technique in the academic literature about cryptocurrencies. However, for more traditional financial series this technique is already quite established as a preprocessing technique to reduce noise and dimensionality from which financial data inherently suffer (Chowdhury, U. ; Chakravarty, S. and Hossain, M. 2018). Moreover (Bouri, E.; Kristoufek, L.; Ahmad, T. et al. 2022) studied the effect of microstructural noise on idiosyncratic volatility in cryptocurrencies which further supports the need for a technique that will mitigate this effect on the predictions.

The research will address the problem of variable selection for different types of predictive models with respect to the analysis of the principle components aiming to reduce the dimensionality and simultaneously increase precision. The second question is whether it is more appropriate to transform the high dimensionality with

PCA into lower dimensionality or simply omit the variables with high multicollinearity from the models. These approaches are fundamentally different and the answer is not clear.

Methodology The data will come from various sources because the aim is to look at all the possible variables even if they might not seem useful at first glance. As already mentioned the dynamic is driven by a lot of completely different effects. Most of it will be collected from: coinmetrics.io, studio.glassnode.com, and for macroeconomic indicators https://fred.stlouisfed.org/. Some of the data might need to be interpolated to daily observations. Lastly, the observations will need to be sliced to different window sizes and shifted by one so that the predictions can be made for the next day with the data available on that day.

Afterward, the multicollinearity in the data will be examined and different approaches to solve it will be used. The two main ones are using only a smaller set of uncorrelated variables (simple dimensionality reduction) and the second being employing PCA transformation to preserve a predefined threshold of variance or directly targeting the number of principal components.

All the different setups will be compared across three models: linear regression, SVM, and LSTM neural network. The hypothesis is that PCA transformation will substantially lower the measured errors for linear regression and SVM although for LSTM it will result in lower performance as it will only decrease the capacity of the model as the model is powerful enough to create such uncorrelated features without the PCA transformation.

Expected Contribution Existing research agrees that financial data and especially cryptocurrency data are significantly affected by noise. The main goal is to extend the research on the topic of variable selection for algorithmic trading models as there are still a lot of unanswered questions. It will most likely become clearer which approach to dimensionality reduction is the most efficient concerning cryptocurrencies.

Also, not only the underlying pricing dynamics will be detected but the results can be used for investors that are trying to lower their risk of loss which is relatively high in crypto markets. The effect of having a more stable and precise model might significantly cut the transaction costs that are associated with more frequent exchanges as the predictions will become less volatile. That is a desirable property needed to maximize profit and increase credibility towards its customers.

Core bibliography

Kukacka, J., & Kristoufek, L. (2023). Fundamental and speculative components of the cryptocurrency pricing dynamics (Vol. 9). Financial Innovation.

Kristjanpoller, W., & Minutolo, M. C. (2018). A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis (Vol. 109). Expert Systems with Applications.

Chowdhury, U. N., Chakravarty, S. K., & Hossain, M. T. (2018). Short-Term Financial Time Series Forecasting Integrating Principal Component Analysis and Independent Component Analysis with Support Vector Regression (Vol. 6).

Bouri, E., Kristoufek, L., , & Shahzad, S. J. H. (2022). Microstructure noise and idiosyncratic volatility anomalies in cryptocurrencies. Springer Link. https://doi.org/10.1007/s10479-022-04568-9

Rea, A., & Rea, W. (2016). How many components should be retained from a multivariate time series PCA?. arXiv preprint arXiv:1610.03588.

Author

Supervisor

Chapter 1

Introduction

Since the introduction of the first cryptocurrency Bitcoin (BTC) associated with the unknown author Satoshi Nakamoto (2008) cryptocurrencies have become part of our everyday life. Their high volatility, futuristic name and alternative nature are of interest to the media and the general public. According to **coinmarketcap.com** the overall cryptocurrency market capitalization peaked at around 2.8 trillion $USD in the year 2022 which makes them a substantial part of the financial sphere. The initial idea of BTC was to establish an alternative to traditional fiat currencies. The BTC whitepaper pointed out the weakness of the current trust-based model that relies on a third-party instance responsible for verifying transactions. A different approach was suggested to validate transactions known as the proof-of-work which utilizes the computational power of miners in the network. The fact that the power is distributed across the network ensures that it becomes exponentially harder with an increasing number of blocks to generate blocks faster than the rest of the miners (Nakamoto 2008, pg. 6). The mining process is interconnected with the creation of new BTCs which is a crucial parameter in all monetary systems. This fact gives researchers such as Kukacka & Kristoufek (2023) the possibility to use various attributes of the network to study the pricing dynamics of cryptocurrencies. On the other hand, there are a couple of substantial drawbacks that make price modeling relatively challenging. Those are non-stationarity of the target prices, relatively short historical window, the limited power of proxies for speculative components and as pointed out by many researchers such as Bouri *et al.* (2022), Dimpfl & Peter (2021), WÄ. . . torek *et al.* (2023) an idiosyncratic noise in volatility. Addressing these issues might potentially lead to better-performing models, especially with longer forecasting periods.

Likewise in other fields, the recent rise of machine learning has also affected the cryptocurrency area where various Machine Learing (ML) and Deep Learing (DL) models are often being used to model the price Khedr *et al.* (2021) or volatility Kristjanpoller & Minutolo (2018).

The main objective of this thesis is to try to tackle the problem of idiosyncratic noise in the high dimensional data used for price and returns modeling across three ML models: Ridge Linear Regression (LR), Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN). We will examine the effect of a method known as Principal Component Analysis (PCA) which was according to Farebrother (2022) developed in 1933 by Harold Hotelling. However, others often refer to the fact that the idea was already introduced before by Karl Pearson in the article *On lines and planes of closest fit to systems of points in space* Pearson (1901). This technique aims to compress data from a higher dimensionality space into a lower space while retaining a maximum amount of variance. It utilizes linear transformation of the covariance matrix to do that. Nevertheless, despite the initial focus on dimensionality reduction different types of PCA are often being used as noise reduction techniques in signal or image processing. Interestingly many studies in recent years have incorporated PCA for time series data as a part of their preprocessing pipeline Chowdhury *et al.* (2018), Kristjanpoller & Minutolo (2018). The idea stems from the fact that removing the most idiosyncratic components might help with capturing clear dynamics that enter the price-making process. We perceive that there is currently a lack of literature that would examine the effects of noise reduction techniques on the performance of other ML based regression techniques for cryptocurrencies. We want to mitigate most of the identified challenges using the currently available academic knowledge and focus exclusively on the effect of noise in the data. Admittedly it is always intricate to establish a ceteris paribus relationship in such a scenario where many variables change, the randomness of the training process using Stochastic Gradient Descent (SGD) plays a crucial role and the size of the dataset is relatively limited. We want to contribute with an alternative approach, especially in the preprocessing pipeline that can be used in future studies to decrease the volatility of predictions. We do not aim to provide a generally applicable approach, as different techniques can produce varying outcomes on different datasets. This phenomenon partially corresponds to the *No Free Lunch Theorem* Wolpert *et al.* (1995) which has turned into a buzzword in the ML community over the years.

The remainder of the thesis is organized as follows: The following chapter introduces the fundamentals of cryptocurrencies and their unique characteristics. It also covers the usage of ML methods in this field and especially focuses on the literature about the usage of PCA in various areas. The data chapter explains in detail which data were used and elaborates on the basic resampling methods that we used. In methodology, we focus on each specific ML method and explain the core concepts that are crucial for understanding the training process. Similarly, we propose our complete forecasting framework. Chapter results and discussion evaluates the findings for each currency-model pair across different settings. We also include a limitations section which is especially crucial for our study where we acknowledge those problematic parts of our approach that might be improved in the future. The conclusion focuses on the overall impact and proposes paths that should be explored in the years to come. All the tables, source codes and visualizations can be found in the appendices.

Chapter 2

Literature Review

2.1 Cryptocurrencies

2.1.1 Bitcoin

In the year 2008, an unknown author with the pseudonym Satoshi Nakamoto introduced the idea of a purely peer-to-peer electronic cash system. Interestingly the author mentions small casual transactions as something that the current model relying on third-party financial institutions fails to deliver because of unavoidable transaction costs Nakamoto (2008). In contrast, from today's perspective, Bitcoin is a relatively slow medium for micro-transactions because technically the receiver has to always wait for a certain amount of blocks to be mined such that it becomes statistically unlikely that double-spending has been committed by the payer Conti *et al.* (2018). This phenomenon can be demonstrated on the data from **coinmetrics.io** which show that the mean size of a BTC transaction ranges in thousands of USD$. Another important aspect is that the miners prioritize transactions with higher fees in the block which introduces considerable costs to each payment. MĂ¶ser & BĂ¶hme (2015) have shown the relationship between the transaction fee and the transaction latency meaning the time it takes for the transaction to be almost surely valid. Even though there has been some divergence from the original idea of small transactions all of the security measures regarding the double spending problem in the original whitepaper have turned out to be relatively well-defined in a medium time horizon.

Despite the fact, that Bitcoin is generally regarded as the first cryptocurrency it relies on many older ideas and technologies that are mostly mentioned in the original whitepaper. First and foremost stands the conference paper *How*

to Time-Stamp a Digital Document Haber & Stornetta (1991) which focuses on the problem of third parties responsible for verification of digital documents. It makes use of an already established family of functions known as hashes that surpass privacy concerns and generally surpass the need for a third party to be involved in the verification process when combined with the correct consensus algorithm. They define a hash function as follows:

Definition 2.1 (Hash). This is a family of functions $h : \{0,1\} \rightarrow \{0,1\}^l$ compressing bit-strings of arbitrary length to bit-strings of a fixed length l, with the following properties:

1. The functions h are easy to compute, and it is easy to pick a member of the family at random.

2. It is computationally infeasible, given one of these functions h, to find a pair of distinct strings x, x' satisfying $h(x) = h(x')$. (Such a pair is called a collision for h)(Haber & Stornetta 1991, see Chapter 4.1)

And suggest hashing documents together with the time of their creation. However, the time-stamping might fail if the users can tweak the time of their machines. Interestingly, the authors have already mentioned that and introduced the idea of chaining the data together with their metadata sequentially in a long chain so that the user can trust that something was not overwritten (Haber & Stornetta 1991, see Chapter 5). This is possible due to the properties of the hash functions. Meaning that we can concatenate arbitrarily long inputs and always produce a fixed-size output. Another significant influence came from b-money which was an idea for an anonymous digital cash system presented in Dai (1998). B-money proposed the concept of Proof of Work (PoW) which is a validation protocol that many cryptocurrencies still use. The idea was to solve computationally challenging puzzles where it can be determined how much effort was used to do so (see Dai 1998, pg. 1). However, certain worries were being raised about how to regulate a system if the computational power of computers is increasing every year (see Dai 1998, pg. 3). This has been addressed by Bitcoin with the regulation of difficulty based on the average time it takes to solve the puzzle rather than the difficulty itself (see Nakamoto 2008, pg. 3).

Since many characteristics of the network are often being used by researchers such as: (Kukacka & Kristoufek (2023), Kristoufek (2023), Kubal & Kristoufek (2022) or Jay *et al.* (2020)) in their research, it is critical to understand the

underlying mechanics that form them. Bitcoin takes a completely adverse approach to the general financial system. Whereas traditionally banks and other institutions try to keep every transaction encrypted Bitcoin makes all the transactions publicly visible and available but hashing the addresses of the sender and receiver. The process of sending Bitcoins to someone else essentially means adding a digital signature to the previous transaction from which you received that money. However, as (see Chapter 2 Nakamoto 2008, pg. 2) suggests this only mitigates the privacy concerns but the double spending risk needs to be dealt with a smarter design. This is solved by the introduction of the PoW algorithm. The idea is that transactions are collected into blocks by the miners who try to solve a computationally difficult task that can only be solved by a brute-force search. It is simply a race to find a hash with a certain amount of leading zeros which is adjusted based on the mining power of the network. The miners are incentivized by BTC price for winning the race and also by the transaction fees that can be added to each transaction as a reward for being prioritized. The leading concept is that the blocks are connected sequentially in a chain through the hash. If we assume that most of the nodes/miners are honest their profit-maximizing behavior should always be working on the longest chain and thus transactions that have already been spent do not get included in the chain. After the puzzle is solved by a node it can be validated by all other nodes in a linear time and they move on to the next block. Despite that, there remains the risk of an attacker forking a malicious block and sending his money back or elsewhere. Nakamoto (2008) claims that the probability of an attacker catching up (or reaching breakeven from the memoryless property of Poisson distribution) drops exponentially with each block if the mining power (probability of solving the puzzle) of the attacker is lower than the power of honest nodes. This mechanism is the root of the Bitcoin security. However, it also implies that there is an implicit tradeoff between security and the desired liquidity of cash. Another important property is that there will ever exist only a limited amount of 21 million of BTCs which makes it inherently a deflationary currency at least after all of the BTCs are mined. Technically there will be less as some wallets do get lost together with their contents. Limiting supply might be an intentional design choice to contrast the traditional model where banks issue money and cause inflation.

As cryptocurrencies are a relatively new phenomenon they are currently a frontier topic of academic research in many different aspects. They are being studied on multiple levels such as law, technology, cryptography, security,

economics or machine learning. Generally, the area of economics and machine learning will be of interest as we want to uncover whether there exist some determinants of the bitcoin price or at least features that can be used to estimate the pricing dynamics. We assume that there are theoretically three simplified possibilities of the pricing model of BTC and other cryptocurrencies. Firstly, it might be a purely efficient market and the price technically follows a random walk process with a potential drift. The second model is that the price is entirely driven by speculative components and lastly, it might be a combination of speculative and fundamental components.

2.1.2 Ethereum

At the time of writing according to the data from **coinmarketcap.com**, Ethereum (ETH) is the second cryptocurrency based on the market capitalization standing at 318 billion USD. Although, there are a lot of similarities with BTC the idea behind ETH is much more profound and builds an entire technological infrastructure on top of blockchain which provides easily scriptable smart contracts. Tikhomirov (2018) defines smart contracts as follows:

> *Turing-complete programs that are executed in a decentralized network and usually manipulate digital units of value.*

We might want to reformulate this definition for the purposes of this thesis to cover a broader meaning.

Definition 2.2 (Smart contract). A program that is running on the blockchain infrastructure as an endpoint with a specific address that other entities on the platform can interact with in order to execute a transaction for a cost proportional to the number of computational steps. The contract usually has a predefined set of rules which is applied to the input data and automatically executes on hold whenever called. This allows developers to build applications on top of blockchain infrastructure that is run by the distributed network and grants them the unique benefits of the blockchain model.

According to the original whitepaper, the main purpose of ETH was not to create another cryptocurrency but build a simple-to-use scripting language that would allow developers to create custom applications running on the blockchain that share the benefits of the distributed nature of the system but reduce the need for hardware as the transactions are handled by the network and also

software as the transactions can be defined in a few lines of code. Notably it describes cryptocurrencies as a state transition system:

Definition 2.3 (Digital currency ledger). From a technical standpoint, the ledger of a cryptocurrency such as BTC can be thought of as a state transition system, where there is a *state* consisting of the ownership status of all existing coins and a *state transition function* that takes a *state* and a *transaction* and outputs a new state *state-new* which is the result.

In this view, we can understand building smart contracts, see Definition 2.2, as creating use case specific *state transition functions* ((Buterin *et al.* 2013, see Chapter Bitcoin As A State Transition System),(Tikhomirov 2018, see Chapter 2)).

We can observe the fundamental difference in philosophy between the BTC and ETH from their respective whitepapers. ETH is much more focused on its role as an application platform whereas BTC was mainly intended to work as a currency. The ETH whitepaper even presents many ideas for future applications that might benefit from this framework. This fact is crucial in understanding the underlying price-making mechanics as investments into ETH might be affected by its perceived potential as a technological product rather than a typical currency that might be valued mostly as a medium of exchange.

Even though there are a lot of similarities in the blockchain architecture, there have been many functional and implementation differences. The most notable is the ETH scripting language which is a Turing-complete counterpart to the BTC scripting language that did not support infinite looping (Buterin *et al.* 2013, see Chapter Scripting). The second fundamental implementation variation is that each Ethereum block saves the entire *state* of the whole network and thus does not need to store the whole history of the blockchain. On the contrary, BTC is doing exactly that, as the blocks are state-unaware and only validate blocks based on the transactions and the wallet software usually calculates the balances (Tikhomirov 2018, see Chapter 2). As Buterin *et al.* (2013) pointed out there is a workaround that makes this implementation roughly equally efficient which utilizes the propagation properties of the Merkle tree with the fact that only a small part of the state changes with each block allowing for efficient state change using pointers to the specific branches and leaves of this data structure, see Figure 2.1. The data are stored at the bottom of this infrastructure and there is an efficient algorithm that allows the ETH software to reference the affected addresses and paths leading to them.

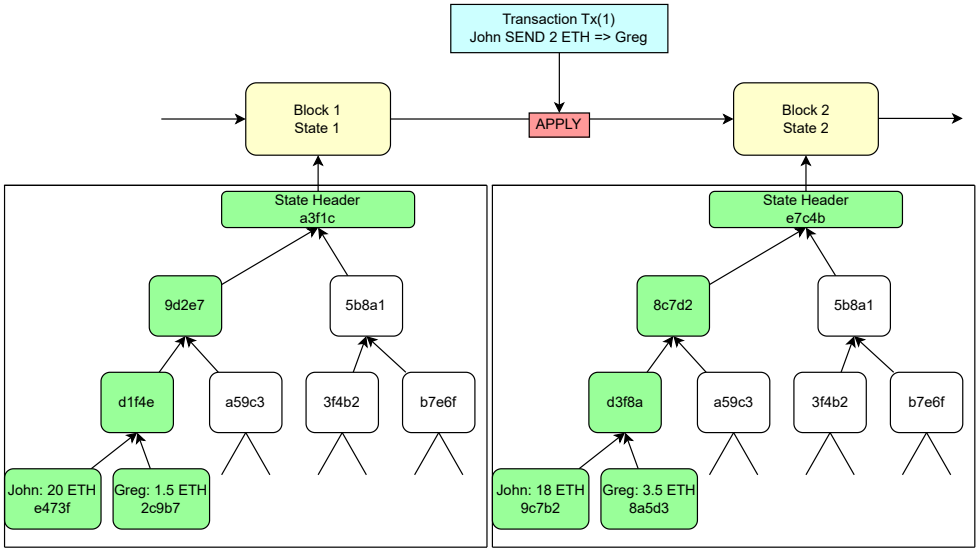


Figure 2.1: Merkel tree with pointers allows efficient state change

Lastly, there is a difference in the supply of new coins. Contrasting the model of BTC where the supply is limited ETH introduces a model of an infinite linear supply of coins to provide incentives for future users to join the network as they might still obtain new coins and thus limit the wealth concentration common in BTC (Buterin *et al.* (2013), Tikhomirov (2018)). Note that ETH already switched to the proof-of-stake model in September 2022 but our dataset does not include this period and thus this fact is not especially relevant to this thesis.

On the outside ETH acts similarly to BTC. It is a ledger that stores the coin balances of accounts where each has a designated unique address. There are two types of accounts: externally owned accounts which are essentially the typical users and contract accounts which are the abstraction on top of which smart contracts can be built with contract code that executes when the account receives a message from another account (externally owned or a different contract account). Because of the presence of infinite loops in the scripting language ETH employs a strategy that prevents users from essentially exploiting the Denial-of-service attack. Nonetheless stands the Halting problem. That can be simplified to the fact that for a Turing-complete model, there is no way of saying ex-ante whether the program will halt or run indefinitely. As Lucas (2021) suggests the Halting problem is usually attributed to Alan Turing's paper (Turing *et al.* 1936, On computable numbers, with an application to

the Entscheidungsproblem), however, the problem was reformulated in various forms by others. This implies that this undecidability also holds for any ETH contract code. This is fixed by introducing a gas currency that acts as a cost of computation and the maximum has to be predefined in each message so that the recipient knows what is at stake Buterin *et al.* (2013). We can think of this as a type of timeout based on a currency. If gas runs out all of the state changes are reverted. This also explains the origin of the term transactions which typically refers to a set of instructions in SQL or other databases that are bundled together and executed in an all-or-nothing fashion Kleppmann (2017). The last important fact, as the whitepaper describes, is that the contract code is run by all of the miners verifying the block which essentially means applying all of the transactions and reverting in case of an error.

2.1.3 Litecoin

In comparison to ETH, the goal of Litecoin (LTC) was pronounced from the beginning as building a better version of BTC that shines where BTC has failed. We might say that LTC is a tweaked version of BTC with different parameters or a hard fork of the BTC protocol. To our best understanding, the only document that is wildly considered the original whitepaper is a transcript of a forum post by the founder Charlie Lee where he suggests reading also the BTC whitepaper. Essentially, two main differences address the problems BTC embodies. The first one is faster confirmation time that allows LTC to be used truly in a fashion that was intended for BTC as digital cash with high liquidity sacrificing a bit of security. Achieving that mostly through four times faster block generation. And second one is a different proof-of-work algorithm. As Padmavathi & Suresh (2018) suggests the intention was most likely since BTC mining was dominated by GPU and ASIC miners which led to a concentration of mining power in pools and thus more centralized distribution of BTC. Despite the initial promises Litecoin has most likely not fulfilled its envisaged role and does not currently belong even to the ten most popular cryptocurrencies by market capitalization.

2.2 Machine Learning Methods for Cryptocurrencies

Thanks to transformer-based architectures that are the backbone of most chatbots. ML and artificial intelligence have gotten a lot of public spotlight during the last two years. However, ML methods have been especially prevalent in research in the last decade. Particularly in data-driven fields of academia, there have been various use cases where ML shines and outperforms traditional statistical models. Forecasting has always been an area of interest in many different fields as the idea of predicting the future based on historical data provides intrinsic value in itself. The field of cryptocurrencies is no exception as the significant volatility is a thought-provoking problem to tackle and an opportunity to win against the rest of the market.

In order to use ML or any other data-driven method we implicitly have to assume that there are some underlying dynamics of which we can make sense. This argument is hard to make without the proper data and might have to be studied separately for different time horizons. Thankfully, Kukacka & Kristoufek (2023) argue that some cryptocurrencies especially BTC are driven by the interaction of speculative and fundamental components and thus provide an incentive for us to untangle and study these relationships. Disturbingly, even though there are many other papers focusing directly on the practical implementation of forecasting frameworks for cryptocurrency prices a significant amount of them do not compare their model to a random walk or other simple statistical model. We believe that their approaches should not be condemned but we strongly emphasize that their results should be interpreted cautiously.

We believe that there is currently an upsetting trend in the studies focusing on modeling the price or returns of cryptocurrencies using ML methods. There seem to be a lot of inconsistencies in terms of splitting the data into training and testing sets, making the results robust using some forms of cross-validation, comparably presenting the results and contrasting the models to a meaningful baseline model. This fact arguably contributes to the fact that the results of most of the studies are relatively underwhelming and have stagnated in the last few years. The other side of the coin, arguably worse, are studies that present overly optimistic results. We propose an idea for future research that would imaginably help the field advance further and stimulate innovation. The suggestion lies in the creation of a standardized dataset with set splits

between train, validation and test data that would allow for the comparison of different approaches and would encourage researchers to make their models generalizable. Taking inspiration from the field of image recognition where the Image-net is a state-of-the-art dataset to compare models for image recognition. This would allow for a competitive environment boosting innovation and development. We acknowledge that there is a fundamental difference between a dataset that consists of an unordered series of images and an especially challenging non-stationary time-series. Some compromises would undeniably have to be introduced in terms of set prediction horizons, different data granularity and artificially set splitting points. However, there is always a place for variation and this dataset could have multiple versions. Even though, this solution is sub-optimal we firmly believe the current scattered state of knowledge in this field makes it extremely difficult for further development to blossom.

Despite that, we would like to provide a brief overview of the methods that are often incorporated in price or returns modeling pipelines. As we already mentioned in order for ML to work as a forecasting algorithm there has to exist a possibility of drawing information about the future from the past. That is a non-trivial assumption because it contradicts The Efficient market hypothesis that was formulated for capital markets by Miller *et al.* (1970). It is thus no oddity that Ren *et al.* (2022) have found out that across 395 scientific articles about the use of ML in cryptocurrencies the most cited article was Urquhart (2017) that studied the efficiency of BTC and concluded that the BTC market is inefficient but may become efficient in the future. Furthermore Ren *et al.* (2022) found out that the keyword inefficiency was the one with the highest burst of emergence.

Khedr *et al.* (2021) pointed out that cryptocurrencies lack seasonal trends which makes them challenging to predict for traditional statistical models. They also distinguish four types of factors influencing cryptocurrency price. Namely: demand and supply, crypto market, macro-economic and political. Our work will cover the first three types of factors as the political factors are hard to quantify. Later Ren *et al.* (2022) suggests that future researchers should study measurement tools for political factors. If you want to find out more about research in this field from 2010-2020 please view Khedr *et al.* (2021) which provides a comprehensive review. However, as we suggested earlier the aspect of comparison between studies is rather limited.

2.3 Principal Component Analysis

2.3.1 PCA in Other Areas

2.3.2 PCA in Time Series

2.4 Web Search Data in Financial Applications

Chapter 3

Data

3.1 Cryptocurrency Specific Technical Data

3.2 Macroeconomical Data

3.3 Web Search Data

3.4 Preprocessing

Text text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text. Text text **?**.

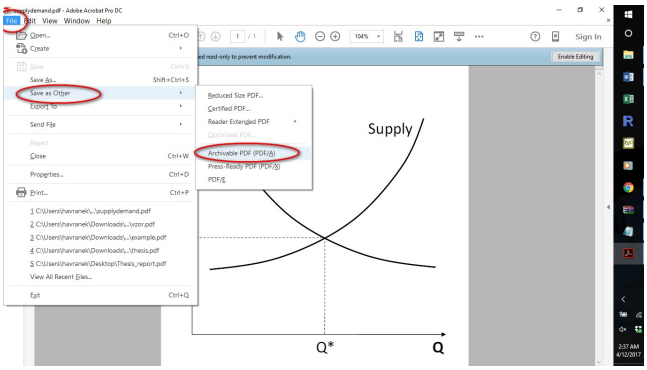
Text text text text text text text text text text text text text text text text. Text text text text text text text (see, *inter alia*, **?**, pg. 10).

Text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text. Text text text text text text. Politicians usualy like inward BTC and an **MNC!** (**MNC!**) appreciates **FDI!** (**FDI!**) subsidies. Are **MNC!**s greedy?

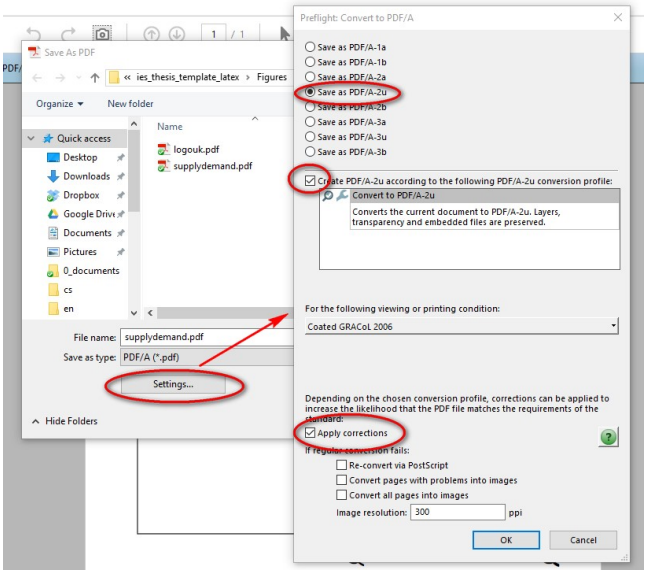
To achieve compatibility with PDF/A 2u, your file must not include links to external fonts, audio, video, or scripts. On the other hand, your file must declare each color environment you use, it must include all the pictures/figures either in jpeg or PDF/A 2u format, used fonts compliant under Unicode (your file cannot use any external fonts), and it must include meta-data in XMP format.

Most troubleshooting comes from the conversion of figures to compliant formats. You can convert from simple PDF using Adobe Acrobat:

- Select File » Save as Other » Archivable PDF (PDF/A)



- Save as PDF/A-2u:



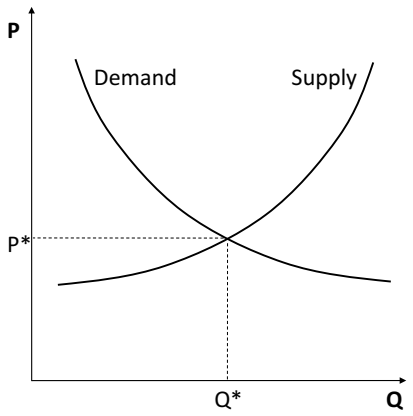
But most of the vector graphics gets distorted to lower quality in Adobe (like pictures in pdfs generated from Stata, unless jpeg is sufficient for you). You can also use GhostScript, the conversion tool is provided by courtesy of the Faculty of Mathematics and Physics at

https://kam.mff.cuni.cz/pdfix/

Text text text text text text.[1] Font of Latin phrases should be consistent: Furthermore, there is no *ex post* price effect, all things being equal (*ceteris paribus*). This is *per se* truth.

[1]Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text.

Figure 3.1: Market equilibrium



Source: **?**.

Look at the Figure 3.1. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

If you use Stata, you might want to check the `sutex`, `outtable`, `outtex`, and `estout` tools, which help you with exporting Stata tables to LaTeX.

Table 3.1: Model's predictions

Case	Y_1	Y_2	τ_1	τ_2	a	n
CR—Slovakia	10.9	10	0.24	0.19	1,000	2.16
CR—Poland	13.3	12	0.24	0.19	1,000	0.38
CR—Hungary	10.4	8	0.24	0.16	1,000	1.10

Source: If the source is author himself (like a calculation output), this line is redundant.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text text text text text.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text text text text text. Let us make a box:

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text.

Figure 3.2: Boxy's example

- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.

Source: **?**

Text text text text text text text text text text. Text text text text text text text text text text.

Definition 3.1 (My original definition). This is a definition.

Assumption 3.1 (My realistic assumption). This is an assumption.

Proposition 3.1 (My clever proposition). *This is a proposition.*

Lemma 3.1 (My useful lemma). *This is a lemma.*

Example 3.1. This is an example.

Proof. This is a proof. □

Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

$$U = \underbrace{\int_0^\infty \frac{1}{1-\sigma}\left(C^{1-\sigma}-1\right)e^{-\rho t}\,\mathrm{d}t}_{\text{meaning of life}}$$

Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

$$U = \int_0^\infty \overbrace{\frac{1}{1-\sigma}\left(C^{1-\sigma}-1\right)}^{\text{instantaneous utility}} e^{-\rho t}\,\mathrm{d}t \tag{3.1}$$

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text text text text.

$$\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C} \tag{3.2}$$

- to literature (**?**, pg. 10) or **?**, pg. 10,
- to Figure 3.1,
- see Table 3.1,
- to **??**,
- to Definition 3.1, to Proposition 3.1, Example 3.1,
- to equations like this: see (3.1).

You can input a source code like this:

```
omega = 1;
syms zeta;
jmn = [1 2*zeta*omega omega^2];
figure(1);
    for zeta = 1E-5 : 0.2 : 1+1E-12
        G = tf(omega^2,subs([1 2*zeta*omega omega^2]));
        bode(G); hold on;
    end
legend('\zeta = 0','\zeta = 0,2','\zeta = 0,4','\zeta = 0,6',');
```

Should you prefer a different font size, redefine file `Styles/Mystyle.sty`.

Usually you should not use the first person singular (I) in your text, write we instead. As a general recommendation, use the first person sparsely, sometimes it can be replaced by a phrase like "This work presents ...."

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text (**?**). Let us make two paragraphs:

Proin Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text. Text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. And a subparagraph:

Velit Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text text text text.

Chapter 4

Methodology

4.1 Machine Learing

4.2 Ridge Linear Regression

4.3 Support Vector Machines

4.4 Long Short-Term Memory Recurrent Neural Networks

4.5 Principal Component Analysis

4.6 Time-Series Specifics

4.7 Proposed Forecasting Framework

Many people use simple n-dash in many occasions – like this –, where however typographic convention—it looks a bit strange at first sight—requires m-dash. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text ?.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text ?. Let us describe the following animals:

Item 1 Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text.

Item 2 Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text.

Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. See what Edmund Burke said about the duties of a Member of Parliament (Speech To The Electors Of Bristol At The Conclusion Of The Poll, November 3, 1774):

> It ought to be the happiness and glory of a representative to live in the strictest union, the closest correspondence, and the most unreserved communication with his constituents. Their wishes ought to have great weight with him; their opinion, high respect; their business, unremitted attention. It is his duty to sacrifice his repose, his pleasures, his satisfactions, to theirs; and above all, ever, and in all cases, to prefer their interest to his own. But his unbiased opinion, his mature judgment, his enlightened conscience, he ought not to sacrifice to you, to any man, or to any set of men living. These he does not derive from your pleasure; no, nor from the law and the constitution. They are a trust from Providence, for the abuse of which he is deeply answerable. Your representative owes you, not his industry only, but his judgment; and he betrays, instead of serving you, if he sacrifices it to your opinion.

Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text.Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text.

(i) The first item, the first item, the first item, the first item, the first item, the first item,

(ii) and the second item.

(a) The first item, the first item, the first item, the first item, the first item, the first item,

(b) and the second item.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text **?**.

Chapter 5

Results and Discussion

5.1 Results Intepretation

5.1.1 Basline Autoregressive Integrated Moving Average

5.1.2 Linear Regression

5.1.3 Support Vector Machines

5.1.4 Long Short-Term Memory

The following checklist should help in avoiding some frequently made mistakes, if any of the following propositions apply for your thesis, there is a problem:

- You have citations in your abstract.
- The introduction does not cover the three parts as described in Chapter 1.
- The introduction contains subheadings.
- You described different aspects than promised in the title.
- You copied some parts of the text from other work without proper referencing and citing.
- You used automatic translation tools to produce text by translating it from another language.
- Your thesis contains many typos and grammatical errors. (Use an electronic spell checker. Please!)

- You used color in your figures and refer to the “blue” line (assume that your readers use a monochrome printer).
- You mainly used websites and other unrefereed material as your sources or you used Wikipedia as your source.
- You refer to something in your conclusion which you have not mentioned before.
- Some forenames in the references are abbreviated, some not.
- Some references miss a publishing date.

5.2 Limitations

Chapter 6

Conclusion

If you write in English, you might find the following hint useful: The indefinite article a is used as an before a vowel sound—for example an apple, an hour, an unusual thing, an (because the acronym is pronounced Em-En-See). Before a consonant sound represented by a vowel letter a is usual—for example a one, a unique thing, a historic chance. Few more tips to follow:

- Don't give orders—don't write in the imperative mood—unless you are training to be a teacher.
- Avoid the use of questions. You may know the answer: does your reader? It's much safer to tell her, or him.
- Do not become entangled in the problems of 'sexist' language. It is much easier to write in the plural. "Students should check their work" is good English. "A student should check—" is also good English, but now the problems begin: "—her work?" "—his work?" Which? You can write "his or her," but that seems clumsy. Stick to the plural.
- If you must refer to yourself, use the third person such as "The present writer would recommend that . . . " may be useful.
- Use the full forms of words and phrases, not contractions like "he's," "don't," etc. Keep the apostrophe to indicate possession—and use it correctly. Academics really sneer at students who use the "Greengrocer's apostrophe."
- Do not despise short, workmanlike, and effective plain English words. If they mean what you want to say. Accurately.

- Avoid the use of humor in academic writing—unless you are very sure of yourself.
- Even when you are not being funny, avoid the use of irony or sarcasm.
- Paragraphs in academic English should contain more than one sentence. (Short paragraphs look as if you are writing for a tabloid newspaper—or a simple Template!) I guess that the average academic book runs to two or three paragraphs per page. Look at the books in your subject, and get a feel for how long your own paragraphs should be when you are imitating the academic style.
- Use the word that more in formal writing than most of us do in speech—particularly after such verbs of utterance as to say, to report, to think etc. It can help to make your writing much clearer.
- Develop an academic vocabulary. The 'long words' you learn in the course of your studies are long usually because they have more precise meanings than their less formal equivalents. They are therefore better when you want to be accurate. (Also they allow you to sound like someone who deserves a degree.)
- Use as few words as you can; but use enough words to express your meaning as fully as you can. Your judgment of what is appropriate here is part of what you should learn throughout your course.
- Avoid lazy words such as "nice". It is usually better to say "acquire" or "obtain" than "get;" and it may be better, if you mean "through the use of money," to say "purchase" or—better still—"buy."
- A short word like "buy" is better than a long one like "purchase"—unless the long one is more accurate. A "statutory instrument" is better than a "rule"—to a lawyer, at any rate.
- Proof-read with care. Ask someone else to help—you may be too close to your work to be able to see your mistakes.
- If in doubt, choose the more formal, or possibly just the more old-fashioned, of two words. For example, say quotation rather than quote whenever you mean the use of somebody else's words.

- You will often sound more academic if you include doubts in your work—and qualifications. Within the scope of this thesis, the current writer cannot hope to cover all the possible implications of the question.ÔŤ

- In this context, the use of litotes sounds very academic. This is the construction where a writer uses a negative with a negative adjective, e.g. it is not unlikely that ... This does not mean the same as it is probable that ... It has a shade of meaning and qualification that can be useful to academic writers.

Bibliography

Bouri, E., L. Kristoufek, T. Ahmad, & S. J. H. Shahzad (2022): "Microstructure noise and idiosyncratic volatility anomalies in cryptocurrencies." *Annals of Operations Research* .

Buterin, V. *et al.* (2013): "Ethereum white paper." *GitHub repository* **1**: pp. 22–23.

Chowdhury, U. N., S. K. Chakravarty, & M. T. Hossain (2018): "Short-term financial time series forecasting integrating principal component analysis and independent component analysis with support vector regression." *Journal of Computer and Communications* **06(03)**: pp. 51–67.

Conti, M., E. Sandeep Kumar, C. Lal, & S. Ruj (2018): "A Survey on Security and Privacy Issues of Bitcoin." *IEEE Communications Surveys amp; Tutorials* **20(4)**: pp. 3416–3452.

Dai, W. (1998): "B-money."

Dimpfl, T. & F. J. Peter (2021): "Nothing but noise? price discovery across cryptocurrency exchanges." *Journal of Financial Markets* **54**: p. 100584.

Farebrother, R. W. (2022): "Notes on the prehistory of principal components analysis." *Journal of Multivariate Analysis* **188**: p. 104814.

Haber, S. & W. S. Stornetta (1991): *How to Time-Stamp a Digital Document*, pp. 437–455. Springer Berlin Heidelberg.

Jay, P., V. Kalariya, P. Parmar, S. Tanwar, N. Kumar, & M. Alazab (2020): "Stochastic neural networks for cryptocurrency price prediction." *IEEE Access* **8**: pp. 82804–82818.

Khedr, A. M., I. Arif, P. R. P V, M. El-Bannany, S. M. Alhashmi, & M. Sreedharan (2021): "Cryptocurrency price prediction using traditional

statistical and machine-learning techniques: A survey." *Intelligent Systems in Accounting, Finance and Management* **28(1)**: pp. 3–34.

Kleppmann, M. (2017): *Designing data-intensive applications*. Beijing: O'Reilley, first edition edition. Hier auch spĂ¤ter erschienene, unverĂ¤nderte Nachdrucke.

Kristjanpoller, W. & M. C. Minutolo (2018): "A hybrid volatility forecasting framework integrating garch, artificial neural network, technical analysis and principal components analysis." *Expert Systems with Applications* **109**: pp. 1–11.

Kristoufek, L. (2023): "Will Bitcoin ever become less volatile?" *Finance Research Letters* **51**: p. 103353.

Kubal, J. & L. Kristoufek (2022): "Exploring the relationship betweenb Bitcoin price and networkâ€™s hashrate within endogenous system." *International Review of Financial Analysis* **84**: p. 102375.

Kukacka, J. & L. Kristoufek (2023): "Fundamental and speculative components of the cryptocurrency pricing dynamics." *Financial Innovation* **9(1)**.

Lucas, S. (2021): "The origins of the halting problem." *Journal of Logical and Algebraic Methods in Programming* **121**: p. 100687.

Miller, C. N., R. Roll, W. Taylor *et al.* (1970): "Efficient capital markets: A review of theory and empirical work." *The journal of Finance* **25(2)**: pp. 383–417.

MĂ¶ser, M. & R. BĂ¶hme (2015): *Trends, Tips, Tolls: A Longitudinal Study of Bitcoin Transaction Fees*, pp. 19–33. Springer Berlin Heidelberg.

Nakamoto, S. (2008): "Bitcoin: A peer-to-peer electronic cash system." *Decentralized business review* .

Padmavathi, M. & R. M. Suresh (2018): "Secure P2P Intelligent Network Transaction using Litecoin." *Mobile Networks and Applications* **24(2)**: pp. 318–326.

Pearson, K. (1901): "Liii. on lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2(11)**: pp. 559–572.

Ren, Y.-S., C.-Q. Ma, X.-L. Kong, K. Baltas, & Q. Zureigat (2022): "Past, present, and future of the application of machine learning in cryptocurrency research." *Research in International Business and Finance* **63**: p. 101799.

Tikhomirov, S. (2018): *Ethereum: State of Knowledge and Research Perspectives*, pp. 206–221. Springer International Publishing.

Turing, A. M. *et al.* (1936): "On computable numbers, with an application to the Entscheidungsproblem." *J. of Math* **58(345-363)**: p. 5.

Urquhart, A. (2017): "Price clustering in bitcoin." *Economics Letters* **159**: pp. 145–148.

Wolpert, D. H., W. G. Macready *et al.* (1995): "No free lunch theorems for search." *Technical report*, Citeseer.

WÄ. . . torek, M., M. SkupieĹ„, J. KwapieĹ„, & S. DroĹĽdĹĽ (2023): "Decomposing cryptocurrency high-frequency price dynamics into recurring and noisy components." *Chaos: An Interdisciplinary Journal of Nonlinear Science* **33(8)**.

Appendix A

Detailed Results Tables

Appendix B

Aditional Contents

All of the source codes and data to reproduce the results are available at https://github.com/Tomas-Barhon/Noise-reduction-and-feature-extraction. Including all the instructions on how to install the necessary dependencies.