Cryptocurrency Correlation Modeling with Multivariate GARCH

Isak Ottosson, Nawa Rauf

Master's thesis
2024:E37

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

This thesis investigates the dynamic correlation between the S&P 500 and Bitcoin returns from 2018 to 2024. To capture potential regime-specific dynamics, a Markov Switching Model is employed to segment each return data into two distinct states characterized by high or low volatility, using a LASSO regression to find the appropriate explanatory variables for the returns.

Univariate GARCH(1,1) models are initially fitted to each regime's return series. Using the parameters from the univariate models, the multivariate models are estimated for each combination of high and low volatility regimes for both S&P 500 and Bitcoin, for a total of four different regimes. A DCC-GARCH(1,1) framework is used for the multivariate model. The univariate model parameters are fixed in the multivariate model, only estimating the remaining parameters a and b for the correlation update equation. The parameter estimates are compared to a GARCH(1,1) and a DCC-GARCH(1,1) estimated on the complete dataset.

This approach gives us four different parameter sets, each corresponding to a specific volatility regime combination. The parameter sets are run on the complete dataset, giving us four different correlation and volatility estimates for the investigated period. Finally, to get a smoother transition between regimes, fuzzy clustering is applied to the volatility estimates.

The resulting model gives a correlation estimate that is similar to a standard DCC-GARCH(1,1) on the dataset but is more responsive like a rolling correlation measure. We note that the correlation is at an elevated level during and after the COVID-19 pandemic, but has dropped to pre-pandemic levels during the end of 2023.

Keywords: Bitcoin, S&P500, Markov Regime Switching, LASSO regression, DCC-GARCH, Volatility, Fuzzy Clustering

Acknowledgments

We would like to express our deepest gratitude to our supervisor Erik Lindström at the Centre for Mathematical Sciences at Lund University, for your guidance, support, and patience during this thesis. You have given us many ideas, insights, and valuable feedback along this journey.

Contents

Abbreviations

AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
BTC	Bitcoin
DCC-GARCH	Dynamic Conditional Correlation GARCH
EMA	Exponential Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
LASSO	Least Absolute Shrinkage and Selection Operator
MA	Moving Average
MGARCH	Multivariate GARCH
MS	Markov Switching
OLS	Ordinary Least Squares
RSI	Relative Strength Index
S&P 500	Standard and Poor's 500
VIX	CBOE Volatility Index

1 Introduction

The introduction of this master thesis report starts by giving a background of the subject. Thereafter, past studies that have been done on this subject are presented followed by the purpose of this thesis work.

1.1 Background

On 3 January 2009, Satoshi Nakamoto created the first BTC network. One year later, the first commercial transaction with BTC was made when two pizzas were purchased for 10 000 BTC. Those bitcoins would be worth approximately 700 million US dollars as of today (5 March 2024). However, the journey of bitcoin has definitely not been a journey without major setbacks and bear markets. For a long time, many investors thought BTC was an extremely volatile instrument with no correlation to the biggest stock market indexes. Today, there are still many people who don't understand what BTC really is and the history of BTC. Also, it is still not fully clear whether or not BTC is correlated to large stock market indices such as the S&P 500. Therefore, this section will give an introduction to both BTC and the S&P 500, along with an explanation of the historical correlation between the two.

In short, BTC is a decentralized digital currency and worldwide payment system with a market capitalization of over 1 trillion US dollars (CoinMarketCap, 2024). The idea behind BTC was to create a currency that, unlike other traditional currencies, such as US dollars or Euros, is not controlled by central authority. This would make it nearly impossible for governments and others to manipulate or intervene. Decentralization implies no single entity controls the currency. Bitcoin transactions are confirmed by computers in the network using cryptography and stored in a public ledger called a blockchain. BTC wallets are used for storing and sending or receiving bitcoins. The main advantage of BTC according to its supporters is that it provides a more efficient and secure payment compared to traditional payments. On the other hand, critics of BTC mean that the currency can be used in illegal activities. BTC has also been criticized for its high energy consumption (Wikipedia, 2024).

While BTC is a unique and non-traditional investment, the S&P 500 is the opposite. The S&P 500 is one of the most common stock market indices and consists of 500 of the largest public companies in the United States of America. The total market capitalization of the S&P 500 is approximately 43 trillion US dollars as of 2024. The 500 companies included in the stock index account for

80% of the total market cap of all public companies in the US. Some of these companies are Apple, NVIDIA, Microsoft, JP Morgan, Meta, etc. Because of the fact that the S&P 500 includes approximately 80% of the total US market, it is seen as one of the most well-diversified portfolios for an investor to place their money in. Because of this, it is common for people to refer to the S&P 500 and other large indexes such as the MSCI World Index as "investing in the market".

BTC and the S&P 500 represent two ends of the investment spectrum. BTC has become known as an asset with high volatility and potentially explosive returns. On the other hand, the S&P 500 is known for its relative stability and consistent, albeit often more modest, returns. This difference in risk and reward has traditionally kept these asset classes separate in investor portfolios. However, in recent years it has been an interesting topic about whether or not Bitcoin's relationship with the traditional market has matured. Historically, BTC's price movements have been weakly related to traditional assets, such as those in the S&P 500. This made it an attractive diversification tool, and investors could use BTC as a hedge against downturns in their stocks or bonds. However, some recent data suggests a possible shift in this dynamic, particularly after COVID-19. If this stronger correlation is not only a short-term phenomenon, it would have significant implications for portfolio diversification. Investors may need to re-evaluate BTC's role in their strategies if it loses its role as an uncorrelated asset.

1.2 Objective

In this thesis, we will examine if the change in correlation during COVID-19 is a temporary phenomenon or if it is a long-term trend. This will be done by utilizing data extending to the beginning of 2024. The objective of this thesis is to derive a dynamic model to estimate the correlation between the crypto asset BTC with the stock market index S&P 500.

This will be done by splitting the historical data series into distinct regimes based on specific characteristics. This allows for the identification of periods with similar correlation dynamics. Within each identified regime for both BTC and the S&P 500, separate DCC-GARCH models will be estimated. GARCH models excel at capturing volatility patterns, leading to more accurate parameter estimates when applied to data with similar characteristics within each regime. To avoid abrupt transitions between regimes with distinct parameter sets, we will employ fuzzy clustering techniques. This allows for smoother

transitions between regimes, reflecting the possibility of data points exhibiting characteristics of more than one regime to a certain degree.

1.3 Previous Work

The relationship between BTC and traditional asset classes like the S&P 500 has attracted significant research interest in the past. This section aims to review existing studies on the correlation between BTC and traditional assets. In particular, the focus will be on what correlation patterns previous studies have identified. The focus will also be on past studies that have used similar approaches as the one in this thesis, where the correlation will be modeled by a DCC-GARCH model.

A new study by the IMF (2022) highlights a growing concern as cryptocurrencies gain wider adoption and their correlation with traditional assets has risen. This reduces the diversification benefits of cryptocurrencies and raises the risk of contagion across financial markets, meaning it increases the chance of problems in one market spreading to others. The research finds that the correlation coefficient between daily movements of BTC and the S &P 500 was negligible, 0.01, between 2017 and 2019. However, this value increased to 0.36 between 2020 and 2021, which indicates a much stronger positive correlation between the assets. The authors in that study calculated the correlation coefficient between BTC and S &P 500 by calculating a 60-day rolling correlation.

In the research done by Nguyen (2021), they examine the relationship between the stock market and BTC during COVID-19 and other uncertainty periods. This is done by using quantile regression to estimate BTC returns on the S&P 500 market during different levels of uncertainty periods. The researchers model the conditional distribution of BTC by applying a VAR(1)-GARCH(1,1) which is supposed to be advantageous for observing the spillover effect from the stock market to BTC. The study shows a significant influence of past stock market returns on BTC returns, especially during times of market stress and after the emergence of COVID. This would mean that BTC becomes more sensitive to stock market movements during higher uncertainty periods. The researchers also identified volatility spillover effects from the stock market to Bitcoin. This was especially seen during the COVID-19 pandemic and other periods of high uncertainty. All of this suggests that the stock market and BTC are more correlated during periods of high uncertainty.

In a more recent study, Terraza et al. (2024) uses another approach, applying a

multivariate VAR-DCC-EGARCH to investigate the dependency between BTC, gold, and the S&P 500, NASDAQ, and Dow Jones indices. They further divide the dataset into two distinct time periods, 2018-2020 and 2020-2022, where they find a significant increase in the correlation between BTC and the S&P 500 in the latter period. The paper also discusses that this behavior of BTC is similar to that of more traditional investment assets.

In another research made by Cortese et al. (2023), they examine which are the key features that drive the return dynamics of the largest cryptocurrencies. However, their approach focused on fitting a more tractable model by segmenting the data into different states or regimes, each with a common underlying structure. This was done by applying a sparse jump model, and it was found that a model with three states was most suitable. These states were interpreted as bear, neutral, and bull markets. First of all, it can be seen in the research that the major cryptocurrency log returns follow a similar pattern as the log returns of BTC. Therefore, it might be reasonable to only look at Bitcoin when analyzing the relationship between cryptocurrencies and traditional stock market indices.

The study identified several key drivers of cryptocurrency returns: first moments of returns, trend and reversal signals from technical analysis, market activity, and public attention. This can be used by investors to identify upward and downward market trends, and when a switch between these states occurs. Furthermore, the research revealed significant differences in cryptocurrency behavior across three distinct states. Notably, volatility was roughly twice as high during bear markets compared to other states. Additionally, correlations between cryptocurrencies were also found to be stronger in bear markets. These observations highlight the limitations of a single-state model and support the benefits of a regime-switching approach, where model parameters are estimated for each unique market state.

1.4 Structure

The rest of this thesis is structured as follows. Section 2 gives an introduction to the theory of uni- and multivariate GARCH models, parameter optimization, technical analysis, regression Markov Switching Models, and fuzzy clustering. The methodology is described in Section 3, while Section 4 presents the results from our empirical analysis. Section 5 concludes the thesis' results, and a further discussion of the results is provided in Section 6.

2 Theory

2.1 Stylized facts

Financial data tends to show some properties that differ from other types of data in nature. These properties can have an impact from a statistical point of view, regarding the choice of model and distributions. Cont (2001) summarizes these facts, citing exchange rate returns, S&P 500 returns, and returns on stocks of large companies, while Lindström et al. (2015) confirms the stylized facts on returns of the OMXS30 index.

2.1.1 Heavy tails

Financial returns exhibit fatter tails than the normal distribution, meaning that large price movements, both positive and negative, are more common than predicted by a standard bell-curve model. This signifies that the normal distribution underestimates the risk of extreme events in financial markets.

2.1.2 Volatility clustering

Unlike the constant volatility observed in linear Gaussian time series, financial data have time-varying volatility. This means that the level of how much the prices fluctuate is not constant, but changes over time. Financial markets can experience periods of relative calm followed by periods of high volatility. This time-varying nature of volatility is a key factor contributing to the heavy tails observed in financial return distributions.

2.1.3 No autocorrelation in returns

Financial returns exhibit minimal linear autocorrelation. This aligns with the efficient market hypothesis. If investors could predict future asset values, they would exploit these to generate profits. This would drive prices towards their fair value, removing opportunities for abnormal returns based on past information. However, there is still some autocorrelation in the first lags, due to trading friction etc.

2.1.4 Significant autocorrelation for squared returns

Even though returns show little to no autocorrelation, which is also supported by the fact that volatility clustering is present, clear dependency is visible when investigating the absolute returns. This persistent autocorrelation is usually long, ranging up to at least 150 days. Recent studies on the matter have been

conducted, suggesting that this phenomenon is due to a lack of stationarity, rather than actual dependence.

2.2 Time Series Modeling

Financial data is generated over time, and in many cases, stock prices, exchange rates, and other economic indicators exhibit trends, seasonality, and varying levels of volatility. For several financial tasks, such as portfolio optimization and risk management, it is very important to understand the dynamics of these time series. This is where time series analysis comes into play and can be very useful. Time series analysis includes several different economic methods and models used to analyze a time series of data. Some examples of these models are the AR process, the MA process, and a combination of these two called the ARMA process. In this section, the theory behind all of these models will be presented and explained.

2.2.1 Moving Average processes

A Moving Average process y_t is defined as per below (Jacobsson, 2021):

$$y_t = e_t + c_1 e_{t-1} + ... + c_q e_{t-q} = C(z)e_t \tag{1}$$

where C(z) is a monic polynomial of order q:

$$C(z) = 1 + c_1 z^{-1} + ... + c_q z^{-q} \tag{2}$$

where $c_q \neq 0$ and e_t is a white noise process with zero mean and variance σ_e^2. The resulting MA(q)-process y_t is always stable, and has the following properties:

$$\mu_y = E[y] = 0 \tag{3}$$

$$\sigma_y^2 = Var(y) = \sigma_e^2(1 + c_1^2 + ... + c_q^2) \tag{4}$$

2.2.2 Autoregressive processes

Another basic linear process is the AutoRegressive process, and as the name suggests the process is a regression in itself. In an AR process, the current value of the time series is influenced by its own past values. This means that predictions of future values are made by looking at past behavior. We define an autoregressive process of order p as the following (Jacobsson, 2021):

$$A(z)y_t = y_t + a_1 y_{t-1} + ... + a_p y_{t-p} = e_t \tag{5}$$

where A(z) is a monic polynomial of order p:

$$A(z) = 1 + a_1 z^{-1} + ... + a_p z^{-p} \tag{6}$$

where $a_p \neq 0$ and e_t once again is a zero-mean white noise process with variance σ_e^2 and is uncorrelated with y_{t-l} for $l > 0$. The AR process has the following properties:

$$\mu_y = E[y] = 0 \tag{7}$$

$$\sigma_y^2 = Var(y) = \frac{\sigma_e^2}{1 - a_1\rho_1 - a_2\rho_2 - ... - a_q\rho_p} \tag{8}$$

where ρ is the autocorrelation.

2.2.3 ARMA processes

The AR process and the MA process can be combined to form an ARMA process. This process takes both the past errors and the past values of itself into account. The ARMA process is defined as below (Jacobsson, 2021):

$$A(z)y_t = C(z)e_t \tag{9}$$

where C(z) and A(z) are defined as per Equation 2 and 6. Time series data can be modeled using AR, MA, or ARMA processes to capture the underlying structure and identify the residuals (errors) of the model. These residuals, which represent the difference between the actual data and the fitted model, are then used to estimate GARCH models. GARCH models specifically focus on modeling the conditional variance, which is the volatility of the residuals over time.

2.3 Technical Analysis

Multiple techniques in the realms of technical analysis have been shown to be drivers of cryptocurrency returns (Cortese et al., 2023). We use this fact to estimate the Markov Switching Model in Section 2.5.

2.3.1 Relative Strength Index

The RSI is a technical analysis indicator used to measure the momentum of price changes in an asset. RSI is a momentum oscillator developed by Welles Wilder Jr. (1978), designed to identify potential overbought or oversold conditions in the market. RSI is a measure between 0 and 100 and is often plotted under the

graph of the asset's price, unlike many other technical analysis techniques such as EMA. Calculating the RSI consists of two steps, where the first step uses the following formula:

$$RSI_{StepOne} = 100 - \left(\frac{100}{1 + \frac{\text{Average Gain}}{\text{Average Loss}}} \right) \tag{10}$$

The average gain or loss used in this calculation is the average percentage gain or loss during a specific look-back period, for example, 7 or 14 days. The formula uses a positive value for the average loss. The second part of the RSI calculation can begin when there are n periods of data available for an RSI-n. Here, n is the look-back period mentioned above. The formula for the second part of the RSI calculation is presented below:

$$RSI_{StepTwo} = 100 - \left(\frac{100}{1 + \frac{\text{Previous Average Gain} \cdot (n-1) + \text{Current Gain}}{\text{Previous Average Loss} \cdot (n-1) + \text{Current Loss}}} \right) \tag{11}$$

Traditionally, an RSI reading of 70 or above indicates an overbought situation. This means the asset is trading at a higher price than its intrinsic value, and should therefore be sold. A reading of 30 or below indicates an oversold condition. This means that the asset is trading at a lower price than its intrinsic value and should therefore be sold (Investopedia, 2024).

2.3.2 Exponential Moving Average

An exponential moving average, often referred to as EMA, is a type of moving average that puts more weight on the most recent values. EMA can be calculated for different lengths, but some of the most common ones are 5, 12, 50, and 200 days. The formula for calculating EMA for n number of days can be found below:

$$EMA_t = \left(P_t \cdot \frac{SF}{1+n} \right) + EMA_{t-1} \cdot \left(1 - \frac{SF}{1+n} \right) \tag{12}$$

where P_t is the closing price at time t. SF is a smoothing factor, often chosen as 2 (Investopedia, 2024).

Calculating EMA requires calculating the Simple Moving Average (SMA) first.

This is because you need to use the SMA as the first observation of EMA_{t-1} from the formula above. For example, for an EMA_7 one needs to calculate the SMA for the first 7 observations. The value of this SMA is then used when calculating the EMA for the 8th day. SMA is easily obtained by using the following formula, once again for a length of n number of days.

$$SMA_t = \frac{\sum_{t=1}^{n} P_t}{n} \tag{13}$$

EMA is used by technical traders as a tool to help them identify when an asset should be bought or sold. However, it is important to know that there are limitations to the EMA. One of these limitations is that recent data can be overweighted compared to older data. This could create a bias that leads to false alarms. Another important limitation of the EMA is the so-called efficient market hypothesis, which means all available information already is reflected in the prices. If the EMH holds, it should not be possible to predict future asset prices by using historical data. (Investopedia, 2024)

2.3.3 Exponentially Weighted Linear Correlation Measures

One way of calculating the exponentially weighted linear correlations is to use an Exponentially Weighted Moving Average (EWMA) model (Hull, 2018). The model shows similarities with the formula for EMA and is formulated as:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2 \tag{14}$$

where λ is a constant between zero and one, and r_t is the return on day t. For simplicity, we identify λ in the same way as for the Exponential Moving Average, meaning that $\frac{SF}{1+n} = 1-\lambda$. The correlation parameter is calculated with the standard formula

$$\rho_{t,xy} = \frac{\text{cov}_{t,xy}}{\sigma_{t,x}\sigma_{t,y}} \tag{15}$$

where $\sigma_{t,x}$ is the standard deviation of asset x at time t that is obtained from Equation 14. In the bivariate case, the covariance between asset x and y is calculated with the following formula:

$$\text{cov}_{t,xy} = \lambda\text{cov}_{t-1,xy} + (1-\lambda)r_{t-1,x}r_{t-1,y} \tag{16}$$

2.4 Linear Regression

One common method for optimizing the parameters in a linear regression is to use OLS. The principle is to minimize the sum of squares of the residual between the observed dependent variable and the output of a linear function of the independent variable. We suppose the linear model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{17}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are $n \times 1$ vectors of the response variables and error terms, and $\boldsymbol{X}$ is an $n \times p$ matrix of regressors. If x_{ij} denotes the i, j:th element of the matrix $\boldsymbol{X}$ and y_i, β_j denote the i:th and j:th element of $\boldsymbol{y}$ and $\boldsymbol{\beta}$ respectively, the OLS estimate is given by:

$$\hat{\beta}_{OLS} = \min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \tag{18}$$

It is also possible to formulate the optimal solution in matrix form, solving the so-called *normal equations* (Goldberger, 1964, p. 158), giving the solution:

$$\hat{\beta}_{OLS} = (\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y} \tag{19}$$

2.4.1 LASSO Regression

LASSO is a regression analysis technique that aims to improve the regular OLS estimate for linear models by adding a penalty to the objective function. The penalty is added to the absolute values of the coefficients in the regression, effectively performing a variable selection and regularization by shrinking the coefficients. The term LASSO was first coined by Tibshirani (1996) and has been extended to account for more general regressions.

When using the LASSO method, we assume that the regression model like the one in Equation 17 contains sparse coefficients, that is some of the covariates are not relevant, $\beta_j = 0$. By removing these, and only keeping relevant covariates, we can improve the model. To achieve this, Tibshirani adds the ℓ_1 penalty to the OLS objective function, resulting in the following problem:

$$\hat{\beta}_{LASSO} = \min_{\beta_0, \beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{20}$$

Here, x is the covariate vector, β_0 is the constant coefficient, β is the coefficient vector and t is a tuning parameter. Another way to formulate the problem is to use the covariate matrix $\boldsymbol{X}$, rewriting the problem in Lagrangian form:

$$\hat{\beta}_{LASSO} = \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} ||y - \boldsymbol{X}\beta||_2^2 + \lambda ||\beta||_1 \right\} \tag{21}$$

where the relationship between t and λ is data dependent. For greater values of λ, the coefficients have a greater penalty, leading to more coefficients being shrunken towards zero.

2.5 Markov Switching Model

The Markov Switching (MS) Model (Hamilton, 1989, 2005), also referred to as the regime-switching model, is one of the most popular non-linear time-series models. This model allows the underlying process to show distinct behaviors, characterized by multiple structures. This corresponds to different regimes or states of the time series. The MS Model is able to capture more complex dynamic patterns within the data thanks to it permitting transitions between the different states.

A fundamental feature of the regime-switching model lies in its switching mechanism controlled by a hidden state variable. This variable follows a first-order Markov chain, implying the Markovian property. The Markovian property says that the current state only depends on its immediate past state, and not on the entire history. Consequently, a particular regime can continue for a random duration before a transition to another regime happens. This makes it possible for the model to effectively capture the time-varying nature of the data by allowing for shifts in its underlying structure.

The MS Model can be represented in a generalized notation in the following way (Perlin, 2014):

$$y_t = \sum_{i=1}^{N_{nS}} \beta_i x_{i,t}^{nS} + \sum_{j=1}^{N_S} \phi_{j,S_t} x_{j,t}^{S} + \varepsilon_t \tag{22}$$

where N_{nS} denotes the number of non-switching coefficients and N_S the number of switching coefficients. S_t is the unobservable state variable and the innovations ε_t are assumed to have some distribution with corresponding parameters

that also may or may not change depending on the state.

As an example, let's assume S_t can take on the values 0 or 1 which symbolize two different states. We observe a simple switching model for the variable y_t:

$$y_t = \begin{cases} \alpha_0 + \beta y_{t-1} + \varepsilon_t, & \text{if } S_t = 0, \\ \alpha_0 + \alpha_1 + \beta y_{t-1} + \varepsilon_t, & \text{if } S_t = 1. \end{cases} \tag{23}$$

where $|\beta| < 1$ and ε_t are iid random variables with zero mean and variance σ_ε^2. By noting that these specifications are stationary AR(1)-processes, the mean μ of these processes is the following for the two states:

$$\mu = \begin{cases} \frac{\alpha_0}{1-\beta}, & \text{if } S_t = 0 \\ \frac{\alpha_0+\alpha_1}{1-\beta}, & \text{if } S_t = 1 \end{cases} \tag{24}$$

Depending on what value S_t takes on and provided that $\alpha_1 \neq 0$, the model has two dynamic structures. The values of y_t are determined by two different probability distributions with two different mean values. The hidden variable S_t controls which distribution governs y_t at any given time.

In the MS model, one assumes that S_t follows a first-order Markov chain with the following transition matrix $\boldsymbol{P}$:

$$\boldsymbol{P} = \begin{pmatrix} P(S_t = 0|S_{t-1} = 0) & P(S_t = 1|S_{t-1} = 0) \\ P(S_t = 0|S_{t-1} = 1) & P(S_t = 1|S_{t-1} = 1) \end{pmatrix} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{25}$$

Within the MS model, the behavior of y_t is shaped by two key elements: the random innovations ε_t and the hidden state variable S_t. Frequent and random changes of model structures are yielded from the Markovian state variable, and the persistence of each state is determined by the transition matrix $\boldsymbol{P}$.

2.6 ARCH Model

Financial market studies frequently show non-constant variance in asset returns. This means fluctuations in the variance, with periods of high and low volatility. To capture this so-called "volatility clustering," Engle (1982) introduced the ARCH model. The ARCH model assumes the return of an asset between times $t-1$ and t equals its conditional expectation plus a random error term as per below:

$$r_t = \sum_{i=1}^{k} b_i x_{i,t} + \varepsilon_t \tag{26}$$

Here, an assumption is made that the error terms follow a Normal distribution conditional on the information available at time $t-1$, but other distributions such as the Student's t-distribution can also be considered. The information is usually denoted $\mathcal{F}_{t-1}$, and for the case of the Normal distribution, the errors are defined as

$$\varepsilon_t | \mathcal{F}_{t-1} \sim N(0, \sigma^2) \tag{27}$$

As per above, the conditional mean of the residuals is zero. Therefore, the unconditional variance of ε can be obtained by taking the expectation of the conditional variance as per below:

$$\sigma^2 \equiv E[\varepsilon^2] = E[E_{t-1}[\varepsilon^2]] = E[\sigma_t^2] \tag{28}$$

which is independent of the time t. Therefore, the unconditional variance is constant over time. The unconditional distribution of ε_t has fatter tails than the Normal distribution because of time-varying variances. In the original specification by Engle (1982), the preferred model for ε_t is given as

$$\varepsilon_t = \sigma_t z_t \tag{29}$$

where

$$z_t \sim IID \quad \text{and} \quad N(0,1) \tag{30}$$

As Equation 30 shows, z_t has a unit variance, and therefore the conditional variance depends on past values of squared errors:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + ... + \alpha_q \varepsilon_{t-q}^2 \tag{31}$$

The equation above shows an ARCH(q) model. One important constraint that needs to be fulfilled for the ARCH model is that ω and α_i have to be greater than or equal to zero (ω, $\alpha_i \geq 0$). In other words, they can not be negative. This constraint makes sure the model gives non-negative variances at all times, which is a fundamental requirement. The larger amount of α-parameters you

have in the model, the longer memory can be captured by the model, meaning it can capture the influence of past shocks for a longer period.

By considering the unconditional variance, we can easily see that if

$$|\alpha_1 + \alpha_2 + ... + \alpha_q| < 1 \tag{32}$$

the unconditional variance for an ARCH(q) is:

$$\begin{aligned} \sigma^2 = E[\sigma_t^2] = E[\omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + ... + \alpha_q \varepsilon_{t-q}^2] = \\ = \omega + \alpha_1 E[\varepsilon_{t-1}^2] + \alpha_2 E[\varepsilon_{t-2}^2] + ... + \alpha_q E[\varepsilon_{t-q}^2] \quad , \end{aligned} \tag{33}$$

and because the unconditional variances are constant over time, we have:

$$\sigma^2 = \omega + \alpha_1 \sigma^2 + \alpha_2 \sigma^2 + ... + \alpha_q \sigma^2. \tag{34}$$

After simplifications, this gives the following equation:

$$\sigma^2 = \frac{\omega}{1 - a_1 - a_2 - ... - a_q}. \tag{35}$$

2.6.1 GARCH Model

Numerous variants of the ARCH model have been proposed, but the most common one is likely to be the Generalized Autoregressive Conditional Heteroskedastic (GARCH) model by Bollerslev (1986). In the GARCH(p, q) model, the conditional variance depends not only on past squared errors ε, but also on the past conditional variances σ:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2. \tag{36}$$

The parameters ω, α, and β in the GARCH model must be greater than or equal to 0 in order to ensure positive variance. To ensure stationarity, the sum of α and β has to be less than 1. The unconditional variance can be written in a similar fashion as for the ARCH model, resulting in the expression:

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{i=1}^{p} \beta_i}. \tag{37}$$

2.6.2 Multivariate Models

For financial applications such as portfolio optimization and risk management, it is important to consider the correlation between assets. When modeling the volatility of assets, it is possible to extend the GARCH framework to a multivariate setting, the so-called MGARCH, by assuming that there is a correlation between the assets and then co-model the variances. Over the years, multiple specifications of MGARCH models have been proposed, beginning with the VEC-GARCH model by Bollerslev et al. (1988), which is a generalized version of the univariate GARCH model. According to Silvennoinen and Teräsvirta (2008), the specification of an MGARCH model should be flexible enough to represent the conditional variances and covariances but also allow for easy estimation and interpretation of model parameters. As the parameters in an MGARCH model often increase rapidly with the number of assets, this often poses a problem.

In the same paper, Silvennoinen and Teräsvirta define the standard multivariate GARCH framework by considering that returns follow a zero-mean stochastic process $\{\boldsymbol{r}_t\} \in S \subseteq \mathbb{R}^N$ with dimensions $N \times 1$. They let $\mathcal{F}_{t-1}$ denote the information set available at time $t-1$ and assume that $\{\boldsymbol{r}_t\}$ is conditionally heteroskedastic:

$$\boldsymbol{r}_t = \boldsymbol{H}_t^{1/2}\boldsymbol{\eta}_t \tag{38}$$

Here, $\boldsymbol{H}_t = [h_{ijt}]$ is an $N \times N$ matrix of the conditional covariance matrix of the returns, and $\boldsymbol{\eta}_t$ is an iid vector of errors, such that $\mathbf{E}[\boldsymbol{\eta}_t\boldsymbol{\eta}_t^{'}] = \boldsymbol{I}$. Silvennoinen and Teräsvirta specifies four different categories of MGARCH models to model the covariance matrix $\boldsymbol{H}_t$. As our objective is to model the correlation, we have chosen the Dynamic Conditional Correlation (DCC-) GARCH model, which is in the category of models that decompose the conditional covariance matrix.

The DCC-GARCH model was introduced by Engle (2002), where the conditional covariance matrix is defined as follows:

$$\boldsymbol{H}_t = \boldsymbol{D}_t\boldsymbol{P}_t\boldsymbol{D}_t \tag{39}$$

where $\boldsymbol{D}_t = \text{diag}(h_{1t}^{1/2}, \ldots, h_{Nt}^{1/2})$ and $\boldsymbol{P} = [\rho_{ij}]$ is positive definite with $\rho_{ii} = 1, i = 1, \ldots, N$. The off-diagonal elements of $\boldsymbol{H}_t$ are defined in the same way as for any covariance matrix:

$$[\boldsymbol{H}_t]_{ij} = h_{it}^{1/2} h_{jt}^{1/2} \rho_{ij}, \quad i \neq j \tag{40}$$

where $1 \leq i, j \leq N$. In this notation, h_{it} is the variance for the process $\{r_{it}\}$, which can be modeled with a GARCH(p, q) model, and ρ_{ij} is the correlation between process i and j. For simplicity, the conditional variance can be written in a vector from:

$$\boldsymbol{h}_t = \boldsymbol{\omega} + \sum_{j=1}^{q} \boldsymbol{A}_j \boldsymbol{r}_{t-j}^{(2)} + \sum_{j=1}^{p} \boldsymbol{B}_j \boldsymbol{h}_{t-j} \tag{41}$$

where $\boldsymbol{\omega}$ is an $N \times 1$ vector, $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ are diagonal $N \times N$ matrices, and $\boldsymbol{r}_t^{(2)} = \boldsymbol{r}_t \odot \boldsymbol{r}_t$. Engle further defined a dynamic process $\boldsymbol{Q}_t$ used to update the correlation matrix $\boldsymbol{P}_t$ as follows:

$$\boldsymbol{Q}_t = (1 - a - b)\boldsymbol{S} + a\varepsilon_{t-1}\varepsilon_{t-1}^{'} + b\boldsymbol{Q}_{t-1} \tag{42}$$

where a is positive and b is a non-negative scalar with the restriction that $a + b < 1$ and $\boldsymbol{S}$ is the unconditional correlation matrix of the errors $\boldsymbol{\varepsilon}_t$. To ensure positive definiteness, $\boldsymbol{Q}_0$ is chosen as positive definite. Lastly, $\boldsymbol{Q}_t$ is rescaled to obtain a valid correlation matrix as follows:

$$\boldsymbol{P}_t = (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2} \boldsymbol{Q}_t (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2} \tag{43}$$

2.7 Parameter optimization

In mathematical statistics, different parameter estimators are available, such as OLS and its generalized version GLS, and the Maximum Likelihood estimator. In the original specification of the ARCH framework by Engle (1982), he proposes the use of Maximum Likelihood to estimate the parameters, which is also used in this thesis. In Maximum Likelihood estimation, the objective is to find the parameter set that maximizes the so-called likelihood function. To achieve this, different optimization techniques such as Stochastic Gradient Decent (SGD) and Quasi-Newton can be used.

2.7.1 Maximum Likelihood

Maximum Likelihood is a good estimation method when the distribution of the underlying observation is known. Consider a sequence of observed variables, $(r_1, \ldots, r_T)$, where each sample has a probability density function $f(r_t; \theta)$, θ

being the parameters which characterize $f(r_t;\theta)$. If the samples are not *iid*, the joint density is given by the product of the conditional marginal densities as

$$\begin{aligned} f(r_1,\ldots,r_T;\theta) &= f(r_1;\theta)\cdot\ldots\cdot f(r_T|r_1,\ldots,r_{T-1};\theta) = \\ &= \prod_{t=1}^{T} f(r_t|r_1,\ldots,r_{t-1};\theta) \end{aligned} \tag{44}$$

where the density function is conditioned on the information known at each time t. Further, we define the likelihood function $L(\theta|r_1,\ldots,r_T)$ to be the joint density function, treated as a function of the parameters given by θ. The MLE is the parameters θ which maximize the likelihood function, i.e.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta|r_1,\ldots,r_T) = \arg\max_{\theta} \prod_{t=1}^{T} f(r_t|r_1,\ldots,r_{t-1};\theta) \tag{45}$$

The argument θ that maximizes this expression is not affected by a log transformation, so the problem can be rewritten as a sum instead of a product:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ln L(\theta|r_1,\ldots,r_T) = \arg\max_{\theta} \sum_{t=1}^{T} \ln f(r_t|r_1,\ldots,r_{t-1};\theta) \tag{46}$$

This log transform is referred to as the log-likelihood. Usually, the normal distribution is used in Maximum Likelihood estimation for GARCH models, as for the original specification (Bollerslev, 1986), but other distributions such as student's t-distribution can be used. We consider a simple model for asset returns:

$$r_t = \mu + \varepsilon_t \tag{47}$$

where $\varepsilon_t = \sigma_t z_t$. The variance σ_t^2 is assumed to follow a GARCH$(1,1)$ specification from Equation 36 and z_t follows a standardized Normal or Student's t-distribution. In the case of z_t following a Normal distribution, the log-likelihood function is written as:

$$\ln L(\theta|r_1,\ldots,r_T) = \sum_{t=1}^{T} \ln \frac{1}{\sigma_t\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{r_t-\mu}{\sigma_t}\right)^2} \tag{48}$$

This expression can easily be rewritten in the following way:

$$\ln L(\theta|r_1,\ldots,r_T) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln\sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(r_t-\mu)^2}{\sigma_t^2} \tag{49}$$

Where the parameters θ are μ, ω, α and β. With the assumption of Student's t-distributed z_t, the log-likelihood function is written in the following manner (MathWorks, 2024):

$$\begin{aligned} &\ln L(\theta|r_1,\ldots,r_T) = \\ &= T\ln\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)} - \frac{1}{2}\sum_{t=1}^{T}\ln\sigma_t^2 - \frac{\nu+1}{2}\sum_{t=1}^{T}\ln\left(1+\frac{(r_t-\mu)^2}{\sigma_t^2(\nu-2)}\right) \end{aligned} \tag{50}$$

where $\Gamma(z)$ is the Gamma function evaluated in the point z.

2.8 Fuzzy Clustering

Fuzzy c-means clustering, also known as soft clustering, was developed by Dunn (1973) and later improved by Bezdek (1981) and is a data clustering technique used to group data points into different clusters. However, unlike traditional K-means clustering, which assigns each data point to a single cluster with a sharp boundary, fuzzy c-means allows a data point to belong to multiple clusters with varying degrees of membership. For example, if there are 4 different clusters to which the data points can belong, the K-means clustering assigns a data point to only one of the clusters. In c-means clustering, each data point is assigned a membership degree between 0 and 1 for each cluster. A value closer to 1 means stronger membership in that cluster. The sum of the membership degrees for a data point must be equal to 1.

Given a data set

$$X = \{x_1, x_2, ..., x_n\} \subseteq \mathbb{R}^n \tag{51}$$

where $c \leqslant n$ and the mean of the partition of X is larger than 1. The aim is to find fuzzy sets on X that fulfill the following property:

$$\sum_{i=1}^{c} u_{ik} = 1, u_{ik} = u_i(x_k) \tag{52}$$

such that the functional

$$J(u, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m ||x_k - v_i||^2 \tag{53}$$

is minimized, where v_i is the center of the i:th cluster:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{54}$$

A simple representation of the differences between K-means clustering and fuzzy clustering can be seen in Figure 1. K-means clustering divides the data into distinct clusters without overlap between them, while fuzzy clustering allows data points to belong to multiple clusters, determined by the calculated membership degree.

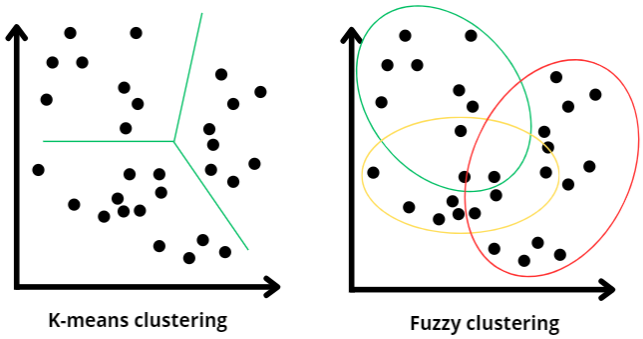


Figure 1: A simple representation of the differences between K-means clustering and fuzzy clustering. In the case of fuzzy clustering, we can see that a single data point can belong to multiple clusters.

3 Methodology

To get a clear overview of the methodology and the steps taken, a short summary is provided. The first step to building the model is to gather all the data needed. The data that is collected are data on returns of BTC, the S&P 500 index, gold and oil prices, and the VIX and Google Trends indices. The next step in the process is to divide the BTC and S&P 500 returns into two distinct states, where some of the other mentioned data sets are used as additional explanatory variables. To decide what data sets are relevant for this step, a LASSO regression is conducted. The states themselves are estimated by utilizing a Markov Switching Model. With the estimated states, four different regimes can be extracted. The third step is to estimate the correlation using a multivariate GARCH framework, namely the so-called DCC-GARCH model. To get sound parameter estimates, we start by estimating standard univariate GARCH models to the states that were first extracted from the Markov Switching Model. We later use these univariate estimates in the multivariate setting to get four different models, one for every regime. The last step is to weigh the four regimes together to get the final model. This is done by using a Fuzzy Clustering algorithm.

3.1 Software

In this thesis, most data treatment, processing, modeling, and visualization are conducted using a mix of **MATLAB** and **R**. **MATLAB** is used for the calculations of the technical analysis, regime estimation, and fuzzy clustering, while **R** is used in the GARCH modeling. To improve the management and execution of the **R** code, `RStudio` is used to write said code.

3.2 Datasets

The full dataset contains daily closing values and volumes denoted in USD of the cryptocurrency BTC and the S&P 500 stock market index from Yahoo Finance (2024). The data spans from 1 November 2017 until 31 January 2024, giving a total of 2283 and 1571 observations of BTC and the S&P 500 respectively. Data on gold and crude oil prices, as well as the VIX index, was also acquired from Yahoo Finance. We use the logarithmic daily returns, or log-difference, of the datasets, defined as

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

where P_t is the closing price on day t and r_t is the return for day t. Of the resulting series of returns, observations between 1 January 2018 and 31 December 2023 are used for GARCH modeling, while the full dataset is utilized for the technical analysis, exponentially weighted correlation, and rolling correlation.

Daily Google Trends index data of the search term "bitcoin" between 1 December 2017 and 31 January 2024 is extracted using the **R** package `gtrendsR` (Massicotte and Eddelbuettel, 2022). Daily trends index data is extracted on a monthly basis, as Google Trends only shows daily observations for periods of one month or less. This gives a trend index from $0 - 100$ for every month. For longer time periods, Google Trends gives a monthly trend index, and the information on a daily basis disappears. As such, the trend index for the entirety of the time period can be extracted without any special methods but gives a trend index on a monthly basis from $0 - 100$.

To calculate the RSI, **MATLAB**'s Financial Toolbox is utilized, while EMA is calculated using Equation 12. The RSI index is calculated on the closing prices of BTC, with a window size $d = 14$, while the EMA is calculated on the returns of BTC with window sizes $d = 7, 14$. The exponentially weighted linear correlations of the BTC returns and gold, oil, and VIX index returns, as well as the log differences of Google Trends index and trading volume, are calculated with half-lives $d = 7, 14$.

3.2.1 Data Preprocessing

As commodities such as gold and oil, and the S&P 500 companies are only traded on weekdays, there is a significant discrepancy in the number of observations between the datasets. To account for this, we use linearly interpolated values for days where the closing price is missing, using the **R** package `zoo` (Zeileis et al., 2023). This yields equal numbers of observations between the series and is used to calculate the logarithmic daily returns. As a result, 2191 observations can be used for the modeling set. The returns of BTC and the S&P 500 after preprocessing are shown in Figure 2.

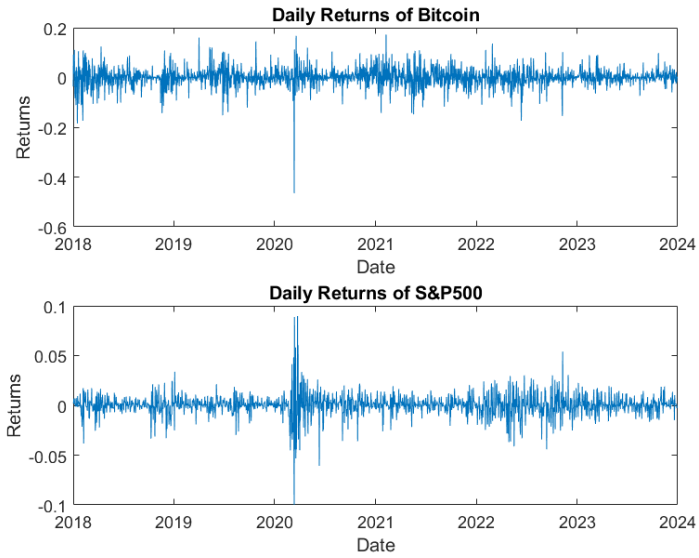


Figure 2: Daily log-returns of Bitcoin and S&P500 respectively. Note the difference in scale.

On 20 April 2020, crude oil futures contracts fell rapidly in price, resulting in negative pricing of this asset (U.S. Energy Information Agency, 2021). As this date is in the modeling set, the negative pricing causes a problem when using the log difference on the price. To solve this issue, we take the log difference between the closest positive closing prices of the affected days, 17 April and 21 April 2020, and average the return over those days.

In order to get a Google Trends index of daily observations over the entirety of the time period, the daily index data is normalized with respect to the monthly index data. The following formula is used to normalize the observations:

$$\text{Adjusted index} = \frac{\text{Daily index} \cdot \text{Monthly index}}{100}$$

where the monthly index is fixed for every observation of the daily index for each month. The resulting adjusted index is shown in Figure 3.

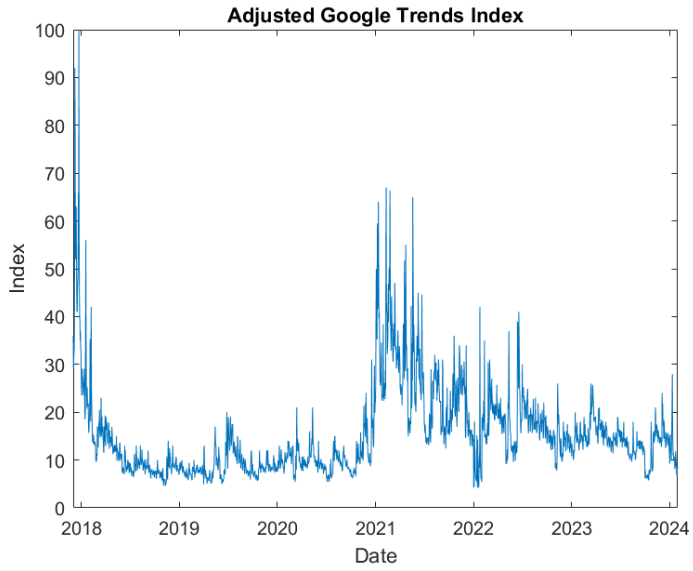


Figure 3: The adjusted Google Trends index which shows the daily trend index between 2017-12-01 and 2024-01-31.

3.2.2 Data Limitations

As previously mentioned, interpolated values are used for days on which no trades occur in the traditional markets. The choice of using interpolated values is made to get more usable data to build the models upon, but also to capture potential weekend effects (Keim, 1984). It is also tested to only use the observations that are available for both datasets, resulting in an approximately 30 % decrease in the number of usable observations. This, however, leads to a multivariate regime division where some of the regimes contain very few observations, making it more difficult to estimate GARCH models on said regimes. Hence, in the rest of the thesis, the models presented are based on datasets containing interpolated values.

3.3 Regime Estimation

Once all the data is preprocessed, we attempt to estimate regimes for the two assets. As Cortese et al. (2023) points out, returns on cryptocurrencies can be explained by external explanatory variables, so the first step in order to estimate the regimes is to find what variables can explain the returns of BTC.

3.3.1 LASSO Regression

We use the LASSO regression analysis method to make a variable selection and regularization. We pick the external explanatory variables that Cortese et al. found to be relevant to BTC returns, as well as other variables that can be thought to have an impact on investors' attitudes towards BTC. The variables that are chosen are returns on gold, oil, the S&P 500 itself, and the VIX index, with a more detailed description in Table 1. The relationship to the general market for the first two has been thoroughly investigated, while the latter two could indicate the state of the general market. These variables were also considered by Cortese et al. but were shown to have no significant effect on BTC returns, as per the assumptions of the LASSO regression.

Table 1: The chosen explanatory variables used for the LASSO regression.

Variable(s)	Code	Description
r_{BTC}	EMA_7	7-day EMA
r_{BTC}	EMA_{14}	14-day EMA
r_{BTC}	RSI_{14}	14-day RSI
r_{BTC}, volume BTC	BTCVol_7	7-day Exponential Linear Correlation
r_{BTC}, volume BTC	BTCVol_{14}	14-day Exponential Linear Correlation
r_{BTC}, Google Trends	BTCGT_7	7-day Exponential Linear Correlation
r_{BTC}, Google Trends	BTCGT_{14}	14-day Exponential Linear Correlation
$r_{BTC}, r_{S\&P500}$	$\text{BTCS\&P500}_7$	7-day Exponential Linear Correlation
$r_{BTC}, r_{S\&P500}$	$\text{BTCS\&P500}_{14}$	14-day Exponential Linear Correlation
r_{BTC}, r_{Gold}	BTCGold_7	7-day Exponential Linear Correlation
r_{BTC}, r_{Gold}	BTCGold_{14}	14-day Exponential Linear Correlation
r_{BTC}, r_{Oil}	BTCOil_7	7-day Exponential Linear Correlation
r_{BTC}, r_{Oil}	BTCOil_{14}	14-day Exponential Linear Correlation
r_{BTC}, VIX	BTCVIX_7	7-day Exponential Linear Correlation
r_{BTC}, VIX	BTCVIX_{14}	14-day Exponential Linear Correlation

The Statistics and Machine Learning Toolbox in **MATLAB** is used to run the LASSO regression. The explanatory variables are normalized, and the regression is done for multiple values of λ on the log scale 10^{-5} to 10^{0}.

3.3.2 Markov Regime Switching Model

To estimate the regimes, the **MATLAB** package MS_Regress by Perlin (2014) is used, which is an implementation of the Markov Regime Switching model (Hamilton, 1989). As with the traditional classification of bull- and bear markets, we make the assumption that both assets can be divided into two different regimes. The returns of BTC are set up to be explained by the trend, residu-

als, and explanatory variables that were found to be relevant from the LASSO regression previously done, according to Equation 22. We make the assumption that all of the variables have the feature of having regime-switching coefficients. The returns of the S&P 500, on the other hand, are set up to be explained by only the trend and the residuals, both having regime-switching coefficients. We estimate the regimes for the assets one by one, with the assumption that the residuals follow both a Normal and a Student's t-distribution.

3.4 GARCH Modeling

Once the regimes are estimated, we extract the data belonging to each regime. This gives us four different series of returns, high and low volatility of BTC and S&P 500 respectively. The data is also categorized in the multivariate volatility setting according to high/high, high/low, low/high, and low/low, where high/high means high volatility for BTC/high volatility for S&P 500, etc. GARCH models are fitted to each of these series, four univariate models and four multivariate models. To compare the model parameters to a benchmark, univariate GARCH(1,1) models and a multivariate DCC-GARCH(1,1) model for the complete dataset of BTC and S&P 500 are also estimated.

3.4.1 Univariate GARCH

To get parameter values for the multivariate model, univariate GARCH(1,1) models are fitted to each regime, with high and low volatility, for both assets. We use the **R** package `rugarch` (Galanos and Kley, 2023) for this purpose, selecting the optimization method "nlminb". As this optimization routine may find a local minimum, even when a better minima is present (Gay, 1990), multiple starting values for the GARCH models are tested. We compare the log-likelihood for each set of starting values for every regime and choose the set of stating values that yield the highest log-likelihood. In one of the states for the S&P 500, the optimal solution found yielded parameters that were far from the optimal parameters of the benchmark model, giving a volatility estimate not commonly seen for assets. In this case, a solution with parameters more similar to the benchmark model was used, together with a visual comparison. The resulting optimized parameters are saved to be used as fixed parameters for the multivariate GARCH models. For the benchmark model, the same procedure is followed, giving a univariate model for both assets, which is used for the multivariate model.

3.4.2 Multivariate GARCH

To estimate a DCC-GARCH for every multivariate regime, the **R** package `rmgarch` (Galanos, 2022) is used. We use the same assumption that the underlying GARCH processes are of the order $p = q = 1$ and use the resulting parameters from the univariate regimes as fixed values. Only the parameters a and b in Equation 42 are estimated, as the parameter estimation tends to be problematic when dividing the regimes even further. This means that in the regime high/high, the parameters from BTC high volatility and S&P500 high volatility are used, etc. As for the univariate case, we choose the optimization method "nlminb". Once the multivariate models are estimated, we use the optimized parameters and run the models on the complete dataset using the "dccfilter" function. This gives us four realizations of the correlation between the assets, as well as four sets of volatility for the assets for all days t in the complete dataset.

3.5 Fuzzy Clustering

To get a smoother transition between the models, we run a fuzzy clustering algorithm on the volatility, trying to divide the volatility estimates into four clusters, each representing the four different multivariate regimes. We use the implementation of this algorithm in the Fuzzy Logic Toolbox in **MATLAB**, using the random number generator seed 1. We achieve this by setting up a time series of daily volatilities, where the volatility on the day t is derived from the model that corresponds to the specific regime that was present on that day. In other words, if day t is in the regime high/high, we use the volatility estimate from day t that is estimated from the MGARCH model based on the regime high/high, etc. Once the time series is set up, we run the algorithm to get probability estimates of each volatility observation belonging to each cluster, as well as the centroids of the clusters.

The final weighted DCC-GARCH model is calculated by using the probabilities that are obtained from the clustering algorithm. A weighted sum is utilized to calculate the final model for each day t. This is done for both the variance and the correlation like below:

$$\sigma^2_{WDG,t} = \sum_{k \in D} = p_{k,t}\sigma^2_{k,t}$$

$$\rho_{WDG,t} = \sum_{k \in D} = p_{k,t}\rho_{k,t}$$

where WDG stands for weighted DCC-GARCH and $D = \{L/L, L/H, H/L, H/H\}$. D is a representation of the different multivariate regimes that are estimated, where H stands for High and L stands for Low regime. To ensure that the correct probability p is used, we compare the centroids of the clusters, which are marked in Figure 4, to figure out which centroid belong to what regime. The probabilities p are also plotted and compared with the estimated regimes to facilitate the categorization. They are presented in the results.

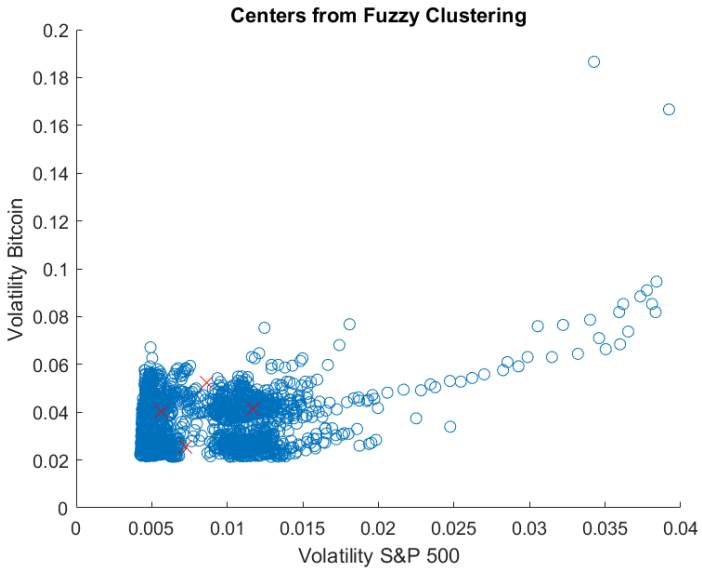


Figure 4: Centers from the Fuzzy Clustering algorithm in relation to the resulting volatilities from the final model.

4 Empirical Results

4.1 Data

Analysis of the returns in Figure 2 reveals that both series exhibit negative skewness, as well as excessive kurtosis compared to what is expected from data following a normal distribution. To verify if the data follows a normal distribution or not, a Jarque-Bera test is conducted, both on all data, but also with outliers outside three standard deviations removed. The **MATLAB** function "jbtest" in the Statistics and Machine Learning Toolbox is used for this purpose. The resulting statistics are presented in Table 2, where numbers in parenthesis denote data with outliers removed. As Table 2 also shows, the hypothesis that the data follows a normal distribution is rejected in both cases. The critical value is determined from Monte-Carlo simulation, as the datasets contain more than 2000 observations.

Table 2: Descriptive statistics of the return series of BTC and S&P 500. Jarque-Bera test is rejected at a 5% significance level, as the critical value is 5.9656.

	BTC	S&P 500
Mean	4.9922e-04 (0.0011)	2.6102e-04 (5.0835e-04)
Standard Deviation	0.0368 (0.0319)	0.0099 (0.0076)
Skewness	-1.0541 (0.0343)	-0.2378 (-0.1174)
Kurtosis	17.2405 (4.7455)	18.8866 (5.2494)
Jarque-Bera Test Statistic	1.8919e+04 (274.9019)	2.3061e+04 (458.2050)

We note that the Jarque-Bera test statistic is remarkably high, but normalization of the data does not significantly change this score. The skewness and kurtosis seems to be highly influenced by the outliers that are present in the data above three standard deviations. Removing these has a great impact on the Jarque-Bera test statistic as well, but the hypothesis of normally distributed data is still rejected.

The parameters from a distribution fitting with the **MATLAB** function "fitdist" for a Normal- and a location-scale Student's t-distribution are presented in Table 3. A visual comparison of the empirical distribution and the two fitted distributions is shown in Figure 5, where the location-scale Student's t-distribution gives a better fit.

Table 3: Resulting parameters from the distribution fitting of the returns series.

	BTC		S&P 500	
Parameter	Normal	Student's t	Normal	Student's t
μ	0.000499	0.000930	0.000261	0.000443
σ	0.036816	0.020021	0.009857	0.003934
ν	-	2.295800	-	1.660570

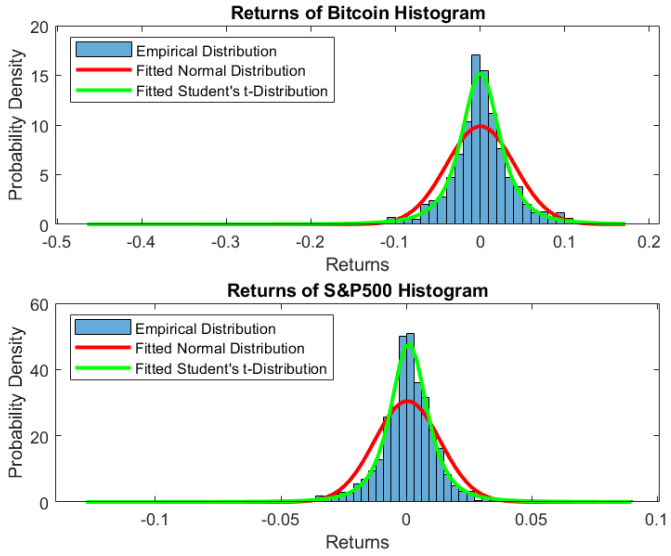


Figure 5: Fitted Normal- and Student's t-distributions to the returns.

The autocorrelation function for BTC shows no dependence in the returns, and a weak dependence in the absolute returns, visualized in Figure 6. The autocorrelation function for the S&P 500 shows some dependence in the returns, and a strong decaying dependency in the absolute returns, visualized in Figure 7. The results are generally in line with the stylized facts, even if some dependency is shown for the weak decaying dependency for the absolute BTC returns and the sprawling dependency in the S&P 500 returns. One thing worth noting is the apparent seven-day dependence for the absolute returns, mostly visible for the S%P 500.

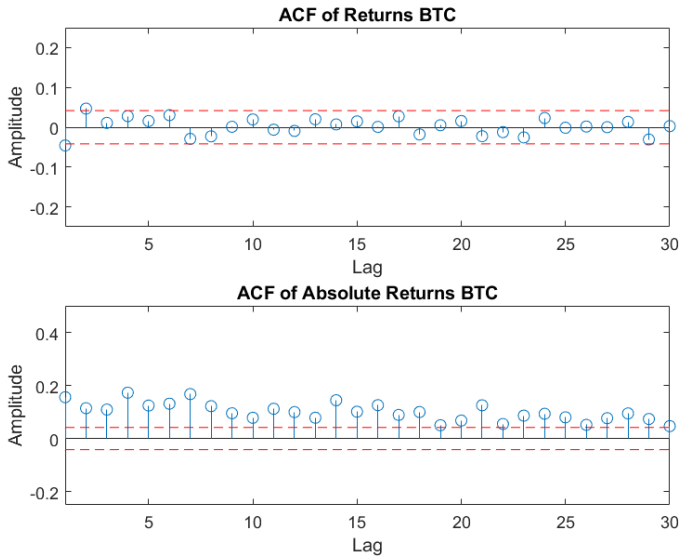


Figure 6: Auto-correlation function of BTC.

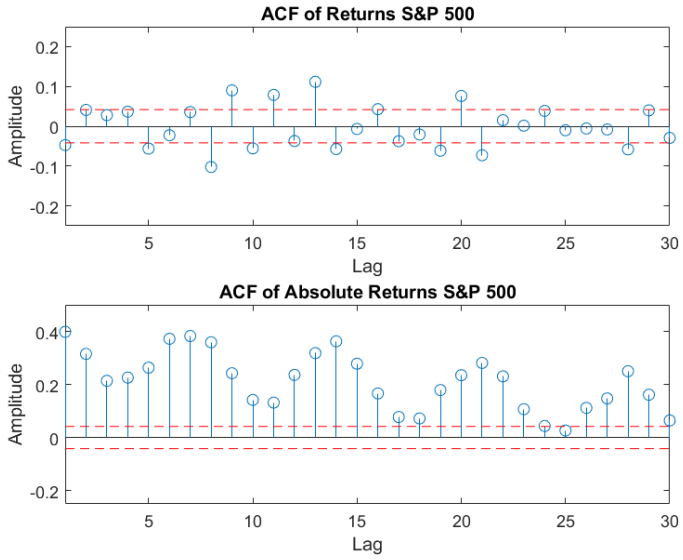


Figure 7: Auto-correlation function of S&P 500.

4.2 Regime Estimation

The resulting LASSO regression can be seen in Figure 8. We choose to include the first seven significant variables, which are EMA_7, EMA_{14}, RSI_{14}, BTCVol_{14}, BTCGT_7, BTCOil_7 and BTCVIX_{14}, to use as explanatory variables for BTC returns.

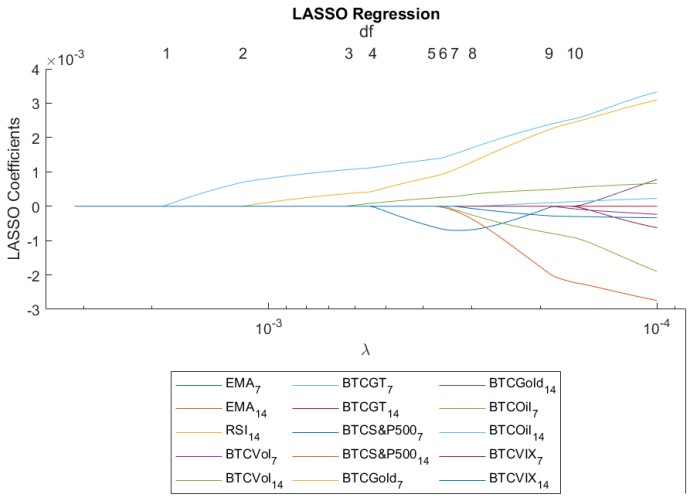


Figure 8: LASSO Regression of explanatory variables for BTC returns, when varying the penalizing parameter λ.

The coefficients in the plot represent the weight or influence of each of the 15 parameters on the target variable, i.e. BTC. In lasso regression, some of these coefficients may be driven to zero, effectively removing those variables from the model. In the figure above, we can see that the first two coefficients differ from zero before any of the others. The last five coefficients that were chosen differ from zero a little bit later, but still clearly before the next coefficient. Therefore, the first seven coefficients are chosen as explanatory variables for BTC returns.

Using the assumption that the returns follow a student's t-distribution from the previous section gives us the following models for the returns:

$$r_{BTC,t} = \mu_{S_t} + \sum_{i=1}^{7} \beta_{i,S_t} x_{i,t} + \varepsilon_t$$

and

$$r_{S\&P500,t} = \mu_{S_t} + \varepsilon_t$$

where i denotes the i:th explanatory variable. The errors ε_t are set up to follow a student's t-distribution with two distinct regimes. Using the MS_Regress package, the following regimes are estimated, as seen in Figure 9. An in-depth analysis of the probabilities of each regime can be found in Figures A.1 and A.2 in the Appendix. An attempt to estimate Markov Switching Models with normally distributed innovations is also made. The resulting regimes can be seen in Figure A.3 in the Appendix. We reject this model due to unreasonably rapid regime switching, in favor of the above-presented student's t-distributed model.

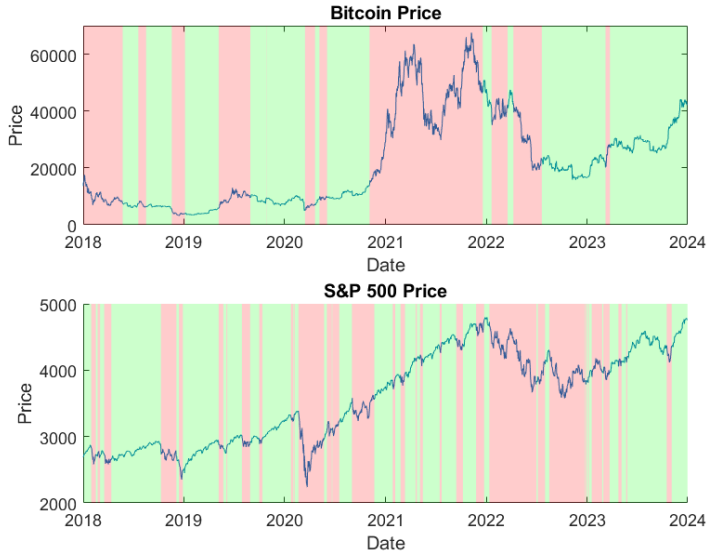


Figure 9: The estimated regimes from the **MATLAB** package MS_Regress. Green indicates a low volatility period, while red indicates a high volatility period.

The results from the MS_Regress package are presented in Tables 4 and 5. As seen by the p-values, no variable except the variance and degrees of freedom are

significant in both regimes for both assets.

Table 4: The estimated parameters for the BTC Markov Switching Model from the MS_Regress package for.

BTC	Regime 1		Regime 2	
Variable	Estimate	P-value	Estimate	P-value
σ^2	0.000160	**0.000000**	0.001223	**0.000000**
ν	2.091605	**0.000000**	4.710794	**0.000011**
μ	-0.006474	0.050397	0.003818	0.662142
EMA_7	0.033563	0.865363	-0.417033	0.142283
EMA_{14}	-0.714545	0.060177	0.467384	0.393693
RSI_{14}	0.000139	**0.023986**	-0.000018	0.910576
$BTCVol_{14}$	-0.144250	0.634354	0.827640	0.138878
$BTCGT_7$	1.257455	**7.1758e-08**	-0.322392	0.412766
$BTCOil_7$	0.124794	**0.004923**	-0.134807	0.366555
$BTCVIX_{14}$	-0.040814	**0.048251**	-0.004323	0.967530

The expected duration of Regime 1 is 95.09 days and Regime 2 is 75.54 days, with the following transition matrix:

$$\boldsymbol{P} = \begin{pmatrix} 0.9895 & 0.0132 \\ 0.0105 & 0.9868 \end{pmatrix}$$

Table 5: The estimated parameters for the S&P 500 Markov Switching Model from the MS_Regress package.

S&P 500	Regime 1		Regime 2	
Variable	Estimate	P-value	Estimate	P-value
σ^2	0.000006	**0.000000**	0.000072	**0.000000**
ν	1.966624	**0.000000**	2.949946	**3.4417e-14**
μ	0.000607	**1.7186e-11**	-0.000209	0.562328

The expected duration of Regime 1 is 26.61 days and Regime 2 is 20.53 days, with the following transition matrix:

$$\boldsymbol{P} = \begin{pmatrix} 0.9624 & 0.0487 \\ 0.0376 & 0.9513 \end{pmatrix}$$

By comparing the two models, we can see that the regimes are longer for BTC, supported by the higher probabilities of staying in the same state in the transition matrix. By observing the variance in the regimes, we notice that in Regime 1, the variance is lower than in Regime 2 for both assets, leading us to classify the

regimes as low- and high-volatility regimes, respectively. After the data is sorted by regime, we get a return series for each regime and asset, shown in Figures 10 and 11. These return series are used to estimate four different GARCH models. In the figures, the volatility clustering pattern is not as visible for BTC high volatility and S&P 500 low volatility as it is for the complete dataset. More clear patterns of volatility clustering can be seen for the other return series. One common observation between the assets is that returns in the low-volatility regimes are generally of lower magnitude than in the high-volatility regimes.

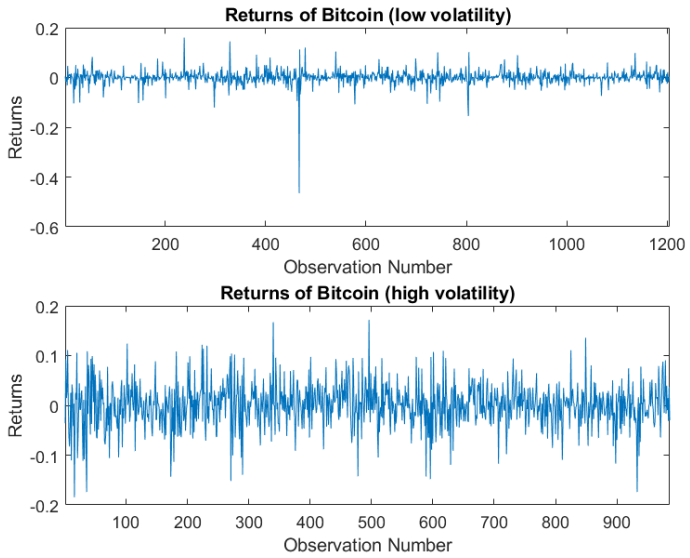


Figure 10: Returns of BTC in the respective regimes.

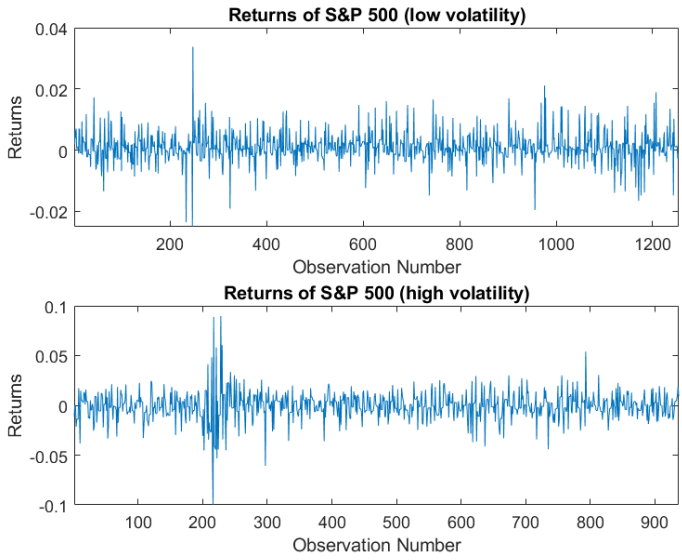


Figure 11: Returns of the S&P 500 in the respective regimes.

4.3 GARCH Modeling

4.3.1 Univariate

We use the result from the previous section that no external explanatory variables were significant in both regimes when estimating the GARCH models. As such, we make a simplifying assumption that the returns only depend on a drift term and an innovation term. The drift term μ is considered to make the return series zero-mean and to compare it with the estimates from the MS_Regress package. The resulting parameters for both Normal and Student's t-distributed innovations are presented in Tables 6 - 9.

Table 6: GARCH parameter estimates from the BTC low volatility regime.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.001274	0.128316	0.000374	0.390149
ω	0.000115	**0.005598**	0.000460	0.128475
α	0.156039	0.206355	0.186374	**0.005283**
β	0.738313	**0.000000**	0.813626	**0.000000**
ν	-	-	2.112606	**0.000000**
Log-likelihood	2613.961		2921.102	

Table 7: GARCH parameter estimates from the BTC high volatility regime.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.001080	0.457632	0.001346	0.279432
ω	0.000106	**0.036816**	0.000081	0.095858
α	0.038664	**0.006604**	0.040373	**0.010011**
β	0.907228	**0.000000**	0.920498	**0.000000**
ν	-	-	5.448768	**0.000000**
Log-likelihood	2776.811		1701.126	

For BTC, we can see that the estimates for μ are not significant in either regime or choice of distribution. The estimates are also different from the Markov Switching Model, suggesting a high uncertainty in the parameter. We can also see that ω is not significant when considering a student's t-distribution, implying that the long-term unconditional volatility would tend towards zero. This implication would not make sense for financial returns, and we are as such sceptical of a rejection of this parameter. In Table 6, we notice a rejection of the α parameter when considering normally distributed innovations, but not in the case of student's t-distributed innovations. The estimates are, however, similar, leading us to speculate that the estimate is correct.

Table 8: GARCH parameter estimates from the S&P 500 low volatility regime.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.000926	**0.000000**	0.000603	**0.000000**
ω	0.000001	**0.000000**	0.000003	**0.000000**
α	0.013780	**0.021064**	0.038681	**0.000000**
β	0.948049	**0.000000**	0.952273	**0.000000**
ν	-	-	2.142758	**0.000000**
Log-likelihood	4959.747		5134.75	

Table 9: GARCH parameter estimates from the S&P 500 high volatility regime.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	-0.000541	0.139060	-0.000314	0.374470
ω	0.000009	**0.000000**	0.000008	**0.000000**
α	0.062370	**0.000000**	0.057006	**0.000000**
β	0.884572	**0.000000**	0.893516	**0.000000**
ν	-	-	4.825818	**0.000000**
Log-likelihood	2776.811		2808.071	

For S&P 500, we notice that μ is not significant in the high-volatility regime. The estimates are, however, similar to those of the Markov Switching Model. The resulting parameters in Table 8 do not give the optimal log-likelihood but are chosen to give a more familiar GARCH structure. The optimal solution yields an estimated β of zero, with a p-value of one, suggesting that the estimate should be rejected. The optimal parameters and the respective volatility plot of the modeling data can be found in the Appendix in Table A.2 and Figure A.4. One interesting observation from the chosen GARCH parameters is that a majority of the p-values are exactly 0, sparking a debate of how robust these estimates are.

By comparing the parameters for both Normal- and Student's t-distributed innovations, we see that the estimates are similar for the two distributions are in large similar, the only exception being the β parameter for the BTC low volatility regime. We also notice that the degree of freedom, denoted ν, is higher for the high-volatility regime for both assets.

4.3.2 Multivariate

When estimating a DCC-GARCH model in each multivariate regime, we fix the parameters from the univariate models and only estimate the a and b parameters. For two of the multivariate regimes, a model could not be estimated when considering a Student's t-distribution, and we have as such resorted to only presenting the Normal model for consistency purposes. The solver failed to converge, even when considering non-fixed parameters and change of starting parameters. The resulting a and b parameter estimates are presented in Table 10 - 13.

Table 10: The a and b parameters of the DCC-GARCH estimated for the Low/Low regime.

	Normal	
Parameter	Estimate	P-value
a	0.049362	**0.005432**
b	0.918722	**0.000000**
Log-likelihood	4607.917	

Table 11: The a and b parameters of the DCC-GARCH estimated for the Low/High regime.

	Normal	
Parameter	Estimate	P-value
a	0.026138	**0.019928**
b	0.967250	**0.000000**
Log-likelihood	2385.13	

Table 12: The a and b parameters of the DCC-GARCH estimated for the High/Low regime.

	Normal	
Parameter	Estimate	P-value
a	0.000895	0.907320
b	0.988993	**0.000000**
Log-likelihood	2927.574	

Table 13: The a and b parameters of the DCC-GARCH estimated for the High/High regime.

	Normal	
Parameter	Estimate	P-value
a	0.019571	**0.029622**
b	0.970640	**0.000000**
Log-likelihood	2178.072	

We observe that a and b differ between the multivariate regimes, giving each regime its own characteristics when it comes to updating the correlation. In the High/Low regime, the starting value of b must be set to at least 0.975 in order to get the presented estimates. If not, the b parameter will be estimated to be around 0.55, as presented in Table A.1 in the Appendix, resulting in a correlation estimate that tends to be noisy. It is in this regime that the only insignificant parameter can be found, a.

In Figures 12 - 15, the resulting volatilities and correlation between the assets are plotted. For the volatility, we can notice a slight difference in magnitude and responsiveness between the low and high volatility regimes. As the GARCH parameters are fixed, the same volatility plot is given for both multivariate regimes that contain the same univariate regime. This means, for example, that both regimes that contain the low volatility state for BTC give the same volatility plot.

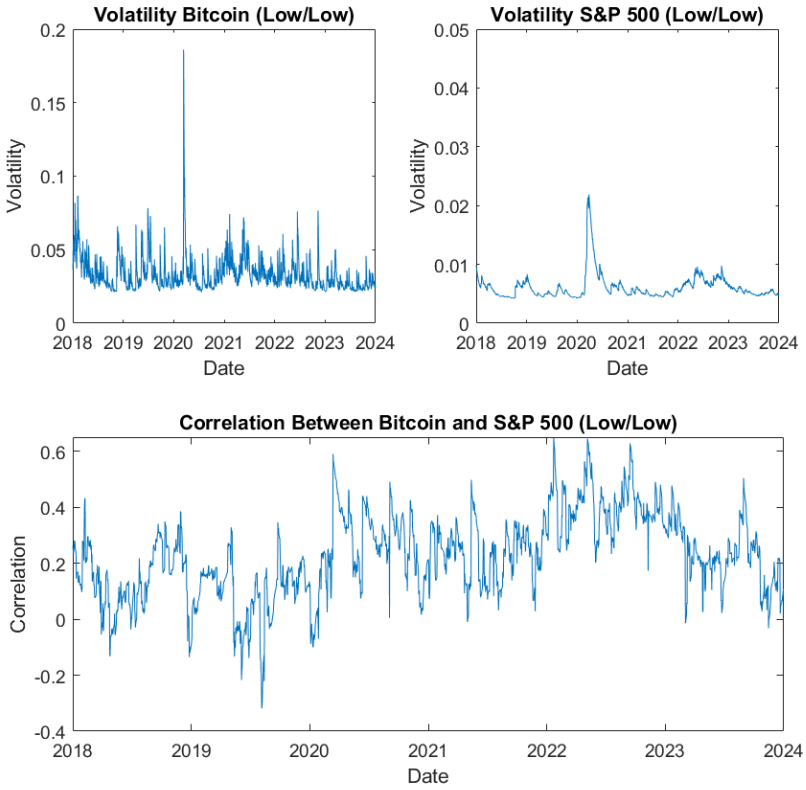


Figure 12: Volatility and correlation estimated for the Low/Low regime when the resulting parameters are used for the complete dataset.

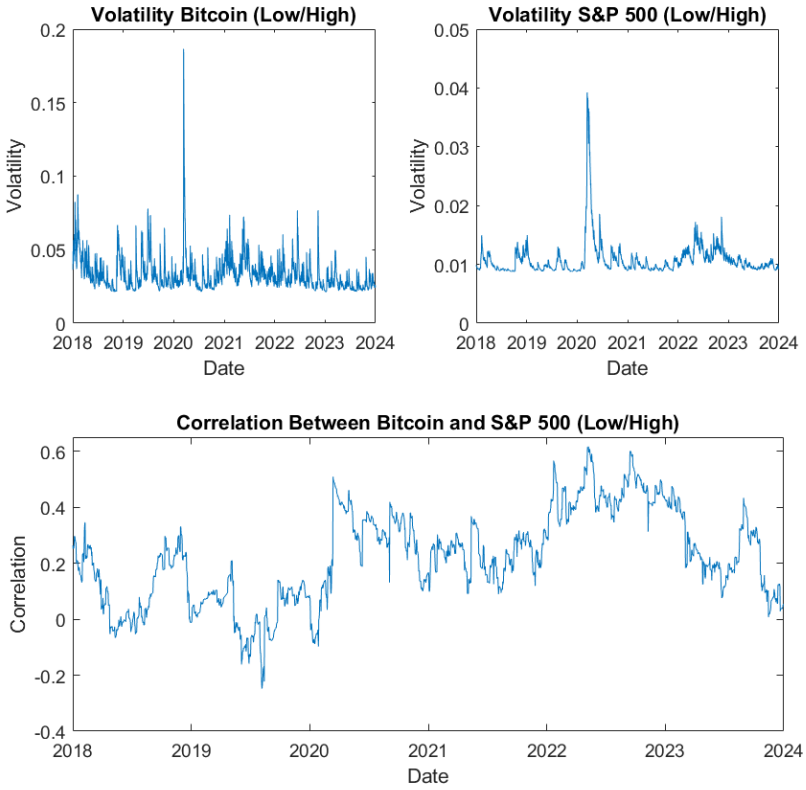


Figure 13: Volatility and correlation estimated for the Low/High regime when the resulting parameters are used for the complete dataset.

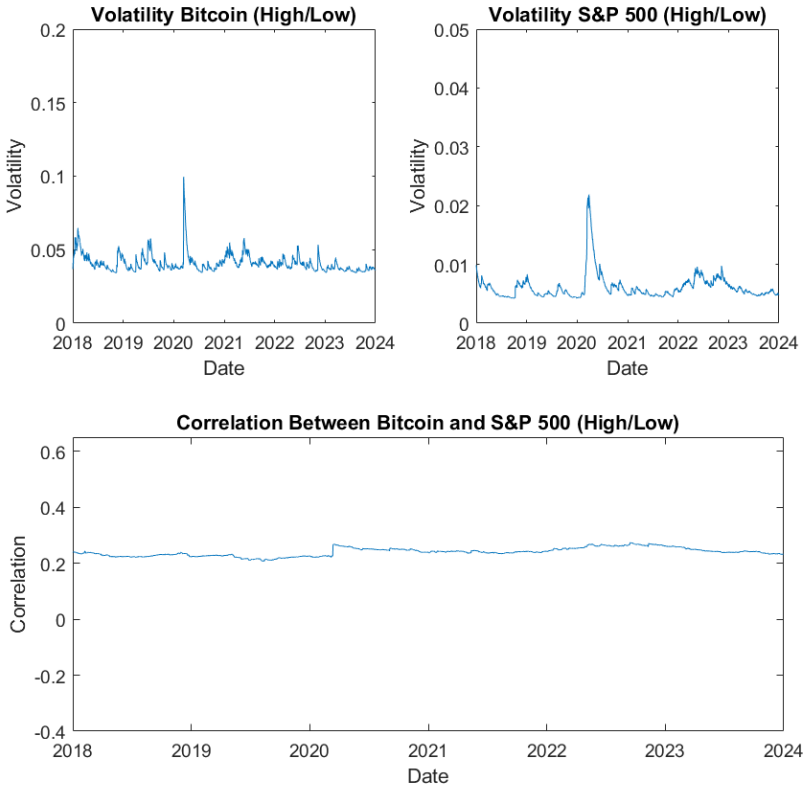


Figure 14: Volatility and correlation estimated for the High/Low regime when the resulting parameters are used for the complete dataset.

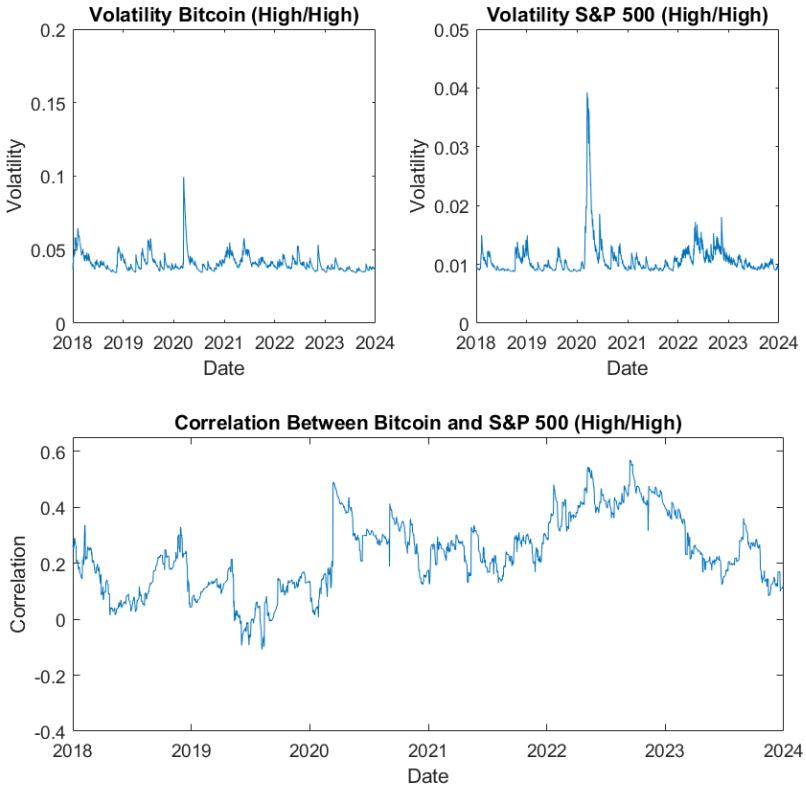


Figure 15: Volatility and correlation estimated for the High/High regime when the resulting parameters are used for the complete dataset.

All but one regime gives a similar correlation estimate over the period. The High/Low state seems to give an estimate which is almost constant over the period, which is seen in Figure 14. If we zoom in on the correlation plot, we can, however, see that it has the same structure as the other regimes. This plot can be seen in Figure A.5 in the Appendix. The behavior of the correlation for this regime can be explained by the estimated parameters for that regime previously presented. The other regimes tend to only differ in how reactive they are to new innovations.

4.3.3 Benchmark Model

A benchmark model is estimated on the complete data set without regime classification. The resulting parameters for the univariate GARCH models are found

in Tables 14 and 15, while the DCC-GARCH parameters are found in Table 16.

Table 14: GARCH parameters for BTC estimated for the complete dataset.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.001268	0.098987	0.000638	0.172110
ω	0.000076	**0.009178**	0.000011	0.477870
α	0.102600	**0.019393**	0.075159	**0.000000**
β	0.848816	**0.000000**	0.924841	**0.000000**
ν	-	-	3.132831	**0.000000**
Log-likelihood	4251.597		4561.014	

Table 15: GARCH parameters for S&P 500 estimated for the complete dataset.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.000643	**0.000045**	0.000570	**0.000000**
ω	0.000002	0.856532	0.000002	0.863510
α	0.111168	0.138795	0.123909	0.312790
β	0.876391	**0.000000**	0.876091	**0.000000**
ν	-	-	2.868181	**0.000000**
Log-likelihood	7520.001		7743.631	

For BTC, we notice the same phenomenon where the μ parameter is rejected and the ω parameter is rejected in the case of Student's t-distributed innovations as for the regime case. With the same argument, we are skeptical of the latter rejection, as ω should be positive. For the S&P 500, we notice a rejection of the ω parameter for both distributions, contrary to the previous results for the regimes. We are, once again, skeptical of this rejection. We can also observe that the α parameter is not significant for either distribution, which is also something not observed for the regime models. The estimates are, however, similar for both distributions, leading us to trust the estimates.

Table 16: DCC-GARCH parameters for the benchmark model.

	Normal	
Parameter	Estimate	P-value
a	0.012879	**0.019148**
b	0.983257	**0.000000**
Log-likelihood	11850.06	

As for some of the regime models, the benchmark model failed to converge when

considering Student's t-distributed innovations. The model is, as such, only presented for normally distributed innovations. The parameter estimates for a and b are both significant, where the relatively large b gives the most weight to the previous observation of the correlation matrix. The resulting volatility and the corresponding correlation over the period between the assets are visualized in Figure 16.

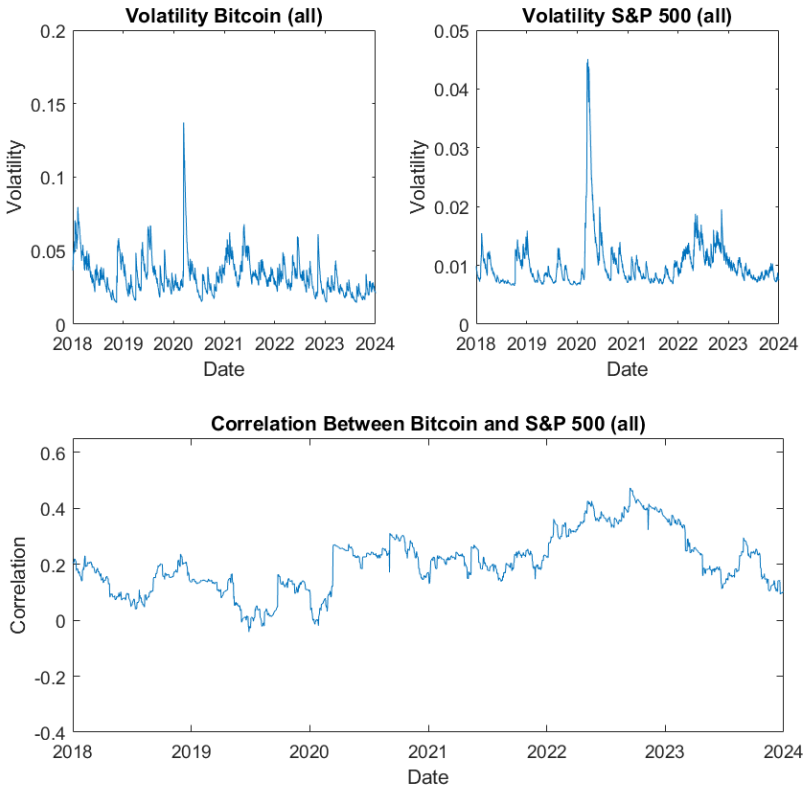


Figure 16: Volatility and correlation estimated for complete dataset.

4.4 Fuzzy Clustering

After running the fuzzy clustering algorithm with four clusters, we get the probability of every observation belonging to each cluster at each time point and the respective cluster center. The probabilities are shown in Figure 17 and the centers are found in Table 4.

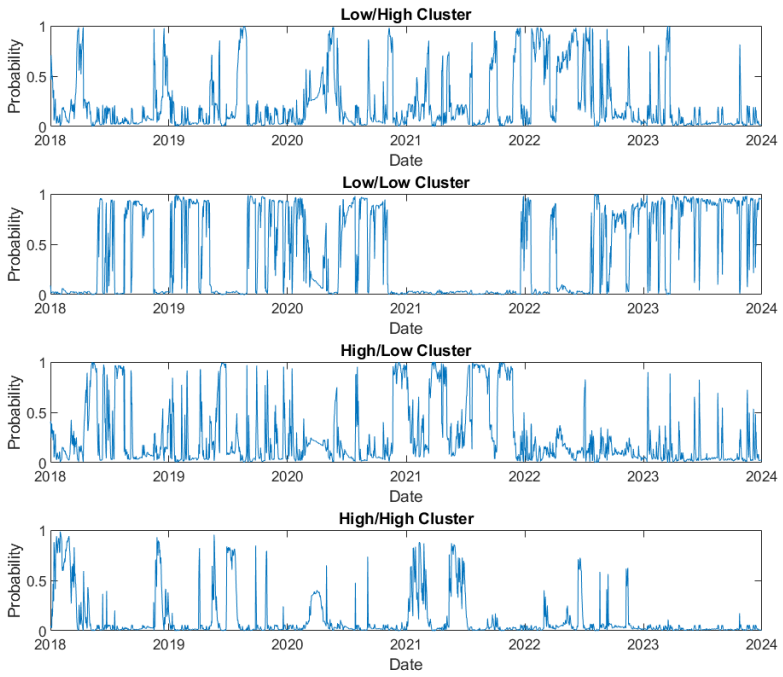


Figure 17: The estimated regimes from the **MATLAB** package MS_Regress. Green indicates a low volatility period, while red indicates a high volatility period.

Table 17: Centers of the clusters from the fuzzy clustering algorithm.

	BTC	S&P 500
Cluster	Volatility Center	Volatility Center
Low/High	0.041608	0.011676
Low/Low	0.025475	0.007215
High/Low	0.040426	0.005634
High/High	0.052246	0.008648

To be able to determine what probabilities belong to what regime, we use the regime division obtained from the MS Model and compare it with the probability plot in Figure 17. We can generally link a high probability in the plot to the obtained regime division, but we also take note of the cluster centers. The centers for regime Low/Low and High/High in Table 17 are determined by being the lowest and highest volatility for both assets simultaneously. The regimes

Low/High and High/Low are more difficult to determine, as the volatility center for BTC is similar in both cases. We make the decision based on the volatility center of the S&P 500 for those regimes, as the difference is more noticeable. We do notice, however, that the categorization of High and Low volatility regimes is hard only based on the cluster centers, as some of the centers are close to each other. For a visual comparison, we refer to Figure 4 in the Methodology Section.

The volatility and correlation from the GARCH models that were estimated on the four different multivariate regimes in Section 4.3 are weighted according to the probabilities visualized in Figure 17. The resulting volatilities and correlation for both assets are shown in Figure 18. The correlation from the weighted model is also shown in Figures 19 - 21, together with the benchmark, and rolling 30- and 60-day correlations between the returns.

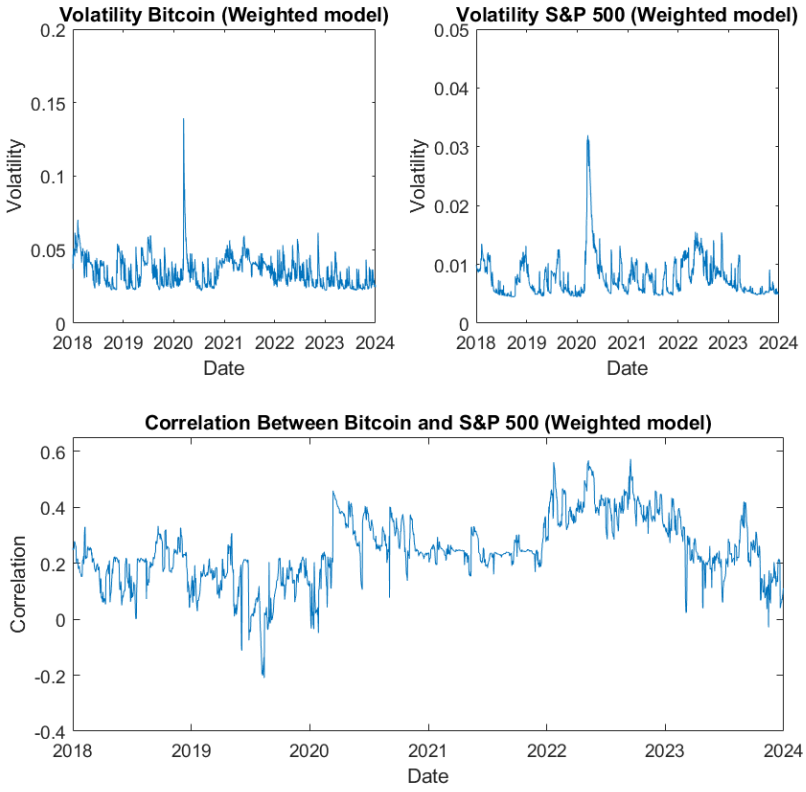


Figure 18: Volatility and correlation estimated for the final weighted model.

The obtained volatility over the period from the final weighted model seems to be reasonable. It has a recognizable GARCH structure, which is also obtained from the benchmark model, and it seems to be able to capture volatility clustering in a good way. We also notice a slightly higher correlation from the beginning of 2020 until the middle of 2023, compared to before 2020.

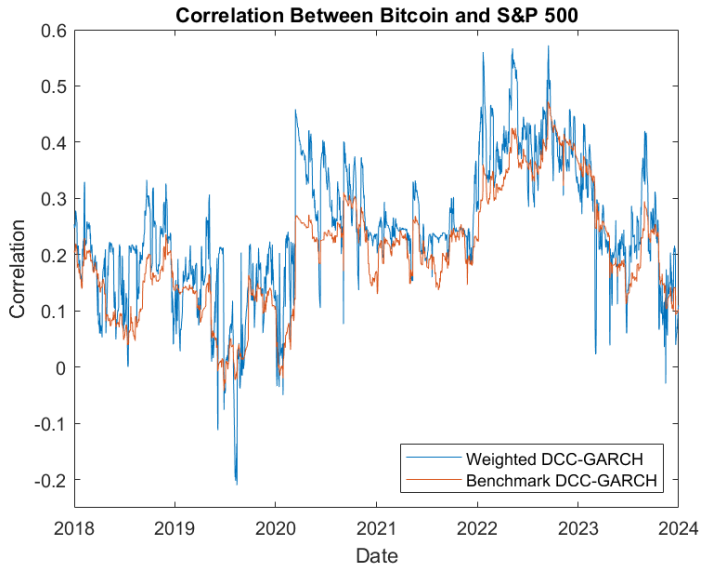


Figure 19: The weighted DCC-GARCH model compared with benchmark DCC-GARCH correlation between the assets.

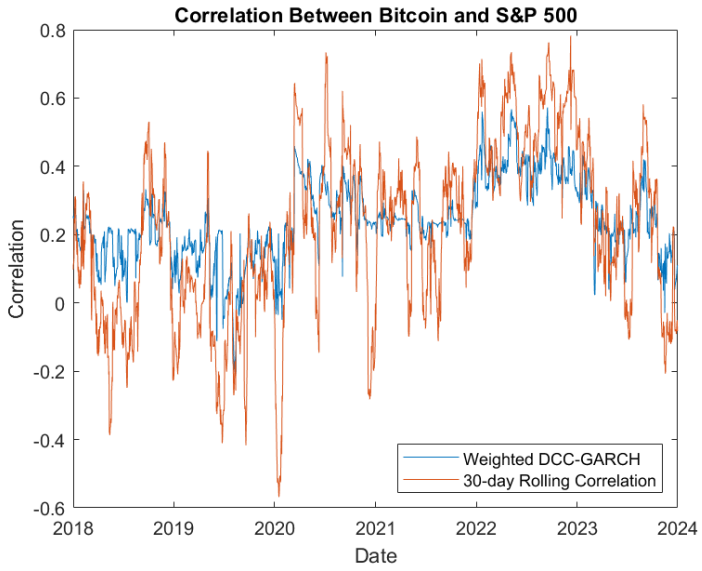


Figure 20: The weighted DCC-GARCH model compared with the 30-day rolling correlation between the assets.

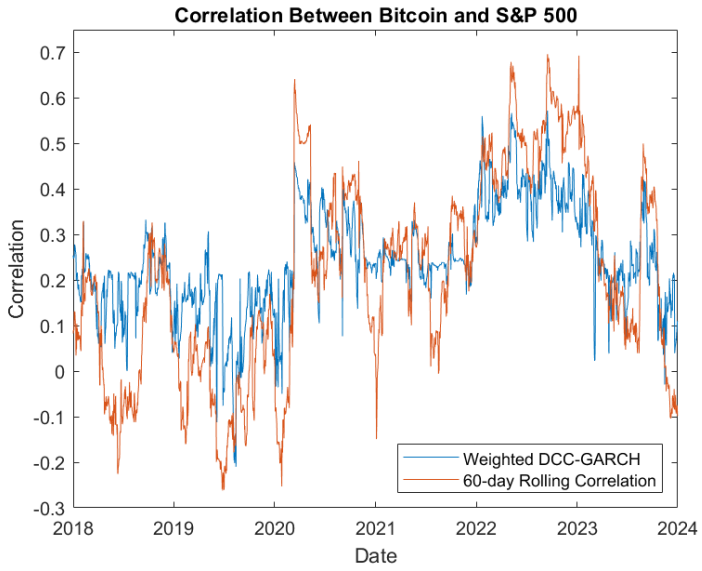


Figure 21: The weighted DCC-GARCH model compared with the 60-day rolling correlation between the assets.

When comparing the obtained correlation over the period of the final weighted model and the above-mentioned measures, there are two main observations to be mentioned. When comparing to the benchmark DCC-GARCH model in Figure 19, we notice that the final model follows the same structure of the correlation throughout the period. As with the DCC-GARCH model, we notice a downward trend in the correlation from 2023 onwards. The correlation from the weighted model is, however, more reactive to changes in correlation than the DCC-GARCH. When comparing with the 30- and 60-day rolling correlations in Figures 20 and 21, we notice that the weighted model is not as reactive as the 30-day correlation measure, but more similar to that of the 60-day correlation. We do notice that the structure and shape of the correlation differs more between these two measures and the weighted model than that of the DCC-GARCH model. From the 30- and 60-day correlation it is relatively easy to notice the 30- and 60-day windows. As an example, at the beginning of 2020, we see a spike in correlation, and after 30 and 60 days respectively, it drops, something that is not present in the weighted model.

5 Discussion

5.1 Datasets

5.1.1 Data filtering

At the beginning of this thesis, we plotted the autocorrelation functions of both assets and observed that the S&P 500 returns were not white noise, as there seemed to be some dependency. For BTC returns, the autocorrelation function did not show any dependency Therefore, we wanted to try using data filtering techniques in order to capture the autocorrelation in S&P500 returns. We specifically used AR(1) and ARMA(1,1) models as these models looked most relevant when observing the autocorrelation functions.

However, after implementing these filtering techniques, we did not observe any significant improvement in the final results of the DCC-GARCH model. In other words, the parameters of the final model did not change significantly after applying any data filtering techniques. This led us to opt for an unfiltered DCC-GARCH model in the final analysis. There were key reasons behind this decision. We wanted to keep the final model as simple as possible. By avoiding unnecessary filtering steps, we aimed to maintain a simpler and more interpretable model structure. While filtering was not beneficial in this specific case, future research could revisit its application by trying other filters than the AR(1) and ARMA(1,1) that were tried in this research.

Although our final model did not include any filtering techniques, the autocorrelation function of the S&P500 in Figures 7 and A.7 in the Appendix revealed an interesting pattern. There seems to be a presence of a 7-day periodicity, which aligns with the well-documented "weekend effect" in stock markets. This effect suggests that returns might be influenced by untraded weekend days, impacting prices when markets reopen on Mondays. This potential dependency in the S&P500 data aligns with the observed periodicity in the autocorrelation function figures. Interested in the weekend effect, we explored relevant research. In the paper written by Keim (1984), they found consistently negative Monday returns, further suggesting the potential benefits of data filtering for capturing periodicities. However, in our case, implementing weekend filtering did not significantly improve the final model's performance. This led us to maintain the unfiltered approach.

5.1.2 Interpolated values

As mentioned in Section 3.2.1, closing prices of the S&P 500 are missing on weekends and on holidays. To address this, we used linear interpolation to estimate the missing values. Using interpolation comes with both benefits and potential problems. While it can offer a practical solution for maintaining a continuous dataset, it's crucial to acknowledge its potential influence on the analysis of the correlation between BTC and S&P 500.

On the one hand, interpolation offers some benefits. First and foremost, it ensures a continuous time series, which makes it possible to apply statistical methods that require a complete time series. Additionally, for short gaps like weekends or holidays, interpolation may introduce minimal errors. This is particularly the case for major stock indices like the S&P 500, which are stable, and daily price movements, in general, are smaller.

On the other hand, interpolation also has its limitations. Firstly, it introduces artificial data points that may not reflect the true market movements. This could potentially affect the accuracy of our correlation estimates, especially in periods with higher volatility where the true price movements might be a lot different than the interpolated values. Secondly, it might be the reason why we observe the strong periodicity of the autocorrelation function in both Figure 7 and A.7 in the Appendix. Lastly, the interpolation method one uses might affect the final results. In this thesis, we used a basic linear interpolation, which may underestimate actual volatility. For future research on this topic, perhaps one could try using other interpolation methods such as polynomial interpolation and spline interpolation, and taking the average of these values.

5.2 Regime Estimation

We used a Markov-switching model to segment the return data for S&P 500 and Bitcoin into distinct states. This approach has several advantages. By segmenting data into regimes with similar characteristics, the model can potentially capture more nuanced relationships within each regime. This can lead to a better overall fit compared to a single-regime model. Identifying distinct regimes allows for a clearer understanding of how the relationship between S&P 500 and Bitcoin returns behaves under different market conditions, providing valuable insights for investors.

However, the application of the Markov-switching model presented some challenges for Bitcoin data. Notably, the model seemed to prioritize volatility over

return direction when segmenting regimes. This resulted in up-trends being classified as "bear" states (red) due to their high volatility, while down-trends were classified as "bull" states (green) due to lower volatility. Meanwhile, it was the opposite for the S&P 500 data, where up-trends were classified as bull states, and down-trends as bear states. This would normally be the general interpretation, but this was not the case for the BTC data.

To address this issue, we attempted two different strategies. First, we incorporated technical analysis variables identified by LASSO regression to potentially influence the regime classification beyond just volatility. These variables are presented in Section 4.2 of this thesis. However, this did not affect the state segmentation that much for the Bitcoin data, and the problem with up-trends being classified as bear states remains. Second, we experimented with a three-state model to potentially capture a wider range of market conditions. However, this increased model complexity and difficulty of interpretation, with some states having limited data points, potentially leading to unreliable parameter estimates when applying GARCH models.

5.3 GARCH Modeling

5.3.1 Choice of Model

Estimating the GARCH models on each respective regime tended to be more difficult than initially anticipated. With a plethora of different types of GARCH models to choose from, and with different types of distributions possible for the innovations, selecting the correct one will be a trade-off between the model's ability to capture the volatility and ease of estimating the parameters. In the multivariate case, the latter becomes harder due to the increase in the number of parameters needed to be estimated. As previously mentioned, we have chosen a standard DCC-GARCH$(1, 1)$ model, and use the assumption that the innovations are Normally distributed. With these assumptions, we ran into trouble when trying to estimate the GARCH model for the S&P 500 low volatility regime, as we needed to choose a suboptimal solution to avoid the parameters of an ARCH model. We conclude that an ARCH model is not satisfying when comparing the resulting volatility estimates from this model with that of the other regimes. For the ARCH model, the estimates are spiky and do not show the classical signs of volatility clustering, where the volatility keeps a higher level for multiple consecutive days. As such, we reject this model in favor of a GARCH model.

To combat the phenomena of estimating an ARCH model, we investigate the

possibility of extreme outliers affecting the result. With outliers removed, we try to re-estimate the model, without success. We also investigate the log-likelihood function, examining which data point affects it the most, and remove those. This yields the same results as before, only changing the original ARCH parameter estimates by a few second decimal points. As previously mentioned, including an AR(1) or ARMA(1,1) mean model does not change the result when removing outliers either, leading us to believe that a GARCH approach for modeling the variance in this regime might be wrong. We have not investigated the possibility of other types of GARCH models and will be left for further research.

5.3.2 Student's *t*-distribution

As expected from financial returns data, the data used exhibits heavy tails, even after dividing the data into different regimes. With this in mind, we consider GARCH models using the Student's t-distribution. The resulting parameter estimates are in large similar to those of using the Normal distribution, with the exception of the BTC low volatility regime, where the β coefficient is significantly higher for the Student's t-distribution. One key difference between the two assumptions is that the parameters tend to go toward an IGARCH model for the low volatility data when assuming a Student's t-distribution. This could be seen as problematic, as it could suggest persistence in the volatility, that the data is not stationary, or that the volatility can not be adequately captured by the simpler GARCH model. As seen from Equation 37, the long-term unconditional variance will also not be defined in the case of an IGARCH model. By examining the autocorrelation function in Figure A.6 in the Appendix, we can see a ringing pattern for the absolute returns of the S&P 500, rather than a decaying one, suggesting either a seasonality or non-stationarity. The same pattern is not shown for BTC, leaving us not able to draw any sound conclusions as to why IGARCH-like parameter estimates are obtained.

A similar problem also arises when trying to estimate the parameters for the S&P 500 low volatility regime when using the Student's t-distribution, where a suboptimal solution must be chosen to get reasonable values. Choosing the solution with the highest log-likelihood gives a model that solely depends on past shocks, giving a noisy and spiky volatility estimate that is not typically seen for financial data, as per the stylized facts. The volatility can be at its highest peak one day while being at its lowest for the dataset the other day, leaving no room for volatility clustering.

A result that is consistent for both assets is that the degrees of freedom are es-

timated to be significantly higher for the high-volatility regimes, suggesting less heavy tails in those regimes. This is to be expected when examining the return series of each regime, as the low-volatility regimes show some heavy outliers which you would not expect from a low-volatility environment. These outliers are then compensated by increasing the degrees of freedom such that they still fit the data properly.

Using the Student's t-distribution assumption proved to be problematic when trying to estimate the multivariate models. When using the data for the multivariate regimes, the solver would not converge for two of the regimes, no matter if the parameters were fixed or not. To combat this, one could try to estimate the DCC-GARCH using multiple starting parameters directly, like previously done for the univariate models. This would, however, be computationally heavy, as one would have to optimize at least 12 parameters, trying different starting values for every iteration. This is left for further research, as time would not allow us to test this.

5.3.3 Difficulty of Estimating the Parameters

When reviewing the estimated parameters, we notice that some of the GARCH parameters are not significant, namely ω and α. It is present both when estimating models for the regimes, seen by the BTC estimates, and for the benchmark model, seen by the BTC and S&P 500 estimates, but also for both distributions considered. We know that from the specification of the standard GARCH framework, these parameters should be significant and positive. The estimates themselves seem reasonable, as similar results are found by e.g. Ardia et al. (2019) and Alfaro and Inzunza (2023), so we speculate that the rejection of these parameters is due to the estimated confidence intervals being too wide.

The selected optimization routine, "nlminb", use a quasi-Newton method (Nash, 2014), meaning that the Hessian matrix is approximated. As both ω and α are relatively small, and should both be positive, we suspect that the log-likelihood function could be skewed or steep around the optima. We speculate that the approximation around this point could therefore lead to confidence intervals that cover beyond the zero point for said parameter, even if the log-likelihood in those points is low. We are unsure of how this problem could be handled, but we suspect that the tolerance of the optimizer could play a role. In **R**, the optimizer allows for the calibration of parameters regarding the convergence, which could be further investigated.

5.4 Final Model

Our final model takes a unique approach to capture dynamic correlations in financial data. It uses a fuzzy clustering algorithm, which allows data points to belong to multiple clusters with varying degrees of membership. The fuzzy clustering process gives probabilities to each observation's membership within the different clusters. These probabilities then become the weights in a standard weighted sum. Essentially, the final model combines the outputs of multiple models, each potentially representing distinct underlying market dynamics identified by the clusters. The models that are weighted together resemble each other, only differing by how much each observation should update the correlation.

An interesting observation regarding the fuzzy clustering can be seen from Figure 4. In this figure, we can quite clearly see four distinct clusters, but the centers which are marked with red crosses do not line up with them. The volatilities that this figure is based on comes from a 100 % weight of the regime that is estimated by the MS Model for each time point, and they clearly show that the volatilities could easily be divided into four clusters. We do have a hypothesis that this phenomenon is due to the random elements in the fuzzy clustering algorithm, as we get slightly different cluster centers every time the algorithm is run. This is also why we set a fixed random seed in **MATLAB** before running the algorithm.

It's worth noting that our final model shows some similarities to the benchmark DCC-GARCH model, especially in the overall shape of the correlation that is estimated. This suggests that the ensemble approach captures the broad trends that are present in the data. However, we believe that our model offers a key advantage, namely that it is quicker to adapt to quick changes in correlation than the DCC-GARCH model. This flexibility allows it to potentially adapt faster to evolving market conditions than the benchmark DCC-GARCH model. Additionally, the final model is shown to be more moderate compared to the 30-day and 60-day rolling correlation. This makes our model good at avoiding exaggerating volatility in the correlation estimates while still offering responsiveness to changing market dynamics.

5.5 Increase in Correlation

As seen by the final model, the benchmark model, and to some extent the rolling correlations, an increased correlation between BTC and the S&P 500 could be seen from 2020 to 2023. A large spike was seen when the COVID-19

pandemic started in March 2020, but the correlation level seems to have been elevated until well after the pandemic was no longer a global threat in May 2023 (World Health Organization, 2024). It is worth noting that this period in general showed a higher volatility, especially for the S&P 500, which can also be seen from the results. We speculate that this phenomenon of higher correlation can be a result of the pandemic stimulus packages that were distributed around the world. In the United States alone, almost $5 trillion was handed out to boost the economy (United States Government, 2024). By introducing this amount of money into the economy, both small and large investors could invest in many types of assets, including cryptocurrency. This could therefore lead to many different types of assets becoming correlated, as money could be invested in everything.

We further speculate that when this money eventually ran out, a phenomenon called flight-to-quality could be seen. This is when investors choose to rebalance their portfolios to contain traditionally "safer" investments (Beber et al., 2008), usually in times of uncertainty. This uncertain time could just as well be that the stimulus money had run out, but no conclusions are drawn in this thesis. Why investors did not choose to only invest in "safer" assets during the pandemic could potentially be explained by the above-mentioned stimulus packages, which relieved some of the economic uncertainty, or that cryptocurrencies were seen as an investment that was decoupled from the traditional market. Most of these explanations are speculative and are worth a further investigation.

6 Conclusion

Our study found an interesting pattern in the connection between BTC and the S&P 500. As mentioned in Section 1.2, the objective of this thesis was to examine if the change in correlation during COVID-19 was a temporary phenomenon or if it was a long-term trend. Just like other research, we saw a stronger correlation during COVID-19, which likely happened because investors wanted diverse investments when the market was uncertain. However, what is interesting is that this stronger correlation didn't disappear completely after COVID. Instead, it remained relatively high for a while, before it dipped in the beginning of 2023. At the end of 2023, the correlation became higher once again, but since then, our presented model suggests that the correlation is on its way down again.

If this weakening connection holds true, it could mean a change in how investors view BTC. When markets get shaky, like during COVID, investors tend to put their money in things they deem "safe". This can make even unrelated investments move together more. But the fact that the correlation seems to be weakening since the end of 2023 might mean investors are starting to see BTC differently again. This could make BTC a good option again for investors who want to diversify their portfolios and potentially protect themselves from risk.

Our new approach proved to be successful. By dividing the data into distinct regimes and estimating separate GARCH models for each, we achieved a more nuanced understanding of the dynamic correlation. The weighted combination of these GARCH models for each regime gave us a result that was both sensitive to changes in correlation and retained the core structure of the benchmark DCC-GARCH model. However, the S&P 500 low volatility regime consistently caused estimation problems for us. Even after trying different techniques, including outlier removal, exploration of AR and ARMA models, and consideration of alternative distributions, this particular regime was still difficult for us to get a clear picture of. This is an area for future research, where perhaps more advanced methods or additional factors could be explored.

6.1 Further Research

During the concluding face of this thesis, multiple ideas for future research have popped into existence. Some of them would try to solve the issues faced during the thesis work, while others would attempt to use the model as an investment tool. The ideas are not pursued, mainly due to time limitations, but also due to limited knowledge of optimization routines.

- Investigate if similar parameters are estimated when using a different dataset that does not contain a financial downturn like the COVID-19 pandemic. This would help confirm that the volatility of different regimes can be modeled with different GARCH parameters. It would be interesting to see if this would help the issue faced in the S&P 500 low volatility regime.

- One could investigate if different GARCH models are able to capture and model the volatility and correlation in a better way. This would include GARCH models that capture the asymmetry, such as the EGARCH model, which would be in line with the asymmetry specified in the stylized facts (Cont, 2001). Another interesting approach could be to use a stochastic volatility model instead.

- A full estimation of the parameters of the DCC-GARCH model, considering different starting values of all parameters simultaneously. With an increasing number of assets, the model parameters quickly grow. Instead of estimating univariate models, fixing these parameters and only estimating the parameters for the correlation updating, it would be possible to estimate the complete model at once, given enough time.

- To test the performance of the model, one could compare it with a simple buy-and-hold strategy. As the regime estimation is built upon a Markov Chain, it could easily be forecasted, and the same can be said with the GARCH framework. If we consider the simple Markowitz mean-variance model (Markowitz, 1952), one could estimate the optimal weights of each asset for every time period, given some transaction costs. The only struggle as we see it is the estimation of the fuzzy clustering probabilities of each state, but other methods might be considered for this.

References

Alfaro, R. and Inzunza, A. (2023). Modeling S&P500 returns with GARCH models. *Latin American Journal of Central Banking*, 4(3).

Ardia, D., Bluteau, K., and Rüede, M. (2019). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29:266–271.

Beber, A., Brandt, M. W., and Kavajecz, K. A. (2008). Flight-to-Quality or Flight-to-Liquidity? Evidence from the Euro-Area Bond Market. *The Review of Financial Studies*, 22(3):925–957.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1st edition.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A Capital Asset Pricing Model with Time-Varying Covariances. *Journal of Political Economy*, 96(1):116–131.

CoinMarketCap (2024). Bitcoin price today. `https://coinmarketcap.com/currencies/bitcoin/` [Accessed 2024-06-10].

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Cortese, F. P., Kolm, P. N., and Lindström, E. (2023). What drives cryptocurrency returns? A sparse statistical jump model approach. *Digital Finance*, 5:483–518.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3).

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. (2002). Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Galanos, A. (2022). *rmgarch: Multivariate GARCH Models*. R package version 1.3-9. `https://cran.r-project.org/web/packages/rmgarch/`.

Galanos, A. and Kley, T. (2023). *rugarch: Univariate GARCH Models*. R package version 1.5-1. https://cran.r-project.org/web/packages/rugarch/.

Gay, D. M. (1990). Usage Summary for Selected Optimization Routines. Technical Report 153, AT&T Bell Laboratories.

Goldberger, A. S. (1964). *Econometric Theory*. New York: John Wiley & Sons.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (2005). Regime-Switching Models. *Palgrave Dictionary of Economics*.

Hull, J. (2018). *Risk Management and Financial Institutions, Fifth Edition*. John Wiley & Sons.

IMF (2022). Crypto Prices Move More in Sync With Stocks, Posing New Risks. https://www.imf.org/en/Blogs/Articles/2022/01/11/crypto-prices-move-more-in-sync-with-stocks-posing-new-risks [Accessed 2024-04-27].

Investopedia (2024). What is EMA? How to Use Exponential Moving Average With Formula. https://www.investopedia.com/terms/e/ema.asp [Accessed 2024-04-22].

Jacobsson, A. (2021). *An Introduction to Time Series Modeling*. Studentlitteratur.

Keim, D. B. (1984). A Further Investigation of the Weekend Effect in Stock Returns. *The American Finance Association*.

Lindström, E., Madsen, H., and Nielsen, J. N. (2015). *Statistics for Finance*. Chapman and Hall/CRC.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 1(7):77–91.

Massicotte, P. and Eddelbuettel, D. (2022). *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.5.1. https://cran.r-project.org/web/packages/gtrendsR/.

MathWorks (2024). Maximum Likelihood Estimation for Conditional Variance Models. https://se.mathworks.com/help/econ/maximum-likelihood-estimation-for-conditional-variance-models.html [Accessed: 2024-04-21].

Nash, J. C. (2014). On Best Practice Optimization Methods in R. *Journal of Statistical Software*, 60(2).

Nguyen, K. Q. (2021). The correlation between the stock market and Bitcoin during COVID-19 and other uncertainty periods. *Finance Research Letters*, 46.

Perlin, M. (2014). MS Regress - The MATLAB Package for Markov Regime Switching Models. `http://dx.doi.org/10.2139/ssrn.1714016`.

Silvennoinen, A. and Teräsvirta, T. (2008). Multivariate GARCH Models. SSE/EFI Working Paper Series in Economics and Finance 669, Stockholm School of Economics.

Terraza, V., İpek, A. B., and Rounaghi, M. M. (2024). The nexus between the volatility of Bitcoin, gold, and American stock markets during the COVID-19 pandemic: evidence from VAR-DCC-EGARCH and ANN models. *Financial Innovation*, 10(22).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1).

United States Government (2024). Covid-19 spending. `https://www.usaspending.gov/disaster/covid-19` [Accessed 2024-05-09].

U.S. Energy Information Agency (2021). Crude oil prices briefly traded below $0 in spring 2020 but have since been mostly flat. `https://www.eia.gov/todayinenergy/detail.php?id=46336` [Accessed: 2024-04-08].

Welles Wilder Jr., J. (1978). *New Concepts in Technical Trading Systems*. Trend Research.

Wikipedia (2024). Bitcoin. `https://en.wikipedia.org/wiki/Bitcoin` [Accessed 2024-06-10].

World Health Organization (2024). Coronavirus disease (COVID-19) pandemic. `https://www.who.int/europe/emergencies/situations/covid-19` [Accessed 2024-05-09].

Yahoo Finance (2024). `https://finance.yahoo.com/` [Accessed: 2024-03-15].

Zeileis, A., Grothendieck, G., Ryan, J. A., Ulrich, J. M., and Andrews, F. (2023). *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. R package version 1.8-12. `https://cran.r-project.org/web/packages/zoo/`.

A Appendix

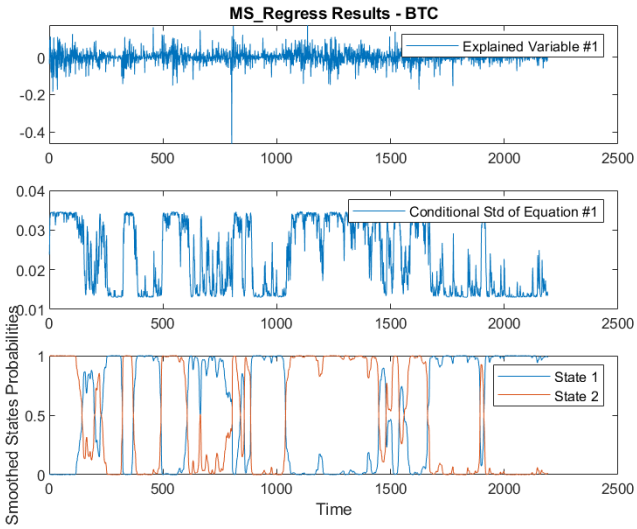


Figure A.1: In-depth analysis of the output from the MS_Regress package for BTC.

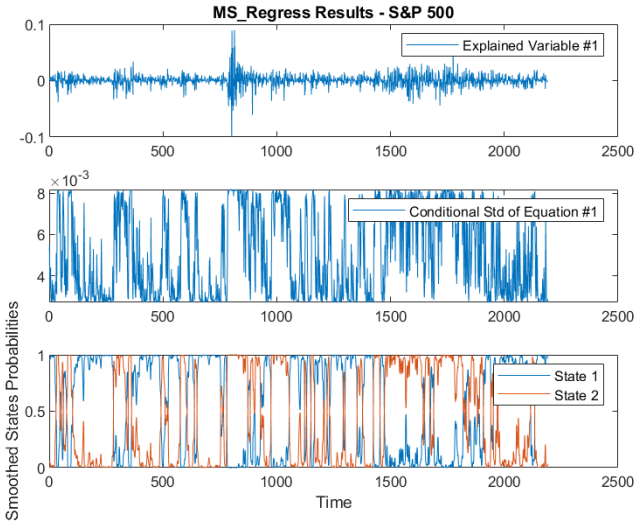


Figure A.2: In-depth analysis of the output from the MS_Regress package for the S&P 500.

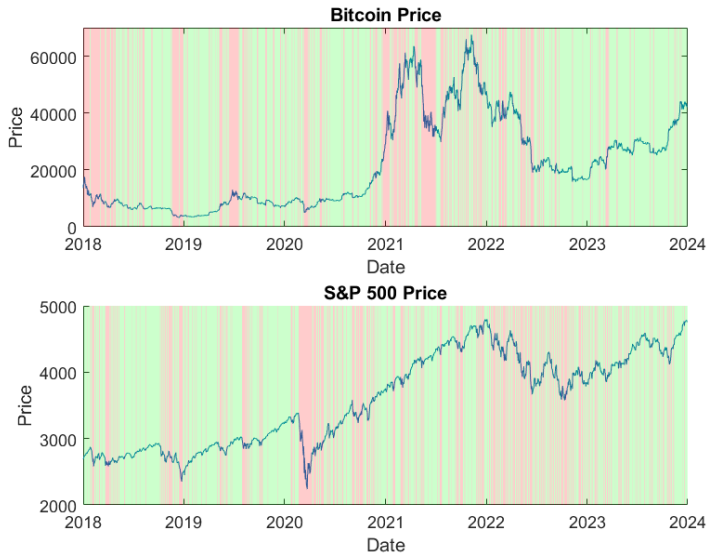


Figure A.3: The estimated regimes from the **MATLAB** package MS_Regress assuming the Normal distribution. Green indicates a low volatility period, while red indicates a high volatility period.

Table A.1: Optimal a and b parameter estimate in the High/Low regime without using starting values.

	Normal	
Parameter	Estimate	P-value
a	0.079174	0.092015
b	0.549524	**0.000942**
Log-likelihood	2928.9	

Table A.2: Optimal GARCH parameter estimates from the S&P 500 low volatility regime.

	Normal		Student-t	
Parameter	Estimate	P-value	Estimate	P-value
μ	0.000911	**0.000000**	0.000632	**0.000000**
ω	0.000018	**0.000000**	0.000062	**0.000201**
α	0.151524	**0.013775**	1.000000	0.255048
β	0.000000	1.000000	0.000000	1.000000
ν	-	-	2.166504	**0.000000**
Log-likelihood	4975.854		5142.754	

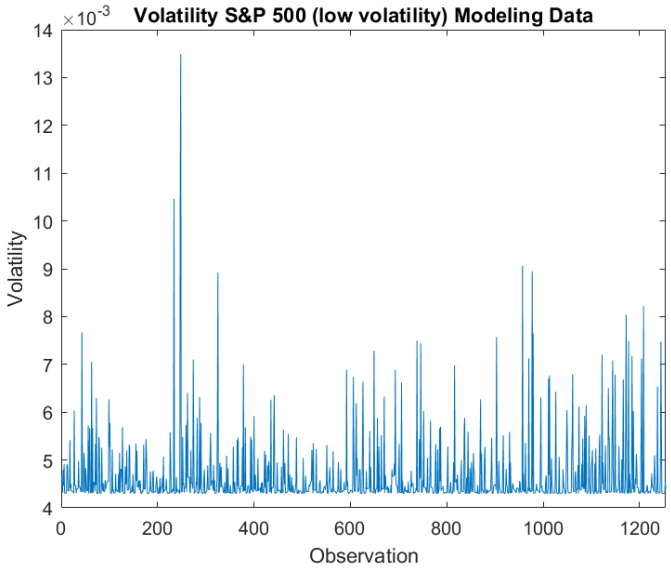


Figure A.4: Estimated volatility with the parameters that yield the highest log-likelihood for the S&P 500 low volatility regime. The volatility is only on the modeling data.

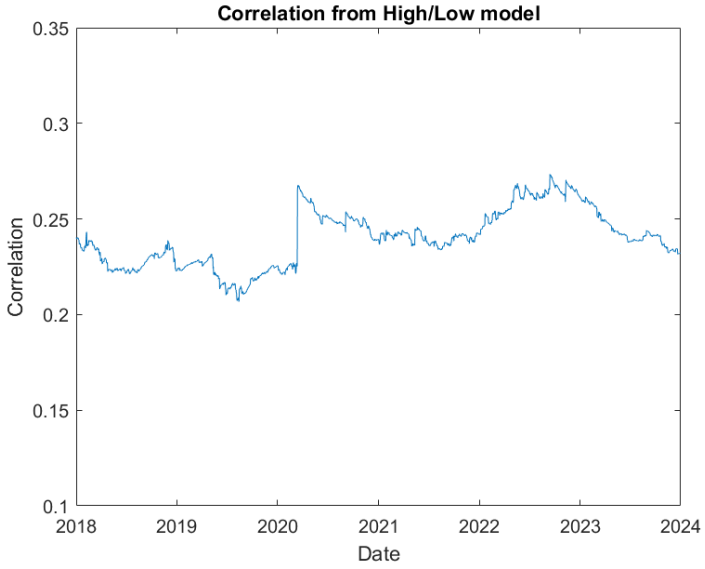
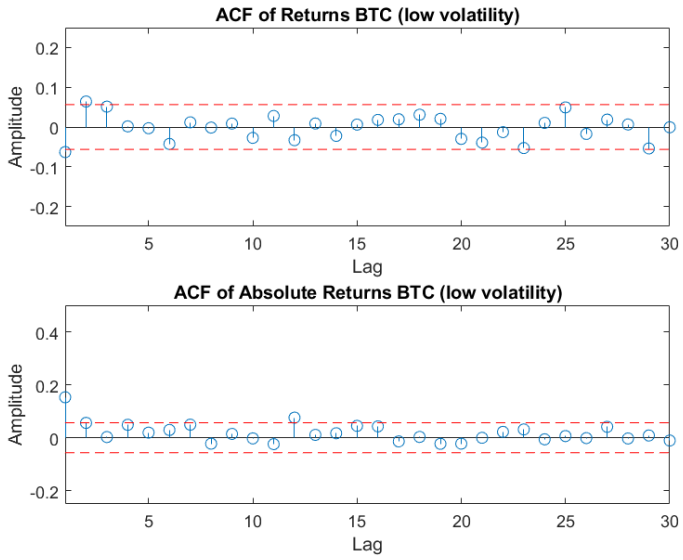
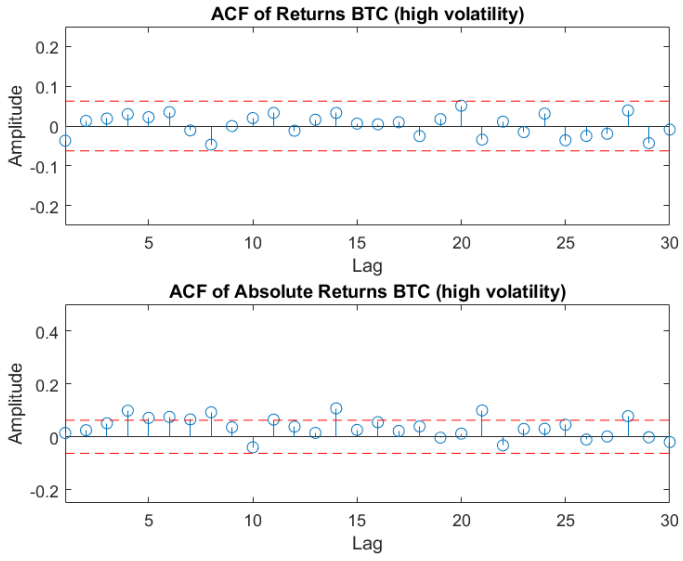


Figure A.5: Zoomed-in plot of the correlation in the High/Low regime. We can observe a similar pattern to the other regimes.

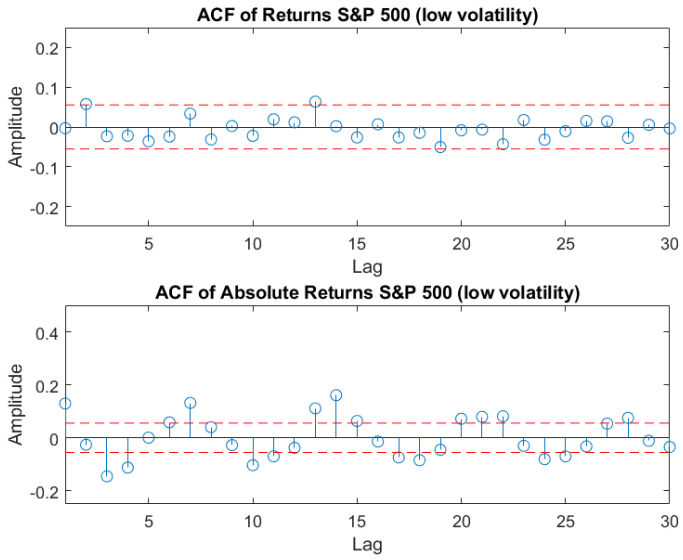


(a) ACF of BTC in the low volatility regime.

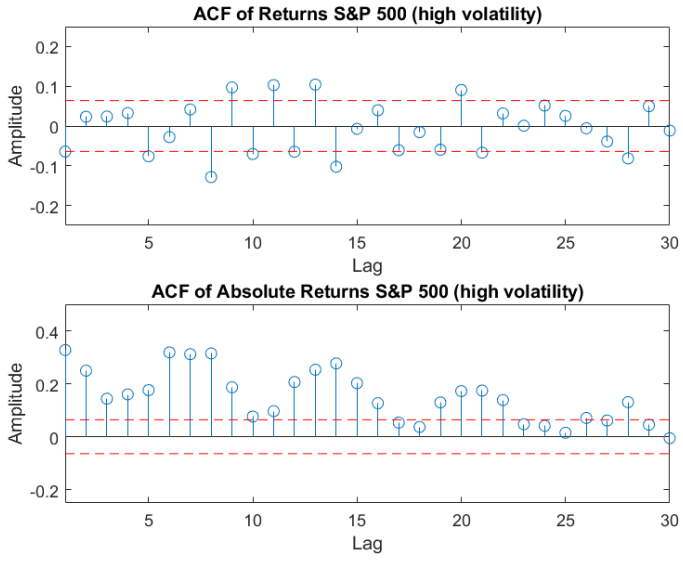


(b) ACF of BTC in the high volatility regime.

Figure A.6: ACF of the returns and squared returns of BTC in the respective regimes.



(a) ACF of the S&P 500 in the low volatility regime.



(b) ACF of S&P 500 in the high volatility regime.

Figure A.7: ACF of the returns and squared returns of the S&P 500 in the respective regimes.

Master's Theses in Mathematical Sciences 2024:E37
ISSN 1404-6342

LUTFMS-3495-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/