Statistical Analysis of Financial Markets

Final Project

23/10 MScFE Capstone Project: G4532

Harsha Chaube	Vivek Verma	Chen Liang
United Kingdom	Canada	United States of America
harshachaube@gmail.com	vivektheintel@gmail.com	landseer92@gmail.com

Abstract

The analysis involves studying time series datasets of 3 major cryptocurrencies - Bitcoin (BTC), Ethereum (ETH), and Binance (BNB), denominated in US Dollars at the Crypto Exchanges, between November 09, 2017, and November 13, 2023. To identify various regions of the market, multiple models such as Clustering, GMM, HMM, and MSM are utilized, which apply a wide range of logic including combination of univariate autoregressive and dependent mixture mode, to classify the behavior of the cryptocurrencies within the 3 hidden state regime - Bull, Bear, Stagnant.

Introduction

Cryptocurrencies, such as Bitcoin, are digital or virtual currency secured by cryptography that first appeared in 2009. Most cryptocurrencies exist on decentralized networks using blockchain technology—a distributed ledger enforced by a disparate network of computers. As of June 2023, there were more than 25,000 other cryptocurrencies in the marketplace, of which more than 40 had a market capitalization exceeding $1 billion [1].

Cryptocurrencies, in contrast to conventional fiat currencies, are not issued or controlled by a single entity. The absence of government regulation has contributed to high levels of volatility in the market value and exchange rates of major cryptocurrencies relative to fiat currencies, including the US dollar and the euro. For instance, a cryptocurrency's price may change by more than 10% in a single day. Opportunities to research and simulate extreme events in financial time series data are presented by this extraordinary price volatility. While forecasting extreme events is a crucial area of study in engineering, climatology, economics, and other fields, comparatively little research has been done on tail risk and regime shift modeling in cryptocurrencies.

High volatility is inherent in the crypto market, which makes it difficult for traditional financial engineering strategies to effectively manage and adapt to the crypto market. Regime shift modeling can aid in development of dynamic volatility management strategies catering to specific market conditions. The latter can also help deal with tail risk management and extreme event mitigation strategies, which can be reverse fed into the existing FE strategies, with some changes. The last pain point from a crypto standpoint would be the limited availability of historical data values, but at the same time, the amount of extreme volatility fluctuations in shorter time frames can help better in understanding regime shifts.

Real life applications of our project idea can be varied, such as, development of adaptive trading strategies, which can have the potential to automatically adjust to different market regimes, leading to efficient buy/sell decisions. Crypto Fund Management can be another outcome, given the fact that not sufficient research is present. Regulators can be a beneficiary as well, as they can use the research to gain insights into the crypto markets dynamics and develop policies that are responsive to the market conditions.

Literature

As the end of 2023 approaches, there has been a lot of attention lately focused on the theory that the world economy is about to enter a recession. When central banks attempt to tighten monetary policy, it's critical to consider factors other than standard macroeconomic downturns. The goal of this study is to examine prior cryptocurrency market data to spot trends and changes in the regime during the previous six years. With this empirical study the aim is to comprehend the underlying trends, movements in the market structure, and volatility dynamics in cryptocurrency markets under various regulatory frameworks. We are focusing on using econometrics and statistical modeling techniques to examine historical variations and contrast the state of the market today with previous patterns. Finding signatures that might point to approaching dramatic price events or changes in the bitcoin markets' regime is the aim. Better risk management and trading tactics may be made possible by the development of forecasting capabilities about market cycles and cryptocurrency volatility.

This survey of the literature examines earlier scholarly works on spotting warning signs. A range of analytical approaches have been employed for cryptocurrency market regime detection and modeling. Common methods include Hidden Markov Models (HMM), various forms of Markov Switching models, Autoregressive Integrated Moving Average (ARIMA) frameworks.

The emergence of virtual currencies known as cryptocurrency during the global financial crisis, are still relatively recent phenomena meaning it is not surprising that the study on cryptocurrencies is still in its mid-early phases of development. Cryptocurrencies are unique assets in terms of how they are used and operated. With the increase in different cryptocurrencies trying to replicate the popularity of bitcoin as well as the lack of regulations around it still makes it a distinct asset in some way. Researchers like Chan et al. (2017) [2], Szczygielski et al. (2020) [3] have examined the behavioral pattern around pricing of a single cryptocurrency, exploring the distributional properties, and understanding the impact of the assumption of a Gaussian distribution for the market.

It is also significant to mention the work that focuses on the characteristics that cause periodic booms. The efficiency side is especially interesting because this undertaking is still relatively new. Nadarajah and Chu (2017) [4], Tiwari et al. (2018) [5] are a few noteworthy studies in this area exploring the inefficiency during weak form.

Theoretical Framework

There are three commonly used methods for regime detection for financial time series data. Murtagh and Legendre (2014) proposed Ward's hierarchical agglomerative clustering method, aiming to produce petitions that minimize the within-group dispersion. [6] When applying towards time series data, the agglomerative clustering labels the latent regimes by clustering similar levels of price changes (merging into a cluster). Therefore, similar trends of price changes form respective clusters, and the stable periods form their own clusters. The algorithm starts with treating each data point as a single cluster, then sequentially measures similarities between pairs of clusters and merges the most similar pair of clusters, until all the clusters are merged into a single cluster. The wald’s methods are applied assessing cluster similarities. A dendrogram is formed during the merging process, and by cutting the dendrogram at selected dissimilarity level, clusters are formed at the cutoff – each cluster will represent one market regime. Song and Chang (2019) applied agglomerative clustering to understand the structure of the cryptocurrency market. [7] As an extension, Najafgholizadeh et al. applied a clustering method towards cryptocurrency regime detection [8].

A second commonly used algorithm is the Gaussian Mixture Model (GMM), which is a popular probabilistic model based on the assumption that data points are generated from a mixture of Gaussian distributions with unknown parameters. Assuming that the financial series data belongs to three regimes, the GMM algorithms estimates the mean and covariance of each gaussian distribution, as well as the proportion of each.

$$P(x) = \sum_{i=1}^{k} w_i N(x; \mu_i, \Sigma_i)$$

$P(x)$ – probability density function
k- number of components in the mixture
w_i- weight of the i-th component
$N(x; \mu_i, \Sigma_i)$- Gaussian distribution of mean μ_i and covariance matrix Σ_i

The parameters are typically estimated using the Expectation-Maximization (EM) algorithm. After estimating the parameters of the gaussian distributions, each data point is classified as belonging to one of the gaussian distributions, and adjacent data points belonging to the same distribution forms a period of market regime.

The third one is the hidden Markov model (HMM), another probabilistic model that inferences the likelihood for a sequence of random states. HMM assumes that conditional on the most immediate past state, the current state is independent from all previous states. Given the sequence of observations and the number of underlying states, HMM estimates the initial probabilities, the transition probability matrix, and the emission probabilities. The transition matrix contains the probability of moving from one state to another, and the emission probability contains the probability of observing a data point given the underlying state.

Initial state probabilities: $\Sigma_{i=1}^{N} \pi_i = 1$

Transition probabilities: $\Sigma_{j=1}^{N} a_{ij} = 1$ for i = 1, …, N

Emission probabilities: $\Sigma_{k=1}^{M} b_j(v_k) = 1$ for j = 1, …, N

, where N = no. of states, M = no. of possible obs. and $V = \{v_1, v_2 \ldots, v_M\}$ i.e. possible observations.

In HMM,

$$P(O, Q \mid \lambda) = \pi_{q1} b_{q1}(o_1) a_{q1q2}\, b_{q2}(o_2) \ldots.\ a_{qT-1qT}\ b_{qT}\ (o_T\ );$$

is probability of sequence observations O = $\{o_1,\ o_2,\ldots o_T\}$ as well as the corresponding sequence states Q = $\{q_1,\ q_2,\ldots q_T\}$. λ represents all 3 probabilities (initial, transition and emission) probability matrix.

The last one is the Markov Switching Model (MSM), which is a type of the Hidden Markov Model, where some of the phenomenon is directly observed via time series regression whilst the rest of it is hidden and is observed via the Markov Model. Estimates involve state determination, estimation of associated params and the probabilities of transition.

$$P(Y_1{:}T, S_1{:}T \mid \theta) = \pi S_1 f S_1(Y_1) \cdot \prod_{t=2}^{T} [\pi_{St-1,St} f_{St}(Y_t)]$$

, where θ represents set of probability parameters $\{\pi, f, \beta\}$, π_{S1} for the initial state, $\pi_{St-1,St}$ for the transition, $f_{St}(Y_t)$ for the emission density and β the regression parameter.

Methodology

For the statistical analysis, we first descriptively summarized the characteristic of the cryptocurrency data for each regime period using standard summary statistics such as mean and standard deviation, test whether there are trends or cycles within each period, as well as run the time series data against ARIMA models to characterize the time series data. We also summarized the segment characteristics such as mean and variance of segment length. Comparing across regimes, we looked for change in the mean and the trend across regime periods, and attempted to understand what, if any events or regulations, contributed to the switch.

Using skewness and kurtosis to understand the data distribution - mainly the asymmetry i.e. skewness and sharpness (kurtosis). This is done to understand the shape of the distribution of daily returns for each variable.

Next, we applied the four regime detection methods, the agglomerative clustering, the GMM, the HMM and the MSM, on the selected time series data of cryptocurrencies. For all methods, we used three underlying regimes of null market, bear market and stagnant. For the data-preprocessing, we started with calculating the moving averages (MA) for each time series, followed by calculating the change of moving averages on the logarithmic-scale. Because the number of regimes/states are fixed, the major changing parameter is the period used to calculate the moving average (n-day average), before supplying to the model for regime identification. For exploratory analysis, we first calculated the minimum, median, max, and average length of the regime periods, against the number of days used to calculate the moving average, from 7 to 30 days, the letter being the convention for calculating MA for cryptocurrency. However, because the length of the periods does not determine the fit of the regimes, we visualized the fitted regimes against cryptocurrency price to find the best fit.

Codebase

https://github.com/crackCodeLogn/WQU-C10-CP

Results

We performed initial analysis on three selected cryptocurrencies - Bitcoin (BTC), Binance Coin (BNB) and Ethereum (ETH) for a time period - Nov 2017 to Nov 2023.

Statistical characteristics

2194 data points for each cryptocurrency (BTC-USD, ETH-USD, BNB-USD).

Volatility - We performed statistical methods on the above three variables including summary and annualized volatility and found Binance coin to have the highest volatility.

BNB-USD 89.937699
BTC-USD 60.174073
ETH-USD 75.768247

Correlation -

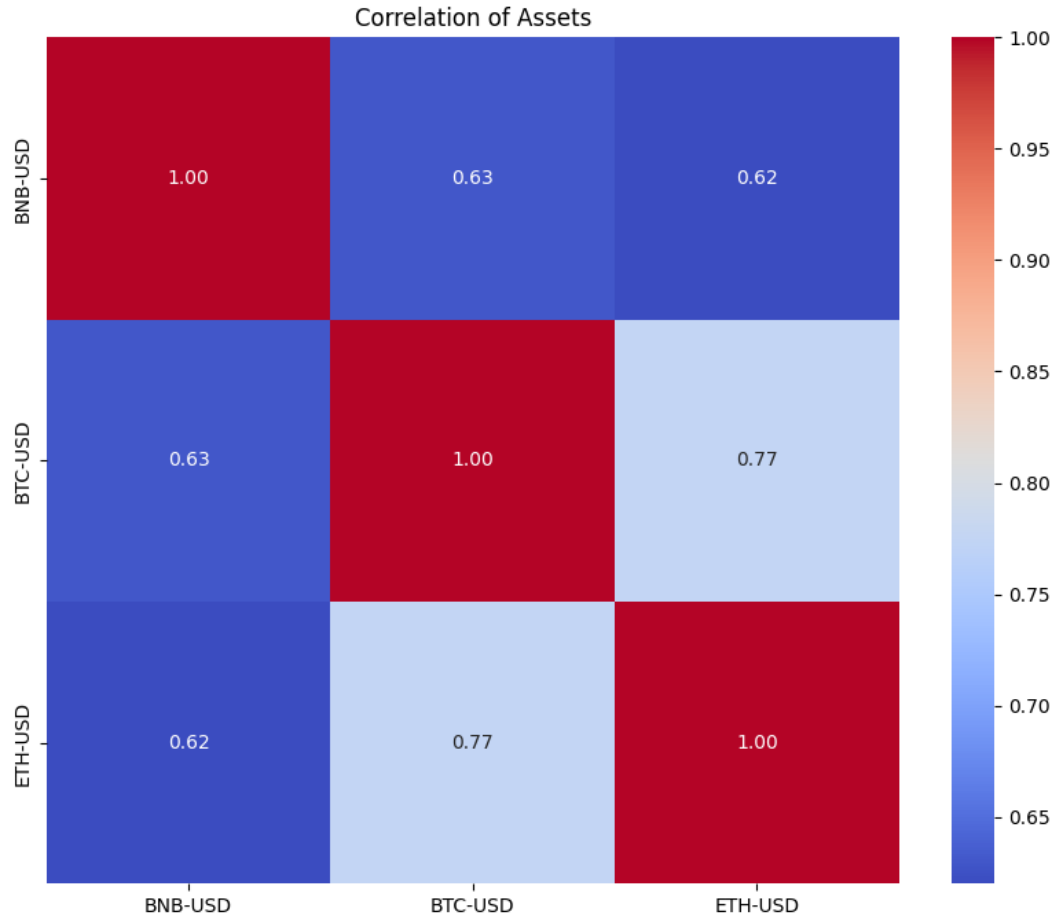

Fig. 1

The initial exploratory analysis suggests a positive correlation among all three crypto variables. This means that an increase in one value, also leads to an increase in others. The correlation between BTC and ETH is comparatively stronger to that of BNB.

Chow Test - We applied Chow test on BTC-USD to assess structural break in the data. We found structural break around early 2018 for BTC.
Skewness & Kurtosis -
Skewness:
[2.02174048 -0.12613035 -0.21650773]

- BNB-USD Skewness is positive (2.02) suggesting positively skewed distribution of daily returns i.e. tendency for larger positive returns than would be expected in a normal distribution.

- BTC-USD Skewness is close to zero (-0.13). Distribution of daily returns is approximately symmetric. No strong skewness towards positive or negative returns.

- ETH-USD Skewness is slightly negative (-0.22) suggest a mild negative skewness, that is a slight tendency for more negative returns than would be expected in a normal distribution.

Kurtosis:

[25.06967797 7.79027473 5.99607921]

- BNB-USD is high (25.07). A high kurtosis value indicates heavy tails and a more peaked distribution compared to a normal distribution. Note - This could imply extreme events (large positive or negative returns) are more likely than in a normal distribution.
- BTC-USD moderate (7.79). Higher than the kurtosis of a normal distribution (which is 3), it's not as extreme as BNB. That means BTC returns have relatively fatter tails and a more peaked distribution than normal.
- ETH-USD moderate (6.00). Returns have a distribution with fatter tails and a more peaked shape compared to a normal distribution.

BNB-USD exhibits the highest skewness and kurtosis, indicating a distribution with a higher likelihood of extreme returns. BTC-USD and ETH-USD show more symmetric distributions with milder deviations from normality

Market Regime -

We used four different methods on the detection of market regimes, namely clustering, HMM, GMM and MSM. For the first three methods, we started with an exploratory analysis on the moving average vs. the segment length. As we can see from figure 2, regardless of the number of days used for calculating the moving average, there are always regime periods of length 1, with more than half of the periods lasting less than five days, that do not constitute real regimes. We also observed more fluctuation for regime detection for ETH, where the resulting regime is very sensitive to the moving average calculation. From the regime period length, it is hard to determine the best moving average to use, and the best moving average seems different for different cryptocurrency. Therefore, we visually exclaimed the best moving average for each method and for each cryptocurrency.

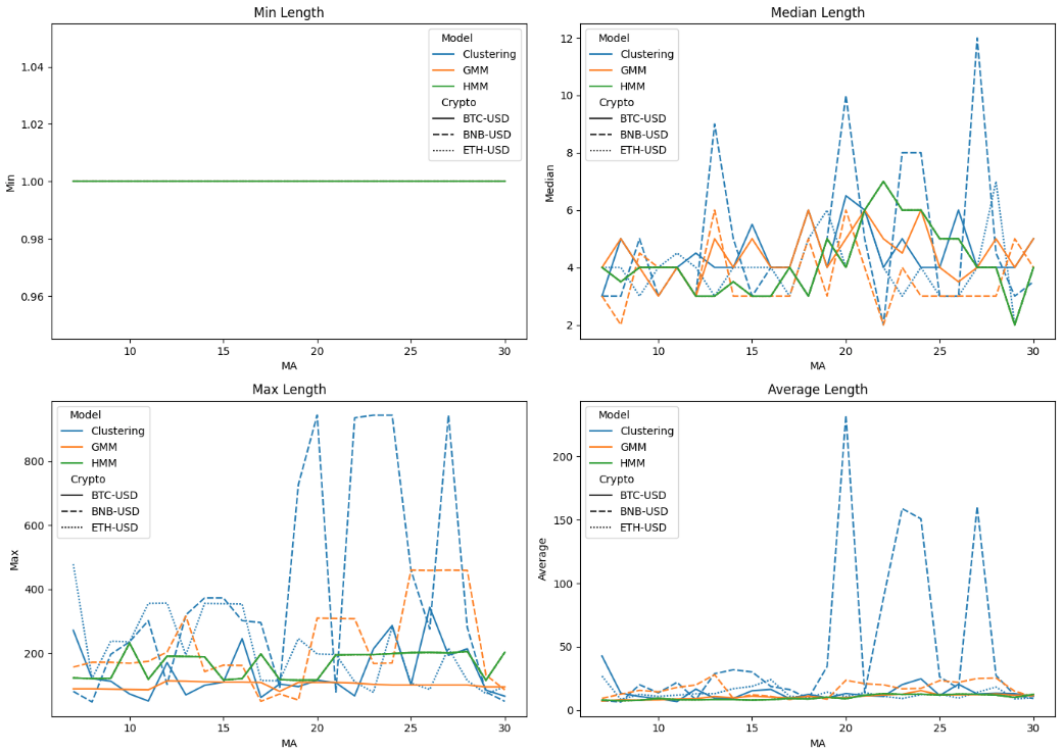

Fig. 2

Clustering

Upon visualizing the data, the optimal moving average for clustering all cryptocurrencies appears to be a 30-day moving average. However, in the context of agglomerative clustering, it's important to note that each cluster (hidden state) doesn't necessarily align with a distinct market regime. Instead, the algorithm identifies segments with similar changing patterns. For BTC, the regime detection seems reasonable with the connected red dots typically represent a bullish market, while blue dots denote a bearish market, and scattered green dots indicate stagnant periods. On the other hand, BNB exhibits a broader range of variation, leading to almost all periods being categorized as stagnant. In the case of Ethereum (ETH), red dots signify a bullish market, blue dots represent a bearish market, and green dots indicate stagnation. Although there is some mixing of bearish and stagnant periods, it remains possible to distinguish between the two.

```
Number of observations for State  0 : 1330
Number of observations for State  1 : 374
Number of observations for State  2 : 122
```

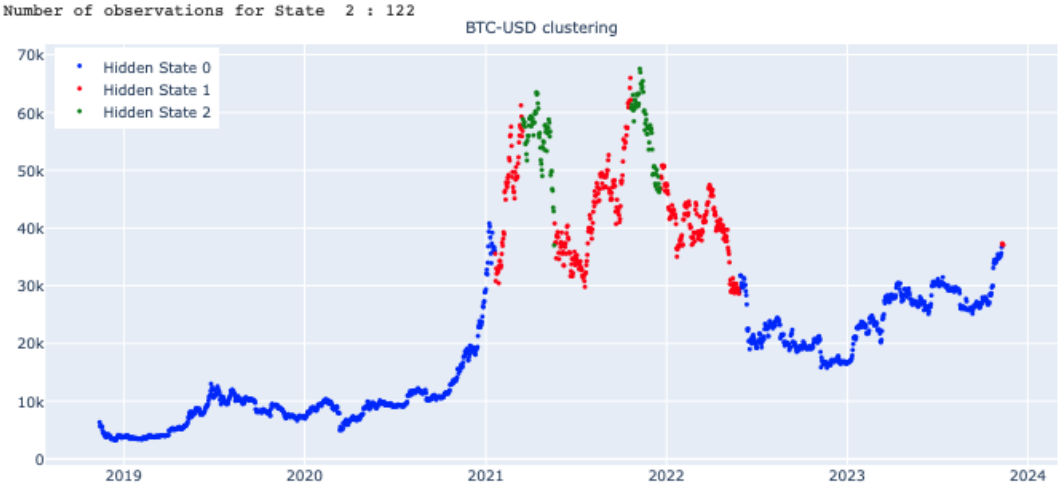

Fig. 3.1

```
Number of observations for State  0 : 1789
Number of observations for State  1 : 5
Number of observations for State  2 : 32
```

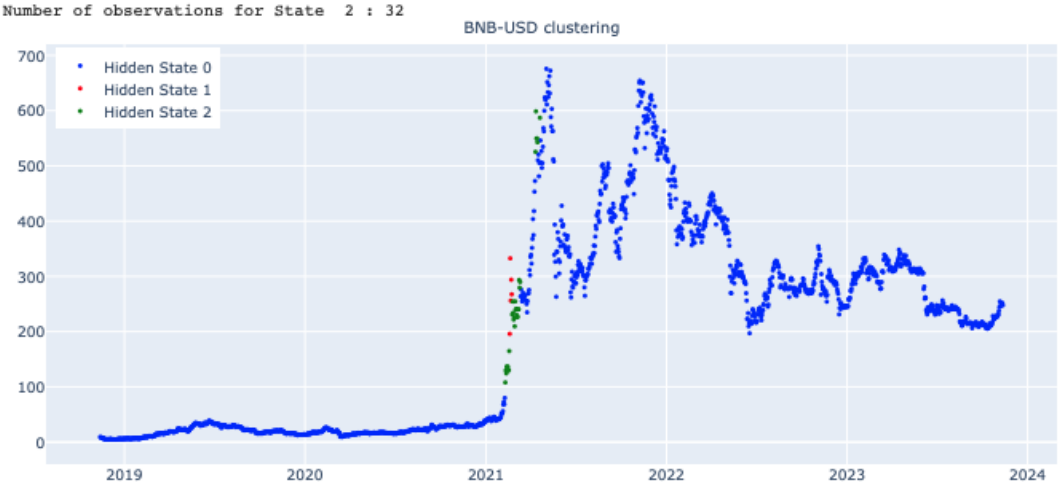

Fig. 3.2

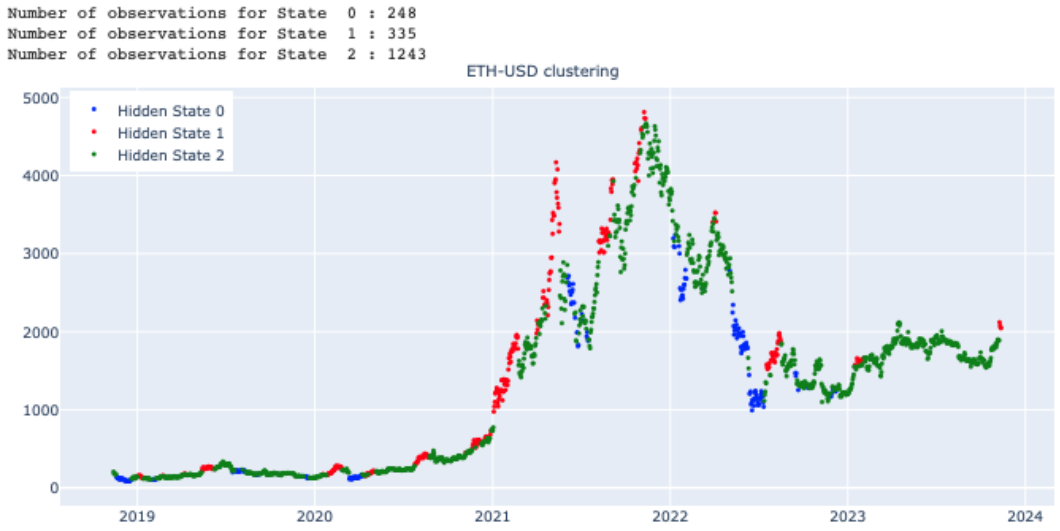

Fig. 3.3

GMM -

On the contrary, the GMM yields distinct results. In the case of BTC, GMM identifies prolonged periods of changes. The green cluster corresponds to the period before 2021, characterized by smaller fluctuations. The blue cluster spans from 2021 to mid-2012, representing the highest fluctuations, while the red clusters correspond to median variations, mainly observed from mid-2022 onward. Interestingly, GMM performs well for BNB, where the red clusters align with bearish patterns and the blue one with the bullish market. Notably, there are fewer days categorized as green, suggesting that the algorithm did not identify a stagnant period. Additionally, the green clusters represent a bullish market in this context. For ETH, GMM also shows promising results. The blue cluster is indicative of a more obvious bull market, the red cluster reflects a more obvious bear market, and the green cluster represents stagnation, potentially with significant fluctuations.

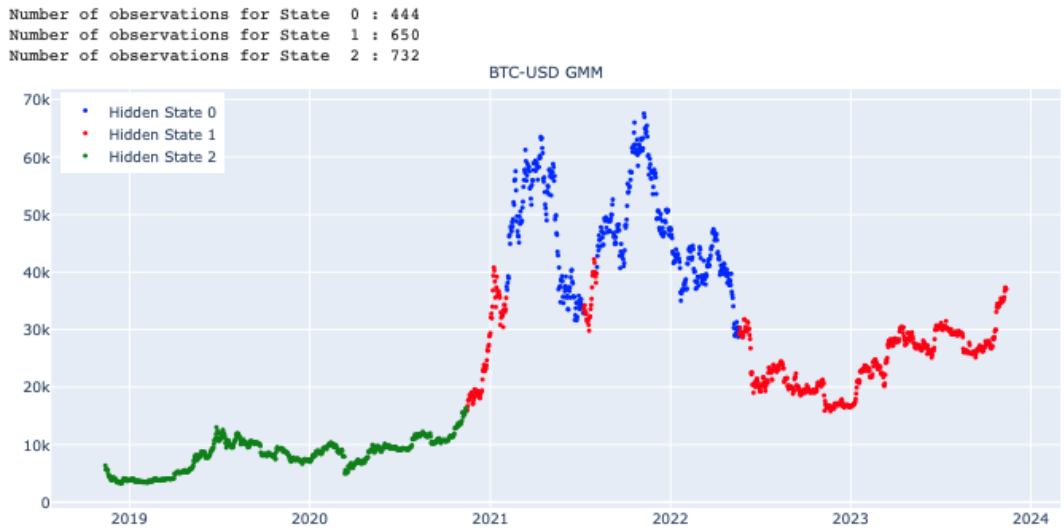

Fig. 4.1

```
Number of observations for State  0 : 927
Number of observations for State  1 : 860
Number of observations for State  2 : 39
```

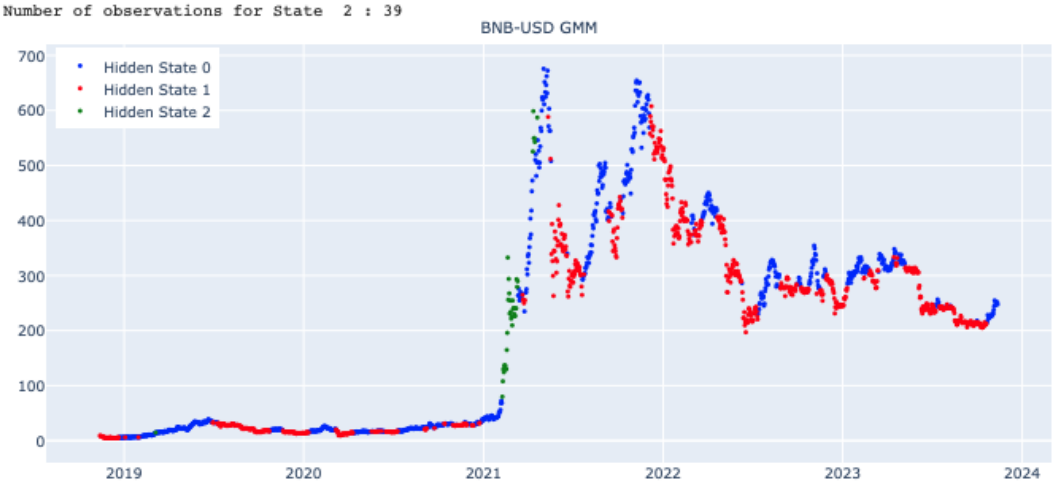

Fig. 4.2

```
Number of observations for State  0 : 489
Number of observations for State  1 : 190
Number of observations for State  2 : 1147
```

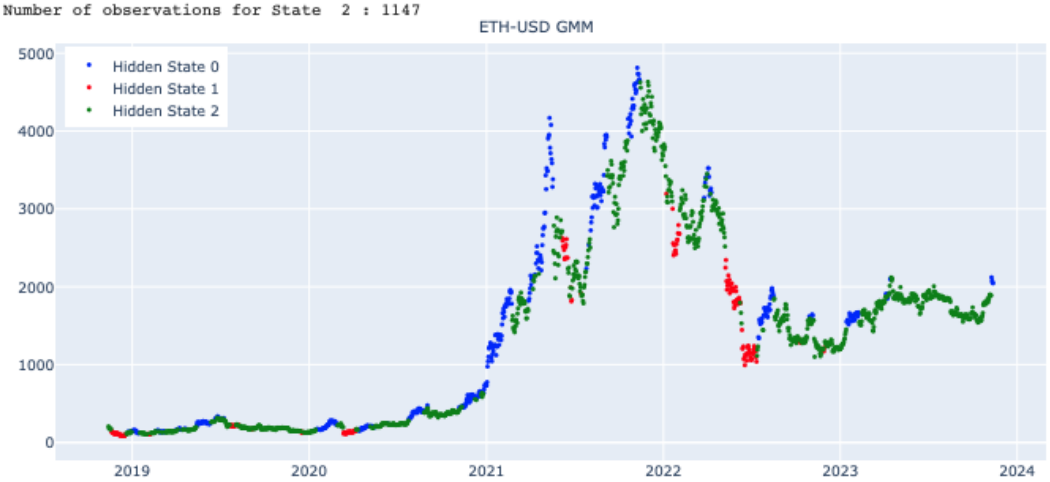

Fig. 4.3

HMM -

In contrast, HMM results exhibit similarities to GMM but with some distinctions. For BTC, the red/blue/green dots represent extended periods characterized by varying fluctuation levels. However, the HMM's depiction is less precise compared to GMM.

In the cases of BNB and ETH, the blue clusters signify periods of larger fluctuations. Interestingly, the green and red clusters are more blended together, potentially indicating smaller ups and downs in the market. The distinctions between stagnant and bearish periods are less clearly defined in the HMM results for these cryptocurrencies when

compared to GMM.

```
Number of observations for State  0 : 735
Number of observations for State  1 : 492
Number of observations for State  2 : 599
```

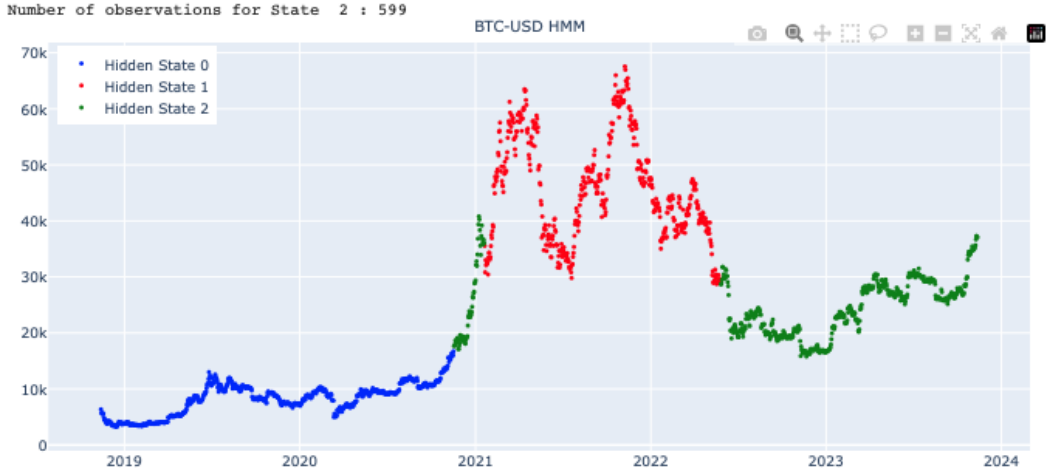

Fig. 5.1

```
Number of observations for State  0 : 269
Number of observations for State  1 : 774
Number of observations for State  2 : 783
```

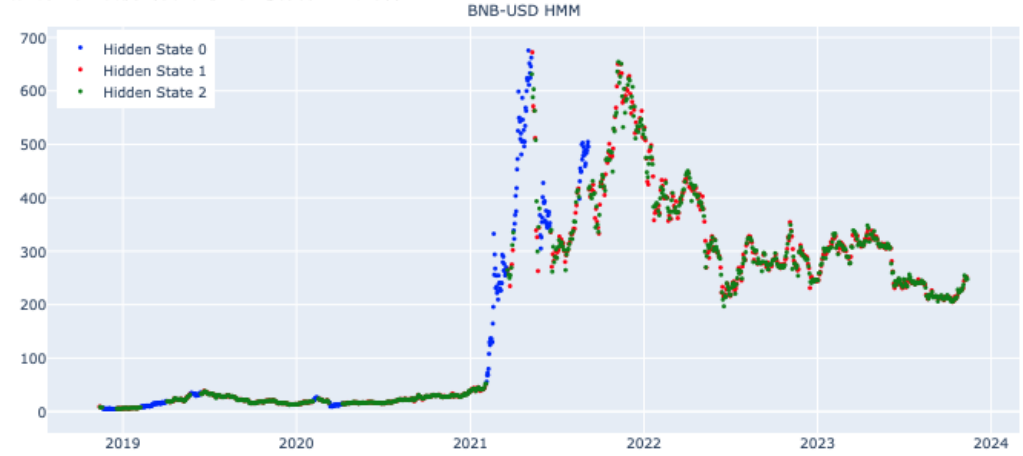

Fig. 5.2

```
Number of observations for State  0 : 317
Number of observations for State  1 : 750
Number of observations for State  2 : 759
```

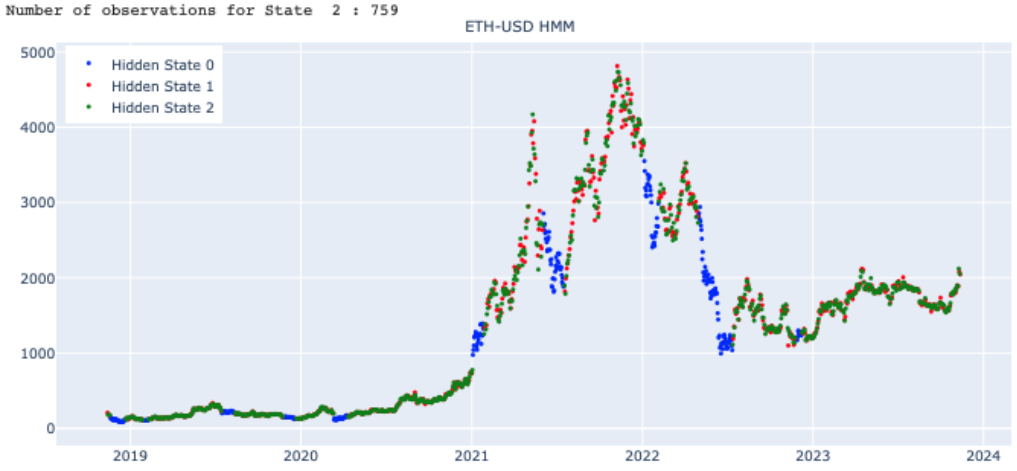

Fig. 5.3

MSM -

The predictions from the MSM model are mixed in nature. For BTC, it seems to have been the most close to actual nature, with few miscategorisations. But overall, the model captured the major details of the 3 market regimes. On the other hand, ETH and BNB data representation were slightly more off the mark. The green clusters signal large areas of volatilities for ETH, and a combination of green and blue for the BNB. Red clusters being almost evenly present int ETH, was largely missing in the BNB representation.

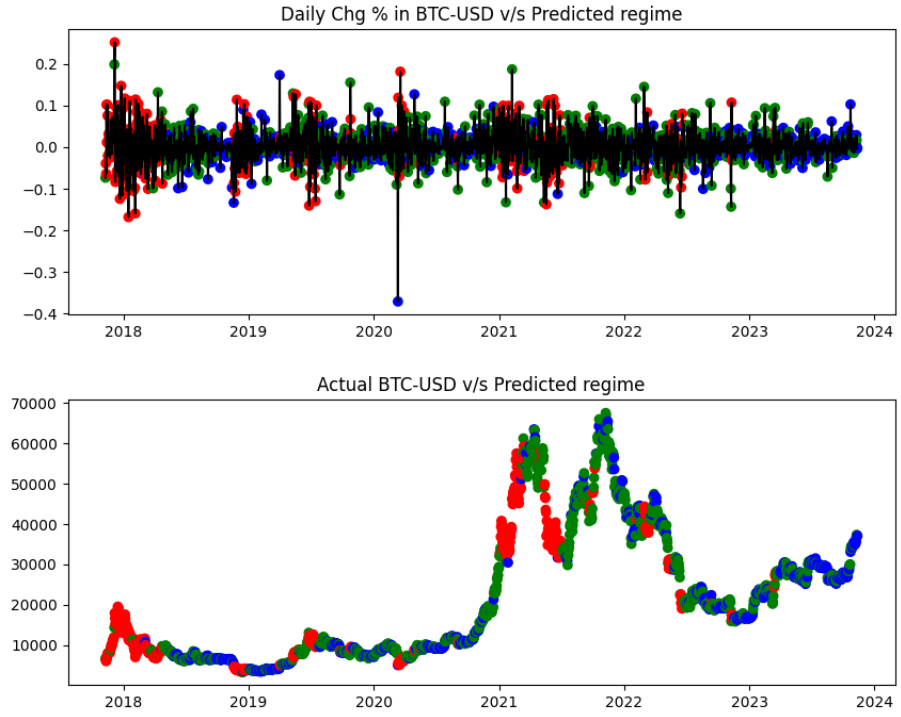

Fig. 6.1

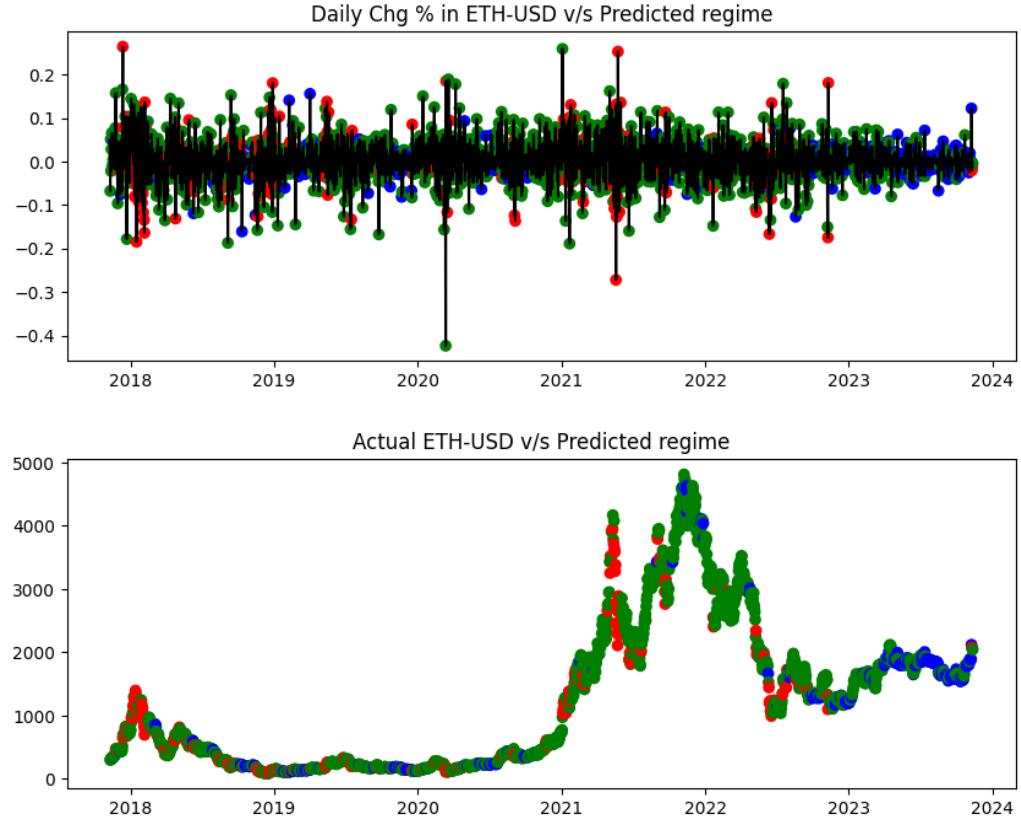

Fig.6.2

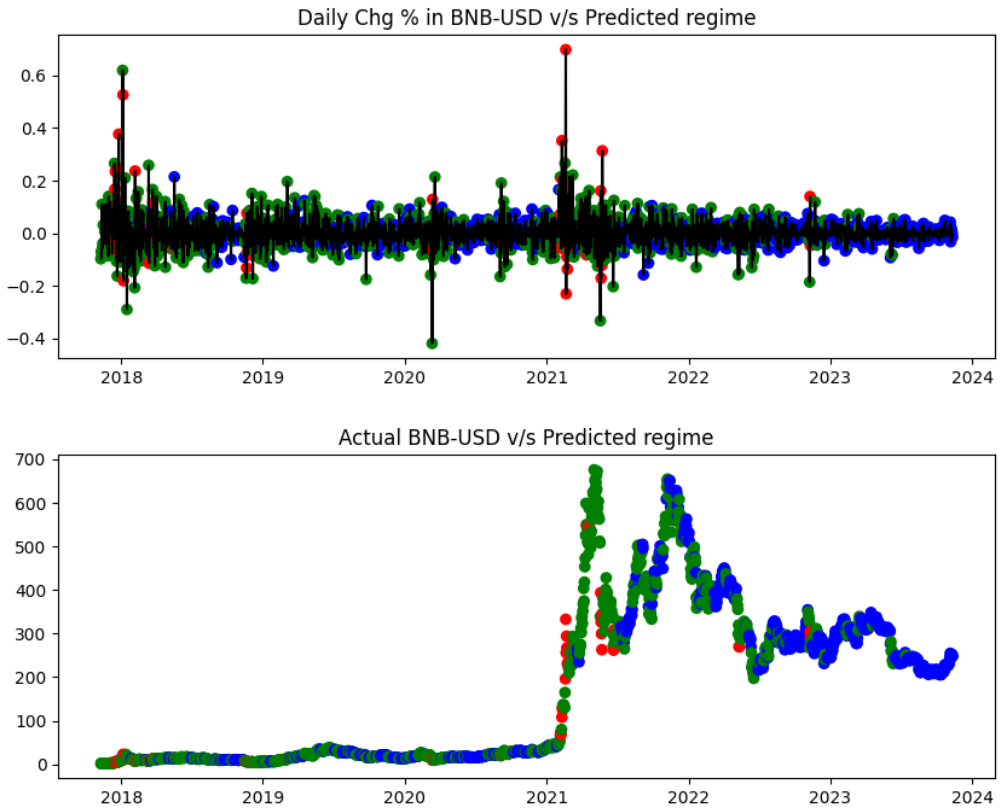

Fig.6.3

Additional Model modifications

In our quest for model enhancements, we delved into three distinct variations. Initially, recognizing the Hidden Markov Model's (HMM) capability to capture a broad spectrum of variations for prolonged periods in Bitcoin (BTC), we implemented a multivariate HMM. This variant considers correlations among the three cryptocurrencies. The resultant model divided the three hidden states into categories representing large changes (green), small upward movements (red), and small downward movements (blue). Although these hidden states align with logical patterns, they don't precisely correspond to our desired classifications of bull, bear, and stagnant markets.

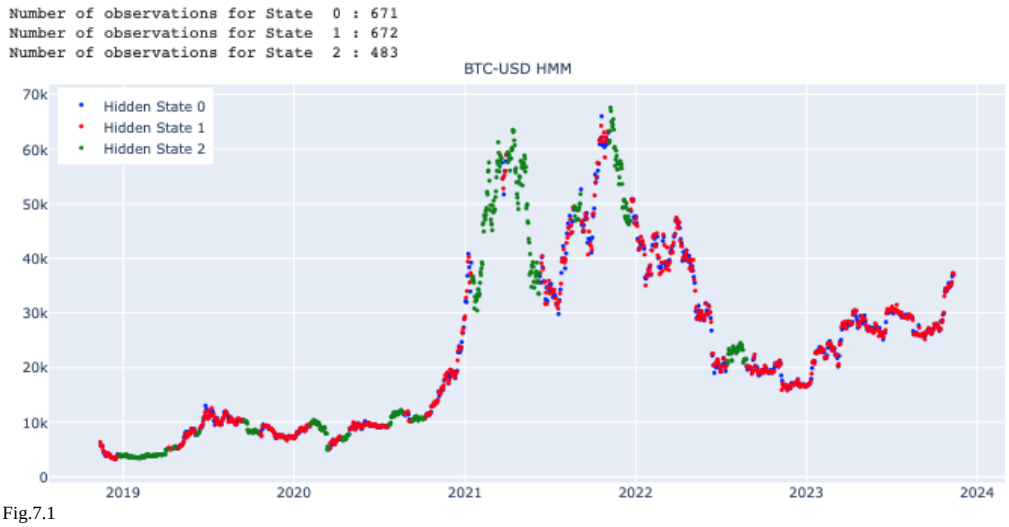

Fig.7.1

```
Number of observations for State  0 : 671
Number of observations for State  1 : 672
Number of observations for State  2 : 483
```

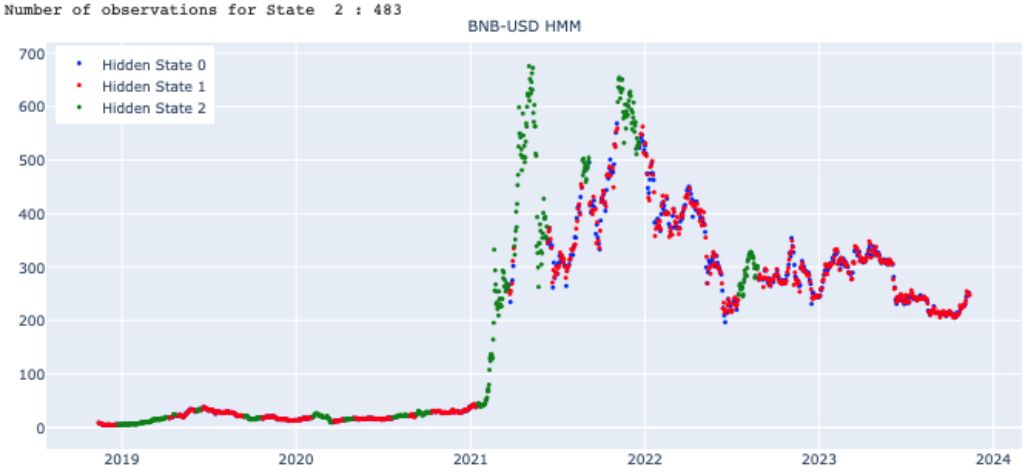

Fig.7.2

```
Number of observations for State  0 : 671
Number of observations for State  1 : 672
Number of observations for State  2 : 483
```

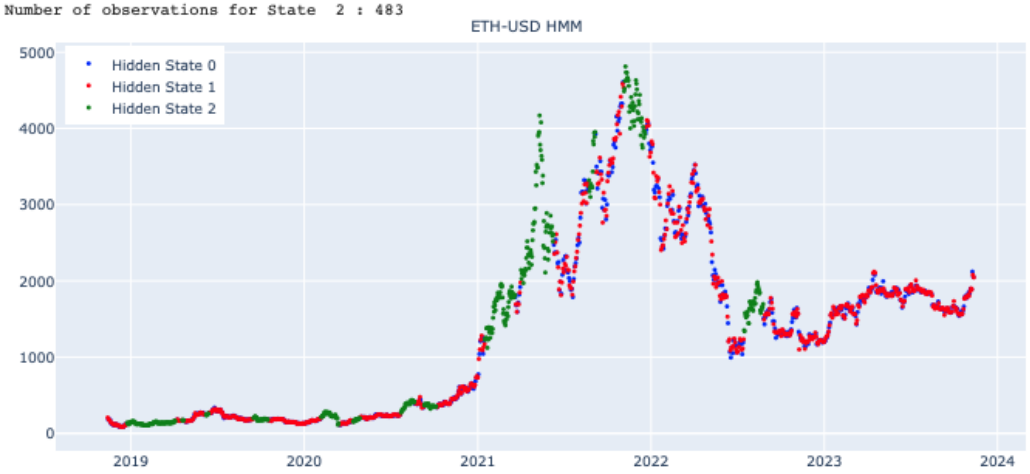

Fig.7.3

GMM also has the capacity to incorporate multivariate time-series data. However, the resulting regimes are less meaningful and therefore not included in this document. The results are in the last segment of the coding files.

Lastly, we divided the whole time series into pre - Oct 2020, Nov 2020 – April 2022 and post April 2022. The GMM worked well for detecting the bear and bull market for the period between Nov 2020 – April 2022, but not so well for the other two with more complex patterns of changes. Figure 8 is a demonstration of regime identification for BTC between Nov 2020 and April 2022

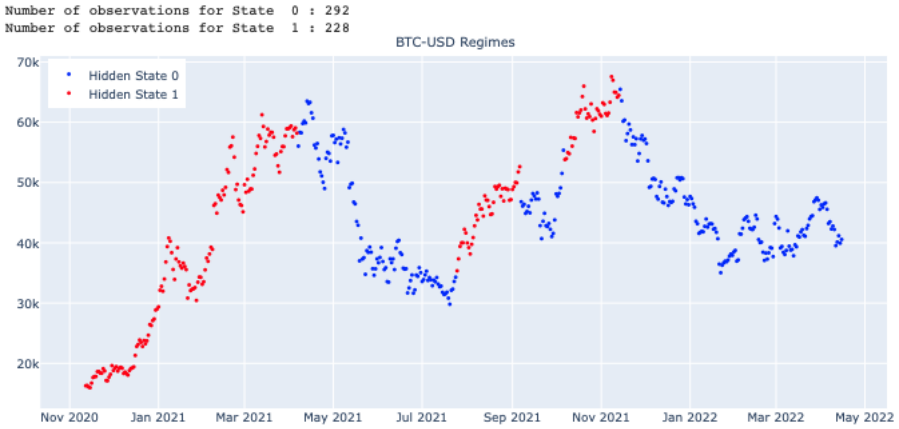

Fig.8

Final model selection

Given all the results above, there is not a single model that worked well for all three cryptocurrencies. The resulting regime estimations significantly changes for different random states, different parameter settings, as well as across the models. Overall, agglomerative clustering seems to provide regimes closest to the bear/bull/stagnant market for BTC, although it failed to identify smaller bear/bull periods. On the other hand, GMM seems to be the better choice for BNB and ETH.

Discussion

We implemented different statistical techniques and models for the initial level of market analysis. This includes structural tests like Chow and mixture models**.**

Our analysis reveals room for improvement in the cryptocurrency market regime detection methods explored. Although there is not a one-fit-all method for detecting market regimes, each method we attempted clusters some pattern of the change in cryptocurrency market. We further compared our results to the ones using multivariate Generalized White Noise (GWN) model [8], with their results referenced in Figure 9 for easier comparison. In the GWN results, the regime detected using either two or three states does not resemble bull/bear/stagnant market as well, and less interpretable compared to ours.

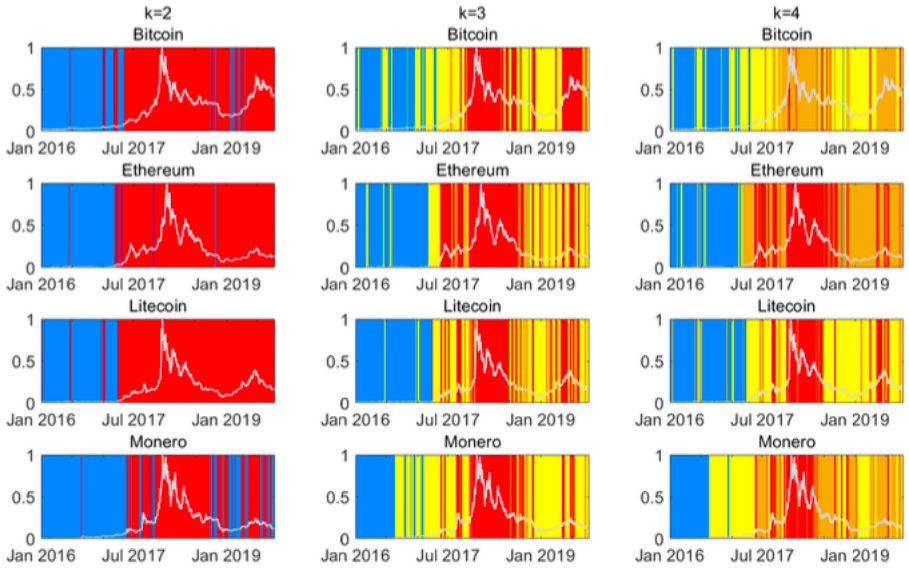

Fig.9

The crypto market's high volatility and lack of guardrails could explain why the techniques still have room for improvement. Unlike equities, no circuit breakers exist to curb explosive swings, enabling huge overnight price spikes and crashes. The free range of fluctuation poses a challenge for the algorithms to distinguish between stagnant, small change and large change, making it a hard task to distinguish them into three stable regimes. Therefore, a possible solution is to assume more underlying hidden states/patterns. On the other hand, the lack of regulation and change of policy imposed more short-time shifts, adding on to the difficulty of regime detection.

Conclusion

The initial analysis including correlation, volatility, distribution, break tests and mix models do help us in determining the price dynamics and relationships. The identification of different regimes can support in making informed trading strategies, by employing more complex methodologies to get more data-driven insights.

Markov Switching models give a strong statistical foundation for modeling regime transitions in the crypto market. Tweaking parameters to improve sensitivity and fit may improve accuracy in detecting bull, bear, and stationary phases. While hidden Markov models has been working well for market detection in equity and other financial instructions, the suboptimal result on cryptocurrency could be related to lack of anchors, circuit breakers and financial regulation – all leading to more liberal changes and less stable patterns.

Nevertheless, beyond the scope of this project and beyond statistical methods, Neural Network techniques such as Long Short-Term Memory (LSTM) models provide strong machine learning capabilities. LSTMs can detect complex time series patterns that standard models overlook. Their adaptable nonlinear structure enables them to track regime shifts even when prices change rapidly over time. Therefore, potential next steps include utilizing machine learning and deep learning methods on cryptocurrency market regime detections.

Our core framework and approach has furnished the building blocks for future upcoming undertakings providing a data-driven view on multilayered probabilistic perspectives on market dynamics to formulate strategic, calculated cryptocurrency strategies. We may robustly test alternative modeling philosophies by supplementing our investigation into bitcoin market dynamics with these sophisticated Markov Switching and LSTM paradigms.

References

[1] Wikipedia contributors. "Cryptocurrency." Wikipedia, Wikimedia Foundation, 18 November 2023, https://en.wikipedia.org/wiki/Cryptocurrency#:~:text=The%20first%20cryptocurrency%20was%20Bitcoin,market%20capitalization%20exceeding%20%241%20billion.

[2] Chan, S., Chu, J., Nadarajah, S., & Osterrieder, J. (2017). "A Statistical Analysis of Cryptocurrencies." Journal of Risk and Financial Management, vol. 10, no. 2, 2017, p. 12, https://doi.org/10.3390/jrfm10020012. Published 31 May 2017.

[3] Schwarz, Gideon. "Estimating the Dimension of a Model." The Annals of Statistics, vol. 6, no. 2, 1978, pp. 461-464. https://doi.org/10.1214/aos/1176344136.

[4] Nadarajah, Saralees, and Jeffrey Chu. "On the inefficiency of Bitcoin." *Economics Letters*, vol. 150, 2017, pp. 6-9. *ScienceDirect*, doi: 10.1016/j.econlet.2016.10.033.

[5] Tiwari, Aviral Kumar, R.K. Jana, Debojyoti Das, David Roubaud. "Informational efficiency of Bitcoin—An extension." *Economics Letters*, vol. 163, 2018, pp. 106-109. ISSN 0165-1765, https://doi.org/10.1016/j.econlet.2017.12.006.

[6] Murtagh, F., Legendre, P. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?". *J Classif*, vol. 31, 2014, pp. 274–295. https://doi.org/10.1007/s00357-014-9161-z.

[7] Song, Jung Yoon, Woojin Chang, Jae Wook Song. "Cluster analysis on the structure of the cryptocurrency market via Bitcoin–Ethereum filtering." *Physica A: Statistical Mechanics and its Applications*, vol. 527, 2019, 121339. ISSN 0378-4371, https://doi.org/10.1016/j.physa.2019.121339.

[8] Najafgholizadeh, A., Nasirkhani, A., Mazandarani, H. R., Soltanalizadeh, H. R., & Sabokrou, M. "Imaging Time Series for Deep Embedded Clustering: a Cryptocurrency Regime Detection Use Case." *2022 27th International Computer Conference*, Computer Society of Iran (CSICC), Tehran, Iran, Islamic Republic of, 2022, pp. 1-6. doi: 10.1109/CSICC55295.2022.9780526.

[9] Pennoni, Fulvia, et al. "Exploring the dependencies among main cryptocurrency log-returns: A hidden Markov model." Econometrics Journal, vol. 24, no. 3, 2021, https://doi.org/10.1111/ecno.12193.

[10] Kodama, Osamu, et al. "Regime Change and Trend Prediction for Bitcoin Time Series Data." CBU International Conference on Innovations in Science and Education, March 22-24, 2017, Prague, Czech Republic, www.cbuni.cz, www.journals.cz, pp. 384. DOI: http://dx.doi.org/10.12955/cbup.v5.954.

[11] Figà-Talamanca, Gianna and Focardi, Sergio M. and Patacca, Marco, Regime switches and commonalities of the cryptocurrencies asset-class (May 15, 2019). Available at SSRN: https://ssrn.com/abstract=3388642 or http://dx.doi.org/10.2139/ssrn.3388642

[12] OpenAI. "Citation Format." *OpenAI Chat*, OpenAI, 18 Dec. 2023, https://chat.openai.com/c/88414c9f-e930-40f6-98cf-91f23bfb9798.