Detecting Financial Market Regimes

CS5831 Final Project: The Mining Avengers

Blake Krouth
College of Computing
Michigan Technological University
Houghton, MI, USA
bjkrouth@mtu.edu

Nandhika Rajmanikandan
College of Computing
Michigan Technological University
Houghton, MI, USA
nrajmani@mtu.edu

Ganesh Vannam
College of Computing
Michigan Technological University
Houghton, MI, USA
gvannam@mtu.edu

***Abstract*—Write this last and summarize the research conducted, the conlusions reached, and the potential implications of those conclusions (up to 250 words)**

***Index Terms*—market regime, clustering, data mining, machine learning**

I. Introduction

Understanding the trend of the financial market is vital to financial investors who wish to intelligently manage their portfolios and assets. Due to the unobservable and complex workings of the financial market, market trend detection is not an easy task. Detecting market regimes is one approach to this problem. In the context of a financial market, a regime refers to a period of similar behavior in the market [1], [2]. By understanding the regime, or overall behavior, of a financial market, an investor can create a better strategy. The goal of this report is to explore various clustering and feature engineering techniques in order to create clusters that accurately and intuitively describe market regimes in way that would be of use to investors.

II. Related Works

Recent attempts to detect market regimes rely on unsupervised learning techniques that assume latent variables (e.g., market returns or macroeconomic indicators) determine the market state. In [3], fuzzy C-Means clustering was applied to a sliding window over multiple time-series data of both synthetic and real-world economic indicators. While the authors were more interested in clustering as a precursor to market forecasting, they were successfully able to create clusters that were fed into their forecasting model. In [4], the financial metric of monthly realized covariances is used to create both statistical (i.e., TVAR, LSTVAR, and MSVAR) and clustering (i.e., AGNES) models. In model evaluation on a synthetic dataset, the clustering model performed the best, and was considered by the authors to be the best performer alongside the LSTVAR statistical model. [2] take a unique approach by first training a k-means clustering model on Federal Reserve Economic Data (FRED) to cluster market regimes and next training various supervised models on both the FRED data and clustered regimes to classify out-of-sample data.

[1] propose a framework for comparing unsupervised models in the context of market regime detection. The framework addresses common problems in market regime detection, such as models not adapting to changing market conditions, ensuring consistent labeling as regimes change over time, and choosing the appropriate number of regimes. Notably, the authors opt for a rolling window retraining, similar to what was done in [3]. Using their framework, [1] evaluated a Hidden Markov Model with Gaussian observation model and a Hidden Markov Model with Gaussian mixture observation model that was fitted to FRED indicators. The process was able to generalize well to other macroeconomic indicators, such as futures and mutual funds.

A frequent drawback of market regime detection models is that they assume a fixed number of regimes that are chosen during training and are unable to adapt to dynamic market conditions in which the number of regimes vary [1], [3]. [3] address this by dividing the data with a sliding window that dynamically chooses the appropriate number of regimes, while [1] describes a thresholding policy. However, [1] only addresses the condition that an additional regime should be added, not that the number of regimes could decrease, which they state is logical since a regime condition does not disappear and there should be historical consistency in the model.

A theme common to [4] and [3] is the use of a synthetic dataset with known, generated regimes to perform a quantitative evaluation on clustering performance. While potentially ideal for model evaluation, synthetic dataset generation is out of the scope of the class and our goals.

III. Data Description

Historical daily time-series stock data of the S&P 500 (ticker ^GSPC) from January 1, 1990 to January 1, 2024 was obtained from Yahoo Finance [5], resulting in 8565 ratio observations containing the daily close, high, low, open, and volume prices.

TODO: what does rubric mean by distribution of predictors variable

IV. Methodology

A. Data Acquisition and Preprocessing

Using the data from [5], the daily return, (1), was calculated using the day's closing price, p_t and the previous day's closing price p_{t-1}.

$$r_t = \frac{p_t}{p_{t-1}} - 1 \tag{1}$$

Due to the noise present in daily time-series financial data, this data was then resampled monthly, with the last value of each month retained. The monthly return was calculated according to (1) using the month's closing price, p_t and the previous month's closing price p_{t-1}.

The compounding cumulative return, (2), was then calculated for each month using the monthly returns up until that point in time.

$$c_t = \prod_{i=1}^{t}(1 + r_i) \tag{2}$$

Finally, the volatility, or standard deviation, (3), of the daily returns was calculated on a 21-day rolling basis. Once again, the motivation behind the rolling basis was to reduce the noise present in the financial data.

$$v_t = \sigma(r_t) \tag{3}$$

At this point, all NA values were dropped to account for the rolling calculations, resulting in the monthly returns and volatilities that can be seen in "Fig. 1" and "Fig. 2," respectively.

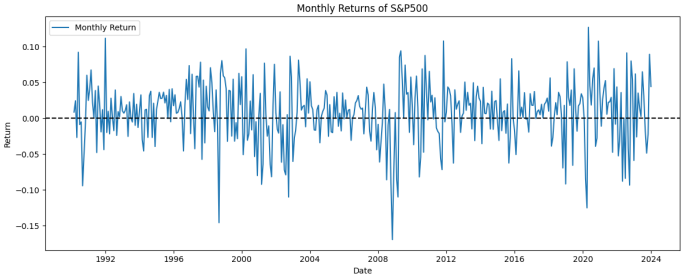

Fig. 1. Monthly returns of S&P 500.

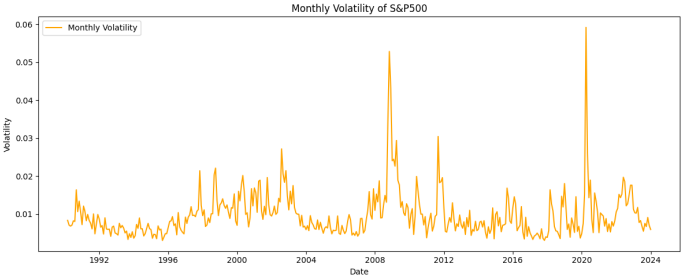

Fig. 2. Monthly volatility of S&P 500.

The distribution of the monthly returns appeared to be roughly normal without significant outliers, as seen in "Fig. 3," indicating that efforts were successful in removing most noise from the data.

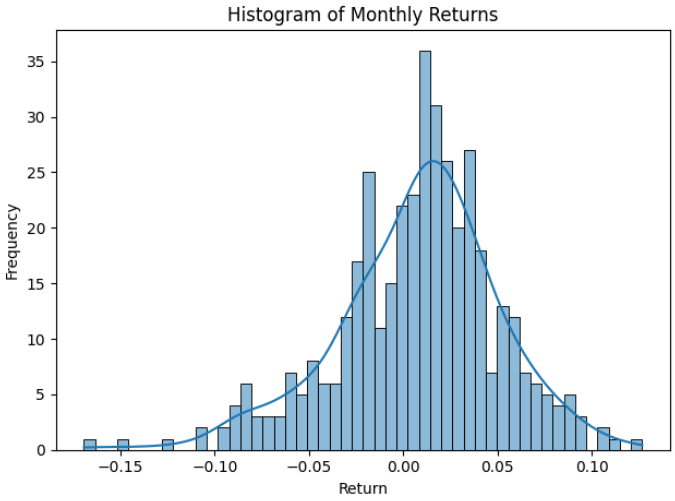

Fig. 3. Distribution of S&P 500 monthly returns.

A second feature set with further reduced noise and additional feature engineering was also created and used for model evaluation. This monthly feature set contained:

- The 6 month moving average of the monthly cumulative return
- The 6 month moving average of the monthly volatility
- The 3-month percent change in monthly closing price
- A z-score representing the month's return in comparison to the rolling average for the monthly returns in the past 12 months

B. Monthly k-means

The first model explored was k-means. Standard scaling was applied to the monthly return, cumulative return, and volatility features calculated previously. Using values of k from 2 to 10, the inertia, silhouette score, Calinski-Harabasz index, and Davies-Bouldin score were all calculated to evaluate the clustering performance, as seen in "Fig. 4."

C. 6-Month Smoothed Cumulative Return

Using the standard-scaled 6 month moving average of the monthly cumulative return obtained from the multivariate smoothed feature set, a k-means model was fit and evaluated with silhouette score. The highest silhouette score of 0.67 was obtained with k=5, as seen in "Fig. 5."

A Hidden Markov Model (HMM) was fit to the standard-scaled smoothed 6-month cumulative return and evaluated with log-likelihood. The highest score of 192.85 was obtained with 6 regimes, as seen in "Fig. 6."

A fuzzy c-means model (FCM) was fit to the standard-scaled smoothed 6-month cumulative return and evaluated with fuzzy partition coefficient (FPC). The highest score of 0.8344 was obtained with 5 regimes, as seen in "Fig. 7."

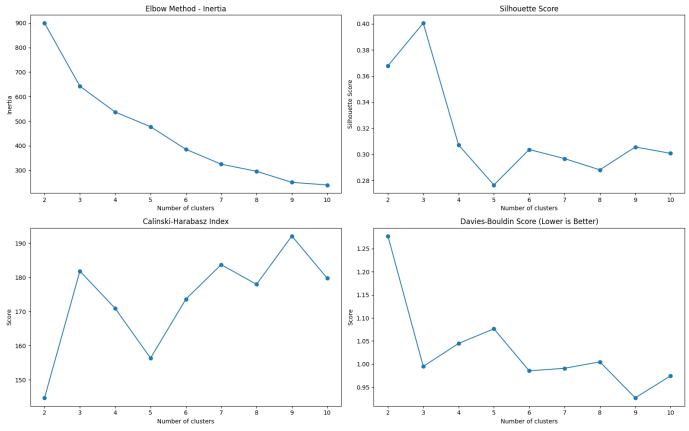

Fig. 4. Evaluation of k-means clustering for k=[2, 6].

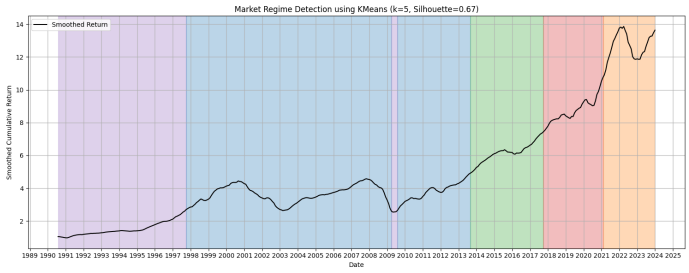

Fig. 5. K-Means Market Regimes Visualized Over 6-Month Smoothed Returns.

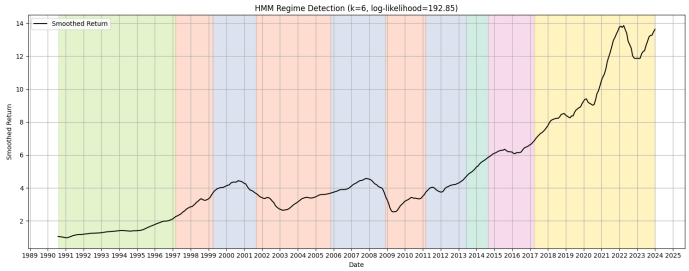

Fig. 6. HMM Market Regimes Visualized Over 6-Month Smoothed Returns.

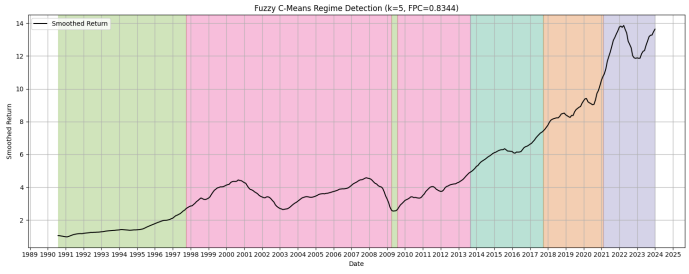

Fig. 7. FCM Market Regimes Visualized Over 6-Month Smoothed Returns.

A Gaussian mixture model (GMM) was fit to the standard-scaled smoothed 6-month cumulative return and evaluated with Bayes Information Criterion (BIC) and log-likelihood. The best BIC of 839.67 was obtained with 5 regimes, with a log-likelihood of -377.86, as seen in "Fig. 8."

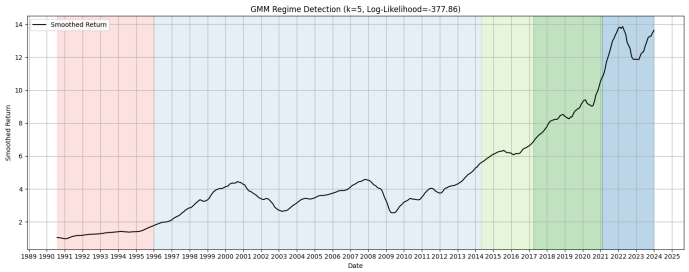

Fig. 8. GMM Market Regimes Visualized Over 6-Month Smoothed Returns.

D. Multivariate Smoothed Feature Set

k-means
hmm
fuzzy c-means
gmm

V. Discussion of Results

A. 6-Month Smoothed Cumulative Return

Initially, it was observed that using a reduced feature set (i.e., only the 6-month moving average of monthly returns) led to higher performance metrics when compared to the results from the higher feature space in "Fig. 4." This motivated the comparison of the various models using the 6-month moving average of monthly returns. While not ultimately successful, this provided insight into behavior of the various models and of the performance metrics.

The k-means and fuzzy c-means detected very similar clusters with strong cluster separation. Visually, the clusters appear to capture general transitions of market trends instead of sharp transitions that one may expect, like a transition from bullish to bearish, for example.

The Gaussian mixture model was the lowest performer, producing clusters that are visually ambiguous with low confidence.

TODO: insert table describing HMM regimes with mean/standard deviation The hidden Markov model produced the most easily interpretable clusters:

- Cluster 0: Moderate but consistent bullish trend
- Cluster 1: Mild growth/recovery with an average return
- Cluster 2: Stable bull with low volatility
- Cluster 3: Strong bull with high return and very low volatility
- Cluster 4: Flat/stagnant with minimal growth
- Cluster 5: Extreme bull or rebound with very high return and variance

Here, the hidden Markov model captures temporal transitions the best. This is intuitive, since other clustering methods used, like k-means for example, simply create spherical clusters that group data based on similar feature sets with no time-series context, while the hidden Markov model captures time-series context with its concept of probabilistic regime changes that depend on previous states.

While using one feature (i.e., the 6-month moving average of monthly returns) seemed to produce better metric scores, it became apparent that the clusters, while mostly high-confidence, were hard to interpret with the exception of those produced by the HMM. Furthermore, with data as complex as financial data, it seems unlikely that trends can be accurately described with only one feature. This motivated further exploration of the full multivariate smoothed feature set.

B. Multivariate Smoothed Feature Set

TODO

VI. Conclusion

Potential ideas to conclude on:
* overall best performer
* the use of a sliding window (literature review)
* out of sample testing
* real-world usage/applicability

References

[1] A. Hirsa, S. Xu, and S. Malhotra,"Robust rolling regime detection (r2-rd): A data-driven perspective of financial markets," To be published. Accessed: Mar. 1, 2025. [Online]. Available at SSRN.

[2] P. Akioyamen, Y. Z. Tang, and H. Hussien, "A hybrid learning approach to detecting regime switches in financial markets," presented at ACM International Conference on AI in Finance, New York, NY, USA, Oct. 15-16, 2020. doi: https://doi.org/10.1145/3383455.3422521

[3] R. Chen, M. Sun, K. Xu, J. Patenaude, and S. Wang, "Clustering-based cross-sectional regime identification for financial market forecasting," in *Database and Expert Systems Applications*, vol. 13427, pp. 3-16, 2022. doi: https://doi.org/10.1007/978-3-031-12426-6_1

[4] A. Bucci and V. Ciciretti, "Market regime detection via realized covariances," presented at 14th International Conference on Computational and Financial Econometrics, virtual, 2020. doi: https://doi.org/10.1016/j.econmod.2022.105832

[5] Yahoo Finance, "S&P 500 (^GSPC) Stock Price" Yahoo Finance. [Online]. Available: https://finance.yahoo.com/quote/%5EGSPC/history/

[6] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. "Good debt or bad debt: Detecting semantic orientations in economic texts," in *Journal of the Association for Information Science and Technology*, vol. 64, 2014. Available: https://huggingface.co/datasets/takala/financial_phrasebank