Week 5: Unsupervised Detection of Latent Market Regimes in Financial Time Series Using Gaussian Hidden Markov Models.

Tishu Verma (20251602024), Mansi Surti (20251602014), Reedam Choudary (20251602020)
Indian Institute of Information Technology Vadodara
Gandhinagar Campus, Gujarat, India

***Abstract*—This report shows Financial time series data is rarely stationary; it exhibits changing behaviors known as" market regimes" (e.g., bull markets, bear markets, or sideways consolidation). This project applies a Gaussian Hidden Markov Model (HMM) to daily log returns of Apple Inc. (AAPL) stock from 2015 to present. By minimizing the Bayesian Information Criterion (BIC), the analysis identified an optimal 4-state model. The results successfully segmented the timeline into distinct periods of low-volatility growth and high-volatility distress, demonstrating the efficacy of HMMs in unsupervised market classification.**

***Index Terms*—Market Regime Detection, Hidden Markov Models (HMM), Time Series Analysis, Volatility Clustering, Apple Inc. (AAPL), Bayesian Information Criterion (BIC).**

I. Introduction

THE CS367/659 Artificial Intelligence course at IIIT Vadodara includes weekly laboratory assignments designed to bridge theoretical concepts with practical implementation. Week 5 Standard financial models often assume that stock returns follow a consistent distribution over time. However, in reality, markets switch between different structural environments. A strategy that works in a calm, trending market may fail catastrophically during a volatile crash.

The objective of this study is to use a probabilistic approach to detect these hidden environments without human bias. We utilize a Hidden Markov Model (HMM), which assumes that the observed stock returns are generated by a set of unobserved (hidden) states. By analyzing the historical price data of Apple Inc., we aim to characterize these states based on their mean returns and volatility. The key objectives are:

1) To Implement a Gaussian HMM: Apply a probabilistic model to the daily log returns of AAPL stock from 2015 to the present.
2) To Characterize Market Regimes: statistically analyze the identified states to define them by their risk (volatility) and reward (mean return) profiles.
3) To Visualize Regime Dynamics: Map the inferred states onto historical price charts to visually validate the model's ability to detect major market shifts, such as the 2020 COVID-19 crash or the 2022 correction.

II. Methodology

1) Data Acquisition and PreprocessingHistorical financial data for Apple Inc. (AAPL) was retrieved using the yfinance API, covering the period from January 1, 2015, to the present. The raw data consisted of daily" Adjusted Close" prices, which account for dividends and stock splits. To prepare the data for statistical modeling, we transformed the raw prices into Logarithmic Returns (log-returns). This step is crucial for two reasons: first, raw prices are non-stationary (they trend upwards), whereas returns tend to be stationary (mean-reverting); second, log-returns are time-additive. The transformation is defined as:

$$r_t = \ln \frac{P_t}{P_{t-1}}$$

where P_t is the price at time t and P_{t-1} is the price at the previous close. Any missing values (NaNs) resulting from the shift operation were removed prior to training.
2) The Gaussian Hidden Markov Model (HMM)An HMM is a probabilistic model used to represent systems that transition between unobserved (hidden) states. In the context of this study, we assume the market exists in one of K distinct "regimes" at any given time, but we can only observe the returns generated by that regime.The model consists of two main components:The Transition Matrix (A): A square matrix where each element a_{ij} represents the probability of switching from state i to state j on the next day.The Emission Probability: The probability of observing a specific return r_t given the current hidden state. Since stock returns generally follow a bell curve, we modeled these emissions using a Gaussian distribution.For a given state k, the returns are assumed to be drawn from:

$$r_t | s_t = k \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

where μ_k is the mean return of that regime (e.g., positive for bull, negative for bear) and σ_k^2 is the variance (volatility).
3) Parameter Estimation (Training)To "fit" the model, we must estimate the unknown parameters: the start probabilities, the transition matrix, and the means and variances for each state. Since the states are hidden, we cannot calculate these directly.We utilized the Baum-Welch algorithm, a special case of the Expectation-Maximization (EM) algorithm. This iterative process works in two steps:E-Step (Expectation): Compute the

probability of being in each state at each time step, given the current parameter estimates.M-Step (Maximization): Update the parameters (means, variances, transitions) to maximize the likelihood of the data based on the probabilities found in the E-Step.This cycle repeats until the model converges (i.e., the likelihood score stops improving).

4) Model Selection and OptimizationDetermining the correct number of hidden states is a critical challenge in unsupervised learning. Too few states (e.g., 2) might oversimplify the market into just "Up" or "Down," missing nuance. Too many states (e.g., 10) would "overfit," detecting random noise rather than real structural patterns.To solve this, we trained separate models with 2, 3, 4, and 5 components. We selected the optimal architecture using the Bayesian Information Criterion (BIC). The BIC introduces a penalty term for the number of parameters, favoring simpler models unless the added complexity significantly improves the fit:

$$BIC = k \ln(n) - 2 \ln(\hat{L})$$

where k is the number of parameters, n is the number of data points, and $\hat{L}$ is the maximized likelihood. The model with the lowest BIC score was chosen for the final analysis.

III. Results

Optimal State Selection The optimization process yielded the following BIC scores:
2 States: -14,769.69
3 States: -14,716.82
4 States: -14,787.37 (Lowest/Best)
5 States: -14,680.55
The 4-state model provided the best trade-off between complexity and accuracy.

a) Regime Characterization The four inferred regimes show distinct risk/reward profiles (based on the output table):
Regime 2 (The "Steady Bull"): Characterized by the lowest volatility(1.08 percent)
Regime 0 (The "Volatile Bull"): Higher mean returns (0.25) percent.
Regime 1 (The "Correction"): Moderate volatility (2.17) percent.
Regime 3 (The "Crash"): The most dangerous state. It exhibits extreme volatility (5.69) percent.
b) Visualization The cumulative performance chart (shaded by regime) visually confirms the model's logic. The "Crash" regime (Regime 3) appears as narrow, distinct bands during known market sell-offs, while the "Steady Bull" (Regime 2) occupies long, continuous stretches of the timeline.
c) Forecasting Using the transition matrix and the posterior probability of the final observation, the model predicts the state for the next trading day.

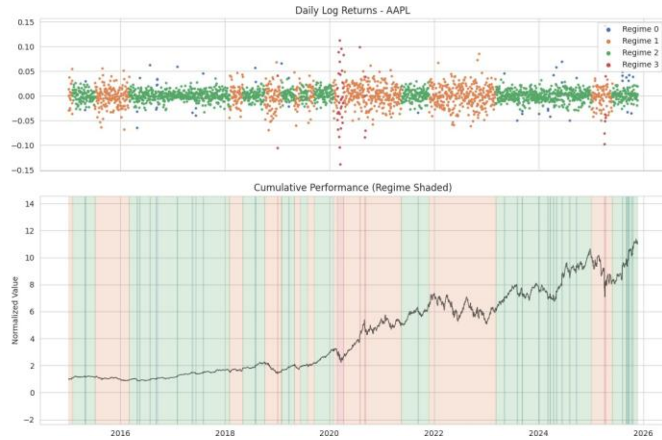

Fig. 1: Market regime segmentation of AAPL returns (2015–Present) using an optimal 4-state Gaussian HMM.

Prediction: There is an 85.5 percent probability that the market will remain in Regime 2 (Steady Bull), with a 10.8 percent chance of shifting to Regime 0.
Figure 1 shows:
The HMM's ability to disentangle market conditions. The top panel reveals distinct volatility clusters: Regime 2 (Green) represents stable, low-variance returns, while Regime 3 (Red) captures extreme outliers. The bottom panel maps these states to the price history, showing a strong correlation between the "Green" regime and steady uptrends, whereas "Red" and "Orange" segments accurately identify periods of stress, such as the 2020 COVID-19 crash and the 2022 correction.

IV. Conclusion

- Optimization: The 4-state Gaussian HMM was statistically verified (via BIC) as the optimal architecture for AAPL returns.
- Regime Classification: The model successfully isolated stable, low-volatility growth periods from high-volatility crash events.
- Real-World Accuracy: It accurately pinpointed historical structural breaks, effectively capturing the 2020 COVID-19 market shock.
- Utility: The study confirms HMMs as robust tools for automated financial regime detection and dynamic risk management.

The work highlights the remarkable foresight of early AI research and the enduring relevance of adaptive learning principles in modern reinforcement learning.

V. GitHub Repository

Complete implementation available at: https://github.com/MansiSurti11/AI_LabManual

REFERENCES

[1] Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1–22.

[2] Hmmlearn: Hidden Markov Models in Python, with scikit-learn like API. GitHub Repository.

[3] yfinance: Yahoo! Finance Market Data Downloader. Python Package Index.