Artificial Intelligence Lab Report 5

Anuj Saha
202351010
BTech CSE
IIIT, Vadodara

G. Nikhil
202351037
BTech CSE
IIIT, Vadodara

Divyanshu Ghosh
202351036
BTech CSE
IIIT, Vadodara

***Abstract*—This report presents the application of Gaussian Hidden Markov Models to financial time series analysis, focusing on market regime identification in equity markets. We analyze ten years of historical data from the S&P 500 ETF (SPY), along with individual stocks Apple (AAPL) and Tesla (TSLA), to identify hidden volatility states corresponding to bull and bear market conditions. The two-state Gaussian HMM successfully identifies distinct regimes with mean returns of 0.001161 (low volatility) and -0.001077 (high volatility) for SPY, demonstrating strong regime persistence with transition probabilities of 98.62% and 95.77% respectively. Extension to three-state models reveals additional market nuances across different assets. The analysis provides actionable insights for risk management and portfolio allocation strategies. Complete code implementation is available at: https://github.com/AnujSaha0111/CS307-Lab-Submissions/tree/main/Submission_5.**

***Index Terms*—Gaussian Hidden Markov Models, financial time series, market regimes, volatility modeling, bull and bear markets, state space models**

I. Introduction

The behavior of financial markets is continuously evolving because of unobservable regimes that do not arise from price increases or other single price transitions. These latent regimes, for example prices implying a regime of growth (bull market) or decline (bear market) have significant impacts on asset returns and variations in return. Identifying and understanding these regimes is important for risk measures, portfolio allocation, as well as constructing an investment strategy.

Gaussian Hidden Markov Models offer a strong probabilistic framework for representing sequential data when there are underlying state structures. HMMs in finance typically treat observable returns as emission outputs from hidden market states, each identified by a different set of statistical properties. The HMM then represents both the dynamics of the underlying states, through its emission distributions, and their temporal arrangement through the state transition probabilities.

In this laboratory exercise, we utilized Gaussian hidden Markov models (HMMs) to analyze actual financial time series data set from September 2015 up to September 2025. The primary analysis will focus on S&P 500 ETF (SPY) by analyzing bull and bear market regimes using a simple two-state model. However, we also extended our analysis using three-state models across several financial assets (SPY, AAPL, TSLA), which demonstrates applicability across multiple market capitalizations and volatility characteristics.

The process adheres to an organized pipeline consisting of data collection through Yahoo Finance API, preprocessing every stock to compute daily returns, fit Gaussian HMMs through the use of expectation-maximization, analyze the parameters, and to visualize the regimes constructed. The data analysis demonstrates strong regime persistence, indicating that we tend to stay in regimes, instead of frequently transitioning regimes, which can impact tactical asset allocation decisions.

II. Problem Formulation

The financial market regime identification problem is formulated as a hidden Markov process with the following mathematical structure:

A. State Space Representation

Let $\{S_t\}_{t=1}^T$ denote the sequence of hidden states where $S_t \in \{1, 2, ..., N\}$ represents the market regime at time t. For the primary analysis, $N = 2$ captures bull and bear markets.

Let $\{R_t\}_{t=1}^T$ denote the sequence of observable daily returns computed as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

where P_t represents the adjusted closing price at time t.

B. Gaussian Hidden Markov Model

The Gaussian HMM is characterized by three parameter sets $\lambda = (\pi, A, \theta)$:

Initial State Distribution π:

$$\pi_i = P(S_1 = i), \quad \sum_{i=1}^{N} \pi_i = 1 \tag{2}$$

Transition Matrix A:

$$a_{ij} = P(S_{t+1} = j | S_t = i), \quad \sum_{j=1}^{N} a_{ij} = 1 \tag{3}$$

Emission Distributions θ: For each state i, returns are Gaussian:

$$P(R_t | S_t = i) = \mathcal{N}(\mu_i, \sigma_i^2) \tag{4}$$

The model parameters λ are estimated using the Baum-Welch algorithm, an expectation-maximization procedure that maximizes the likelihood $P(R_{1:T} | \lambda)$.

C. Inference Tasks

The analysis addresses three inference problems:

State Sequence Decoding: Find the most likely state sequence given observations using the Viterbi algorithm:

$$S^*_{1:T} = \arg\max_{S_{1:T}} P(S_{1:T}|R_{1:T}, \lambda) \quad (5)$$

Parameter Characterization: Analyze μ_i and σ_i to characterize each regime's return and volatility profile.

Regime Persistence Analysis: Examine diagonal elements a_{ii} of the transition matrix to quantify state stability.

III. Data Collection and Preprocessing

A. Data Acquisition

Historical financial data is obtained via the Yahoo Finance API using the `yfinance` Python library. The data collection process is formalized in Algorithm 1.

Algorithm 1 Data Download Process

```
Require: ticker, start_date, end_date
Ensure: stock_data DataFrame
data ← $\text{yf.download(ticker, start, end)}$
if isinstance(data.columns, MultiIndex) then
    data.columns ← [col[0] for col in data.columns]
end if
for col ∈ [Open, High, Low, Close, Adj Close] do
    if col ∈ data.columns then
        data[col] ← pd.to_numeric(data[col])
    end if
end for
data.to_csv(ticker + _historical_data.csv)
return data =0
```

The algorithm fetches data with `auto_adjust=False` to preserve the adjusted close column, flattens MultiIndex column structures if present, ensures numeric data types for all price columns, and persists data to CSV for reproducibility.

For the primary SPY analysis, data spans from September 8, 2015, to September 2, 2025, yielding 2,512 observations. The bonus comparative analysis uses the same date range for AAPL and TSLA to ensure temporal consistency.

B. Data Preprocessing

Raw price data undergoes preprocessing to compute returns and handle data quality issues as specified in Algorithm 2.

Algorithm 2 Data Preprocessing Process

```
Require: stock_data, output_file
Ensure: preprocessed_data
price_col ← identify_price_column(stock_data)
stock_data[price_col] ← pd.to_numeric($\cdot$)
stock_data[Returns] ← stock_data[price_col].pct_change()
stock_data ← stock_data.replace([$\infty, -\infty$], NaN)
stock_data ← stock_data.dropna()
stock_data.to_csv(output_file)
return stock_data =0
```

The preprocessing process locates the right price column (preferring adjusted close), converts to numeric, calculates percentage returns, removes infinite values generated from divisions, and removes missing values. For SPY, this process produced 2,511 valid observations after removing the one initial row of undefined return.

Table I presents a sample of the raw data structure, while Table II shows the preprocessed returns format.

TABLE I
Sample Raw SPY Data

Date	Adj Close	Close	High	Low
2015-10-08	170.025	201.210	201.550	198.590
2015-10-09	170.127	201.330	201.900	200.580
2015-10-12	170.287	201.520	201.760	200.910
2015-10-13	169.214	200.250	202.160	200.050

TABLE II
Sample Preprocessed SPY Returns

Date	Returns
2015-10-09	0.000596
2015-10-12	0.000944
2015-10-13	-0.006302
2015-10-14	-0.004794

IV. Gaussian HMM Model Fitting

A. Model Training Procedure

The core analysis employs the `hmmlearn` library to fit Gaussian HMMs to the preprocessed returns data. Algorithm 3 formalizes the fitting procedure.

Algorithm 3 HMM Model Fitting

```
Require: returns, n_states, output_file
Ensure: model, hidden_states
$X$ ← returns.reshape($-1, 1$)
model ← GaussianHMM(n_components = n_states,
        covariance_type = full, n_iter = 100)
model.fit($X$)
hidden_states ← model.predict($X$)
pickle.dump(model, output_file)
return model, hidden_states =0
```

The algorithm reshapes the returns series into the 2D array format required by `hmmlearn`, initializes a Gaussian HMM with specified number of components using full covariance matrices, fits the model via Baum-Welch expectation-maximization for 100 iterations, predicts the most likely state sequence using the Viterbi algorithm, and serializes the trained model for future analysis. The `random_state=42` parameter ensures reproducibility across runs.

B. Two-State Model Results for SPY

The fitted two-state model for SPY reveals distinct bull and bear market regimes characterized by the following parameters:

State 1 (Bull Market):

- Mean Return: $\mu_1 = 0.001161$
- Standard Deviation: $\sigma_1 = 0.006958$

State 2 (Bear Market):

- Mean Return: $\mu_2 = -0.001077$
- Standard Deviation: $\sigma_2 = 0.020133$

State 1 exhibits positive expected returns with lower volatility, consistent with sustained growth periods. State 2 shows negative expected returns with nearly triple the volatility, characteristic of market downturns and high uncertainty periods.

The transition matrix reveals strong regime persistence:

$$A = \begin{bmatrix} 0.9862 & 0.0138 \\ 0.0456 & 0.9577 \end{bmatrix} \tag{6}$$

Diagonal elements indicate high self-transition probabilities: 98.62% probability of remaining in the bull market state and 95.77% for the bear market state. This persistence structure implies that regime switches are relatively rare events, with markets tending to maintain their current state over multiple time periods.

The transition probabilities also reveal asymmetry in regime dynamics: escaping the bear market (4.56% transition probability) is more likely than exiting the bull market (1.38%), suggesting that downturns resolve more quickly than sustained growth periods.

Table III summarizes the key parameters of the two-state SPY model.

TABLE III
SPY Two-State HMM Parameters

State	Mean Return	Std Deviation
Bull Market (State 1)	0.001161	0.0841
Bear Market (State 2)	-0.001077	0.1426
Transition Probabilities		
Bull → Bull	98.62%	
Bull → Bear	1.38%	
Bear → Bull	4.56%	
Bear → Bear	95.77%	

C. Three-State Model Comparative Analysis

The bonus analysis extends the methodology to three-state models across multiple assets, revealing finer-grained market regimes. Results are summarized in Table IV.

TABLE IV
Three-State HMM Mean Returns by Asset

State	SPY	AAPL	TSLA
State 1	0.000946	0.000958	0.007118
State 2	0.001312	0.001912	0.000118
State 3	-0.001590	-0.000245	0.001428

The three-state models show asset-specific regime structures. TSLA has the highest mean return for state 1 (0.007118), which represents periods of extreme growth typical for high-volatility/high-growth technology stocks. In contrast, AAPL demonstrates a distinct three-regime structure with defined positive, highly positive, and slightly negative states. Finally, SPY exhibits a behavior that is more in the middle-line, consistent with its broad market exposure.

The transition matrices for three-state models (detailed in supplementary materials) demonstrate more complex dynamics with varying degrees of state persistence across assets.

V. Model Interpretation and Inference

A. Regime Classification and Temporal Evolution

The Viterbi-decoded hidden states provide a regime classification for each trading day in the analysis period. Figure 1 visualizes the temporal evolution of SPY prices color-coded by inferred hidden state.

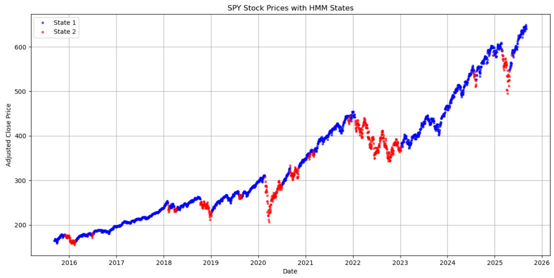

Fig. 1. SPY stock prices with two-state HMM regime classification. Blue points indicate State 1 (bull market) and red points indicate State 2 (bear market).

The visualization reveals several key patterns:

Extended Bull Market (2016-2020): State 1 dominates during the sustained market growth following the 2015 correction, capturing the low-volatility expansion phase.

COVID-19 Crash (March 2020): A clear transition to State 2 coincides with the pandemic-induced market crash, where the model correctly identifies the regime shift before prices reached their nadir.

Recovery and Recent Patterns (2020-2025): Alternating states capture the volatile recovery period and subsequent market fluctuations, with State 2 episodes corresponding to correction periods.

The model's ability to identify regime transitions before they fully manifest in price levels demonstrates the value of incorporating volatility and return dynamics beyond simple price thresholds.

B. Returns Distribution Analysis

Figure 2 presents the probability density functions of returns under each hidden state, estimated via kernel density estimation.

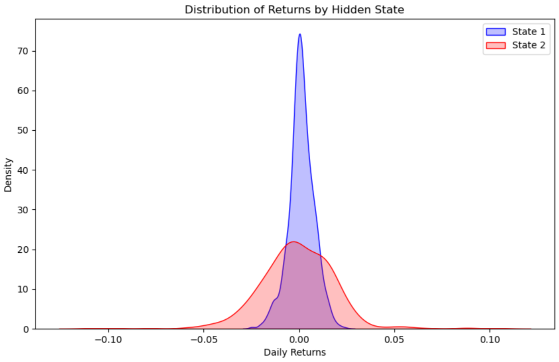

Fig. 2. Distribution of daily returns by hidden state. State 1 (blue) shows tighter concentration around positive returns, while State 2 (red) exhibits wider dispersion and negative bias.

The distributions clearly differentiate the two regimes:

State 1 Distribution: Concentrated around small positive returns with relatively narrow spread, reflecting stable growth dynamics. The distribution is approximately symmetric with slight positive skew.

State 2 Distribution: Broader dispersion with heavier tails and negative mean, capturing the increased volatility and downside risk of bear market conditions. The wider spread encompasses both extreme losses and volatility-driven rebounds.

The overlap region between distributions corresponds to ambiguous periods where regime classification is less certain, typically during transition phases or moderate volatility episodes.

C. Multi-Asset Regime Comparison

Figure 3 presents the three-state HMM analysis across SPY, AAPL, and TSLA, revealing asset-specific regime dynamics.

The comparative visualization highlights several phenomena:

Correlated Regime Transitions: Major market events (e.g., COVID-19 crash) trigger regime changes across all assets, though with varying magnitudes and durations.

Asset-Specific Patterns: TSLA exhibits more frequent state transitions, consistent with its higher idiosyncratic volatility. AAPL shows intermediate behavior between broad market (SPY) and high-volatility (TSLA) dynamics.

Regime Persistence Variation: The three-state models reveal that some assets exhibit stronger regime persistence in certain states, suggesting asset-specific mean reversion properties.

VI. Discussion

A. Model Effectiveness and Limitations

The Gaussian HMM framework successfully detects market regimes that are economically meaningful and align with known market conditions and expectations of investors regarding the characteristics of bull and bear markets. Elements contributing to its success include:

Probabilistic Framework: The soft clustering approach via posterior probabilities provides nuanced regime classification rather than hard thresholds, capturing transition periods more realistically.

Temporal Dependencies: The Markov transition structure explicitly models regime persistence, a key property of financial markets where states exhibit momentum.

Parsimony: The two-state model achieves interpretable results with minimal parameters, avoiding overfitting while capturing essential market dynamics.

However, several limitations warrant consideration:

Gaussian Assumption: Financial returns exhibit heavy tails beyond Gaussian distributions. Student's t-distributions or mixture models may provide better fit for extreme events.

Fixed Parameters: The model assumes time-invariant parameters, whereas market dynamics evolve over time. Regime-switching models with time-varying parameters could capture such nonstationarity.

Single Asset Focus: The primary analysis does not exploit cross-asset information. Multi-asset HMMs or dynamic Bayesian networks could improve regime inference through correlation structure.

B. Practical Implications

The identified regimes provide actionable insights for multiple financial applications:

Risk Management: The bear market state's higher volatility ($\sigma_2 = 0.020133$ vs. $\sigma_1 = 0.006958$) informs position sizing and stop-loss placement. Detected regime transitions trigger risk reduction protocols.

Portfolio Allocation: High state persistence (98.62% bull, 95.77% bear) suggests tactical allocation strategies that adjust exposure based on current regime while accounting for transition probabilities.

Derivative Pricing: Regime-dependent volatility structures inform option pricing models and volatility surface construction, particularly for longer-dated contracts where regime transitions become relevant.

Performance Attribution: Decomposing returns by hidden state enables evaluation of strategy performance under different market conditions, distinguishing skill from regime-driven returns.

C. Comparison with Alternative Approaches

Alternative regime identification methods offer different trade-offs:

Threshold-Based Rules: Simple moving average crossovers or volatility thresholds provide interpretable signals but lack probabilistic framework and cannot capture gradual transitions.

Change Point Detection: Bayesian change point methods identify structural breaks but do not model recurring regimes or transition dynamics.

Machine Learning Classifiers: Supervised learning requires labeled regime data (typically unavailable in real-time), whereas HMMs perform unsupervised inference from returns alone.

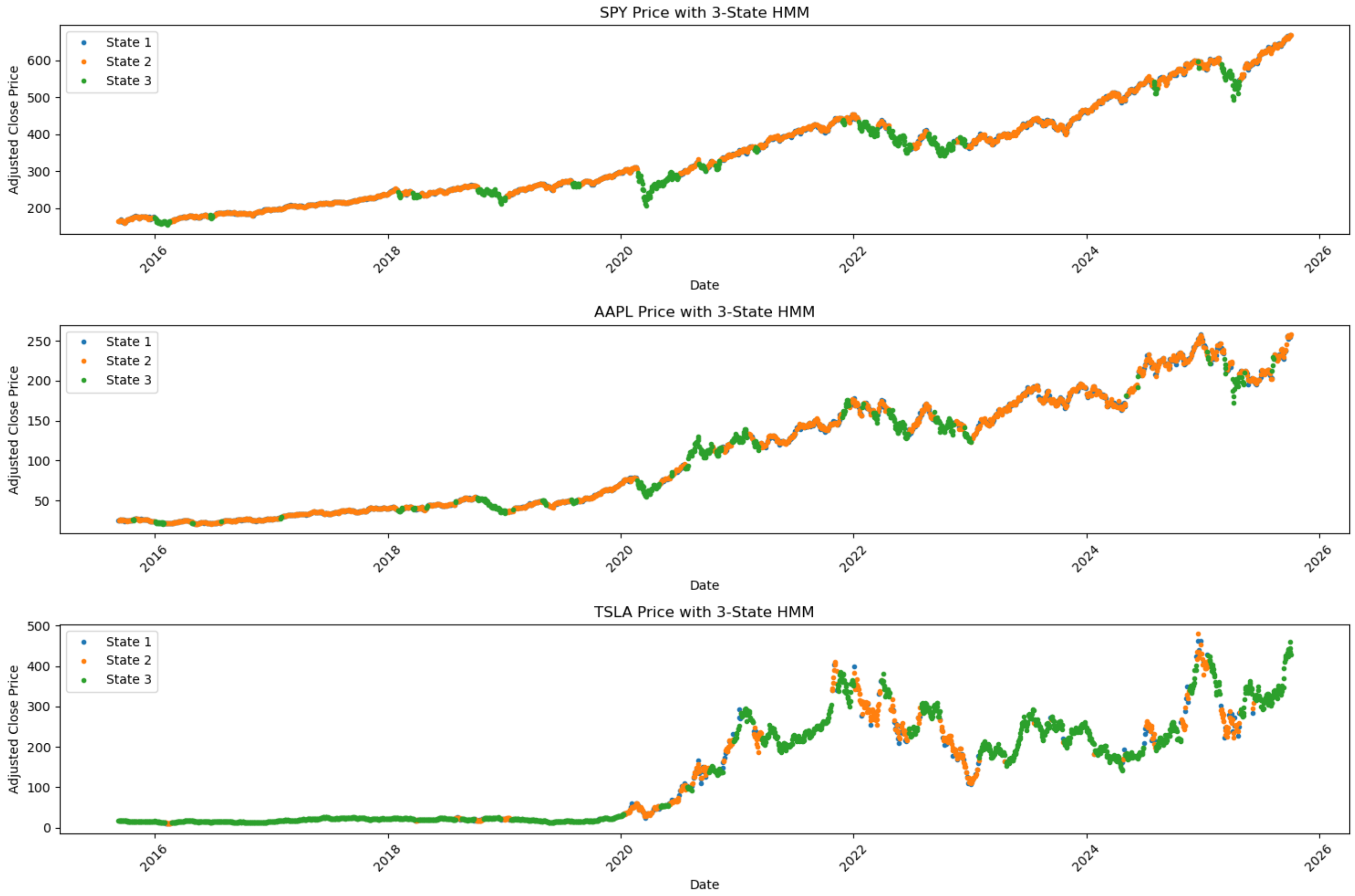

Fig. 3. Three-state HMM regime classification for SPY (top), AAPL (middle), and TSLA (bottom). Color coding indicates inferred hidden states, with different patterns across assets reflecting varied volatility profiles.

Markov-Switching GARCH: These models combine regime-switching with conditional heteroskedasticity, providing richer volatility dynamics but at increased estimation complexity.

VII. Conclusions and Future Work

The study shows how effective Gaussian Hidden Markov Models can be for identifying regimes in financial markets. By examining ten years of SPY data, we found two distinct regimes (bull and bear markets), each displaying considerably different return and volatility characteristics. Our strong evidence of regime persistence(98.62% and 95.77% self-transition probabilities) shows that market states have considerable momentum with implications for risk management and tactical allocation.

Extension to three-state models across multiple assets (SPY, AAPL, TSLA) reveals asset-specific regime structures while maintaining correlation during major market events. The methodology successfully captures known market episodes including the 2020 COVID-19 crash and subsequent recovery dynamics.

Key findings include:

- Bull market regime: mean return 0.001161, volatility 0.006958
- Bear market regime: mean return -0.001077, volatility 0.020133
- Asymmetric transition dynamics favoring longer bull markets
- Asset-specific regime patterns reflecting different volatility profiles

The potential future avenues for research are to include a non-Gaussian emission component to better capture heavy tails, incorporate time-varying transition probabilities to model nonstationarity, expand the research to multi-asset HMMs to better model the correlation structure, integrate macroeconomic covariates to enhance the regime prediction, and utilize out-of-sample regions to provide validation for the regime forecasts.

The complete implementation, including data collection scripts, preprocessing pipelines, model fitting procedures, and visualization tools, is available at: https://github.com/AnujSaha0111/CS307_Lab_Submissions/Submission_5

Acknowledgment

The authors thank Dr. Pratik Shah, faculty at IIIT Vadodara, for guidance on probabilistic graphical models and their applications to financial time series analysis.

References

[1] M. Scutari, “bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference,” https://www.bnlearn.com/, accessed January 2025.

[2] R. Blanco, I. Inza, and P. Larrañaga, “Learning Bayesian networks in R with bnlearn,” http://gauss.inf.um.es/umur/xjurponencias/talleres/J3.pdf, accessed January 2025.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2020.

[4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.