Chapter 2

Machine Learning-Driven Market Regime Analysis in Equity Markets: A Gaussian Hidden Markov Model Approach

Cemal Öztürk[1]

Abstract

This research develops a data-based system which reveals concealed market patterns through its identification of separate market regimes that produce unique return, volatility, and risk characteristics. The current financial models fail to recognize the intricate relationships which exist between different asset classes, including stocks and bonds, interest rates, commodities, and economic data indicators. The research employs a multivariate Hidden Markov Model (HMM) with improved data preprocessing techniques and principal component analysis (PCA) to process data from 2010 to 2025. The developed system detects nine separate market states which match actual economic and financial market situations. The market expansion phases produce strong investment returns at 20-30% annual rates while keeping market volatility at 12% but the contractionary phases lead to dangerous market conditions and negative investment results. The market transitions between different states occur at a slow pace because market conditions tend to stay stable instead of experiencing abrupt changes. The model shows a approximately 100% probability that the market operates under Regime 8 which produces stable returns with a 1.6 Sharpe ratio during November 2025 while showing limited market volatility. Overall, the results highlight the cyclical yet persistent nature of market behavior and provide practical tools for improving risk management, policy assessment, and data-informed investment decisions.

[1] Res. Asst. Dr., Iğdır University, Department of Economics, cemal.ozturk@igdir.edu.tr, ORCID ID: 0000-0003-3850-7416

https://doi.org/10.58830/ozgur.pub1120.c4538

1. Introduction

Financial markets experience recurring transitions between different market regimes, including bull and bear markets, correction periods, and high-volatility phases (Hamilton, 1989; Ang & Bekaert, 2002). These regimes cannot be directly observed from the available high-dimensional data, which integrates information from equities, fixed income, credit, commodities, FX, and macroeconomic indicators. Different market regimes exhibit distinct patterns in returns, volatility, correlations, and risk exposure, all of which influence asset allocation, risk management strategies, and macro-financial oversight. The main challenge for researchers and practitioners lies in identifying these latent market dynamics from complex, noisy data while maintaining both statistical robustness and economic interpretability.

The Hidden Markov Model (HMM) is a widely used framework for modeling unobserved market states through Markov chain transitions, allowing return distributions to depend on the current state (Hamilton, 1989; Krolzig, 1997). In its basic form, the HMM typically distinguishes between two volatility regimes or business-cycle phases using a small number of latent factors. However, modern financial markets require richer and higher-dimensional signal structures, as simple models often fail to capture regime shifts driven by multiple financial indicators, including cross-asset relationships, credit spreads, yield-curve movements, and macroeconomic variables. Expanding HMMs to include many correlated predictors often leads to overfitting and computational instability, reducing the reliability of state estimation (Bishop, 2006; Murphy, 2012).

This study develops a multivariate Hidden Markov Model (HMM) framework designed to identify latent market regimes from a comprehensive panel of financial and macroeconomic variables. The research design integrates three core components. First, it constructs an extensive feature set that combines financial indicators from equities, volatility, credit, rates, commodities, and FX markets with macroeconomic variables such as the yield curve slope, unemployment rate, and high-yield bond spreads. Second, it applies robust scaling and principal component analysis (PCA) to extract a low-dimensional set of orthogonal factors that capture the dominant sources of variation in the data while mitigating the effects of outliers and multicollinearity. Third, it estimates a Gaussian HMM on these factors, determines the optimal number of states using information criteria, and evaluates the identified regimes through a series of return-based performance and risk metrics.

The proposed framework bridges traditional regime-switching models with modern high-dimensional feature engineering, enabling the inference of market regimes from a broad cross-section of assets and macroeconomic indicators rather than from a single index. Methodologically, it demonstrates how robust preprocessing and dimensionality reduction can be integrated with HMM estimation to yield stable, economically interpretable classifications, even in the presence of noisy, highly correlated predictors. Empirically, the study documents clear differences among the detected regimes in terms of returns, volatility, drawdowns, tail risk, and regime persistence—findings that carry important implications for tactical asset allocation and risk management.

The remainder of the paper is structured as follows. Section 2 describes the dataset and methodological framework, including feature construction, robust scaling, PCA, and the HMM specification. Section 3 presents the empirical results, highlighting regime characteristics, performance metrics, and state transition dynamics. Finally, Section 4 concludes with a discussion of the main findings, their limitations, and potential directions for future research

2. Data and Methodology

2.1. Data Description

The empirical analysis is based on a daily panel that combines market-based asset returns with macro-financial indicators. The sample spans from 1 January 2010 to 20 November 2025. Daily price data are obtained from Yahoo Finance via the *yfinance* Python library (Aroussi, 2019), while macroeconomic and financial time series are retrieved from the Federal Reserve Bank of St. Louis via the *FRED* API (Mehyar, 2014). Lower-frequency FRED series are converted to daily frequency by forward-filling the latest available observation until the next data release.

Missing values are handled using forward- and backward-filling to ensure a continuous, balanced daily panel suitable for time-series modeling. The inclusion of diverse asset classes and macro-financial indicators enables the identification of multiple latent market regimes that reflect both financial and macroeconomic dimensions of market dynamics. Table 1 provides an overview of the financial and macroeconomic variables included in the regime-switching analysis, detailing their sources, frequencies, and roles within the model.

Table 1. Overview of variables used in the regime-switching analysis

Group	Asset / Series	Ticker / ID	Source	Frequency	Role in model
U.S. equity index	S&P 500 Index	SPX (^GSPC)	Yahoo Finance	Daily	Core equity market level
U.S. equity breadth	Russell 2000 Index	RUT	Yahoo Finance	Daily	Small-cap risk / market breadth
Volatility	CBOE Volatility Index	VIX	Yahoo Finance	Daily	Market-implied volatility
Credit – high yield	iShares iBoxx $ High Yield Corporate Bond ETF	HYG	Yahoo Finance	Daily	Credit risk / risk-on proxy
Credit – investment grade	iShares iBoxx $ Investment Grade Corporate Bond ETF	LQD	Yahoo Finance	Daily	Investment-grade credit conditions
Rates	iShares 20+ Year Treasury Bond ETF	TLT	Yahoo Finance	Daily	Long-term risk-free benchmark
Sector equity	Financials, Energy, Technology, Health Care, Utilities, Consumer Staples, Consumer Discretionary, Industrials, Materials	XLF, XLE, XLK, XLV, XLU, XLP, XLY, XLI, XLB	Yahoo Finance	Daily	Sector rotation and cross-section
Commodities	Gold, Silver, Crude Oil, Agriculture	GLD, SLV, USO, DBA	Yahoo Finance	Daily	Real asset and inflation hedges
FX	U.S. Dollar Bullish ETF; Euro FX Trust	UUP, FXE	Yahoo Finance	Daily	Dollar strength / EUR–USD proxy
Int’l equity	Developed ex-US; Emerging Markets	EFA, EEM	Yahoo Finance	Daily	Global risk-on / risk-off conditions
Labour market	Unemployment Rate	UNRATE	FRED	Monthly → Daily (FF)	Business-cycle slack
Monetary policy	Effective Federal Funds Rate	FEDFUNDS	FRED	Daily (via FF)	Short-term policy stance
Yield curve slope	10-year minus 2-year Treasury yield	T10Y2Y	FRED	Daily (via FF)	Term-structure / recession indicator

Credit spread	ICE BofA US High Yield Index Option-Adjusted Spread	BAMLH0A0HYM2	FRED	Daily (via FF)	Systemic credit risk
Oil price	WTI Crude Oil Price	DCOILWTICO	FRED	Daily (via FF)	Commodity & inflation expectations
FX (macro)	USD/EUR Exchange Rate	DEXUSEU	FRED	Daily (via FF)	External value of USD
Volatility (macro)	VIX Index (FRED)	VIXCLS	FRED	Daily (via FF)	Macro-level VIX measure

Notes: Yahoo Finance series are obtained through yfinance, and FRED series through fredapi. Lower-frequency macro series are merged into the daily panel by forward-filling the latest available observation.

2.2. Methodology

The empirical strategy consists of three main steps: (i) feature engineering from the raw price and macro data; (ii) robust scaling and dimensionality reduction via Principal Component Analysis (PCA); and (iii) estimation of a multivariate Gaussian Hidden Markov Model (HMM) to infer latent market regimes (Hamilton, 1989; Rabiner, 1989; Jolliffe, 2002).

All computations are conducted in Python using the scientific stack: NumPy for numerical arrays (Harris et al., 2020), pandas for data handling (McKinney, 2010), SciPy for optimization routines (Virtanen et al., 2020), and scikit-learn for scaling and PCA (Pedregosa et al., 2011). All features are scaled using the RobustScaler class from scikit-learn, which standardizes variables by their median and interquartile range (IQR), reducing the impact of outliers. Regime models are estimated using the hmmlearn package (hmmlearn developers, 2015), which provides a Gaussian HMM implementation consistent with the scikit-learn API. Visualizations are produced with matplotlib (Hunter, 2007) and seaborn (Waskom, 2021).

2.2.1. Feature engineering

Let p_t denote the vector of prices at time t for all traded assets listed in Table 1. The corresponding vector of daily returns is

$$\mathrm{r}_t = \left(r_{1,t},\ldots,r_{N,t}\right)^{\top}, r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}.$$

where $P_{i,t}$ denotes the closing price of the asset i on day t. Volatility and higher-order moments are constructed from rolling windows of returns—including 21-, 63-, and 126-day realized volatility, drawdowns, and momentum—providing the HMM with a richer description of the prevailing risk–return environment.

The realized volatility is computed as:

$$\sigma_{i,t}^{(h)} = \sqrt{252 \frac{1}{h} \sum_{j=0}^{h-1} r_{i,t-j}^{2}},$$

where the scaling factor 252 annualizes the measure under the standard trading-day convention.

From these returns, the feature-engineering step constructs several classes of predictors:

- Level and volatility features: rolling realized volatility, Parkinson range-based volatility, Bollinger band positioning, maximum drawdowns, and drawdown durations, capturing stress and recovery cycles across each market segment.
- Momentum and trend indicators: short- and medium-horizon cumulative returns (e.g., 1-, 3-, 6-month) for equity indices and sector ETFs as cross-sectional "risk-on/risk-off" signals, and trend-following oscillators such as RSI, MACD, Stochastic %K/%D, ADX, and Money Flow Index (MFI) to capture technical sentiment.
- Volume and breadth indicators: On-Balance Volume (OBV), sector breadth ratios, and volume-trend indicators reflecting participation strength across sectors.
- Macro-financial indicators: yield-curve slope (T10Y2Y), high-yield spread (BAMLH0A0HYM2), unemployment rate (UNRATE), Federal Funds Rate (FEDFUNDS), oil price (DCOILWTICO), volatility index (VIXCLS), and USD/EUR exchange rate (DEXUSEU)—slow-moving variables that inform the underlying business-cycle phase.
- Cross-market linkages: co-movement among equities (SPX, RUT, EFA, EEM), volatility (VIX, VIXCLS), credit (HYG, LQD), rates (TLT), commodities (GLD, SLV, USO, DBA), and FX (UUP, FXE, DEXUSEU). These linkages are incorporated through the HMM's

full-covariance structure, ensuring that regime identification reflects multivariate dependence across asset classes.

The engineered features span six categories: core market, sectoral, international, commodity/FX, macroeconomic, and technical indicators. This step produces a high-dimensional feature vector $x_t \in \mathbb{R}^p$ for each day t, encompassing returns, volatilities, drawdowns, macro variables, and technical signals.

After feature engineering and data cleaning, 224 model-ready features were retained, spanning equity, credit, rates, commodities, FX, and macroeconomic domains. Among these, 125 engineered predictors are reported in Table 2, grouped into core market, sectoral, international, commodity/FX, macroeconomic, and technical categories, while the remaining variables consist of underlying FRED series, seasonal dummies, and auxiliary transformations that are not individually listed. The resulting dataset comprises 3,994 valid daily observations from 2010-01-04 to 2025-11-20, totaling approximately 16 years of trading history.

Table 2. Summary of engineered feature categories and dataset characteristics

Feature Category	Count
Core features	38
Sector features	18
Commodity & FX features	16
International features	7
Advanced technical features	17
Time-series features	10
FRED macro features	19
Total features	**125**

Note: Table 2 reports only the 125 engineered features. However, the full modeling dataset used for PCA and HMM includes 224 features, combining engineered predictors with raw market series, FRED macroeconomic variables, and seasonal transformations.

The feature correlation matrix (Figure 1) shows how the engineered variables move together across markets and macroeconomic dimensions. Each cell, color-coded from -1 (red) to +1 (green), represents the strength of the linear relationship between two features.

Clusters along the diagonal reveal groups of variables that tend to move in sync—for example, daily price metrics (close, high, low, and open) within sector ETFs. In the macro layer, indicators such as UNRATE (the

U.S. unemployment rate), FEDFUNDS (the effective federal funds rate that reflects the Federal Reserve's policy stance), and BAMLH0A0HYM2 (the high-yield credit spread, measuring the extra yield investors demand for risky corporate bonds) are nearly perfectly correlated with their derived counterparts. This coherence confirms that the macroeconomic variables were successfully aligned and forward-filled.

Altogether, 1,366 feature pairs show correlations above $|0.8|$—a reminder that financial data is naturally multicollinear. To handle this redundancy, the next step applies Principal Component Analysis (PCA) to compress the information into fewer uncorrelated latent factors before estimating the Hidden Markov Model.

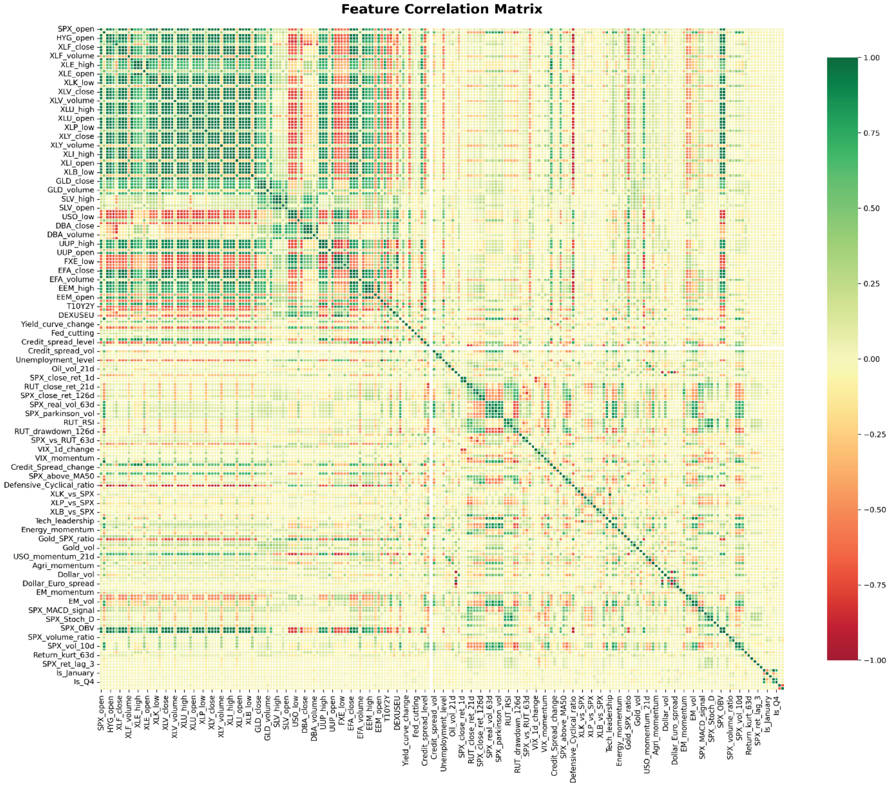


Figure 1. Feature Correlation Matrix

2.2.2. Robust scaling and principal component analysis

To mitigate the influence of outliers and heavy-tailed return distributions, all continuous features are first transformed using a robust scaler, which subtracts the median and rescales by the interquartile range (IQR):

$$\tilde{x}_{j,t} = \frac{x_{j,t} - \text{median}\left(x_j\right)}{\text{IQR}\left(x_j\right)},$$

where $x_{j,t}$ is the j-th raw feature at time t, and the median and IQR are computed over the training sample. This transformation reduces the impact of extreme observations while preserving the relative ordering of data points.

Given the large number of correlated predictors, the next step is to apply PCA to the robust-scaled features. PCA finds an orthogonal linear transformation $\mathrm{z}_t = \mathrm{W}^\top \tilde{\mathrm{x}}_t$ such that the first few components capture most of the variance in $\tilde{\mathrm{x}}_t$ (Jolliffe, 2002). Specifically, PCA solves the eigenvalue problem

$$\Sigma = \text{Var}\left(\tilde{\mathrm{x}}_t\right) = \mathrm{Q}\Lambda\mathrm{Q}^\top,$$

where Σ is the sample covariance matrix, Q is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The k -dimensional principal-component representation is then

$$\mathrm{z}_t = \mathrm{Q}_k^\top \tilde{\mathrm{x}}_t,$$

with Q_k containing the first k eigenvectors. The number of components k is chosen such that a pre-specified fraction of total variance (95%) is retained, trading off parsimony against information loss.

In practice, PCA is implemented using scikit-learn (Pedregosa et al., 2011), applied to the in-sample observations only. Out-of-sample data are projected using the loadings estimated from the training period to avoid look-ahead bias.

The Principal Component Analysis (PCA) process for simplifying the extensive and strongly related feature set appears in Figure 2. The scree plot on the left shows which components explain the most variation in the data. The data signal becomes most prominent in the initial components but subsequent components add less and less information until the pattern reaches a steep decline followed by an extended flat section. The right section of the plot shows how information quantity grows when researchers add more elements to their research. The curve starts with a rapid ascent before reaching a plateau after the addition of 54 components, which enables the capture of more than 95% of total variance (as shown by the dashed lines). The dataset contains 224 features, but only needs 54 principal components to represent its most important variations.

The analysis benefits from dimensionality reduction because it allows researchers to study the most important market and macroeconomic factors while eliminating unneeded data points. The uncorrelated components from this process produce stable data which enables better Hidden Markov Model (HMM) estimation for improved model understanding and operational performance.

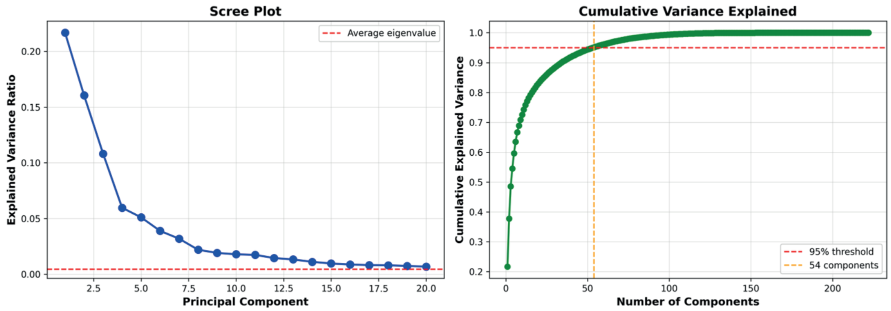


Figure 2. Principal Component Analysis (PCA): Scree Plot (left) and Cumulative Variance Explained (right)

Figure 3 summarizes which features most strongly shape the dataset's underlying structure after applying Principal Component Analysis (PCA). The left panel lists the top 20 most influential variables, showing that short-term oil volatility (Oil_vol_21d) is by far the most significant driver of common variation across assets. It is followed by Fed_rate_change_63d, capturing shifts in monetary policy, and SPX_MACD, a measure of equity market momentum. Several volatility-based indicators—such as Financial_vol, Tech_vol, and RUT_real_vol_21d—also rank highly, emphasizing how fluctuations in risk and liquidity conditions dominate the market's latent structure.

The right panel visualizes these relationships by plotting feature loadings on the first two principal components. Most features cluster near the origin, suggesting limited standalone influence, while a few—especially Oil_vol_21d and Financial_vol—anchor the axes, representing broad volatility and policy-driven factors.

Overall, the PCA results show that market-wide volatility and monetary dynamics are the primary forces driving asset co-movement, providing a compact and interpretable basis for the subsequent regime-switching analysis.

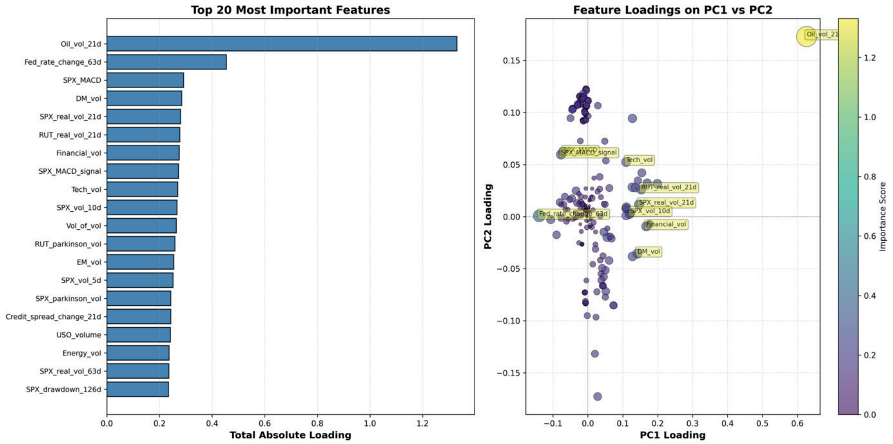


Figure 3. PCA feature importance (left) and feature loadings on PC1 vs PC2 (right)

2.2.3. Hidden Markov Model specification

Let $\{z_t\}_{t=1}^{T}$ denote the sequence of dimension-reduced feature vectors obtained from PCA. The latent market regime at time t is represented by a discrete-valued random variable $S_t \in \{1, \ldots, K\}$. The HMM assumes that:

The latent state process $\{S_t\}$ follows a first-order Markov chain with transition probabilities

$$a_{ij} = \Pr\left(S_t = j \mid S_{t-1} = i\right), i, j = 1, \ldots, K,$$

collected in the $K \times K$ transition matrix A, and the initial state distribution π, where $\pi_i = \Pr\left(S_1 = i\right)$.

Conditional on the current state, the observation z_t follows a multivariate Gaussian distribution with state-specific mean and covariance,

$$z_t \mid \left(S_t = k\right) \sim N\left(\mu_k, \Sigma_k\right), k = 1, \ldots, K.$$

The joint likelihood of observations and states is then

$$\Pr\left(\mathbf{z}_{1:T}, S_{1:T}\right) = \pi_{S_1} \prod_{t=2}^{T} a_{S_{t-1}S_t} \prod_{t=1}^{T} N\left(\mathbf{z}_t \mid \boldsymbol{\mu}_{S_t}, \boldsymbol{\Sigma}_{S_t}\right),$$

Where $z_{1:T} = \left(z_1, \ldots, z_T\right)$ and $S_{1:T} = \left(S_1, \ldots, S_T\right)$.

Estimation proceeds via maximum likelihood using the Expectation–Maximization (EM) algorithm, implemented through the Baum–Welch procedure (Rabiner, 1989; Hamilton, 1989). The forward–backward algorithm yields smoothed state probabilities $\Pr\left(S_t = k \mid z_{1:T}\right)$, and the Viterbi algorithm is used to obtain the most likely state path $\hat{S}_{1:T}$.

The number of regimes K is treated as a model-selection problem. Candidate models $K \in \{3, \ldots, 9\}$ are estimated, and the preferred specification is chosen based on the Bayesian Information Criterion (BIC),

$$\text{BIC} = -2\log\hat{L} + p\log T,$$

and Akaike Information Criterion (AIC),

$$\text{AIC} = -2\log\hat{L} + 2p,$$

where $\hat{L}$ is the maximised likelihood, p is the number of free parameters, and T is the number of time points. Preference is given to models with lower BIC and AIC; in practice, BIC is used as the primary criterion to penalise over-parameterisation more strongly in high-dimensional settings.

2.2.4. Out-of-sample validation

To assess the stability and predictive usefulness of the inferred regimes, the sample is split chronologically into an 80/20 train–test partition. Let $T_{\text{train}} = 0.8T$. The HMM is estimated using observations $\{z_t\}_{t=1}^{T_{\text{train}}}$ only, yielding parameter estimates $\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \{\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k\}_{k=1}^{K}$. These parameters are then held fixed, and regime probabilities for the remaining out-of-sample period $t = T_{\text{train}} + 1, \ldots, T$ are computed via the forward recursion:

$$\alpha_t(j) = \left[\sum_{i=1}^{K} \alpha_{t-1}(i)\,\hat{a}_{ij}\right] \mathrm{N}\left(\mathbf{z}_t \mid \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\right), j = 1, \ldots, K,$$

with normalisation to obtain filtered probabilities

$$\Pr\left(S_t = j \mid z_{1:t}\right) = \frac{\alpha_t(j)}{\sum_{k=1}^{K} \alpha_t(k)}.$$

This procedure allows the paper to (i) examine whether regimes identified in-sample persist out-of-sample, and (ii) relate regime probabilities to subsequent asset-class performance, volatility, and drawdowns without re-estimating the model each time new data arrive.

2.2.5. Software implementation

All computations are implemented in Python 3.12. Data collection employs yfinance (Aroussi, 2019) for market series and fredapi (Mehyar, 2014) for macroeconomic indicators. Numerical operations rely on NumPy and SciPy, data manipulation on pandas, preprocessing and PCA on scikit-learn, and regime modeling on hmmlearn (hmmlearn developers, 2015). Visualization and diagnostics are performed with matplotlib and seaborn.

2.2.6. Model Robustness

To enhance robustness, each HMM configuration is fitted with 10 random initializations, and both BIC and AIC are recorded. The full-covariance emission assumption enables the model to capture multi-asset dependencies, while the multiple initialization strategy reduces the likelihood of converging to suboptimal local maxima. This ensures stable and interpretable regime detection across estimation runs.

3. Results

The Gaussian HMM was estimated using the 54 principal components from the preceding PCA step, which together capture 95% of the variance in the original 224 model-ready features. This reduced representation preserves the essential structure of market and macroeconomic dynamics while ensuring numerical stability in model estimation.

The results of the model selection process are reported in Figure 4, which compares the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) across models with varying numbers of latent states. As the number of states increases from 3 to 9, both criteria decline monotonically, indicating improved model fit. However, because the BIC imposes a more substantial penalty for model complexity, its minimum value provides a more conservative and statistically rigorous benchmark for determining the optimal number of regimes.

The right-hand panel of Figure 4 depicts the corresponding log-likelihood trajectory, which increases steadily with the addition of states, reflecting a progressive increase in explanatory power. The convergence of both information criteria at their minima for nine states suggests that this specification offers the best balance between parsimony and flexibility.

This nine-state configuration provides sufficient granularity to capture the heterogeneity of financial market behavior over time. It allows for the identification of distinct latent regimes corresponding to periods of low-volatility expansion, elevated market stress, and intermediate transition phases. The resulting structure thus forms the empirical foundation for the subsequent regime characterization and transition analysis presented in the following sections.

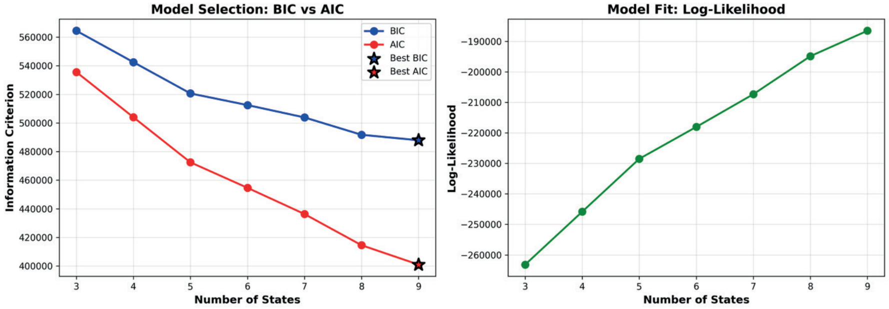


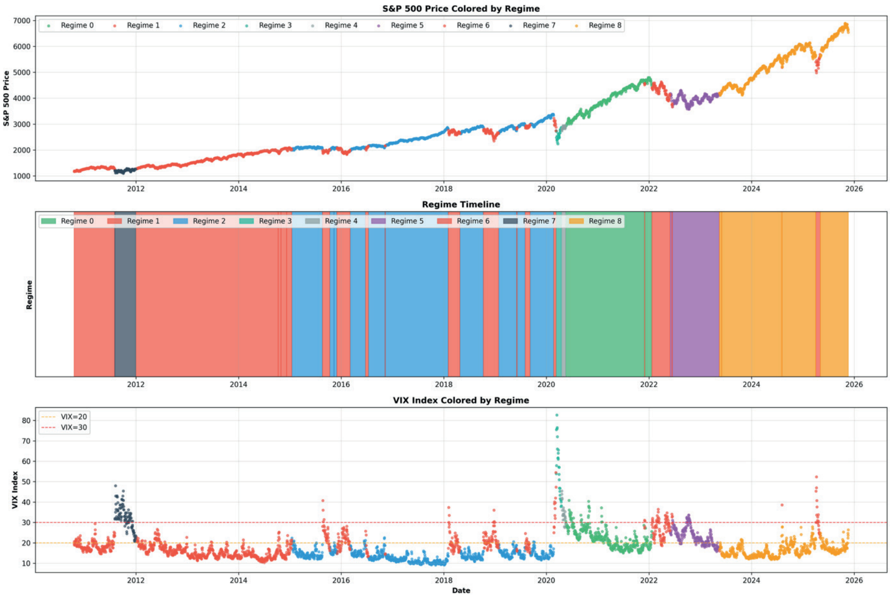
Figure 4. Hidden Markov Model selection results: information criteria comparison (left) and log-likelihood evolution (right)

Figure 5 presents the temporal structure of the nine latent regimes identified by the Hidden Markov Model (HMM) and their relationship with major market dynamics over the 2010–2025 period. The visualization combines three complementary panels that together provide an intuitive interpretation of how the model's inferred regimes align with observable shifts in market behavior and volatility conditions.

The top panel plots the S&P 500 Index level, colored by regime classification. Distinct color segments correspond to periods when the market exhibited consistent statistical characteristics captured by the HMM. For example, extended green and orange intervals reflect prolonged low-volatility expansions. At the same time, intermittent red and gray segments correspond to turbulence and risk-off phases, such as those surrounding the 2020 COVID-19 crash and subsequent normalization.

The middle panel displays the regime timeline, summarizing transitions between latent states over time. This view highlights the persistence and recurrence of specific regimes. Some states, such as Regime 0 and Regime 8, dominate extended portions of the sample, indicating periods of stability, whereas others appear as short-lived episodes, capturing temporary disruptions or transitional phases between bull and bear environments.

The bottom panel plots the VIX Index, a widely used measure of implied market volatility, also colored by regime. Horizontal dashed lines at VIX = 20 and VIX = 30-mark conventional thresholds between normal and elevated volatility conditions. The alignment of high VIX clusters with red and gray regimes confirms that the HMM successfully isolates periods of market stress and uncertainty. Conversely, calm, low-VIX intervals correspond to expansionary, risk-on phases.


S&P 500 Price Colored by Regime

Regime Timeline

VIX Index Colored by Regime

Figure 5. Regime Timeline and Market Context: S&P 500 price (top), regime timeline (middle), and VIX dynamics (bottom)

Figure 6 shows the likelihood that the market will remain in or switch between the nine identified regimes. Each cell represents the probability of moving from one regime (on the y-axis) to another (on the x-axis). The diagonal dominance is striking: most regimes have over a 95% chance of persisting day to day, and several exceed 99%, meaning that once a market condition is established, it tends to last. For instance, Regime 1 and Regime 0 are the most stable, typically lasting around 8–9 months, while Regimes 3, 4, and 6 are much shorter-lived, often representing brief but intense volatility or transition periods. Transitions between very different regimes are rare — the market doesn't jump chaotically from one environment to another. Instead, shifts happen gradually, often through intermediate states.

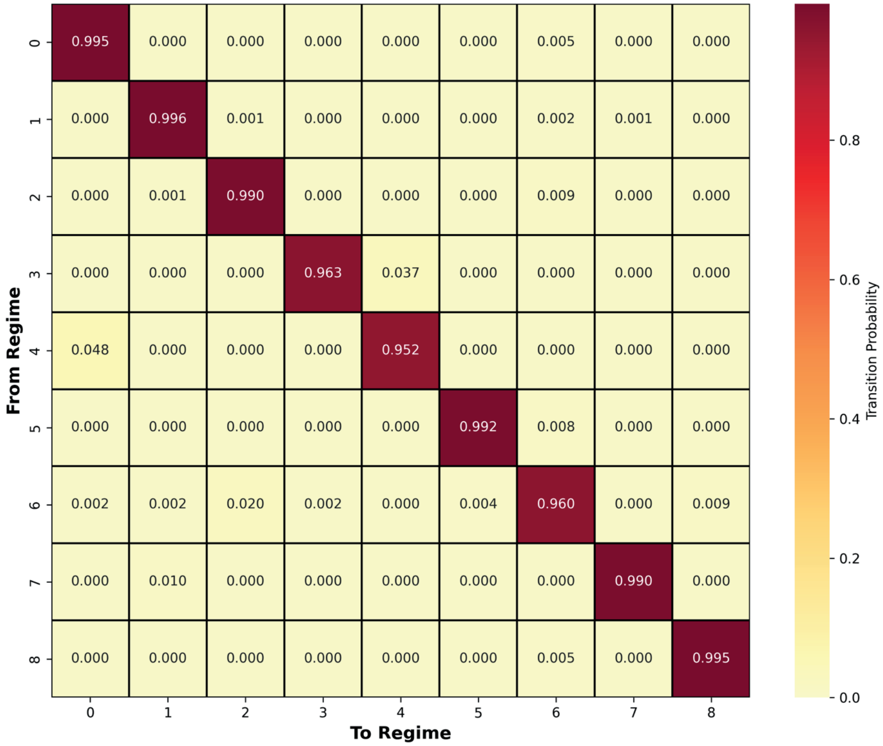


Figure 6. Transition probabilities across the Hidden Markov Model regimes

Figure 7 compares how each of the nine market regimes behaves across several key dimensions — returns, volatility, risk-adjusted performance, drawdowns, win rates, and average volatility levels (VIX). The first panel shows that only a few regimes delivered strong positive returns, reflecting periods of steady market growth. Others, such as Regime 6, represent clear downturn phases with sustained negative performance.

Volatility (top middle) varies sharply between regimes: some display the calm, low-volatility environment typical of bull markets, while others show extreme spikes — around 80% annualized volatility — consistent with crisis or panic conditions. The Sharpe ratio panel highlights the contrast between high-efficiency regimes (in green) that deliver attractive returns per unit of risk and low-efficiency regimes (in red) where risk dominates reward.

In the lower panels, average drawdowns deepen significantly during volatile periods, while win rates generally hover around 50–55%, suggesting alternating cycles of optimism and caution. The last panel connects each regime to its typical VIX level, showing that stress regimes coincide with volatility above 30, while stable regimes stay below the 20 thresholds. Overall, Figure 7 shows that each regime captures a distinct and realistic market environment—from tranquil expansion phases to short-lived risk-off episodes and full-blown market corrections—mirroring how financial cycles unfold in practice.

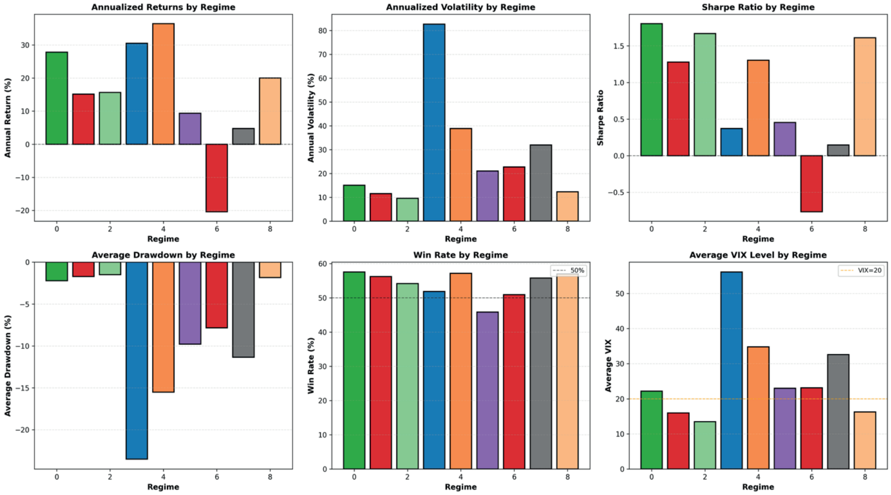


Figure 7. Comparative summary of performance, risk, and volatility dynamics across market regimes

The model presents its current market evaluation through Figure 8. The Hidden Markov Model shows approximately 100% probability that the market operates under Regime 8. The market provides 20% annualized returns with 12% volatility while maintaining a 1.6 Sharpe ratio which confirms its risk-efficient market status. The market shows a 57%-win rate which confirms its stable condition while maintaining an upward direction. The market has experienced a significant decline of -28.2% annualized during the last 21 trading days yet volatility and VIX levels stay near their historical norms. The bottom section of the graph shows how current market data compares to the typical values of this market regime. The present market trend shows that the ongoing market decrease will be short-term and does not indicate any shift in the broader market environment. The market data in Figure 8 shows both low market volatility and a steady bullish market direction. The market experiences short periods of market weakness before stabilizing until market volatility reaches extreme levels or outside factors trigger market disruptions.

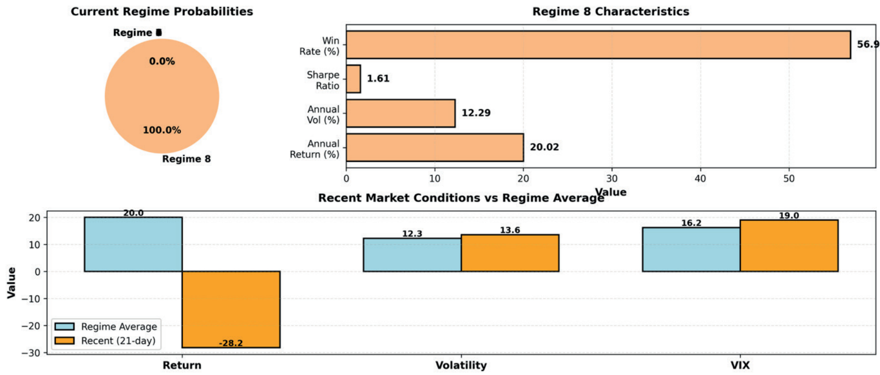


Figure 8. Current market regime and comparison with recent 21-day market conditions

4. Conclusion

This paper develops a comprehensive, data-driven framework for detecting latent market regimes using a multivariate Gaussian Hidden Markov Model (HMM) applied to a high-dimensional panel of financial and macroeconomic variables. By integrating robust preprocessing, principal component analysis, and full-covariance regime modeling, the study successfully identifies nine distinct market states that correspond to empirically meaningful phases of market behavior.

The empirical results demonstrate that the proposed HMM framework captures the essential structure of market dynamics over the 2010–2025 period. The model's nine-state configuration achieves the optimal balance between model fit and parsimony, as evidenced by the minimum Bayesian Information Criterion (BIC). The estimated regimes exhibit clear economic interpretation: several states correspond to stable, low-volatility expansionary phases (e.g., Regime 0 and Regime 8), while others capture transitional or crisis environments characterized by heightened volatility and negative returns (e.g., Regime 6). The strong diagonal dominance in the transition probability matrix indicates that market conditions are highly persistent, with regimes typically lasting several months before transitions occur gradually rather than abruptly.

From a performance standpoint, the Sharpe ratio, volatility, and drawdown profiles across regimes confirm that risk-adjusted efficiency varies substantially between states. Expansionary regimes deliver annualized returns exceeding 20% with volatility near 12%, while downturn regimes are marked by sharp drawdowns and poor risk–reward trade-offs. The current market assessment (Figure 8) indicates that the model assigns a

100% probability to Regime 8 — a stable, low-volatility phase characterized by moderate returns and a 1.6 Sharpe ratio. Despite short-term corrections, volatility and VIX levels remain close to historical averages, implying that recent market weakness is cyclical rather than structural.

These findings highlight the model's ability to differentiate between persistent macro-financial conditions and transient market noise, providing a valuable decision-support tool for asset allocation, tactical positioning, and systemic risk monitoring. By fusing high-dimensional data with interpretable probabilistic structure, the framework bridges the gap between econometric regime-switching models and modern machine learning techniques.

However, several limitations should be acknowledged. First, the Gaussian emission assumption may not fully capture the heavy tails and skewness inherent in financial data. Future work could extend the model to Student-t or mixture-of-Gaussian distributions to improve robustness under extreme conditions. Second, while the PCA step effectively mitigates multicollinearity, it reduces interpretability at the feature level; future research could incorporate sparse or supervised dimensionality reduction techniques to retain more economic meaning. Third, the analysis is confined to U.S. and global developed markets—expanding the dataset to include cross-country or sectoral perspectives would allow for a richer examination of global contagion and spillover effects.

Future research could also explore dynamic model averaging or time-varying transition probabilities to account for structural breaks and evolving policy regimes. Incorporating macroeconomic forecasting components or real-time market sentiment data may further enhance predictive accuracy and responsiveness.

Overall, this study demonstrates that a well-calibrated multivariate Hidden Markov Model provides a powerful and interpretable framework for identifying and characterizing market regimes in complex, high-dimensional environments. The results underscore the persistence and cyclical nature of financial regimes and highlight the potential of probabilistic state-space models as a foundation for data-driven risk management and strategic investment decisions.

References

Ang, A., & Bekaert, G. (2002). *International asset allocation with regime shifts*. *Review of Financial Studies, 15*(4), 1137–1187. https://doi.org/10.1093/rfs/15.4.1137

Aroussi, R. (2019). *yfinance* [Computer software]. *PyPI*. https://pypi.org/project/yfinance/

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Hamilton, J. D. (1989). *A new approach to the economic analysis of nonstationary time series and the business cycle*. *Econometrica, 57*(2), 357–384. https://doi.org/10.2307/1912559

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). *Array programming with NumPy*. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

hmmlearn developers. (2015). *hmmlearn: Hidden Markov models in Python* [Computer software]. *Read the Docs*. https://hmmlearn.readthedocs.io/

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98835

Krolzig, H. M. (1997). *The Markov-switching vector autoregressive model*. In *Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis* (Vol. 454, pp. 9–46). *Lecture Notes in Economics and Mathematical Systems*. Springer. https://doi.org/10.1007/978-3-642-51684-9_2

McKinney, W. (2010). *Data structures for statistical computing in Python*. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

Mehyar, M. (2014). *fredapi: Python API for Federal Reserve Economic Data* [Computer software]. *GitHub*. https://github.com/mortada/fredapi

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rabiner, L. R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*. *Proceedings of the IEEE, 77*(2), 257–286. https://doi.org/10.1109/5.18626

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). *SciPy 1.0: Fundamental* *algorithms for scientific computing in Python*. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). *seaborn: Statistical data visualization*. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021