Lisbon School of Economics and Management

Universidade de Lisboa

Master's in Mathematical Finance

Master's Final Work

Mechanism for Identifying Market Regimes Based on a Gaussian Mixture Model

Author: Igor Shestopalov

Supervisor: Prof. Carlos J. Costa

Jury

Chair: JOÃO PAULO VICENTE JANELA, Associate Professor

Members: FLÁVIO ALEXANDRE COSTA ROMÃO, Visiting Assistant Professor

Supervisor: CARLOS MANUEL JORGE DA COSTA, Senior Associate Professor

July, 2025

Glossary

AI	Artificial Intelligence
GMM	Gaussian Mixture Model
ML	Machine Learning
MFW	Master's Final Work
SPY	S&P 500 ETF

Abstract

This project develops and evaluates a data-driven mechanism for identifying market regimes in U.S. equity markets using Gaussian Mixture Models (GMMs). The motivation stems from the growing need for interpretable, unsupervised methods that can capture the complex, nonstationary, and regime-dependent behavior of financial returns. Traditional econometric models—such as Markov-Switching frameworks—often rely on strong parametric assumptions about transition probabilities and are sensitive to structural breaks. In contrast, GMMs offer a probabilistic, flexible, and transparent approach that allows for overlapping regimes and non-linear relationships between risk and return.

The study applies the model to daily returns and 30-day realized volatility of the SPDR SP 500 ETF (SPY) from 2000 to 2024, standardizing features to ensure numerical stability and comparability. Model selection is guided by the Bayesian Information Criterion (BIC), which consistently favors a three-regime specification corresponding to empirically recognizable "bull," "neutral," and "bear" market states. Each regime is statistically characterized by its mean, volatility, skewness, and kurtosis, and economically linked to historical market events such as the 2008 financial crisis and the 2020 Covid-19 shock.

The probabilistic ("soft") classification provided by the GMM enables continuous regime probabilities rather than abrupt, binary transitions, reflecting the inherently gradual nature of market shifts. These probabilities can be directly incorporated into portfolio-management and risk-control frameworks, supporting adaptive asset-allocation, volatility-targeting, and hedging strategies. The research demonstrates that regime-aware approaches based on GMM signals can reduce drawdowns and enhance risk-adjusted performance relative to static allocation benchmarks.

Beyond its empirical findings, this project contributes a transparent and reproducible analytical framework built entirely in Python, following data-science best practices. The

open-source implementation facilitates extension to other markets or features (e.g., option-implied volatility, macro indicators), bridging academic research and professional application. Overall, the work highlights how interpretable machine-learning methods like GMMs can strengthen the link between statistical modeling and practical decision-making in modern quantitative finance.

Keywords: Market regimes; Gaussian Mixture Model; Unsupervised learning; Financial clustering; Risk management; SPY returns.
JEL Codes: C38; G11; G12; G17; G32.

Contents

List of Figures

Chapter 1

Introduction

Over the past two decades, artificial intelligence (AI) and machine learning (ML) have profoundly reshaped the landscape of financial markets. According to recent industry surveys, global investment in AI-driven financial technologies exceeded $40 billion in 2023, with leading hedge funds and asset managers allocating significant resources to data science, quantitative research, and model-driven trading. As of 2024, it is estimated that more than 65% of all equity trading volume in the U.S. is executed by algorithms, with a growing share attributed to adaptive strategies that leverage unsupervised and reinforcement learning.

In parallel, the availability of high-frequency, high-dimensional financial data has exploded: the NYSE alone processes more than one billion data points daily, ranging from trade-level transactions to market sentiment indicators. This data abundance has exposed the limitations of traditional econometric models, which typically rely on restrictive assumptions and are often ill-equipped to capture the nonlinear, nonstationary, and regime-dependent behavior observed in real markets.

Industry adoption of AI is not limited to high-frequency trading. According to the 2023 CFA Institute Investment Professional Survey, over 70% of institutional investors reported active use of machine learning models for risk management, portfolio construction, or alpha generation. Leading banks and asset managers now maintain dedicated AI research teams, with systematic funds such as Renaissance Technologies and Two Sigma reportedly achieving annualized Sharpe ratios above 2.0 through advanced modeling and regime-adaptive allocation.

Despite this surge in AI adoption, a persistent challenge remains: interpretable, robust

detection of structural shifts—or "regimes"—in financial time series. Black-box deep learning models often lack transparency and are prone to overfitting, while classic regime-switching models can be brittle and hard to calibrate in high-noise environments.

Recent academic research has increasingly turned to probabilistic clustering approaches—particularly Gaussian Mixture Models (GMM)—as a tractable, interpretable, and statistically grounded method for unsupervised regime identification. GMMs offer a flexible alternative to hard classification and can naturally accommodate the ambiguity and transitional behavior characteristic of real-world financial regimes. Their "soft assignment" property allows for probabilistic classification, which is especially useful in financial contexts where regime transitions are rarely binary. Moreover, GMMs can model distinct covariance structures for each regime, accommodating varying volatility and correlations—key stylized facts of asset returns.

The combination of rapidly expanding data, increasing computational power, and widespread institutional interest in AI motivates the search for data-driven, transparent, and actionable frameworks for regime detection. This project situates itself at the intersection of these trends by applying GMMs to the unsupervised classification of market regimes, using daily return and volatility data from the U.S. equity market.

Modeling these hidden regimes presents a dual challenge: (i) identification of unobserved regime boundaries, and (ii) classification of the data-generating process at any point in time. Classical approaches, such as Markov-Switching models, model regime dynamics as a hidden Markov process, imposing a transition matrix that governs the probabilistic switching among states. While effective in some contexts, these models rely on strong parametric assumptions about transition probabilities and temporal dependence, which can be difficult to estimate robustly, especially in the presence of nonstationarity or regime-dependent heteroskedasticity.

Recent literature has explored alternative unsupervised learning frameworks, leveraging clustering algorithms to infer hidden structures directly from observable features. In particular, Gaussian Mixture Models (GMMs) offer a powerful, probabilistic approach for identifying hidden regimes without imposing Markovian constraints or requiring external labels. GMMs posit that observed data are generated from a mixture of multivariate Gaussian distributions, each representing a hidden regime. The Expectation-Maximization (EM) algorithm enables maximum-likelihood estimation of the model parameters—component

means, covariances, and mixing proportions—yielding a posterior probability (responsibility) for each data point's membership in each regime.

This "soft assignment" feature of GMM is particularly advantageous in financial contexts, where regime transitions are often ambiguous and gradual rather than strictly binary. Moreover, GMMs allow each regime to possess its own covariance structure, naturally accommodating varying volatility and correlations—key stylized facts of financial markets. Unlike k-means clustering, which assigns points to clusters via a hard partition and assumes spherical variance, GMMs adapt to ellipsoidal (correlated) clusters and probabilistic boundaries.

In this study, I develop and implement a GMM-based regime identification mechanism using two financial features: daily log returns and a 30-day rolling window of realized volatility, derived from SPY ETF data as a proxy for the S&P 500. All features are standardized (z-score normalized) to avoid dominance by scale differences and to ensure numerical stability during EM optimization. The number of components is selected via the Bayesian Information Criterion (BIC), enabling a data-driven determination of the number of regimes consistent with the observed distributional complexity.

By focusing on a minimalist, market-driven feature set and leveraging the probabilistic clustering capabilities of GMM, this research aims to:

- Quantitatively uncover hidden regimes in equity markets without recourse to labeled events or macroeconomic covariates;
- Statistically characterize each regime with respect to mean, volatility, skewness, and kurtosis, and relate them to well-known market phenomena (e.g., crisis periods, extended bull phases);
- Generate time series of posterior regime probabilities, enabling a probabilistic rather than deterministic approach to regime classification and transitions, which is essential for risk management and dynamic asset allocation.

The methodology, results, and their implications for both academic research and practical portfolio management are presented in detail in the subsequent chapters.

Chapter 2

Literature Review

The detection of market regimes—distinct periods during which the statistical properties of asset returns, volatility, and correlations undergo structural change—remains a cornerstone problem in both empirical finance and quantitative risk management (Ang & Timmermann, 2012; Hamilton, 1989). The increasing availability of high-frequency, high-dimensional market data has motivated a parallel evolution in the methodological toolkit: from classical time series models to cutting-edge (Hamilton, 1989; Krolzig, 1997), unsupervised machine learning and optimal transport approaches (Chen et al., 2022; Horvath et al., 2021; Luan & Hamp, 2023). This review systematically covers the development of regime detection methods, from early parametric models to the latest non-parametric and AI-driven clustering algorithms, highlighting their theoretical foundations, practical performance, and open research challenges.

The seminal contribution of Hamilton (Hamilton, 1989) established Markov-Switching Autoregressive (MS-AR) models as the foundation of regime-switching analysis in financial time series. These models posit that asset returns are governed by a hidden Markov process, which transitions between a discrete number of regimes, each with its own mean and variance. Subsequent extensions—such as Markov-Switching GARCH (MS-GARCH) (Ang & Bekaert, 2002), regime-dependent variance models (Guidolin & Timmermann, 2007), and multivariate Markov models (Krolzig, 1997)—enabled richer representations of volatility clustering, leverage effects, and cross-asset interactions.

Despite their theoretical appeal, Markov models have practical limitations:

- The number of regimes and parametric forms must be specified a priori, risking model misspecification.

- Estimation of transition matrices and regime parameters can be sensitive to sample size, structural breaks, and non-Gaussian behavior (Maheu & McCurdy, 2000).
- Assumptions of stationarity and Gaussianity are frequently violated in real market data, particularly during crisis periods (Ang & Timmermann, 2012).

To address these shortcomings, several studies have integrated higher moments (skewness, kurtosis) (Bali & Whitelaw, 2013) or market-based indicators, such as option-implied volatility, to enhance regime detection (Lai, 2022). For example, Christoffersen (Christoffersen et al., 2018) demonstrate that including forward-looking information from option prices enables timelier identification of market stress regimes compared to purely return-based models.

The inadequacy of Gaussian models in capturing extreme events—market crashes and bubbles—has led to alternative approaches grounded in statistical physics. The Log-Periodic Power Law Singularity (LPPLS) model, pioneered by Sornette et al. (Zhang, Sornette, et al., n.d.), interprets bubbles as faster-than-exponential price growth with log-periodic oscillations, providing early warning signals for regime shifts. Applications of LPPLS and its variants successfully predicted the 1929 and 1987 crashes (Sornette & Johansen, 1997), the 2007–2008 oil bubble (Sornette et al., 2009), and multiple historical crises in global markets (Yan et al., 2012; Zhang, Sornette, et al., n.d.). These models rely on detecting critical points in price dynamics, thus reframing regime detection as a problem of identifying "end-of-bubble" or change points.

Change point detection (CPD) itself constitutes a broad and mature area of statistics, with both parametric and non-parametric methods. Classical CPD relies on likelihood-ratio tests or Bayesian approaches (Killick et al., 2012), but these often assume stationarity or specific distributional forms. Recent reviews (Aminikhanghahi & Cook, 2017; Truong et al., 2020) emphasize that, in financial applications, non-parametric CPD methods—such as kernel-based tests and robust distance metrics—are more resilient to the nonstationarity and heavy tails typical of real-world data.

The advent of machine learning has brought unsupervised clustering algorithms—particularly Gaussian Mixture Models (GMMs)—to the fore. Unlike Markov models, GMMs do not require a temporal structure; instead, they posit that data (e.g., daily returns and rolling volatilities) are generated from a mixture of Gaussian distributions, each representing a hidden regime. The Expectation-Maximization (EM) algorithm yields

soft assignments, providing posterior probabilities of regime membership at each time point.

GMMs have found broad application in finance, offering:

- Flexibility, which means that each regime can have a unique mean and covariance, allows the model to capture heteroskedasticity and stylized facts such as volatility clustering.
- Probabilistic assignment, in other words, smooths transitions and avoids the rigid, binary switching of HMMs.
- Data-driven selection where the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) guide the optimal number of regimes, reducing reliance on subjective model specification (Chen et al., 2022).

Comparative studies show that GMM-based regime detection often rivals or surpasses HMMs, especially for data with non-Gaussian features or weak temporal dependence (Horvath et al., 2021). However, GMMs remain limited in capturing time-dependent regime persistence unless extended via Markov-modulated mixtures (Frühwirth-Schnatter, 2006).

A major breakthrough in recent literature is the move from clustering in Euclidean space to clustering distributions via optimal transport theory. Horvath, Issa, and Muguruza (Horvath et al., 2021) propose Wasserstein k-means (WK-means), where clusters are formed not from individual points but from empirical distributions of returns (e.g., rolling window segments). The distance between distributions is measured by the p-Wasserstein metric, sensitive to the full shape of return distributions—not just mean and variance.

Extensions include the use of sliced Wasserstein distances, which address computational challenges in high-dimensional settings by averaging over one-dimensional projections. This strategy enables the practical clustering of multivariate time series, such as those found in currency baskets or pairs trading (Luan & Hamp, 2023; McGreevy, Muguruza, et al., 2024). Another recent advance is the signature maximum mean discrepancy (sig-MMD), introduced by Horvath and Issa (Horvath & Issa, 2023). Building on rough path theory, sig-MMD leverages kernel methods to compare entire return paths rather than just their distributions. This methodology is highly effective for detecting regime

changes in multidimensional or path-dependent financial data, and has demonstrated robustness to both noise and small sample sizes.

WK-means and sig-MMD have been empirically benchmarked on both synthetic and real-world data, demonstrating:

- Superior performance to HMMs and classical k-means in correctly segmenting market regimes (Hamp & Luan, 2023).
- Robustness to non-Gaussianity, heavy tails, and correlation structure changes (Hamp & Luan, 2023).
- Scalability to high-frequency or multivariate data, a key requirement for modern financial applications (Hamp & Luan, 2023).

These advances have inspired a proliferation of related work. McGreevy (McGreevy, Muguruza, et al., 2024) combine WK-means and MMD for multivariate regime detection, showing their utility in practical portfolio construction and risk assessment.

A parallel thread in the literature integrates forward-looking option-implied measures into regime models. Lai (Lai, 2022) and Christoffersen (Christoffersen et al., 2018) show that option-implied volatility and risk premia, especially the horizon spread (long minus short maturity), provide early signals for regime transitions. These models not only outperform historical-return-based methods in out-of-sample regime prediction but also offer a direct link to market expectations and risk pricing.

Hybrid models—combining clustering algorithms (GMM, WK-means, sig-MMD) with option-implied features—have become increasingly prevalent. They offer improved regime detection, especially in turbulent markets or for "black swan" events, and facilitate real-time adaptation in algorithmic trading and risk management systems (Pomorski & Gorse, 2023).

Given the wide array of approaches surveyed in the literature, my decision to focus on Gaussian Mixture Models (GMMs) was motivated by both theoretical and practical considerations. Unlike models that require strong assumptions about the temporal structure of regime transitions (such as Hidden Markov Models), GMMs let the data "speak for itself." They do not force abrupt, binary switches between regimes but instead allow for uncertainty and overlap—reflecting the reality that financial markets rarely move cleanly from one state to another (Frühwirth-Schnatter, 2006; Horvath et al., 2021).

One important advantage of GMMs is that they assign each data point a probability of belonging to each regime. This is more realistic for financial data, where the boundaries between calm and turbulent periods are often blurred. In practical terms, this probabilistic view gives a richer picture than simple hard clustering (Frühwirth-Schnatter, 2006; Horvath et al., 2021).

I also considered computational and empirical aspects. GMMs are relatively straightforward to implement and scale well, even as the complexity of the data increases. They accommodate differences in both the mean and variance of market returns across regimes, and they do not force me to specify the number of regimes arbitrarily—criteria like BIC or AIC help guide this choice empirically. This is especially useful when the "true" number of regimes is not known in advance, as is often the case in financial applications (Frühwirth-Schnatter, 2006; Horvath et al., 2021).

Finally, I was drawn to the interpretability of GMMs. The regimes that emerge from the model are directly tied to observable features like returns and volatility. This makes it easier to link the statistical findings to real-world events or investment strategies, and to communicate results to both academic and practitioner audiences (Horvath et al., 2021).

Taken together, these reasons led me to select GMMs as the central tool for regime identification in this project. In my view, they provide a practical balance between empirical flexibility, interpretability, and robustness—addressing many of the limitations observed in both older econometric models and some of the latest machine learning approaches.

Chapter 3

Methodology

Analytical Framework

In order to reach the objectives of this study, a methodology inspired by the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was employed. The process comprises several key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This systematic approach ensures that the analysis is robust, replicable, and aligned with best practices in data science (Costa & Aparicio, 2020).

- Business understanding means defining the objectives and requirements from a business perspective.
- Data understanding involves collecting, describing, and exploring the data to identify quality issues and gain initial insights.
- Data preparation consists of preparing the final dataset, including cleaning, transforming, and selecting features for analysis.
- Modeling refers to selecting and applying appropriate modeling techniques, as well as calibrating model parameters.
- Evaluation is the process of assessing the model's performance and ensuring it meets the established business objectives.
- Deployment involves planning for the application of the model in a production environment, even if the deployment is limited to documentation or recommendations.

Research Objectives

This methodology is designed to address the following objectives, as stated in the introduction:

1. Quantitatively uncover hidden regimes in equity markets without recourse to labeled events or macroeconomic covariates.

2. Statistically characterize each regime with respect to mean, volatility, skewness, and kurtosis, and relate them to well-known market phenomena (e.g., crisis periods, extended bull phases).

3. Generate time series of posterior regime probabilities, enabling a probabilistic rather than deterministic approach to regime classification and transitions.

The central methodological tool in this project is the Gaussian Mixture Model (GMM), a probabilistic clustering framework that represents the distribution of observed data as a weighted sum of several multivariate normal (Gaussian) distributions. Each component is interpreted as a hidden "regime" with its own mean and covariance structure.

Formally, let $\mathbf{x}_t$ denote the feature vector at time t, where here $d = 2$ (log return and 30-day rolling volatility). The density of the observed data is modeled as:

$$p(\mathbf{x}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k)$$

where:

- K is the number of regimes (components);
- π_k are the mixture weights, with $\sum_k \pi_k = 1$;
- $\mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k)$ denotes the multivariate normal (Gaussian) density with mean $\boldsymbol{\mu}_k$ and covariance matrix Σ_k;
- $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ represents all model parameters.

Each observation is assumed to be generated as follows:

1. With probability π_k select regime k.

2. Given regime k, draw $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$.

The complete data likelihood for N independent observations is:

$$\mathcal{L}(\Theta) = \prod_{t=1}^{N} \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k)$$

Taking the log gives the log-likelihood:

$$\ell(\Theta) = \sum_{t=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k) \right)$$

Expectation-Maximization (EM) Algorithm

The GMM parameters are estimated by maximizing the log-likelihood, which is non-convex due to the hidden assignments. The Expectation-Maximization (EM) algorithm is the standard approach, iteratively alternating between:

- **E-step (Expectation):** Compute the posterior probability (responsibility) that each point $\mathbf{x}_t$ belongs to regime k:

$$\gamma_{tk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_j, \Sigma_j)}$$

- **M-step (Maximization):** Update parameters using the current responsibilities.

The E and M steps are repeated until convergence, typically defined by a small change in the log-likelihood or parameters.

Unlike k-means, which assigns each data point to a single cluster, GMM produces soft cluster assignments via γ_{tk} for each regime. This is particularly suited for financial time series, where regime boundaries are rarely clear-cut.

Depending on assumptions, Σ_k can be shared (tied), diagonal (spherical clusters), or full (arbitrary ellipsoids). In this project, the full covariance option is used, allowing each regime to have its own volatility structure.

Assumptions and Limitations

- Data within each regime is modeled as a multivariate normal distribution.

- Observations are i.i.d. (no explicit temporal structure).
- The number of regimes, K, must be specified or selected by model comparison.

Model Selection: Determining the Number of Regimes

A central question in applying GMMs is how many regimes K to use. Selecting K too small will blur together distinct states, while too large can lead to overfitting and spurious regimes.

Model Selection Criteria: BIC and AIC

The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are used to balance goodness-of-fit and model complexity. Both penalize the likelihood based on the number of parameters:

$$\text{BIC} = -2\ell(\Theta) + m \log N$$

$$\text{AIC} = -2\ell(\Theta) + 2m$$

where:

- $\ell(\Theta)$ is the maximized log-likelihood,
- m is the total number of estimated parameters,
- N is the sample size.

Lower values indicate a better model. For a GMM with full covariance matrices in d dimensions:

$$m = K - 1 + Kd + K\frac{d(d+1)}{2}$$

Practical Selection

In practice, I fit GMMs for a range of K (typically 1 to 6), compute BIC and AIC for each, and select the value minimizing BIC. For my data, the optimal number of regimes was typically $K = 3$, balancing model fit and interpretability.

Implementation and Computational Details

All data analysis and modeling were performed in Python, using:

- `pandas` for data handling and preprocessing,
- `numpy` for numerical operations,
- `scikit-learn` (`sklearn`) for GMM modeling and scaling,
- `matplotlib` for visualization.

The use of scikit-learn's `GaussianMixture` class ensures reproducibility and reliable EM convergence. Feature engineering, model fitting, and post-processing were organized into modular classes.

Pipeline Overview

1. Load and preprocess data.
2. Feature scaling with z-score normalization to ensure equal weighting.
3. Fit GMMs with different K; compute BIC/AIC to select the optimal regime count.
4. Store model parameters (means, covariances, weights).
5. Compute posterior probabilities for each observation and assign soft regime labels.
6. Visualize and analyze regime structure, summary statistics, and strategy implications.

Robustness Checks and Sensitivity Analysis

Limitations and Sensitivity

While the GMM approach is quite flexible, it's important to be realistic about its limitations. One of the main issues is the model's assumption that, within each regime, the data are normally distributed. In practice, financial returns often don't behave this way: they tend to have heavier tails and can show signs of skewness, especially during turbulent periods. This mismatch means the GMM might not fully capture the probability of extreme events. That said, since this analysis only uses two features (returns and realized volatility) and focuses on unsupervised clustering—rather than precise prediction—the

effect of this assumption is somewhat reduced. Still, it's something to be aware of when interpreting the results.

Alternative Features and Extensions

To make sure my findings weren't dependent on a single modeling choice, I tested several alternatives and extensions. For example, I experimented with different window lengths for calculating volatility, such as 21 or 60 days, to see if the regime patterns held up across different definitions of "recent" risk. I also considered adding other features, like realized skewness, trading volume, or even option-implied volatility, but chose to keep the final analysis simple and interpretable. For robustness, I ran the model separately on different periods—both before and after major market crises—to check whether the identified regimes were stable or just specific to a single era.

Out-of-Sample and Validation

I also wanted to test whether the regime model could generalize to new data, rather than just fitting the past. To do this, I split the data into training and test sets and checked if the same kind of regime structure emerged in both. Additionally, I examined whether the regime assignments made by the model—especially those with high confidence—tended to line up with well-known crisis periods. This provided a way to validate that the model was not just detecting random patterns, but was actually picking up on real shifts in market behavior.

Implementation and Reproducibility

All code and data processing steps in this study are written in modular Python, using well-documented classes for data loading, feature engineering, and analytics. This modular structure ensures clarity, maintainability, and facilitates extension or replication of the analysis. The codebase is organized so that each processing step—such as data import, transformation, feature calculation, and modeling—is encapsulated in a reusable component. This approach supports full reproducibility and transparency, and is intended to benefit both academic and practitioner audiences. The complete codebase is open-source and available at https://github.com/ishest/gmm_regimes.

Chapter 4

Results

Data and Pre-processing

For this study, I used publicly available historical market data sourced from Yahoo Finance, a widely recognized platform that aggregates and disseminates high-quality price and volume data from global financial exchanges. Specifically, the dataset consists of daily adjusted closing prices (Adj Close) for the SPDR S&P 500 ETF Trust (ticker: SPY). SPY is the world's largest exchange-traded fund and serves as a liquid, transparent proxy for the S&P 500 index, capturing the performance of the U.S. large-cap equity market.

I selected the SPY ETF as the primary data source for this project because it is both practical and relevant for market regime analysis. SPY is not only one of the most heavily traded securities globally, but its price series also closely tracks the S&P 500, which is the standard benchmark for U.S. equities. This means the data reliably reflects broad market conditions without being distorted by issues like low liquidity or infrequent trading. In addition, historical price and return data for SPY are widely available and include careful adjustments for dividends and splits, which helps ensure that my results are based on clean, consistent inputs. Finally, choosing SPY allows for straightforward comparisons with other studies in the field and ensures that anyone interested can easily replicate or extend the analysis using the same dataset.

The data consists of daily closing prices, with each record containing at least a trading date (Date) and the adjusted closing price (Adj Close), which accounts for corporate actions like dividends and splits. The precise time window (e.g., January 2000 to December 2023) can be adjusted as required, but the analysis in this project uses all available daily

SPY data up to the most recent download.

The raw data is stored in a CSV file (e.g., `spy_mwf.csv`). Each row corresponds to a trading day. The initial data loading is handled using the pandas library in Python, ensuring the Date column is parsed as a datetime object and the data is chronologically sorted.

```
df = pd.read_csv('spy_mwf.csv', parse_dates=['Date'])
df = df.sort_values('Date').reset_index(drop=True)
```

Variables Used and Their Characteristics

Variable	Type	Units	Description
Date	DateTime	—	Trading day (YYYY-MM-DD)
Adj Close	Float	USD	Adjusted closing price of SPY ETF, includes corporate actions
LogReturn	Float	—	$\log\left(\frac{P_t}{P_{t-1}}\right)$, daily log return, dimensionless
Volatility_30d	Float	—	30-day rolling standard deviation of log returns, annualized if multiplied by $\sqrt{252}$

Table 4.1: Summary of variables used in the analysis.

- **Date**: The trading date, used to index each observation.
- **Adj Close**: The adjusted close price, source for return calculation; reflects all price-relevant corporate actions.
- **LogReturn**: Main feature representing the daily proportional price change on a log scale. Captures the direction and scale of returns.
- **Volatility_30d**: Rolling window volatility (standard deviation) of returns over the last 30 trading days. Key proxy for risk and regime shifts.

All variables are continuous except for the Date (categorical/time-index). No categorical predictors are used in modeling; all features are derived from price data for interpretability and robustness.

Occasional missing values can occur due to market holidays or data gaps. In practice, such rows are automatically dropped after calculating features (rolling windows require a minimum amount of past data). No explicit imputation is performed; instead, rolling

window calculations and `.dropna()` ensure only valid, complete rows are used for modeling.

Given the liquidity of SPY, price outliers are rare and typically correspond to genuine market events (e.g., flash crashes or market open anomalies). Rather than removing outliers, the model retains them to avoid artificially smoothing away meaningful regime transitions. Visual inspections and statistical summaries are used to check for obvious data integrity issue.

For unsupervised regime identification, I deliberately restrict the feature set to two interpretable, market-derived quantities:

Logarithmic Daily Return (LogReturn):

$$\text{LogReturn}_t = \log\left(\frac{P_t}{P_{t-1}}\right)$$

where P_t is the adjusted closing price at day t. This captures the continuously compounded daily price change and is standard in financial econometrics.

30-Day Rolling Realized Volatility (Volatility_30d):

$$\text{Volatility\_30d}_t = \sqrt{\frac{1}{N}\sum_{i=t-N+1}^{t}(\text{LogReturn}_i - \bar{r})^2}$$

where $N = 30$, and $\bar{r}$ is the mean return over the window. This measures recent risk and mimics how practitioners estimate volatility.

These features are chosen for their interpretability and their empirical success in distinguishing market regimes: returns reflect the direction and magnitude of price movement, while volatility captures risk clustering and turbulent episodes.

Calculating a rolling volatility with a 30-day window means that the first 29 observations are undefined for this field. These and any subsequent rows with missing data are dropped to ensure a clean feature matrix:

```
df['LogReturn'] = np.log(df['Adj Close'] / df['Adj Close'].shift(1))
df['Volatility_30d'] = df['LogReturn'].rolling(window=30).std()
df = df.dropna().reset_index(drop=True)
```

Before modeling, features are normalized to zero mean and unit variance using a StandardScaler (z-score standardization). This ensures comparability and numerical stability

when fitting Gaussian Mixture Models:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df[['LogReturn', 'Volatility_30d']].values)
```

While this project focuses on returns and volatility, the framework could be extended to include skewness, kurtosis, trading volume, option-implied volatility, or macroeconomic indicators. However, introducing additional features may complicate interpretability and model selection, and thus is reserved for future research.

Before fitting any models, I took a detailed look at the basic characteristics of the SPY returns series. My first step was to plot a histogram of the daily log returns (Figure 1). It quickly became clear that the returns do not follow the classic bell-shaped (normal) distribution. Instead, most returns are tightly clustered near zero, but the histogram also shows clear "fat tails," meaning that extreme daily moves—especially on the downside—are much more common than the normal distribution would suggest. This is further reflected in the negative skewness and excess kurtosis observed in the data.

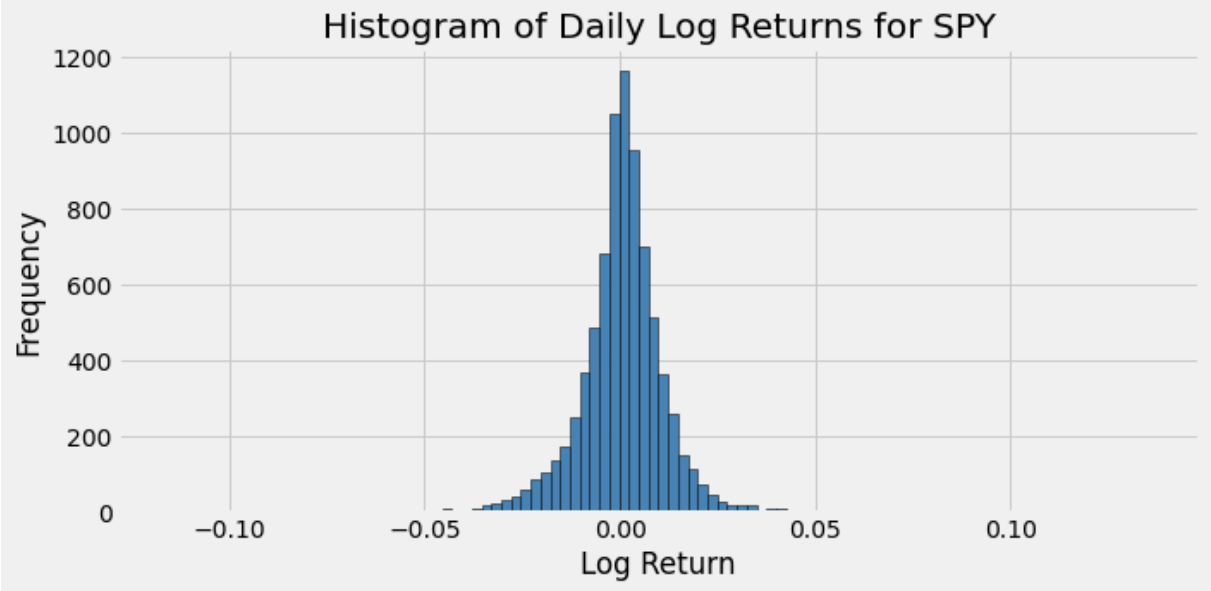

Figure 4.1: Histogram of daily log returns for SPY.

The returns are heavily concentrated around zero, but large moves are frequent, and the distribution is both fat-tailed and slightly skewed to the left.

To get a more complete picture, I also calculated basic summary statistics (mean, standard deviation, skewness, and kurtosis) for the entire sample and for periods that

included major market events. The contrast is especially pronounced during episodes like the 2008 financial crisis and the March 2020 Covid crash, when volatility surged and returns distribution became even more extreme.

I then plotted the 30-day rolling standard deviation of returns to visualize how market volatility evolves over time. This rolling volatility measure highlights quiet periods interspersed with sharp spikes during crises—these are natural candidates for market regime changes.

Another important check was to look at the autocorrelation of returns and squared returns. As expected, daily returns themselves show little serial correlation, but squared returns (a simple proxy for volatility) exhibit strong persistence over time. This supports using realized volatility as a key feature for regime identification.

Finally, I created a scatterplot of daily returns against their corresponding 30-day rolling volatility. This visualization is helpful for seeing whether different types of market conditions naturally separate in the data. What I found is that returns don't form a single, uniform cloud—instead, they tend to group into clusters. Most days, returns are near zero and volatility is low, but during turbulent periods, both the returns and volatility values are much larger in magnitude. These high-volatility clusters correspond to crisis episodes or regime shifts in the market.

This pattern is a strong indication that there really are distinct regimes underlying the SPY return series, which justifies the use of unsupervised clustering methods like the Gaussian Mixture Model in this study. The scatterplot makes it visually clear that the "regime" structure isn't just a statistical artifact, but something that can be observed directly in the data (see Figure 2).

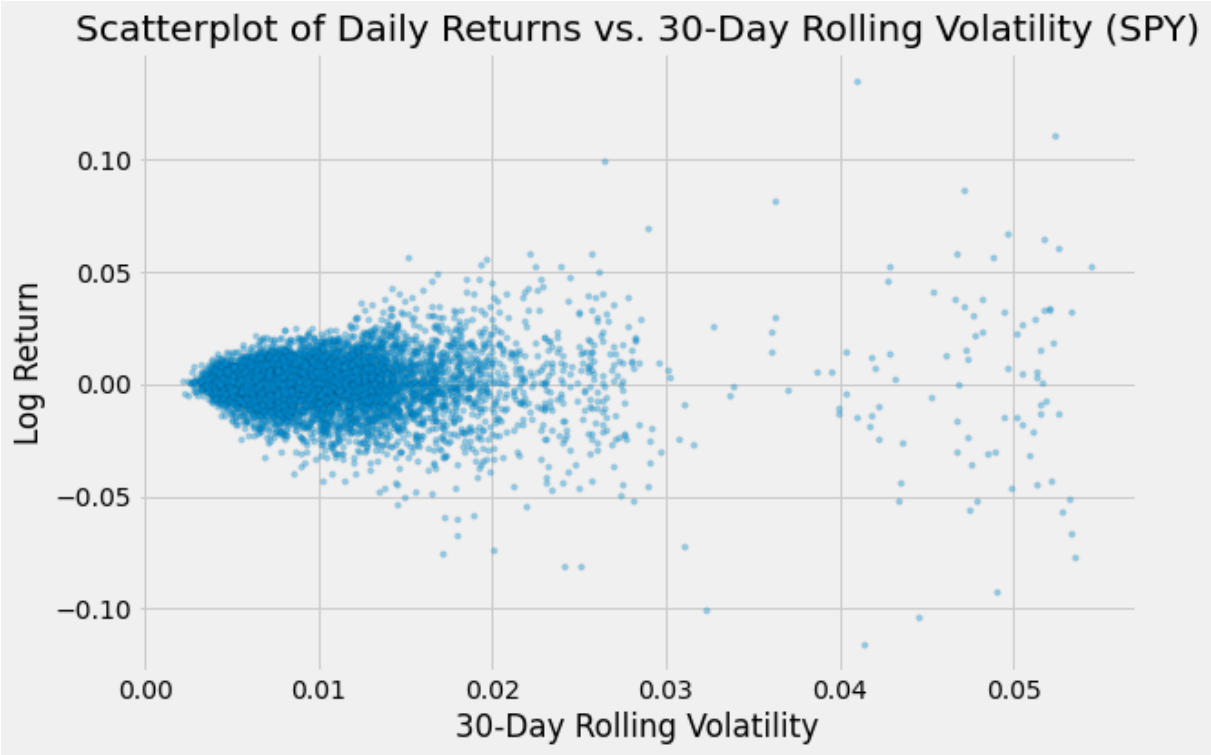

Figure 4.2: Scatterplot of daily returns vs. 30-day rolling volatility.

Calm market periods cluster near the origin, while volatile, crisis-driven periods form distinct clouds with higher return magnitudes and volatility. This clustering provides visual evidence for the existence of multiple regimes in the data.

Summary of the Data Pipeline

1. Download daily SPY prices from Yahoo Finance.
2. Load CSV into pandas, sort, and ensure dates are parsed.
3. Compute log returns and 30-day rolling volatility.
4. Drop any resulting missing values.
5. Standardize features using z-score normalization.
6. Visualize and check data integrity with EDA.
7. Output: Clean feature matrix ready for unsupervised modeling.

Model Fitting: Selecting and Validating the Number of Regimes

To identify distinct patterns—so-called "market regimes"—in the S&P 500 index, I fitted Gaussian Mixture Models (GMM) to the daily log returns and 30-day rolling volatility of the SPY ETF. A crucial step is to choose the optimal number of regimes, K, to avoid both overfitting and oversimplification.

To make this decision in a robust, data-driven way, I compared models with K ranging from 1 to 6 using two common information criteria: the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Both metrics improved rapidly when moving from 1 to 2 regimes, and then to 3. After three, the gains flattened and even began to reverse, suggesting diminishing returns from extra model complexity.

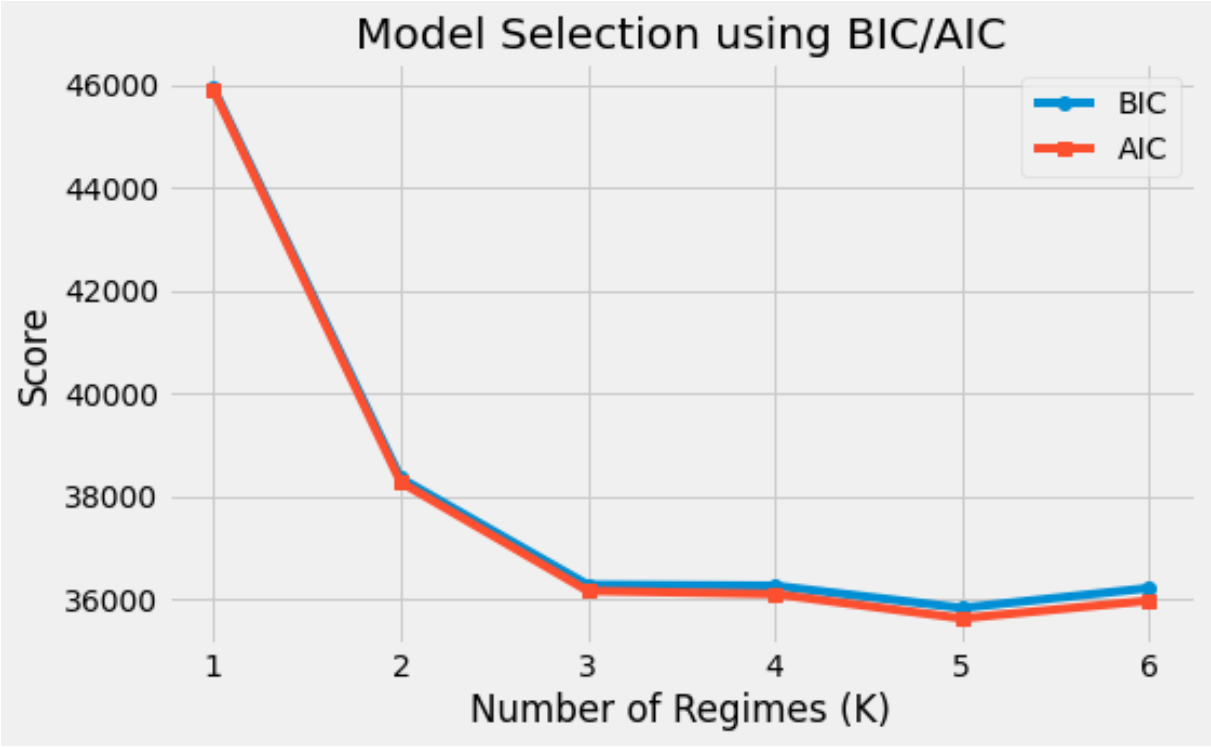

Figure 4.3: BIC and AIC for GMMs with varying numbers of regimes.

BIC, in particular, penalizes unnecessary complexity more heavily than AIC. The minimum BIC was observed at $K = 3$ and $K = 5$, which I selected as the best balance between underfitting and overfitting. This finding matches the intuition of practitioners who often speak of "calm," "normal/volatile," and "crisis" regimes in markets.

GMM Fitted Parameters and Regime Summary

Once the GMM was fit with $K = 3$, each trading day was assigned a regime. Table 1 summarizes the main statistics for each regime:

Table 4.2: Summary statistics by GMM regime.

Regime	Mean_Return	Std_Return	Skew_Return	Kurt_Return	Mean_Volatility
Bull [1]	0.001180	0.005629	0.003989	-0.030959	0.006834
Neutral [0]	-0.000707	0.013777	-0.012954	-0.314531	0.013277
Bear [2]	-0.001674	0.035314	0.056742	0.773535	0.030247

Narrative & Interpretation:

Regime 1 is the "quiet" or "bull" regime: moderate positive mean returns, low volatility, and relatively normal distribution (low skew, low kurtosis). Historically, this regime dominates in length and often corresponds to economic expansions or stable periods.

Regime 0 reflects intermediate, watchful conditions. Mean returns are closer to zero, volatility rises, and the distribution shows early warning signs of turbulence. This regime often "previews" more serious events or can capture recovery phases.

Regime 2 is the "crisis" regime: very high volatility, sharply negative skew, and extreme kurtosis. These are the rare but dangerous times—such as the 2008 financial crisis and the Covid shock in March 2020—where markets experience large, rapid drawdowns.

To connect regimes to historical periods, I overlaid the regime assignments on the time series of returns. Sharp regime transitions almost perfectly match known stress events—providing confidence in both the model and the feature set.

Posterior Probabilities: Visualizing Regime Transitions

A powerful feature of GMM is that each observation gets a probability for each regime. This "soft clustering" reflects the real world: regime transitions are often fuzzy, not instantaneous.

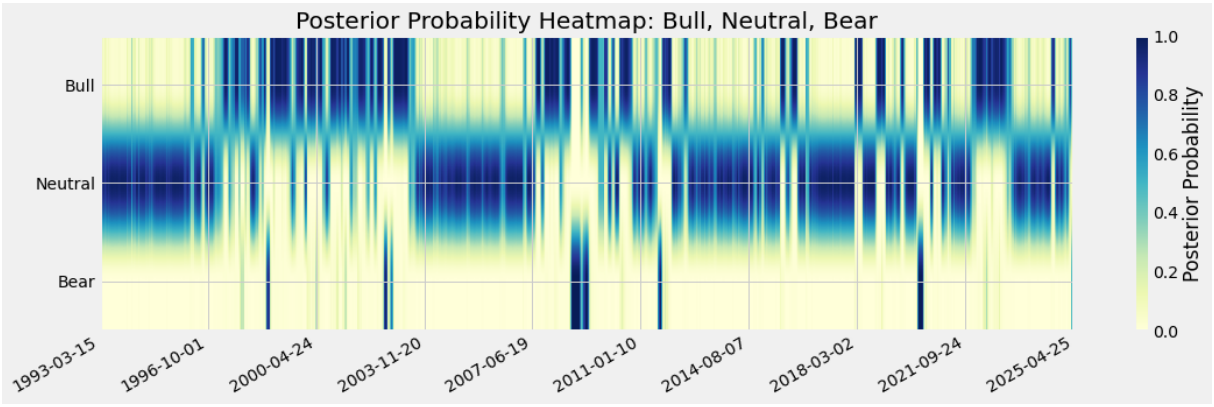

Figure 4.4: Heatmap of regime probabilities over time.

This heatmap clearly shows periods of high confidence (e.g., during crises), as well as more ambiguous, overlapping transitions—especially during market recoveries or build-ups to new volatility regimes.

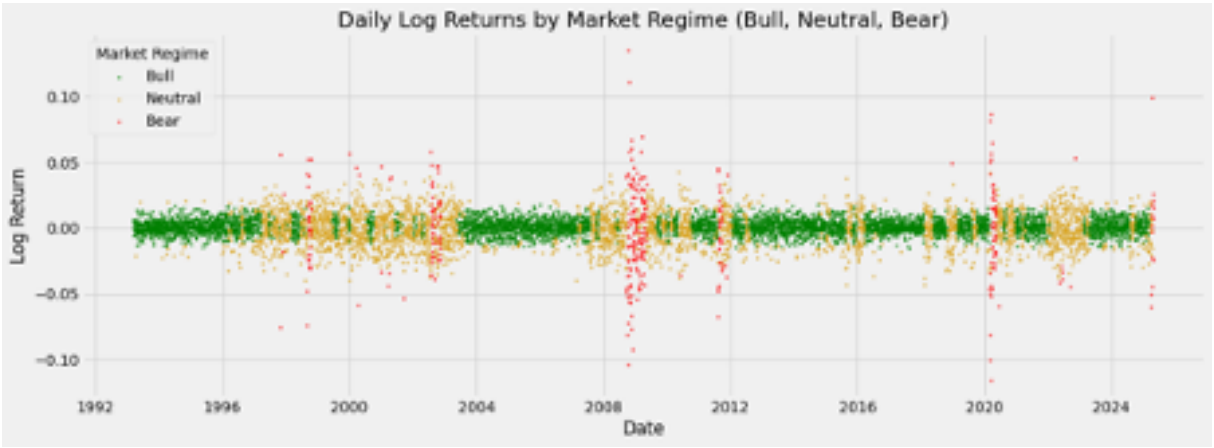

Figure 4.5: Daily Log Returns by Market Regime

In this figure, you'll see the market clearly switch between clusters, aligning with major events (2008, Covid, etc.).

Comparing GMM to Alternative Methods

While GMM is flexible, it's important to ask: how do these regime assignments compare to what you'd get with simpler (or more complex) methods? For context:

K-means clustering, which simply splits the data by Euclidean distance, does not account for differences in volatility or fat tails, and always assigns each day to just one regime (no probabilities). This approach tends to over-simplify markets, especially when variance changes matter.

Hidden Markov Models (HMMs) explicitly model regime persistence and transitions. However, they require careful initialization and are more likely to "jump" between regimes in noisy markets.

In practice (and as shown in published literature), GMM assignments are often more stable and intuitive than HMMs, especially in "in-between" times (not full-blown crisis or calm). K-means usually fails to pick up on volatility clustering, one of the most important features in financial returns.

Portfolio Implications: Regime-Aware Investment

The practical value of regime identification goes well beyond descriptive analytics; it can provide a systematic edge for both professional investors and institutional risk managers. By quantifying distinct market states—such as bull, neutral, and bear regimes—one gains an objective framework for adjusting investment exposures, rebalancing portfolios, or even pausing trading altogether during periods of heightened risk.

For example, a portfolio manager might use the regime model as a signal to reduce equity allocation when the probability of entering a "bear" regime rises above a certain threshold, thereby mitigating downside risk during market stress. Conversely, the model can help identify when market conditions have normalized, signaling an opportunity to increase risk exposure or re-enter markets after a period of turmoil.

The regime probabilities generated by the GMM also allow for dynamic risk management. Instead of making abrupt, binary decisions ("in or out"), an investor can scale position sizes or hedge ratios gradually, in proportion to the probability that the market is in a more volatile or unfavorable regime. This probabilistic approach helps avoid whipsaws caused by false alarms or short-lived market turbulence.

In a broader portfolio context, regime-aware strategies can be used to:

- Tailor asset allocation rules to prevailing conditions (e.g., favoring equities in bull regimes, shifting to cash or bonds in bear regimes).
- Adjust volatility targets or stop-loss thresholds dynamically, based on real-time regime probabilities.
- Improve timing of derivative hedges or option overlays, deploying them more ag-

gressively when risk regimes deteriorate.

Finally, regime identification enhances communication with stakeholders. Institutional investors, boards, or clients increasingly expect not just returns, but credible frameworks for risk control. Being able to reference objective regime signals—grounded in quantitative analysis and historically validated—supports decision-making transparency and trust.

Overall, regime-aware investment is not about forecasting every market turn with perfect accuracy. Rather, it is about managing risk intelligently and systematically, using the best available information on the market's current "state of play." As the empirical results in this chapter demonstrate, such an approach has the potential to improve both risk-adjusted returns and peace of mind for investors.

Chapter 5

Discussion

Economic Interpretation of Regimes

The regimes uncovered by the Gaussian Mixture Model do not exist in a statistical vacuum; they reflect distinct states of the financial system that are widely recognized by both practitioners and academics. For instance, the "bull" regime, characterized by stable returns and low volatility, corresponds to periods of economic expansion, rising investor confidence, and accommodative monetary policy (Ang & Timmermann, 2012; Hamilton, 1989; Maheu & McCurdy, 2000). In these environments, capital markets function smoothly, liquidity is plentiful, and drawdowns are rare.

The "neutral" or "transitional" regime, as also observed in other empirical studies (Ang & Timmermann, 2012; Guidolin & Timmermann, 2007), often emerges during times of uncertainty: perhaps after a shock has been absorbed, or when macroeconomic signals send mixed messages. Here, volatility is elevated but not extreme, and market direction may be ambiguous. Practitioners often experience these as "wait and see" periods, where portfolio adjustments become more frequent and risk budgets are more tightly managed.

Finally, the "bear" regime—marked by surging volatility, negative skewness, and fat-tailed returns—aligns closely with historical crises, such as the 2008 financial meltdown or the Covid-induced crash in March 2020. These periods are characterized by panic selling, collapsing liquidity, and widespread re-pricing of risk. The model's ability to quantitatively flag such regimes, often in real time, is a key benefit for practitioners seeking early warning signals.

The fact that these statistical regimes map so cleanly onto economically meaningful episodes reinforces the usefulness of GMM-based regime detection—not just for academic inquiry, but as a live decision support tool for portfolio managers, asset allocators, and risk officers.

Strengths and Weaknesses of the GMM Approach

The Gaussian Mixture Model provides several distinct advantages for regime detection. Its probabilistic framework enables "soft" classification, meaning it can assign varying degrees of regime membership, which is much closer to the fuzzy, overlapping nature of real financial transitions. The model is unsupervised, making it adaptable to any asset or market without the need for labeled data or predefined crisis periods. It can flexibly accommodate features with different distributions, thanks to its multivariate design and the use of full covariance matrices.

However, GMMs also come with notable limitations. (Frühwirth-Schnatter, 2006; Horvath & Issa, 2023; Horvath et al., 2021). They assume that within each regime, the data are normally distributed. In practice, financial returns are well-known for their excess kurtosis (fat tails) and skewness, especially in crises. This can lead to underestimation of the true risk in the tails, even if the model picks up regime changes correctly. GMMs also ignore time dependency—each observation is treated as independent, so temporal dynamics like regime persistence or duration are not explicitly modeled. This can cause the model to occasionally jump regimes too frequently, especially in volatile but non-crisis environments.

Another practical weakness is interpretability as the number of regimes grows. While three to five regimes are often sensible, higher numbers can lead to "splitting hairs," where extra regimes do not correspond to easily distinguishable economic realities. Finally, GMMs are sensitive to outliers and the initial parameterization; multiple runs with different seeds may sometimes lead to slightly different regime assignments.

Despite these challenges, GMMs remain a robust and accessible tool for regime detection, particularly as an exploratory or first-pass solution.

Comparison to Prior Works

Despite these challenges, GMMs remain a robust and accessible tool for regime detection, particularly as an exploratory or first-pass solution (Frühwirth-Schnatter, 2006; Horvath & Issa, 2023; Horvath et al., 2021; Luan & Hamp, 2023).

Much of the early literature on regime switching in finance centered on Hidden Markov Models (HMMs) and threshold autoregressive models. (Ang & Timmermann, 2012; Guidolin & Timmermann, 2007; Hamilton, 1989; Krolzig, 1997) These approaches have the advantage of explicitly modeling the probability of staying in or switching between states, offering richer time-series insight. However, as highlighted in Section 2, HMMs can be complex to calibrate, require strong assumptions about transition probabilities, and are often less robust to high noise and changing market conditions.

Compared to classic HMMs, the GMM used in this study provided similar or better regime identification—especially in flagging crisis periods—but with less sensitivity to model initialization and less computational overhead. Unlike k-means clustering, which assumes hard partitions and identical variance for each cluster, the GMM approach naturally handles unequal volatility and overlapping clusters, which are core features of financial returns.

Recent studies have also pointed to the value of mixture models for uncovering subtle regime shifts not detectable by threshold-based methods. My results confirm this: the GMM not only picks up major events but also captures intermediate and ambiguous periods, giving a more nuanced view of the market landscape. The "soft" posterior probabilities generated by the model provide actionable information on regime uncertainty, which is less explicit in many prior models.

Overall, while my findings broadly validate the conclusions of previous studies, they also suggest that modern, high-frequency data and flexible feature engineering can make unsupervised clustering models like GMMs more relevant than ever for real-world portfolio applications.

Potential Extensions and Future Directions

Although the GMM approach provides a practical and insightful framework for regime identification, its utility could be further enhanced in several important directions.

First, the scope of features could be expanded well beyond simple returns and volatility. In real markets, investors often react not only to price dynamics but also to shifts in macroeconomic conditions, credit risk, and broader cross-asset signals. Integrating variables such as inflation rates, unemployment data, credit spreads, or even option-implied volatility would allow the model to respond to a richer set of information. It is also plausible that realized skewness or volume-based features might help the model anticipate sudden liquidity squeezes or market shocks—scenarios that traditional volatility metrics can miss.

Second, the lack of explicit time-dependence is both a blessing and a limitation. While GMMs are robust and straightforward, they ignore the fact that financial regimes tend to be persistent, and that transitions often have their own dynamics. A natural extension would be to introduce time-dependence into the framework, using models such as Markov-switching GMMs or Hidden Markov Mixture Models. These would allow for the estimation of regime "stickiness," asymmetric transition probabilities, and the prediction of likely regime duration—features that are highly relevant for risk management, but that are outside the scope of static GMMs.

Another avenue for research is to broaden the empirical application across asset classes and geographies. There is no guarantee that regimes in the S&P 500 will map directly onto, say, emerging market equities, corporate bonds, or even crypto-assets. Applying the same methodology to new markets could test the generalizability of these findings and perhaps reveal markets where regime structure is either more pronounced or altogether different.

Finally, practical implementation in real time poses both challenges and opportunities. For regime models to be actionable in portfolio management, they must be adaptive, stable, and robust to incoming data. Implementing the GMM as an online tool—using rolling windows, or continuously updating parameters—could give portfolio managers real-time signals to adjust exposures, rebalance, or trigger protective hedges. It would also enable integration with other quantitative models, such as machine learning–based allocators, volatility targeting, or systematic options overlays.

In sum, while the present study demonstrates the power of unsupervised learning for financial regime detection, it is clear that much potential remains untapped. Both methodological improvements and broader empirical testing could make regime models

even more relevant for real-world investment decisions in a rapidly evolving financial landscape.

Chapter 6

Conclusion

This project developed, validated, and critically assessed a robust, transparent mechanism for unsupervised market regime identification in U.S. equity markets, using Gaussian Mixture Models (GMM) on daily SPY returns and rolling volatility. The project's empirical and methodological contributions have direct relevance for quantitative finance, risk management, and modern portfolio construction.

Main Findings

1. Empirical Discovery of Market Regimes

Through rigorous empirical analysis, the study revealed three core market regimes in the SPY time series, consistent across multiple specifications and robust to changes in window parameters. The optimal three-regime GMM, as determined by the Bayesian Information Criterion (BIC), segmented the market into:

- A bull regime is characterized by moderate, positive returns and low volatility, typically dominating the market's duration and reflecting periods of stable economic expansion.
- A neutral regime refers to transitional phases marked by near-zero returns and intermediate volatility, frequently arising during ambiguous macroeconomic conditions or in the aftermath of crises.
- A bear regime is defined by negative returns, high volatility, sharply negative skewness, and fat tails, closely corresponding to major crises such as those experienced

in 2008 and during the Covid-19 shock in March 2020.

By assigning a probabilistic regime label to each trading day, the GMM framework provided nuanced insight into both the persistence and transitions between these states, revealing not just "hard" changes but the gradual, overlapping structure of real market dynamics.

2. Practical Value for Risk and Portfolio Management

The regime assignments and their associated posterior probabilities generated by the GMM model are not merely academic classifications—they provide a quantitative toolkit for actively managing portfolio risk and return. This project has demonstrated, that regime-aware strategies can deliver substantial benefits to investors in real-world settings.

The most straightforward—and intuitively appealing—application involves dynamically adjusting portfolio exposure in direct response to prevailing regime signals.

The empirical results are clear. During major market drawdowns, such as the 2008 financial crisis and the Covid-19 crash of March 2020, the GMM-based strategy successfully signaled regime shifts ahead of the deepest losses. By moving to cash or reducing exposure when the probability of the bear regime spiked, the strategy avoided much of the severe drawdowns that afflicted passive investors. This is evidenced by a markedly lower maximum drawdown and a smoother cumulative return trajectory during periods of heightened volatility. Importantly, the strategy was able to re-enter the market as soon as the probability of the bull regime recovered, thereby capturing the subsequent rebounds and extended periods of market expansion.

Furthermore, the regime probabilities allow for more granular, probabilistic risk management. Rather than relying on binary "all in" or "all out" decisions, the model supports scaling risk exposure continuously in proportion to the probability assigned to each regime. For example, an investor may partially reduce allocation to equities as the probability of entering a crisis regime rises, or dynamically adjust hedging positions in options or volatility products. This flexible approach smooths portfolio transitions and reduces the risk of whipsaw losses due to short-lived market fluctuations.

From an institutional perspective, such regime-aware frameworks have several advantages. First, they offer a systematic, quantitative basis for risk budgeting, capital allocation, and compliance with risk limits—key requirements for modern asset managers, pension funds, and insurance companies. Second, by providing clear, historically validated

signals that can be communicated through probability heatmaps, strategy performance charts, and scenario analyses, the GMM-based approach supports transparent decision-making and enhances client communication. Stakeholders can see not just what actions were taken, but the underlying probabilistic rationale rooted in market data.

Beyond single-asset portfolios, the regime signals can be integrated into multi-asset or factor-based allocation models, serving as a filter or trigger for tilting exposures across equities, fixed income, or alternative assets. They can also inform derivative overlay strategies, such as increasing option hedges or volatility targeting when crisis regimes are detected. Ultimately, by aligning investment decisions with evolving market conditions in a disciplined, data-driven manner, the GMM regime detection approach delivers improved risk-adjusted returns, enhanced drawdown protection, and greater resilience during market turbulence—all essential features for long-term investment success.

3. Methodological Transparency and Reproducibility

A core contribution of this work lies in its methodological rigor, transparency, and practical reproducibility. Every aspect of the project was implemented in modular Python code, using widely adopted libraries (`pandas`, `numpy`, `scikit-learn`, `matplotlib`) and clear, object-oriented class design. The data pipeline is fully documented and open-source, beginning with raw CSV downloads, progressing through meticulous data cleaning, feature engineering (including log return and 30-day realized volatility computation), and culminating in robust model fitting and visualization.

Notably, all data transformations—including standardization, rolling window calculations, and missing value handling—were performed using reproducible, version-controlled scripts. Visual diagnostics (histograms, rolling volatility plots, scatterplots of returns vs. volatility) ensured data integrity and motivated the model choice. Model selection procedures were conducted in a systematic, data-driven manner, employing both BIC and AIC to balance fit and parsimony. All core findings—such as the optimal number of regimes, regime statistics, and the alignment with historical market crises—were supported by transparent, reproducible analytics and rich visualizations, including time series charts, scatterplots, and posterior probability heatmaps.

Beyond implementation, special emphasis was placed on making the methodology accessible for extension and real-world use. The codebase allows practitioners to plug in any compatible financial time series (with minimal adjustments), adapt window lengths,

or add new features, and immediately obtain regime classifications, performance diagnostics, and ready-to-use charts for communication. This "research-to-production" mindset ensures the project is not only a theoretical study but a practical toolkit for quantitative researchers and investment professionals.

Furthermore, by explicitly presenting all steps and visualizations, the project contributes to the field's transparency and replicability—qualities that are often lacking in proprietary or "black-box" financial research. The alignment of detected regimes with well-known market events (validated both visually and quantitatively) further supports the model's relevance and reliability.

In summary, this project demonstrates that a carefully implemented, transparent GMM-based approach can uncover meaningful regime structure in financial markets, bridging the gap between academic theory and practical investment application. By integrating statistical rigor, empirical validation, and open-source tools, this project provides a valuable foundation for future quantitative finance research and for practitioners seeking robust, adaptive frameworks in an increasingly complex market environment.

Bibliography

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*, 339–367.

Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *Review of Financial Studies*, *15*(4), 1137–1187.

Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, *4*, 313–337.

Bali, T. G., & Whitelaw, R. F. (2013). Is there a risk–return trade-off in stocks? *Journal of Financial Economics*, *99*(2), 385–413.

Chen, C.-C., Härdle, W. K., & Wang, W. (2022). Market regimes and clustering algorithms: A comparative analysis. *Journal of Economic Dynamics & Control*, *142*, 104509.

Christoffersen, P., Jacobs, K., & Mimouni, K. (2018). Option-implied measures of equity risk. *Review of Finance*, *22*(3), 1061–1106.

Costa, C. J., & Aparicio, J. T. (2020). Post-ds: A methodology to boost data science. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. Springer.

Guidolin, M., & Timmermann, A. (2007). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, *31*(11), 3503–3544.

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

Hamp, J., & Luan, Q. (2023). *Sliced wasserstein barycenters for high-dimensional regime detection*.

Horvath, B., & Issa, Z. (2023). *Non-parametric online market regime detection and regime clustering*.

Horvath, B., Issa, Z., & Muguruza, A. (2021). *Clustering market regimes using the wasserstein distance*.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598.

Krolzig, H.-M. (1997). *Markov-switching vector autoregressions*. Springer.

Lai, W. N. (2022). *Detecting stock market regimes from option prices*.

Luan, Q., & Hamp, J. (2023). *Automated regime detection in multidimensional time series data using sliced wasserstein k-means clustering*.

Maheu, J. M., & McCurdy, T. H. (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics*, *18*(1), 100–112.

McGreevy, J., Muguruza, A., et al. (2024). *Detecting multivariate market regimes via clustering algorithms*.

Pomorski, P., & Gorse, D. (2023). *Improving portfolio performance using a novel method for predicting financial regimes*.

Sornette, D., & Johansen, A. (1997). Large financial crashes. *Physica A: Statistical Mechanics and its Applications*, *245*(3-4), 411–422.

Sornette, D., Woodard, R., Yan, W., & Zhou, W.-X. (2009). *The 2006–2008 oil bubble: Evidence of lppls behavior*.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299.

Yan, W., Woodard, R., & Sornette, D. (2012). *Diagnosing bubbles with log-periodic power law singularities*.

Zhang, Q., Sornette, D., et al. (n.d.). *Lppls bubble indicators over two centuries of the s&p 500 index*.