Hidden Market Regimes in Equity Returns via Hidden Markov Models

Charlie Kushelevsky, Parth Paliwal, Max Zhang, Cole Carter

1 Milestone 1: Project Plan

1.1 Problem Description

Financial markets appear noisy from day to day, yet empirical evidence suggests that price movements are driven by a small number of persistent latent "regimes," such as low-volatility growth periods, high-volatility downturns, or transitional sideways phases (e.g. bull trends, bear markets, high volatility). These regimes are not directly observable, but they affect return distributions, volatility clustering, and risk.

The goal of this project is to use a Hidden Markov Model (HMM) to uncover latent market regimes from historical equity return data. We treat the daily market return as the observed variable and the underlying regime as a hidden state that evolves slowly over time. Using probabilistic reasoning, expectation maximization (Baum-Welch), and sequence inference (Viterbi), we want to characterize how many regimes best describe the behavior of the market and how persistent these regimes are. This problem uses the course topics of latent variable modeling, EM-based learning, and probabilistic inference.

We seek to answer questions such as: (1) How many distinct regimes best capture the distributional structure of daily market returns? (2) What statistical properties (mean, variance) define each regime? (3) How persistent are different regimes, and how frequently do transitions occur? (4) Do inferred high-volatility states align with known market stress periods, such as the 2008 crisis or the 2020 COVID crash?

1.2 Dataset

We will use daily adjusted closing prices of the SPDR S&P 500 ETF (ticker: SPY), obtained from publicly available historical data through the `yfinance` Python API. The dataset will span 1958 through 2025, covering multiple market cycles including the dot com aftermath, the 2008 global financial crisis, the long 2010-2019 expansion, the COVID crash, and the 2022-2023 inflationary period.

Preprocessing consists of: (1) restricting the dataset to valid trading days; (2) computing daily log returns,

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t is the adjusted closing price; (3) dropping missing values from holidays and non-trading days; and (4) optionally standardizing returns for numerical stability. No additional feature engineering is required for the baseline HMM.

This dataset is well suited for sequence models because it provides a long, continuous, high quality return time series, enabling reliable estimation of persistent hidden dynamics.

1.3 Methodology

We will model the return sequence $\{r_t\}$ using a K-state Hidden Markov Model with Gaussian emissions. Each hidden state $S_t \in \{1, \ldots, K\}$ represents an unobserved market regime, while the

emission distribution $r_t \mid S_t = k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ captures the characteristic return behavior of that regime. The transition matrix encodes the persistence and switching behavior between regimes, which we expect to be highly skewed toward self-transitions.

Model parameters (initial state distribution, transition probabilities, and emission parameters) will be learned with the Baum-Welch algorithm (EM). We will experiment with different values of K (e.g., $K = 2, 3, 4$) to compare fit, interpretability, and stability. After learning, we will apply the Viterbi algorithm to infer the most likely regime sequence over time and examine whether high-volatility regimes correspond to known historical events.

Evaluation will include log-likelihood, qualitative inspection of decoded regimes, and analysis of transition probabilities and expected regime durations. If time permits, we may extend the model to incorporate additional observable variables such as realized volatility or trading volume.