Financial Regime Detection

For classifying markets into specific regimes using unsupervised learning techniques, we have used many methods, including feature engineering, dimensionality reduction, clustering, HMM modelling, and visualisation techniques.

Feature Engineering

1. Mid Price: The mid-price is the average of the best bid and ask prices. This is a neutral estimate of an asset's fair value at any given time.

2. Spread: A narrow spread indicates a liquid market with high participation. Wide spreads signify uncertainty, lower liquidity, and higher trading costs, it is vital for liquidity analysis.

3. Spread Percentage: This normalises the raw spread by the mid-price

4. Imbalance Level: (BidQtyL1 - AskQtyL1) / (BidQtyL1 + AskQtyL1)

This measures order book pressure at the top level. A high positive value indicates stronger buying interest. A negative value indicates sell-side dominance. Imbalance can act as a predictor of short-term price movements due to supply/demand pressures. Deeper levels in the order books are also observed for hidden liquidity and iceberg orders.

5. Imbalance Derivative: A rapid change may signal an impending price movement or regime shift. It's a dynamic momentum signal on liquidity.

6. Imbalance Lag5: The 5-tick lag of imbalance. It incorporates short-term memory into the regime analysis. The market often reacts to past pressure changes, which helps detect recurring micro-patterns.

7. Microprice: microprice = (Bid * AskQty + Ask * BidQty) / (BidQty + AskQty)

It weights prices by opposing side quantities, thus incorporating the strength of opposing interests. It's particularly useful in the presence of imbalance.

8. Cumulative bid/ask quantity and Volume: Sum of the volumes across the top 20 levels on each side. They measure overall depth and potential for market absorption. Useful for liquidity regime tagging. It represents demand pressure and interest in the market.

9. Slope and Convexity of bid/ask depth : Steep positive/negative slopes show how liquidity changes with depth. A convex shape often means concentrated liquidity at top levels, while a concave shape implies strategic order placements deeper.

10. Log Returns The log return is the preferred way to measure percentage change because it’s time-additive. It forms the basis for volatility and regime tagging.

11. Rolling Volatility and Skewness: These features quantify local market turbulence, essential for volatility regime classification. Captures asymmetry and fat-tailed behaviour in recent return distributions. High kurtosis or skew indicates abnormal return events or one-sided market behaviour.

12. VWAP Shift: Change in the VWAP (Volume Weighted Average Price). It reflects average trade price evolution and helps detect trend shifts.

13. Market Maker Participation: These features measure how active market makers are and how much they dominate the trading activity. Important for liquidity characterisation.

14. Responsiveness: Responsiveness captures how price reacts to volume shocks. A more reactive market implies low liquidity or strong momentum.

15. Liquidity Score: Composite feature combining spread, book slope, and MM activity. Designed to capture overall trading ease.

16. Trade Intensity: Rolling trade count. Represents trading activity frequency.

Normalisation and Dimensionality Reduction

Feature Normalisation using StandardScaler

Dimensionality Reduction using **Principal Component Analysis (PCA)** and **Autoencoder Encoding**, using a neural network to compress and reconstruct the data. The middle layer learns a compact representation that can capture intricate, non-linear interactions between features. Useful for high-dimensional microstructure data where relationships may not be linear.

Optimal K Detection

Choosing the right number of clusters (K) for K-Means and GMM. It uses:

- **Inertia**: Measures compactness of clusters. Lower is better. Knee Point is found.
- **Silhouette Score**: Captures how well-separated clusters are. Ranges from -1 to 1.
- **Davies-Bouldin Index**: Lower values suggest better clustering (less overlap).
- **Calinski-Harabasz Index**: Higher values indicate well-separated clusters.

Clustering: We use a variety of unsupervised clustering techniques to detect latent regimes:

- **K-Means**: A basic partitioning method. Assumes spherical clusters and equal variance.
- **Gaussian Mixture Model (GMM)**: Uses soft assignment and models clusters as Gaussian distributions. Better for elliptical clusters. We extend it by Bayesian GMM with a Dirichlet Process prior. Automatically reduces unused clusters.
- **HDBSCAN**: A hierarchical, density-based algorithm. Doesn't require a fixed number of clusters and handles noise well. Good for financial data, which can be noisy and irregular. It reassigns outlier labels using nearest neighbours to ensure continuity.
- **Deep Clustering:** It trains an autoencoder where the bottleneck layer is constrained to be a softmax activation with the number of units equal to the desired number of clusters. After training, the final regime label is assigned as the `argmax` of the softmax layer output. This method combines dimensionality reduction and clustering in a single neural model. It is particularly adept at identifying non-linear boundaries and structures.
- **Hidden Markov Models:** Models the time-dependent structure of market regimes using a Gaussian Hidden Markov Model (HMM). Unlike static clustering algorithms, HMM treats regimes as hidden states evolving. The function chooses the number of hidden states using BIC (Bayesian Information Criterion), balancing model fit and complexity.

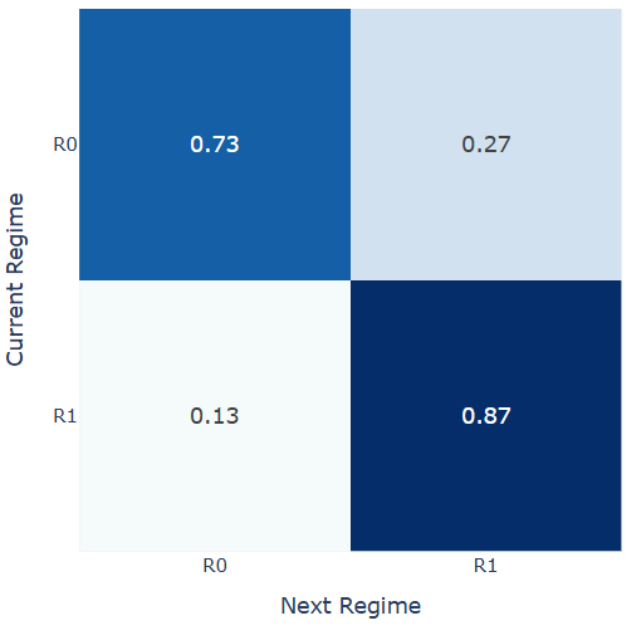
Auto-labelling

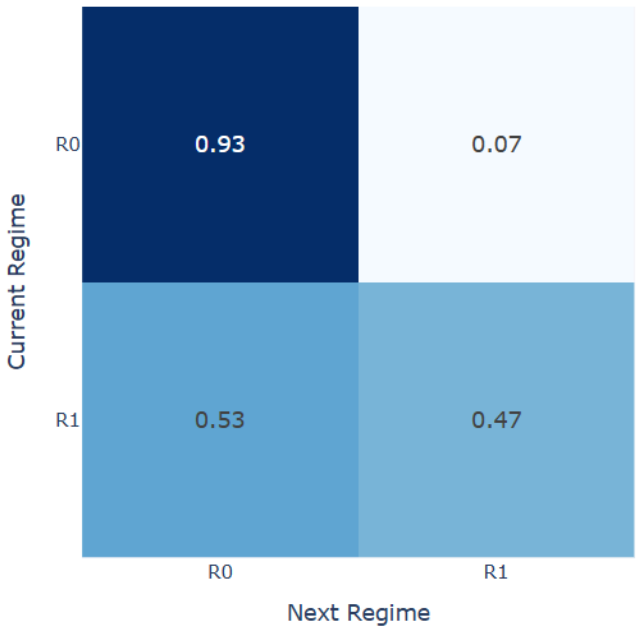
We heuristically assign semantic labels ("Volatile + Trending + Illiquid") to each regime using key features like volatility, return, spread, and volume. It calculates thresholds using quantiles and assigns categorical tags:

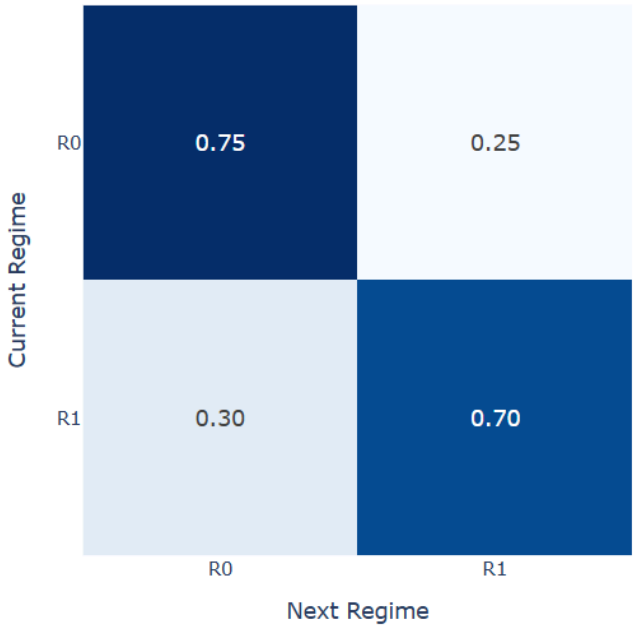
- Volatility: High, Moderate, Low
- Trend: Trending, Mean-Reverting, Neutral
- Liquidity: Illiquid, Mid-Liquidity, Liquid

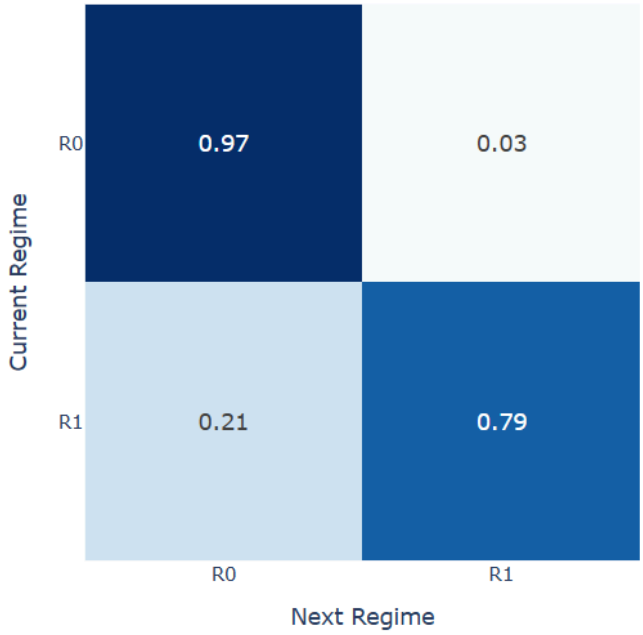
Results

Clustering Method	Regime	Key Features and Insights
K-Means (2 Regimes)	Regime 0: Volatile + Trending + Illiquid	Higher spread (0.0689) and lower MM participation (0.4482) indicate illiquidity. Negative responsiveness (−130.1) suggests price drops. Depth slopes are shallow, and kurtosis (4.28) reflects fat tails for volatile periods.
	Regime 1: Stable + Mean-Reverting + Mid-Liquidity	Tighter spread (0.0481), higher MM participation (0.6201), and positive responsiveness (265.3) highlight liquidity and stability. Low skewness and kurtosis imply normal price behaviour, managed and balanced order flow.
GMM (2 Regimes)	Regime 0: Stable + Mean-Reverting + Mid-Liquidity	Lowest responsiveness (−887.4) despite moderate volume. This could indicate high internalisation or pegged liquidity. Medium spreads and high MM participation reflect efficiency.
	Regime 1: Volatile + Trending + Illiquid	Spread increases (0.0594), volume spikes (10.5), and responsiveness jumps to 8268 — a sign of hyper-reactive prices. High kurtosis (4.21) implies extreme movement potential. Driven by news or aggressive flow.
Bayesian GMM (20 Regimes)	Regime 4: Volatile + Mean-Reverting + Illiquid	Very high volatility (0.0001589), wide spread (0.0695), and high kurtosis (8.6). MM participation is low (0.3431). This regime captures uncertainty and potential price reversals in low liquidity conditions.
	Regime 15: Volatile + Trending + Liquid	High return (0.0001115), massive responsiveness (51,310), and high volume (15.94). Slope features show balance, but behaviour suggests breakouts or momentum surges likely from directional flows.
	Regime 5: Stable + Neutral + Illiquid	Widest spread (0.0834), extremely negative ask slope (−1.285), and moderate responsiveness (55.4). Illiquidity on the ask side indicates vulnerability to sell-side pressure.
	Regime 14: Moderate Volatility + Neutral + Liquid	Highest cumulative volume (16.63), tight spread, and average MM participation. Suggests a high activity, low impact environment, possibly driven by two-way institutional flow.
HDBSCAN (2 Regimes)	Volatile + Mean Reverting + Illiquid	Spread = 0.0585, MM participation = 0.2171, responsiveness = −731.6. Despite low volatility, the market reacts poorly to trades, shallow liquidity.
	Stable + Trending + Mid-Liquidity	Tighter spread (0.0506), MM activity is high (0.9768), and responsiveness = 1169. High MM presence stabilises trend-following behaviour. Likely represents structured market-making zones.
Neural Net (2 Regimes)	Regime 0: Volatile + Mean Reverting + Liquid	Balanced MM presence (0.5616), average spread, low skew and moderate kurtosis. Responsiveness is positive (118.8), indicating that price does respond to flow, but in a controlled fashion
	Regime 1: Stable + Trending + Illiquid	Wide spread (0.0799), lower depth slopes, high kurtosis (5.8). Suggests slower-moving prices dominated by sparse yet impactful trades.
HMM (8 Regimes)	Stable + Trending + Mid-Liquidity	High ask slope (1.566) implies strong resistance above. Moderate MM involvement and responsiveness indicate a slow accumulation phase.
	Stable + Trending + Illiquid	Highest spread (0.0718), low volume, and strong negative ask slope (−2.125). Illiquid upside resistance and potential for trapping trends.
	Vollatile + Mean-Reverting + Illiquid	High kurtosis (7.02), lower MM activity, and high responsiveness — classical signature of a panic regime or flash crash recovery zone.









For K-Means, GMM, HDBSCAN, and Neural Networks,, the Transition Probabilities are plotted