Regime Detection via Unsupervised Learning from Order Book and Volume Data

Problem Statement

The goal is to segment the market into distinct behavioral regimes using unsupervised learning, based on real-time order book and volume data. The segmentation aims to capture:

- Trending vs. Mean-reverting behavior
- Volatile vs. Stable conditions
- Liquid vs. Illiquid states

This is achieved by extracting and clustering meaningful features from the order book and volume, allowing us to identify and interpret different market regimes.

Feature Engineering

To capture the microstructure dynamics of the market, several hand-crafted features were engineered:

- Liquidity & Depth Features:
  - Bid/Ask Spread: spread=ask1−bid1spread=ask1−bid1
  - Order Book Imbalance (Level 1): imbalance_lvl1=bid_qty1−ask_qty1bid_qty1+ask_qty1imbalance_lvl1=bid_qty1+ask_qty1bid_qty1−ask_qty1
  - Microprice: microprice=bid1×ask_qty1+ask1×bid_qty1bid_qty1+ask_qty1microprice=bid_qty1+ask_qty1bid1×ask_qty1+ask1×bid_qty1
  - Cumulative Depth: Summing bid and ask quantities across all levels (e.g., cum_bid_qty, cum_ask_qty)
- Volatility & Price Action:
  - Rolling Mid-Price Return: log(midt/midt−1)log(mid*t*/mid*t*−1)
  - Price Volatility: Standard deviation of returns over short windows (e.g., 10s, 30s)
- Volume Features:
  - Volume Imbalance: Difference between buy and sell volumes
  - Cumulative Volume: Aggregated over recent time windows
  - VWAP Shift: Change in VWAP over short windows
- Derived Features:
  - Sloped Depth: Measures how quickly liquidity decays away from the top of the book
  - Trade Wipe Level: Average order book levels wiped by trades over short durations

These features were computed at each timestamp, providing a comprehensive snapshot of market conditions.

Clustering Approach

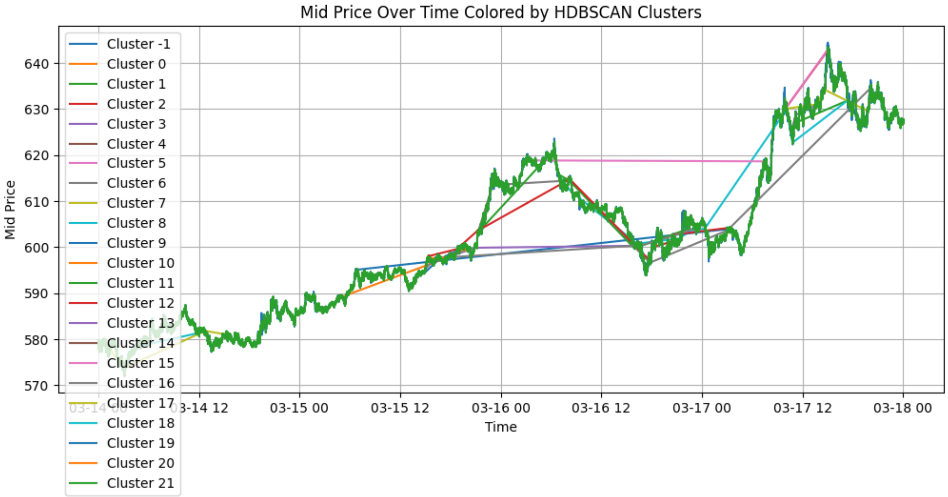
- **Clustering Algorithm:** HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) was used due to its ability to find clusters of varying densities and handle noise/outliers.
- **Dimensionality Reduction:** UMAP (Uniform Manifold Approximation and Projection) was applied for visualization and to facilitate clustering in a lower-dimensional space.
- **Clustering Metric:** The clustering was performed on the engineered features, capturing both price action and liquidity/volume structure.

- **Clustering Results & Visualizations**

1. Market Regimes Over Time

[Mid Price Over Time Colored by HDBSCAN Clusters]

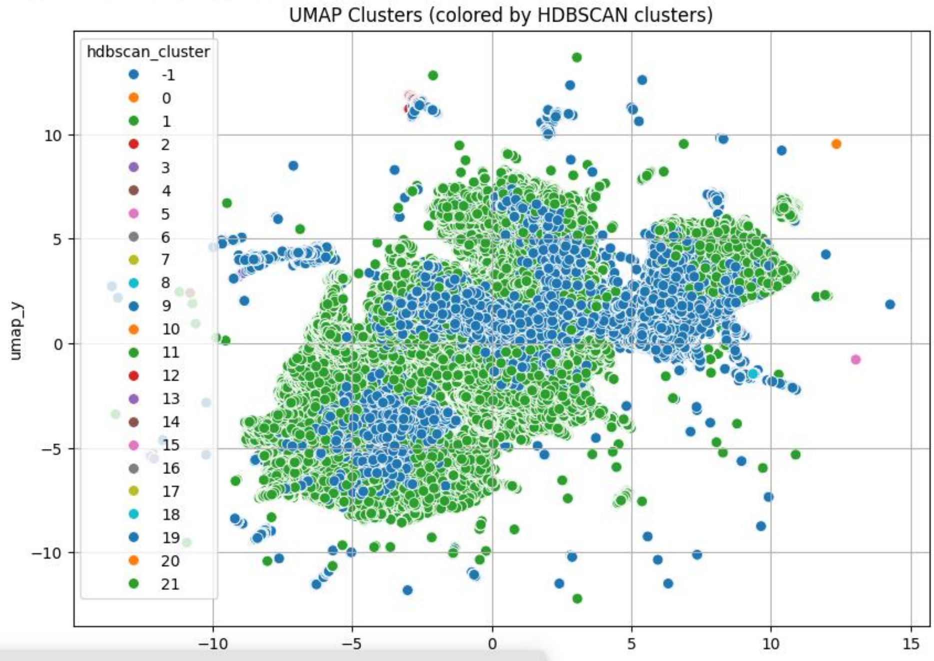
- The time series plot shows the mid-price trajectory, colored by detected HDBSCAN clusters.
- Distinct segments correspond to different regimes, reflecting changes in price behavior and liquidity.
- Transitions between clusters often align with shifts in price trend, volatility, or liquidity conditions.



2. Cluster Structure in Feature Space

[UMAP Clusters (colored by HDBSCAN clusters)]

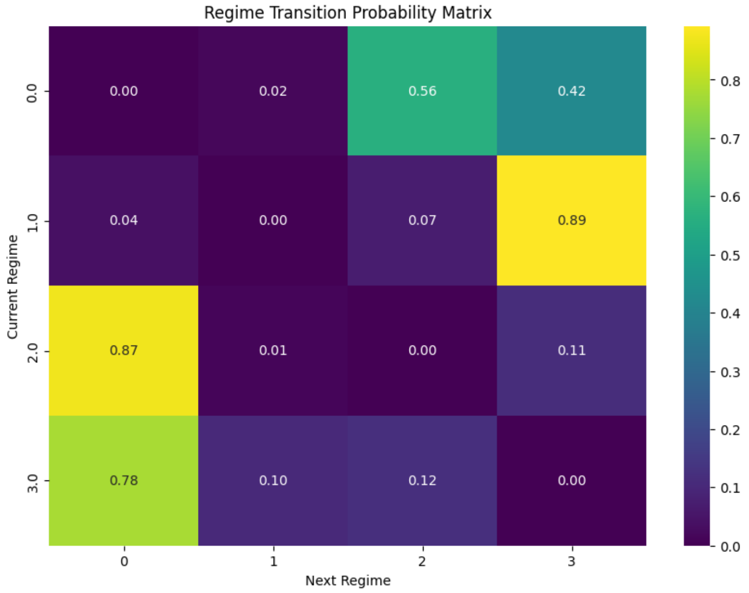
- The UMAP projection visualizes the distribution of data points in the engineered feature space.
- Clusters are well-separated, indicating that the features capture meaningful distinctions between regimes.
- The largest clusters (e.g., clusters 0 and 1) likely represent the dominant market regimes, while smaller clusters may correspond to rare or transitional states.



3. Regime Transition Dynamics

[Regime Transition Probability Matrix]

- The regime transition matrix quantifies the probability of moving from one regime to another.
- High diagonal values (e.g., 0.89 for regime 1) indicate persistence, meaning the market tends to remain in the same regime.
- Off-diagonal probabilities reveal which transitions are most likely, offering insight into regime switching behavior (e.g., regime 2 often transitions to regime 0).



- **Regime Insights**

- **Trending vs. Mean-Reverting:** Some clusters align with trending price movements (sustained up/down moves), while others coincide with mean-reverting or range-bound behavior.
- **Volatile vs. Stable:** Clusters with higher volatility features correspond to more erratic price action, while stable regimes show tighter spreads and lower return variance.
- **Liquid vs. Illiquid:** Regimes with high cumulative depth and low spread are identified as liquid, while those with thin order books and wide spreads are illiquid.

Custom Features Impact:
Features like microprice, order book imbalance, and cumulative depth were crucial in distinguishing between liquid/illiquid and trending/mean-reverting regimes. The inclusion of rolling volatility and volume imbalance enabled the identification of volatile versus stable periods.

- **Conclusion**

By engineering microstructure-informed features and applying HDBSCAN clustering, we successfully segmented the market into interpretable regimes. These regimes capture the essential axes of market behavior—trend, volatility, and liquidity—providing a robust foundation for downstream tasks such as trading strategy adaptation or risk management.