Regime Detection via Unsupervised Learning from Order Book and Volume Data

1. Introduction

This report summarizes the methodology and insights from identifying market regimes using unsupervised learning. We used both order book depth and trade volume data to uncover hidden states such as trending, mean-reverting, volatile, stable, liquid, and illiquid market conditions.

---

2. Feature Engineering

Features were extracted from:

- **Trade Volume Files**
- **Order Book (depth20) Snapshots**

Feature	Description
log_return_x/y	Log returns of mid-price
volatility_30s_x/y	Rolling volatility over 30s
volume_10s	10-second rolling trade volume
volume_imbalance	Net bid-ask volume imbalance
obi_5	Order Book Imbalance over top 5 levels
spread	Bid-ask spread
microprice	Volume-weighted average of top bid/ask
rolling_skew	Skewness of recent returns
rolling_kurt	Kurtosis of recent returns

---

3. Clustering Pipeline

Dimensionality Reduction:

- **PCA** reduced feature space while retaining ~95% variance.

Clustering Models:

- **KMeans**: Silhouette Score = 0.2058
- **GMM**: Silhouette Score = 0.0236
- **HDBSCAN**: Silhouette Score = 0.3210 ✅ (Best)

HDBSCAN was selected due to its superior clustering quality.

---

4. Regime Interpretation

Regimes were labeled by analyzing the average feature profiles for each cluster:

Regime	Label	Traits
0	Trending, Illiquid, Stable	Low volume, stable price, wide spread
1	Mean-Reverting, Liquid, Stable	Tight spread, low volatility
2	Trending, Liquid, Volatile	High microprice action, high volatility
3	Mean-Reverting, Illiquid, Volatile	High spread, noisy movement

---

5. Visual Insights

A. Regime Over Time

- Plotted regime labels against time to show how the market evolves.

B. Price and Volatility Overlay

- Regime labels overlaid on microprice and volatility series to validate clustering alignment with real behavior.

C. Regime Transition Matrix

- First-order transition matrix revealed regime shift probabilities:
    - Stable regimes persist (highs self-probability)
    - Volatile-illiquid regimes occasionally precede stable-liquid phases

---

6. Conclusion

This project applied clustering techniques on engineered features to uncover latent market regimes. HDBSCAN effectively separated market conditions, with transitions and labeling offering insights into the dynamics of financial microstructure. This framework is useful for building adaptive trading strategies or real-time monitoring tools.