Hybrid Wasserstein–HMM Regime Inference:
A Probabilistic–Distributional Framework for Market Structure Detection

Jeevan Jonas

November 2025

Abstract

Financial markets exhibit non-stationary and regime-dependent behavior that cannot be captured by purely statistical or purely distributional models alone.

The **Hybrid Wasserstein–HMM framework** combines the temporal probabilistic inference of Hidden Markov Models (HMM) with the shape sensitivity of Wasserstein distance clustering to identify evolving market regimes such as Trending, Range-bound, and Choppy states.

The approach unites two complementary paradigms: HMM models the hidden temporal dynamics of market states, while Wasserstein clustering measures distributional similarity across short sliding windows. Together, they form a robust regime inference engine that adapts to both slow structural transitions and abrupt behavioral changes, offering smoother state continuity than naive clustering and sharper boundary recognition than classical HMMs.

1. Introduction

Market regimes—broadly categorized into trending, ranging, or volatile (choppy) phases—are fundamental to tactical trading and risk allocation. Detecting such regimes in real time remains challenging because markets are non-stationary, heteroscedastic, and path-dependent.

Traditional methods fall into two camps: statistical models such as HMMs and Kalman filters that learn temporal transitions but assume Gaussian stationarity; and unsupervised clustering methods such as k-means that detect structural patterns but ignore time continuity.

The proposed **Hybrid Wasserstein–HMM (WHMM)** method bridges these domains by combining temporal likelihood modeling with distributional awareness, yielding a more stable and interpretable market state framework.

2. Theoretical Basis

2.1 Hidden Markov Model (HMM) Layer

The HMM is trained on sequential feature vectors $X_t \in \mathbb{R}^d$ representing market microstructure indicators—returns, volatility, ADX, ATR, and range ratios. Each HMM

component S_i models a distinct hidden regime via a Gaussian mixture, with transition probabilities

$$A_{ij} = P(S_t = j | S_{t-1} = i)$$

and emission probabilities computed through Gaussian densities.

Inference is performed via the forward-backward algorithm, yielding posterior probabilities:

$$\gamma_t(i) = P(S_t = i | X_{1:T})$$

and corresponding entropy and confidence measures used for filtering ambiguous segments. The HMM captures temporal persistence and soft regime boundaries, ensuring continuity in the inferred states.

2.2 Wasserstein Clustering Layer

While the HMM operates on sequential probabilities, the Wasserstein clusterer analyzes distributions of short-term feature windows. For each local buffer $W_t = \{X_{t-k}, \ldots, X_t\}$, the first-order Wasserstein distance is:

$$\text{dist}(a, b) = \mathbb{E}[|F_a^{-1}(u) - F_b^{-1}(u)|]$$

where F_a^{-1} is the quantile function. Distributions are grouped based on shape similarity, allowing the system to recognize volatility compression, symmetric oscillations, or noisy bursts.

Clusters are ordered by variance and mean slope, yielding semantic mapping:

- Cluster 0 → Trending-like (low variance, consistent slope)

- Cluster 1 → Balanced/Range

- Cluster 2 → Noisy/Choppy

3. Training Pipeline

3.1 Feature Engineering

The trainer extracts nine standardized features across a moving window:

- Return, Slope, R^2 – directional momentum and regression quality

- ADX, ATR, Range ratio – strength and magnitude of move

- RangeVol, Volatility, ATR normalized – distributional spread

All features are scaled via `StandardScaler` for numerical stability and transformed into stationary representations (log returns, normalized ratios).

3.2 HMM Training

The Gaussian HMM is trained on historical intraday data using Expectation-Maximization (EM). Number of hidden states $K = 3$ represents Trending, Range, and Choppy regimes.

Training minimizes negative log-likelihood:

$$\mathcal{L} = -\sum_t \log P(X_t|\theta)$$

where θ represents transition and emission parameters.

3.3 Wasserstein Clustering

Simultaneously, feature distributions over short windows are clustered using the `WassersteinClusterer`. Cluster centroids are refined iteratively until convergence. Final centroids are stored for inference-time matching.

4. Inference Logic

4.1 Hybrid Posterior Integration

During live inference:

1. The HMM produces posterior probabilities $P(S_t|X_t)$.

2. Entropy and confidence are computed to determine label quality.

3. Wasserstein cluster label C_t is computed from local feature distributions.

4. Both are fused through rule-based logic:

Examples:

- If HMM = Range and $C_t = 0 \rightarrow$ Mild-Uptrend

- If HMM = Trending and $C_t = 2 \rightarrow$ Transitional

- If HMM = Choppy and $C_t = 1 \rightarrow$ Range

This combination preserves the temporal coherence of HMM and the structural awareness of Wasserstein clustering.

4.2 Multi-Scale Smoothing

The inference engine employs three-tier temporal smoothing—short (30m), medium (1h), long (2h)—with exponential weighting:

$$P_t^{smooth} = \alpha P_t + (1 - \alpha)P_{t-1}^{smooth}$$

and regime decisions made hierarchically by the RegimeGovernor enforcing persistence (minimum hold time). This prevents noise-driven state oscillation.

5. Strengths

- **Temporal stability:** HMM ensures regime persistence across time.

- **Distributional adaptability:** Wasserstein distance captures shape changes.

- **Interpretable states:** Regime labels align with intuitive price patterns.

- **Multi-scale coherence:** Integrating short and long horizons improves robustness.

- **Modularity:** Trainer and inference modules are decoupled for reuse.

6. Limitations

- **Data dependency:** Requires clean, continuous intraday data.

- **Model rigidity:** Gaussian assumptions may under-fit extreme volatility.

- **No causal inference:** The model identifies patterns, not drivers.

- **Detection latency:** Multi-scale smoothing introduces lag.

- **Static clusters:** Wasserstein centroids require periodic retraining.

7. Empirical Observations

In backtests and live sessions on NIFTY futures:

- The hybrid model produced fewer regime reversals than HMM-only baselines.

- Wasserstein overlay improved sensitivity to volatility transitions.

- Enhanced regime persistence during trending hours.

- Subjective validation showed regime boundaries aligned with major price inflections.

8. Conclusion

The Hybrid Wasserstein–HMM framework represents a probabilistic–distributional synthesis combining sequential inference with statistical geometry. It balances continuity and sensitivity—smooth enough to avoid whipsaws, sharp enough to detect emergent structure. Its ability to interpret, modular design, and compact footprint make it ideal for educational, research, or semi-automated strategy selection frameworks. While not a trading signal in itself, it provides an invaluable layer of market context awareness.

Disclaimer

This model is provided strictly for **educational and research purposes**. It is not financial advice nor a substitute for professional investment judgment. The author assumes no responsibility for financial losses incurred from trading decisions derived from this methodology.