Market Regime Detection with MMD^2

Elijah Whitham-Powell

Portland State University

December 9, 2025

Overview

Kernel Mean Embedding: Definition

Given a kernel K defined on a topological set $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$, the mean embedding of a *Borel* probability distribution $\mathbb{P}$ on $\mathcal{X}$ is the function $\mu_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}$ defined as:

$$\mu_{\mathbb{P}}(y) = \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)]$$

- For any x, $x' \in \mathcal{X}$,

$$K(x, x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$$

- The kernel trick: For any $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$K(x, x') = \langle f, K_x \rangle_{\mathcal{H}}$$

- For any Borel measure $\mathbb{P}$ and $\mathbb{Q}$,

$$\mathbb{E}_{(X,Y) \sim \mathbb{P},\mathbb{Q}}[K(X, Y)] = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}$$

- The kernel trick for expectations: For any $f \in \mathcal{H}$ and Borel measure $\mathbb{P}$,

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$$

Kernel Mean Embedding: Expectation evaluation in an RKHS

- Expectations of all RKHS functions in $\mathcal{H}$ can be evaluated using the result of the kernel trick.

$$\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$$

- The kernel mean embedding: $\mu_{\mathbb{P}}(y) = \mathbb{E}_{X \sim \mathbb{P}}[K(X, y)]$
- The kernel trick: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$
- The kernel mean embedding can be estimated using the empirical mean of N samples from $\mathbb{P}$:

$$\hat{\mu}_{\mathbb{P}}(x) = \frac{1}{N} \sum_{i=1}^{N} K(X_i, x), \quad X_i \overset{iid}{\sim} \mathbb{P}$$

Kernel Mean Embedding: Does it exist?

If $\mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{K(X,X)}\right] < \infty$, then there exists a unique $\mu_{\mathbb{P}} \in \mathcal{H}$ such that:

$$\mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$$

Let $T_{\mathbb{P}}f = \mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right]$. By assumption, $T_{\mathbb{P}}f$ is bounded:

$$\begin{aligned} |T_{\mathbb{P}}f| &= |\mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right]| \\ &\leq \mathbb{E}_{X\sim\mathbb{P}}\left[|f(X)|\right] \\ &= \mathbb{E}_{X\sim\mathbb{P}}\left[|\langle f, K_x\rangle_{\mathcal{H}}|\right] \\ &\leq \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{K(X,X)}\right] \|f\|_{\mathcal{H}} \end{aligned}$$

By Riez's theorem, there exists $\mu_{\mathbb{P}} \in \mathcal{H}$ such that $T_{\mathbb{P}}f = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$.

Intuition Behind Kernel Mean Embeddings

- **Compact Representation:** Maps a distribution $\mathbb{P}$ to a single point $\mu_{\mathbb{P}}$ in the RKHS.
- **Efficient Computation:** Use the kernel trick to compute in high-dimensional spaces implicitly.
- **Key Characteristics:** The embedding $\mu_{\mathbb{P}}$ captures the essential features or "fingerprint" of $\mathbb{P}$.
- **Comparison Ready:** This sets the stage for comparing distributions (e.g., using MMD).

What is MMD?

The maximum mean discrepancy (MMD) is the distance between mean embeddings,

$$
\begin{aligned}
MMD^2 &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 \\
&= \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \underbrace{\mathbb{E}_{X,X'\sim\mathbb{P}}\left[k(X,X')\right]}_{\text{(i)}} + \underbrace{\mathbb{E}_{Y,Y'\sim\mathbb{Q}}\left[k(Y,Y')\right]}_{\text{(i)}} - \underbrace{2\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}\left[k(X,Y)\right]}_{\text{(ii)}}
\end{aligned}
$$

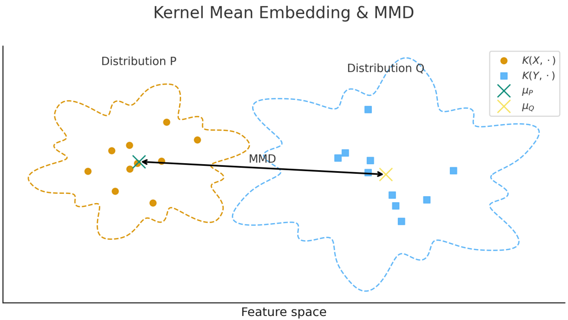
(i) within-distribution similarity

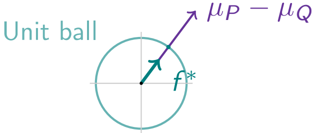
(ii) between-distribution similarity

Intuition Behind MMD

- **Fingerprint Distance:** MMD measures the distance between the "fingerprints" (kernel mean embeddings) of two distributions.
- **Interpreting MMD:** A small MMD implies that the distributions are similar, a large MMD implies they are not.
- **Witness Function:** $f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$ is the direction that maximally distinguishes P from Q
- **Nonlinear Comparison:** The kernel trick allows MMD to capture complex, nonlinear differences.
- **Characteristic Kernels:** With a characteristic kernel, MMD is zero if and only if the distributions are identical.
    - Exponential, Gaussian, and others that can be proven to be characteristic if the mapping is injective.

Visual Representation of MMD





$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

f^* maximally distinguishes P from Q

Estimating MMD

Given samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$, the empirical MMD is:

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = \frac{1}{n^2}\sum_{i,j} k(X_i, X_j) + \frac{1}{m^2}\sum_{i,j} k(Y_i, Y_j) - \frac{2}{nm}\sum_{i,j} k(X_i, Y_j)$$
$$= \frac{1}{n^2}\sum_{i,j} k(X_i, X_j) + \frac{1}{m^2}\sum_{i,j} k(Y_i, Y_j) - \frac{2}{nm}\sum_{i,j} k(X_i, Y_j)$$

The empirical MMD is a biased estimator of the true MMD. The bias can be corrected by using the unbiased estimator:

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = \frac{1}{n(n-1)}\sum_{i \neq j} k(X_i, X_j) + \frac{1}{m(m-1)}\sum_{i \neq j} k(Y_i, Y_j) - \frac{2}{nm}\sum_{i,j} k(X_i, Y_j)$$

Applications of MMD

Two-sample testing:

- Test whether two samples come from the same distribution
- MMD as test statistic + permutation test for significance

Generative models:

- Evaluate quality of generated samples
- MMD as training loss (MMD-GANs)

This project: Sliding-window two-sample tests for regime detection

Motivation: Why Detect Regime Changes?

- Financial markets exhibit **non-stationary behavior**
- Periods of qualitatively different dynamics: bull markets, crashes, recovery
- Traditional approaches assume parametric models (e.g., HMM with Gaussian emissions)

Kernel methods approach:

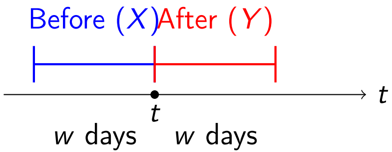
- Compare full distributions before and after each time point
- Detect changes *without specifying* what form they take
- Nonparametric: let the data speak

Sliding Window MMD

At each candidate change point t:

1. Extract **before** window: $X = \{x_{t-w}, \dots, x_{t-1}\}$
2. Extract **after** window: $Y = \{x_t, \dots, x_{t+w-1}\}$
3. Compute $\widehat{\text{MMD}}^2(X, Y)$
4. Test significance via permutation test

Significant MMD $\Rightarrow$ distributional shift at time t



Slide window by step size s
Repeat for all t

Permutation Test for Significance

Under H_0: distributions before and after are identical ($P = Q$)

Procedure:

1. Pool samples: $Z = X \cup Y$
2. For $b = 1, \ldots, B$:
   - Randomly permute Z
   - Split into pseudo-samples X', Y'
   - Compute $\widehat{\text{MMD}}_b^2(X', Y')$
3. Compute p-value: $\hat{p} = \frac{1}{B} \sum_{b=1}^{B} 1\left[\widehat{\text{MMD}}_b^2 \geq \widehat{\text{MMD}}_{\text{obs}}^2\right]$

Alternative metric: Standard deviations from null mean

$$z = \frac{\widehat{\text{MMD}}_{\text{obs}}^2 - \bar{\mu}_{\text{null}}}{\hat{\sigma}_{\text{null}}}$$

More informative when p-values cluster near zero.

Feature Representation

Input: Daily OHLCV data for SPY (S&P 500 ETF), 2020–2024

Features per day:

- Log prices: $\log(\text{Open}), \log(\text{High}), \log(\text{Low}), \log(\text{Close})$
- Log volume: $\log(\text{Volume})$

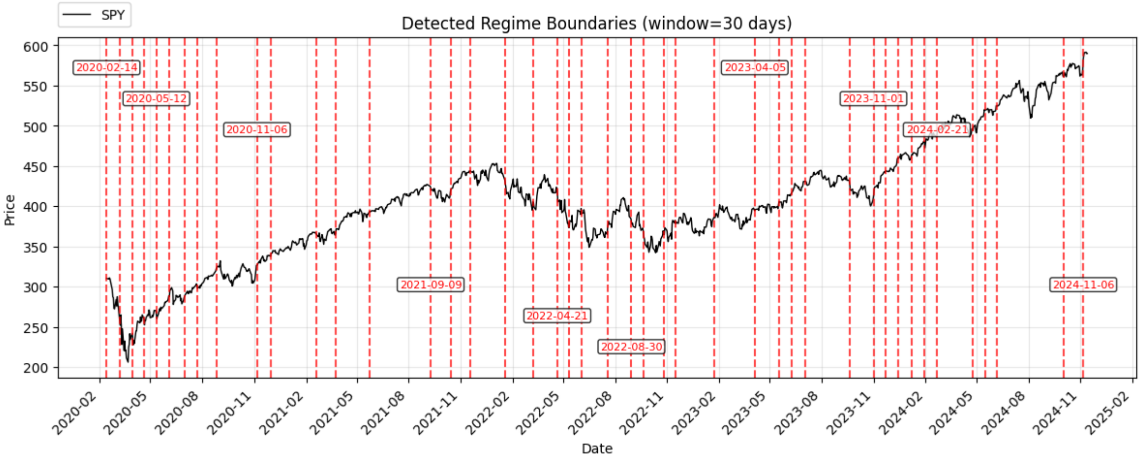
Preprocessing:

- Standardize each feature (zero mean, unit variance)
- Prevents high-magnitude features (e.g., volume) from dominating kernel distances

Kernel: RBF with median heuristic bandwidth

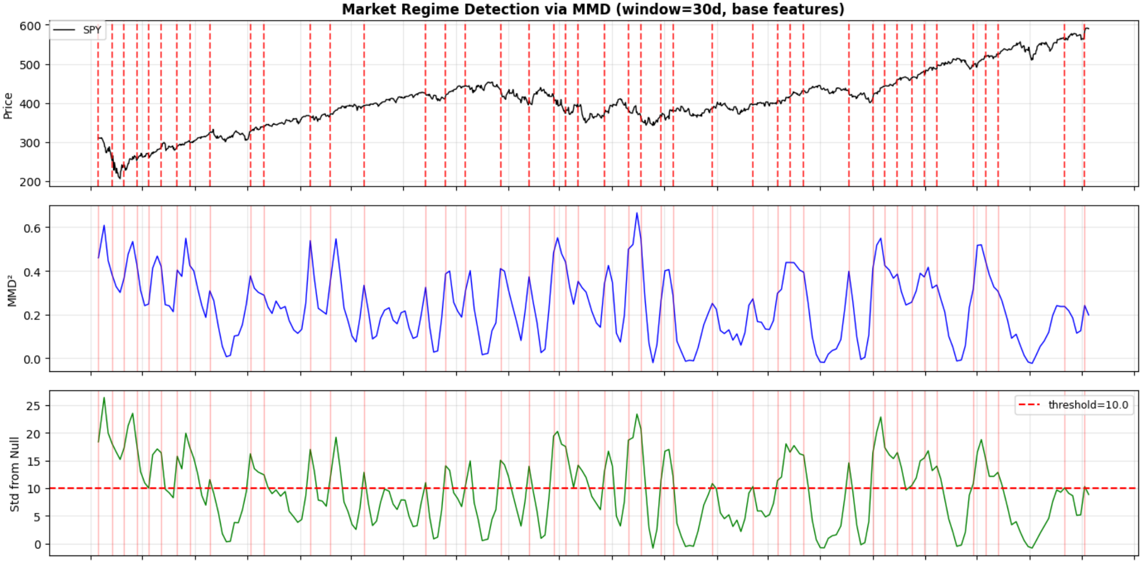
$$k(x, y) = \exp(-\gamma\|x - y\|^2), \quad \gamma = \frac{1}{2\sigma^2}, \quad \sigma = \text{median}(|X - \text{median}(X)|)$$

Detected Regime Boundaries



Window = 30 days, Step = 5 days, Threshold = 10 std from null

MMD Statistic Over Time



Validation Against Known Events

Detected Boundary	Market Event
Feb 2020	COVID-19 crash onset
Mar–Apr 2020	Fed intervention / recovery begins
Jan 2022	Start of 2022 bear market
Jun 2022	Mid-2022 volatility spike
Oct 2022	2022 market bottom
Oct–Nov 2023	Bull market acceleration

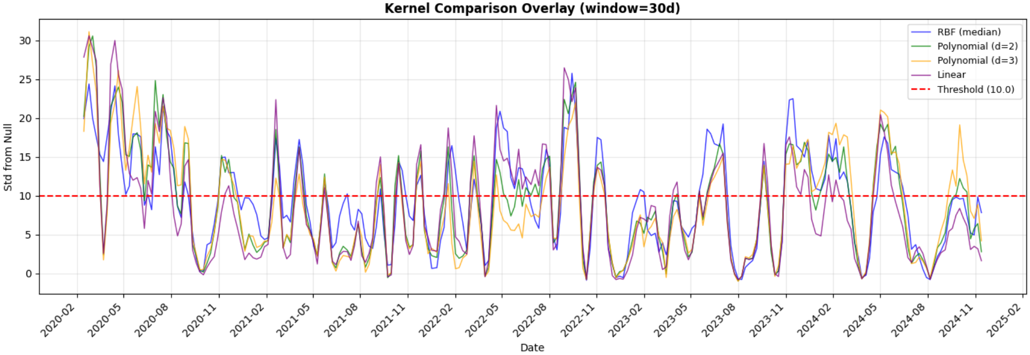
Detected boundaries correspond to **genuine market events**, not spurious statistical artifacts.

The Three Knobs

Parameter	Affects	Recommendation
Kernel	What differences are captured	RBF + median heuristic
Window size	Statistical power	30–60 days
Step size	Resolution, smoothing, runtime	5 days

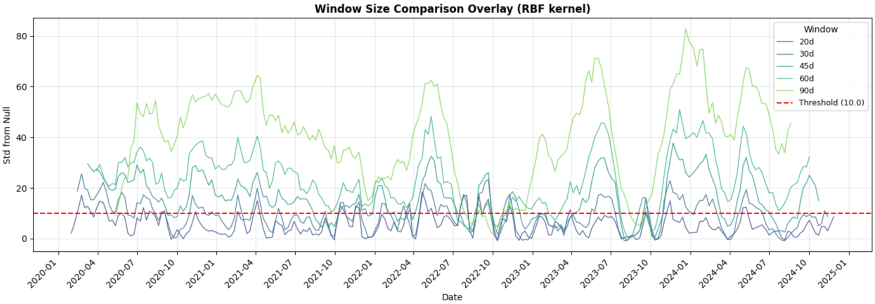
- **Kernel choice:** All kernels detect major events (COVID crash)
- **Window size:** Larger windows $\Rightarrow$ more statistical power $\Rightarrow$ more detections
- **Step size:** Larger steps $\Rightarrow$ implicit smoothing $\Rightarrow$ fewer detections, faster

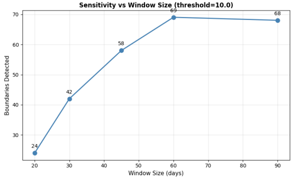
Kernel Comparison



RBF, Polynomial (d=2, d=3), Linear — all detect COVID crash

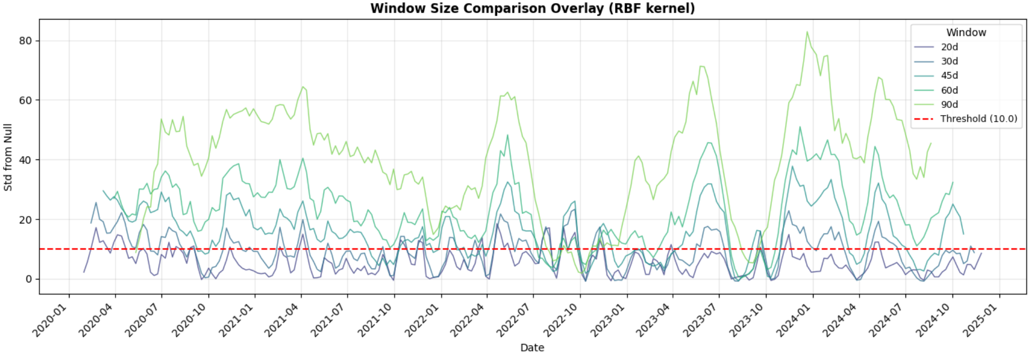
Window Size Effect





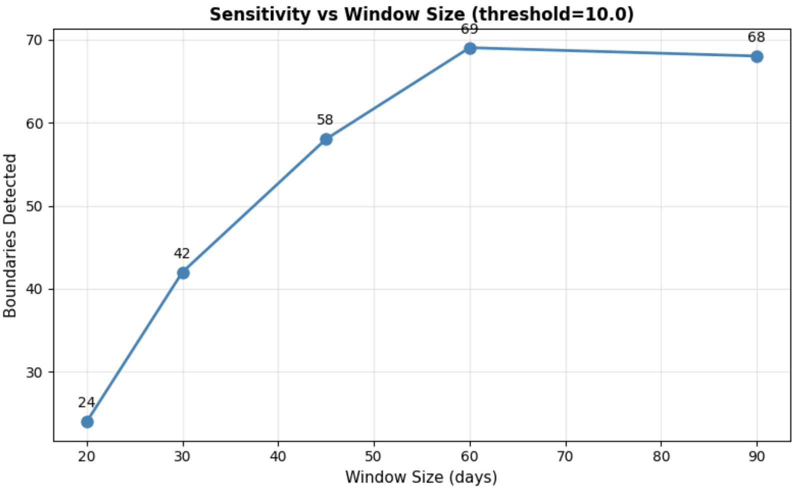
Larger windows have more statistical power (tighter null distribution)

Window Size Effect



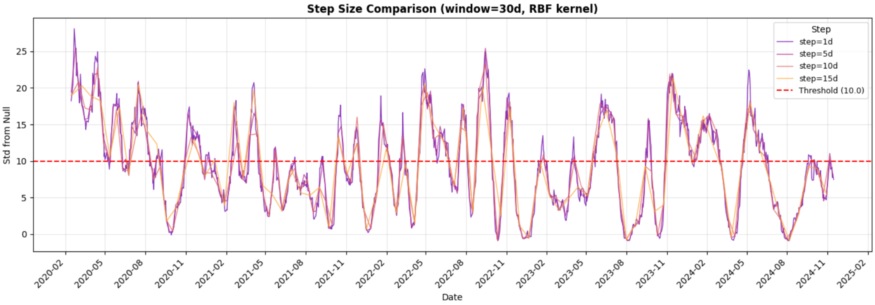
Larger windows have more statistical power (tighter null distribution)

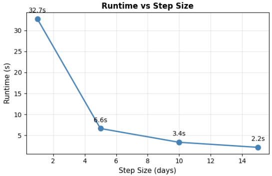
Window Size Effect

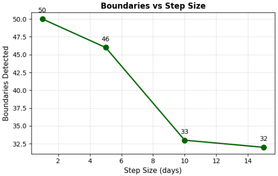


Larger windows have more statistical power (tighter null distribution)

Step Size Effect

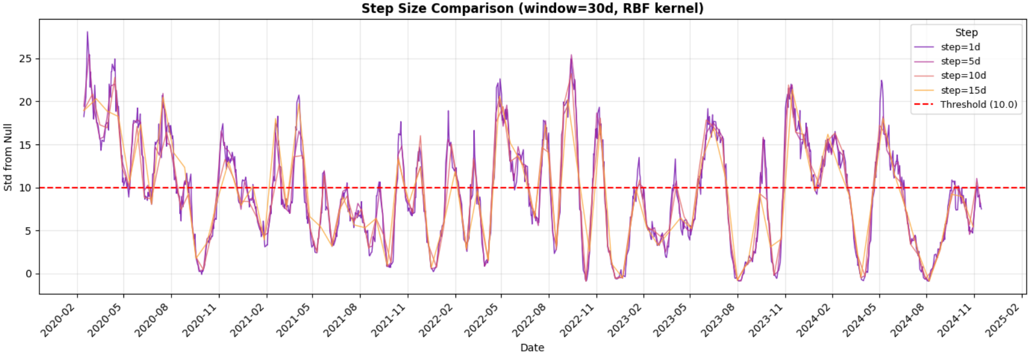






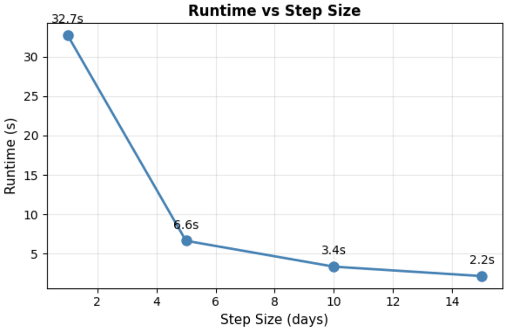
Step size is a runtime vs. resolution trade-off (also reduces noise)

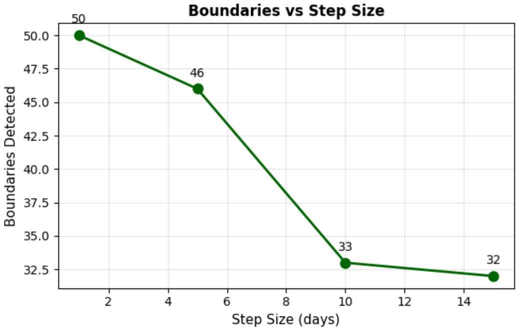
Step Size Effect



Step size is a runtime vs. resolution trade-off (also reduces noise)

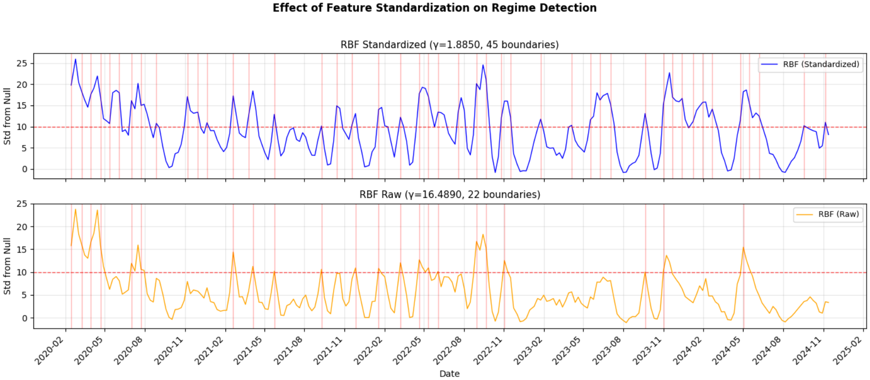
Step Size Effect





Step size is a runtime vs. resolution trade-off (also reduces noise)

Effect of Feature Standardization



- Without standardization: volume (∼18–20) dominates price features (∼5–6)
- With standardization: all features contribute equally
- **Recommendation:** Always standardize for multi-feature inputs

Future Work: Which Features Matter?

Once regimes are detected, which features **discriminate** them?

Kernel-Target Alignment (KTA):

$$\text{KTA}(K, Y) = \frac{\langle K, YY^\top \rangle_F}{\|K\|_F \cdot \|YY^\top\|_F}$$

- Measures how well kernel matrix aligns with regime labels
- Optimize **ARD kernel** bandwidths via gradient ascent on KTA:

$$k(x, y) = \exp\left(-\sum_{d=1}^{D} \frac{(x_d - y_d)^2}{2\sigma_d^2} \right)$$

- Features with small σ_d are most discriminative

Goal: Identify whether volatility, momentum, or price structure best characterizes detected regimes.

Future Work: Regime-Specific Prediction

Question: Do predictive relationships change across regimes?

Approach:

1. Fit global model: Kernel Ridge Regression on all data
2. Fit regime-specific models: Separate KRR per detected regime
3. Compare prediction error on held-out data

If regime-specific models outperform:

- Evidence that predictive structure genuinely differs across regimes
- Motivation for adaptive/switching models in practice

Conclusion

MMD for Regime Detection:

- Nonparametric: detects distributional shifts without specifying the form
- Validated: detected boundaries correspond to known market events
- Tunable: window size, step size, threshold control sensitivity

Key Findings:

- All kernels detect major events (robustness)
- Larger windows $\Rightarrow$ more statistical power
- Feature standardization is essential

Code: `github.com/whitham-powell/mmd-regime-change`

References I

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2012). *A Kernel Two-Sample Test*. JMLR, 13:723–773.

Muandet, K., Fukumizu, K., Sriperumbudur, B., & Schölkopf, B. (2017). *Kernel Mean Embedding of Distributions: A Review and Beyond*. Foundations and Trends in ML, 10(1–2):1–141.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). *On Kernel-Target Alignment*. NIPS 2001.

Harchaoui, Z. & Cappé, O. (2007). *Retrospective Multiple Change-Point Estimation with Kernels*. IEEE Workshop on Statistical Signal Processing.

Questions

Questions?

Code: `github.com/whitham-powell/mmd-regime-change`