Boundary-Aware Phase–Amplitude Decomposition for the Sign-Aware Jaccard Distance and a Jaccard Correlogram for Signals (with Multi-State Extension)

Anonymous Authors

September 25, 2025

Abstract

We present a boundary-aware generalization of the sign-aware Jaccard ("peak") similarity and its associated distance that decomposes disagreement into *boundary-side* (state) mismatches and *amplitude* imbalances. Unlike classical correlation that conflates phase opposition with magnitude differences, our framework provides an exact, interpretable decomposition at each lag. The construction yields a bounded, scale-invariant *Jaccard correlogram* in one and two dimensions, addressing key limitations of semivariograms. We extend the framework to a *multi-state* setting with strict partitioning (e.g., hurricane categories, credit ratings) where signals belong to exactly one state at each point. The resulting distance decomposes exactly into a category mismatch component and a within-category magnitude imbalance component. We demonstrate spatial isotropy/anisotropy diagnostics and provide kernel constructions based on the negative-type property of the distance.

1 Introduction and Motivation

1.1 The Problem with Classical Distances

Classical distance measures conflate fundamentally different types of disagreement. Consider two scenarios:

- **Scenario A**: Two temperature sensors both read below freezing (-5°C and -10°C)
- **Scenario B**: One sensor reads +5°C while the other reads -5°C

The Euclidean distance is 5°C in both cases, yet these represent qualitatively different situations. Scenario A shows magnitude disagreement within the same state (frozen), while Scenario B shows a state disagreement (frozen vs. unfrozen). This distinction matters critically in many domains:

- **Finance**: Gains vs. losses relative to a benchmark
- **Neuroscience**: Neural excitation vs. inhibition relative to baseline
- **Climate**: Wet vs. dry periods relative to seasonal averages
- **Image Analysis**: Dark vs. bright regions relative to background
- **Audio Processing**: Sound vs. silence, or loudness categories

1.2 Our Contribution

We develop a boundary-aware Jaccard distance that:

1. **Decomposes exactly** into interpretable components: boundary-side (state) mismatch and amplitude imbalance
2. **Remains bounded** in [0,1], enabling cross-dataset comparisons
3. **Exhibits scale invariance**, focusing on patterns rather than units
4. **Extends naturally** to multiple states with strict partitioning
5. **Generates valid kernels** through its negative-type property

1.3 Intuitive Preview

Before diving into formalism, here's the key insight: We measure how much two signals "overlap" when considering their excursions above and below boundaries separately.

Example 1 (Simple illustration)**.** *Consider daily stock returns for two stocks relative to zero (the natural boundary):*

- *Stock A: [+3%, +5%, -2%, -4%, +1%]*
- *Stock B: [+2%, +4%, +1%, -5%, +2%]*

At day 3, Stock A is down 2% while Stock B is up 1%. This is a state mismatch *– they're on opposite sides of zero. At day 1, both are positive (+3% vs +2%), showing only* magnitude imbalance *within the same state. Our framework quantifies these distinct types of disagreement separately.*

2 Mathematical Framework

2.1 Notation and Setup

Let $A, B \in \mathbb{R}^n$ be signals sampled at points $t_1, \ldots, t_n$. Let θ denote a *boundary* (threshold):

- In 1D: $\theta \in \mathbb{R}$ (constant) or $\theta(t)$ (time-varying)
- In 2D: $\theta(x, y)$ (spatial field)

Define boundary-centered signals:

$$\widetilde{A} := A - \theta, \qquad \widetilde{B} := B - \theta. \tag{1}$$

Define positive and negative parts relative to the boundary:

$$\widetilde{A}^+ = \max\{\widetilde{A}, 0\} \quad \text{(excursions above } \theta\text{)} \tag{2}$$

$$\widetilde{A}^- = \max\{-\widetilde{A}, 0\} \quad \text{(excursions below } \theta\text{)} \tag{3}$$

Note that $|\widetilde{A}| = \widetilde{A}^+ + \widetilde{A}^-$ and $\widetilde{A}^+ \cdot \widetilde{A}^- = 0$ pointwise.

2.2 Boundary-Aware Jaccard Similarity

Definition 1 (Intersection and Union)**.** *The boundary-aware intersection and union are:*

$$N_\theta(A,B) = \sum_{i=1}^{n} \Big(\min\{\widetilde{A}_i^+, \widetilde{B}_i^+\} + \min\{\widetilde{A}_i^-, \widetilde{B}_i^-\} \Big), \tag{4}$$

$$U_\theta(A,B) = \sum_{i=1}^{n} \Big(\max\{\widetilde{A}_i^+, \widetilde{B}_i^+\} + \max\{\widetilde{A}_i^-, \widetilde{B}_i^-\} \Big). \tag{5}$$

Intuition: We compute overlap separately for excursions above and below θ, then combine. If both signals are above θ by similar amounts, they contribute to intersection. If one is above and the other below, there's no intersection at that point.

Definition 2 (Jaccard Similarity and Distance)**.**

$$J_{\text{peak}}^{(\theta)}(A,B) = \begin{cases} \dfrac{N_\theta(A,B)}{U_\theta(A,B)}, & U_\theta(A,B) > 0, \\ 1, & U_\theta(A,B) = 0, \end{cases} \qquad d_{\text{peak}}^{(\theta)}(A,B) = 1 - J_{\text{peak}}^{(\theta)}(A,B). \tag{6}$$

3 Exact Decomposition: The Core Result

3.1 Partitioning by Boundary Side

Partition indices based on whether signals are on the same or opposite sides of θ:

$$S_{\text{same}}^{(\theta)} = \{i : \ \widetilde{A}_i \, \widetilde{B}_i \geq 0\} \quad \text{(same side of } \theta\text{, including zeros)} \tag{7}$$

$$S_{\text{opp}}^{(\theta)} = \{i : \ \widetilde{A}_i \, \widetilde{B}_i < 0\} \quad \text{(opposite sides of } \theta\text{)} \tag{8}$$

3.2 Component Contributions

At each index:

- If $i \in S_{\text{opp}}^{(\theta)}$: intersection = 0, union = $|\widetilde{A}_i| + |\widetilde{B}_i|$
- If $i \in S_{\text{same}}^{(\theta)}$: intersection = $\min\{|\widetilde{A}_i|, |\widetilde{B}_i|\}$, union = $\max\{|\widetilde{A}_i|, |\widetilde{B}_i|\}$

Theorem 1 (Exact Decomposition)**.** *The distance decomposes exactly as:*

$$d_{\text{peak}}^{(\theta)}(A,B) = \pi_{\text{state}}^{(\theta)}(A,B) + \pi_{\text{mag}}^{(\theta)}(A,B), \tag{9}$$

where

$$\pi_{\text{state}}^{(\theta)} = \frac{1}{U_\theta} \sum_{i \in S_{\text{opp}}^{(\theta)}} \big(|\widetilde{A}_i| + |\widetilde{B}_i| \big) \quad \textit{(boundary mismatch fraction)} \tag{10}$$

$$\pi_{\text{mag}}^{(\theta)} = \frac{1}{U_\theta} \sum_{i \in S_{\text{same}}^{(\theta)}} \big| |\widetilde{A}_i| - |\widetilde{B}_i| \big| \quad \textit{(amplitude imbalance fraction)} \tag{11}$$

Proof. Using $\max\{a,b\} - \min\{a,b\} = |a-b|$ for $a, b \geq 0$:

$$U_\theta - N_\theta = \sum_{i \in S_{\text{same}}^{(\theta)}} \big| |\widetilde{A}_i| - |\widetilde{B}_i| \big| + \sum_{i \in S_{\text{opp}}^{(\theta)}} \big(|\widetilde{A}_i| + |\widetilde{B}_i| \big). \tag{12}$$

Dividing by U_θ yields the result. $\square$

Example 2 (Temperature Anomalies)**.** *Two weather stations measure daily temperature anomalies relative to seasonal average ($\theta = 0$):*

- *Station A: [+3°C, +5°C, -2°C, -4°C, +1°C]*
- *Station B: [+2°C, +4°C, -3°C, -5°C, +2°C]*

Analysis:

- *Days 1,2,3,4,5: All same-side (both positive or both negative)*
- *State mismatch contribution:* $\pi_{\text{state}} = 0$ *(no opposite-side days)*
- *Magnitude imbalance:* $|3-2| + |5-4| + |2-3| + |4-5| + |1-2| = 5$
- *Total union:* $(3+2) + (5+4) + (2+3) + (4+5) + (1+2) = 31$
- *Distance:* $d_{\text{peak}} = 5/31 \approx 0.16$ *(entirely from magnitude imbalance)*

This tells us the stations agree perfectly on warm vs. cold days but differ slightly in magnitudes.

4 Correlogram Extension (1D and 2D)

4.1 1D Jaccard Correlogram

For lag analysis, define the shifted similarity:

$$J_{AB}^{(\theta)}(\tau) := J_{\text{peak}}^{(\theta)}(A, T_\tau B), \quad \text{where } (T_\tau B)_i = B_{i-\tau} \tag{13}$$

The centered correlogram (analogous to autocorrelation):

$$C_{AB}^{(\theta)}(\tau) := 2J_{AB}^{(\theta)}(\tau) - 1 \in [-1, 1] \tag{14}$$

At each lag, the decomposition remains exact:

$$d_{\text{peak}}^{(\theta)}(A, T_\tau B) = \pi_{\text{state}}^{(\theta)}(\tau) + \pi_{\text{mag}}^{(\theta)}(\tau) \tag{15}$$

Interpretation:

- $C_{AB}^{(\theta)}(\tau) \approx 1$: Strong same-side agreement at lag τ
- $C_{AB}^{(\theta)}(\tau) \approx -1$: Strong anti-phase behavior at lag τ
- $C_{AB}^{(\theta)}(\tau) \approx 0$: No clear relationship at lag τ

4.2 2D Extension for Spatial Fields

For spatial fields $A(x, y)$, $B(x, y)$ with spatial lag $\boldsymbol{h} = (h_x, h_y)$:

$$J_{AB}^{(\theta)}(\boldsymbol{h}) = \frac{\iint [\min\{(A-\theta)^+, (T_{\boldsymbol{h}}B - \theta)^+\} + \min\{(A-\theta)^-, (T_{\boldsymbol{h}}B - \theta)^-\}]\, dx\, dy}{\iint [\max\{(A-\theta)^+, (T_{\boldsymbol{h}}B - \theta)^+\} + \max\{(A-\theta)^-, (T_{\boldsymbol{h}}B - \theta)^-\}]\, dx\, dy} \tag{16}$$

The decomposition $d_{\text{peak}}^{(\theta)}(\boldsymbol{h}) = \pi_{\text{state}}^{(\theta)}(\boldsymbol{h}) + \pi_{\text{mag}}^{(\theta)}(\boldsymbol{h})$ extends directly. Isotropy/anisotropy is diagnosed via level sets of $C_J^{(\theta)}(\boldsymbol{h})$ and component maps.

5 Multi-State Extension with Strict Partitioning

5.1 Motivation and Strict Partitioning Requirement

Many real-world applications involve multiple discrete states rather than binary above/below classifications:

Example 3 (Hurricane Categories). *Wind speeds define strict hurricane categories:*

- *Tropical Storm: [39, 74) mph*
- *Category 1: [74, 96) mph*
- *Category 2: [96, 111) mph*
- *Category 3: [111, 130) mph*
- *Category 4: [130, 157) mph*
- *Category 5: [157, ∞) mph*

A wind speed of 115 mph belongs to Category 3 and only *Category 3. The mathematics requires this exclusivity – a hurricane cannot be "partially Category 3 and partially Category 4."*

Definition 3 (Strict State Partitioning). *Let $\boldsymbol{\tau} = (\tau_1 < \tau_2 < \cdots < \tau_{K-1})$ be boundaries that partition $\mathbb{R}$ into K disjoint, exhaustive states:*

$$\mathcal{S}_1 = (-\infty, \tau_1) \tag{17}$$

$$\mathcal{S}_k = [\tau_{k-1}, \tau_k) \quad \text{for } k = 2, \ldots, K-1 \tag{18}$$

$$\mathcal{S}_K = [\tau_{K-1}, \infty) \tag{19}$$

Critical property*: For any value $v \in \mathbb{R}$, there exists* exactly one *$k \in \{1, \ldots, K\}$ such that $v \in \mathcal{S}_k$.*

5.2 Multi-State Construction

For signal value A_i, let $\text{State}_A(i) \in \{1, \ldots, K\}$ denote its state. Define a magnitude function $m(x) = |x - \text{ref}|$ where ref is a reference point (often 0 or a baseline).

Create one-hot channel representations:

$$m_{A,s}(i) = \begin{cases} m(A_i), & \text{if State}_A(i) = s \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

Define multi-state intersection and union:

$$N_{\boldsymbol{\tau}}(A, B) = \sum_{s=1}^{K} \sum_{i=1}^{n} \min\{m_{A,s}(i), m_{B,s}(i)\} \tag{21}$$

$$U_{\boldsymbol{\tau}}(A, B) = \sum_{s=1}^{K} \sum_{i=1}^{n} \max\{m_{A,s}(i), m_{B,s}(i)\} \tag{22}$$

Proposition 1 (Multi-State Exact Decomposition)**.** *The multi-state distance decomposes exactly as:*

$$d_{\text{peak}}^{(\boldsymbol{\tau})}(A,B) = \pi_{\text{state}}^{(\boldsymbol{\tau})}(A,B) + \pi_{\text{mag}}^{(\boldsymbol{\tau})}(A,B), \tag{23}$$

where

$$\pi_{\text{state}}^{(\boldsymbol{\tau})} = \frac{1}{U_{\boldsymbol{\tau}}} \sum_{i:\text{State}_A(i)\neq\text{State}_B(i)} [m(A_i) + m(B_i)] \quad \textit{(state mismatch)} \tag{24}$$

$$\pi_{\text{mag}}^{(\boldsymbol{\tau})} = \frac{1}{U_{\boldsymbol{\tau}}} \sum_{i:\text{State}_A(i)=\text{State}_B(i)} |m(A_i) - m(B_i)| \quad \textit{(within-state imbalance)} \tag{25}$$

Example 4 (Audio Loudness Categories)**.** *Consider two audio signals with loudness states based on dB levels:*

- *Silent:* < 30 *dB*
- *Quiet:* $[30, 50)$ *dB*
- *Moderate:* $[50, 70)$ *dB*
- *Loud:* $[70, 90)$ *dB*
- *Very Loud:* ≥ 90 *dB*

For two recordings of the same event:

- *Recording A: [25, 45, 65, 85, 95] dB*
- *Recording B: [28, 55, 62, 75, 92] dB*

Analysis:

- *Position 2: State mismatch (Quiet vs. Moderate)*
- *Positions 1, 3, 4, 5: Same states with magnitude differences*
- *The decomposition tells us: "The recordings disagree 20% due to different loudness categories and 15% due to volume differences within the same categories."*

6 Properties and Advantages

6.1 Key Mathematical Properties

Proposition 2 (Scale Invariance)**.** *For any* $c > 0$*:* $J_{\text{peak}}^{(\theta)}(cA, cB) = J_{\text{peak}}^{(\theta)}(A, B)$

Proposition 3 (Metric Properties)**.** $d_{\text{peak}}^{(\theta)}$ *is a metric on the space of signals:*

1. *Non-negativity:* $d_{\text{peak}}^{(\theta)}(A, B) \geq 0$
2. *Identity:* $d_{\text{peak}}^{(\theta)}(A, B) = 0 \iff A = B$ *(a.e.)*
3. *Symmetry:* $d_{\text{peak}}^{(\theta)}(A, B) = d_{\text{peak}}^{(\theta)}(B, A)$
4. *Triangle inequality:* $d_{\text{peak}}^{(\theta)}(A, C) \leq d_{\text{peak}}^{(\theta)}(A, B) + d_{\text{peak}}^{(\theta)}(B, C)$

6.2 Kernel Construction

Theorem 2 (Negative Type and Kernel Validity)**.** *The distance $d_{\text{peak}}^{(\theta)}$ is of negative type. Therefore, for any $\lambda > 0$:*

$$K(A, B) = \exp\{-\lambda d_{\text{peak}}^{(\theta)}(A, B)\} \tag{26}$$

is a positive semidefinite kernel.

Corollary 1 (Spatial Covariance)**.** *For a spatial field $Z(\boldsymbol{s})$, the function*

$$K(\boldsymbol{h}) = \exp\{-\lambda d_{\text{peak}}^{(\theta)}(Z(\cdot), T_{\boldsymbol{h}} Z(\cdot))\} \tag{27}$$

is a valid covariance model, yielding positive semidefinite covariance matrices for kriging and Gaussian processes.

7 Comprehensive Comparison with Classical Methods

8 Practical Considerations

8.1 Choosing Boundaries

The choice of boundary θ (or boundaries $\boldsymbol{\tau}$) is application-specific:

1. **Domain knowledge**: Use natural thresholds
   - Finance: zero (gains/losses), moving averages, or volatility bands
   - Climate: freezing point, drought thresholds, seasonal averages
   - Biology: baseline activity, clinical thresholds
2. **Data-driven**: Statistical approaches
   - Median or mean for centering
   - Quantiles for multi-state (e.g., terciles, quartiles)
   - Clustering algorithms for natural breaks
3. **Adaptive**: Time-varying or spatially-varying boundaries
   - Moving averages for non-stationary signals
   - Local baselines for spatially heterogeneous fields

8.2 Boundary Sensitivity Analysis

Since states are strictly partitioned, values near boundaries can change states with small perturbations. This is a feature, not a bug – it reflects genuine uncertainty at transitions. For robustness:

1. **Stability analysis**: Compute $d_{\text{peak}}^{(\theta+\epsilon)}$ for small ϵ
2. **Boundary proximity**: Report fraction of data within δ of boundaries
3. **Conservative boundaries**: Use fewer, more widely-spaced boundaries if sensitivity is problematic
4. **Bootstrap confidence**: Resample to assess decomposition stability

8.3 Computational Aspects

The algorithm is straightforward and efficient:

1. Center signals: $O(n)$
2. Partition into positive/negative parts: $O(n)$
3. Compute min/max for intersection/union: $O(n)$
4. Calculate ratios and decomposition: $O(1)$

Total complexity: $O(n)$ for signals of length n, fully vectorizable for parallel computation.

9 Applications and Examples

9.1 Financial Time Series

Compare portfolio returns relative to a benchmark:

- State component: How often do portfolios have opposite performance (one beats benchmark, other doesn't)?
- Magnitude component: When both beat/miss benchmark, by how much do they differ?

9.2 Climate Data

Analyze precipitation relative to seasonal norms:

- State component: Wet vs. dry disagreement between locations
- Magnitude component: Severity differences within wet or dry periods

9.3 Neural Recordings

Compare neural activity across brain regions:

- States: Inhibition, baseline, weak excitation, strong excitation
- Decomposition reveals synchrony vs. magnitude coupling

9.4 Image Analysis

Compare image patches relative to local background:

- 2D extension with spatially-varying $\theta(x, y)$
- Anisotropy analysis reveals directional patterns

Table 1: Classical semivariogram vs. boundary-aware Jaccard correlogram

Property	**Semivariogram** $\gamma(\boldsymbol{h}) = \frac{1}{2}\,\mathbb{E}[(Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h}))^2]$	**Jaccard correlogram** $C_J^{(\theta)}(\boldsymbol{h}) = 2\,J_{\text{peak}}^{(\theta)}(A, T_{\boldsymbol{h}}A) - 1$
Output range	Unbounded above, $\gamma(\boldsymbol{h}) \geq 0$	Bounded, $C_J^{(\theta)}(\boldsymbol{h}) \in [-1, 1]$
Interpretation at $\boldsymbol{h} = 0$	$\gamma(0) = 0$ (perfect similarity)	$C_J^{(\theta)}(0) = +1$ (perfect overlap)
Anti-phase behavior	Large γ (via squared differences)	$C_J^{(\theta)} \approx -1$ (explicit opposite-side penalty)
Scale invariance	No (units matter)	Yes: $J_{\text{peak}}^{(\theta)}(cA, cB) = J_{\text{peak}}^{(\theta)}(A, B)$ for $c > 0$
Robustness to outliers	Sensitive (squared differences)	Robust (bounded overlap ratios)
Handles sign/state structure	Implicit via squaring, not interpretable	Explicit: separates state vs. magnitude
Decomposition	None canonical	Exact: $d_{\text{peak}}^{(\theta)} = \pi_{\text{state}}^{(\theta)} + \pi_{\text{mag}}^{(\theta)}$
Symmetry	$\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$ always	Auto: $C_J^{(\theta)}(\boldsymbol{h}) = C_J^{(\theta)}(-\boldsymbol{h})$; Cross: $C_{AB}^{(\theta)}(\boldsymbol{h}) = C_{BA}^{(\theta)}(-\boldsymbol{h})$
Isotropy/anisotropy	Level sets (circular vs. elliptical)	Level sets plus component maps $\pi_{\text{state}}^{(\theta)}$, $\pi_{\text{mag}}^{(\theta)}$
Periodicity	Via oscillatory growth/plateau	Bounded oscillations; negative lobes for anti-phase
Normalization	None; depends on variance	Natural ratio in $[0, 1]$; centered to $[-1, 1]$
Metric properties	$\sqrt{2\gamma}$ relates to ℓ_2	$d_{\text{peak}}^{(\theta)}$ is a metric; components are not
Kernel/covariance link	$C(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h})$ (stationary)	$K(\boldsymbol{h}) = \exp(-\lambda d_{\text{peak}}^{(\theta)}(\boldsymbol{h}))$ is PSD
Multi-state extension	Not natural	Natural via strict partitioning
Boundary sensitivity	N/A	Sensitive near θ; analyze stability
Missing data	Standard imputation methods	Exclude from both N_θ and U_θ
Computational cost	$O(n)$ simple arithmetic	$O(n)$ with partitioning step
Interpretability	Abstract squared differences	Concrete: "20% state mismatch, 15% magnitude imbalance"
Cross-dataset comparison	Difficult (unbounded, scale-dependent)	Easy (bounded, scale-invariant)
Negative values permitted	No (always ≥ 0)	Yes in centered form; indicates opposition
Use in kriging/GP	Standard practice	Via PSD kernel $\exp(-\lambda d_{\text{peak}}^{(\theta)})$
Physical interpretation	Energy/variance based	Overlap/agreement based