README: Dynamic SINDy-to-Simplicial-Complex Pipeline

August 17, 2025

Purpose

This document describes the end-to-end pipeline that converts multichannel time series (e.g. EEG) into a *time-indexed simplicial complex*, one complex per window. The pipeline is robust to outliers, produces *unweighted* complexes per window (with closure), and then applies *fixed global thresholds* to enable consistent dynamics analysis across time.

1 Data Preparation

Smoothing and derivatives. Let the raw data matrix be $R \in \mathbb{R}^{T\times n}$ (samples × channels). We apply Savitzky–Golay (SG) smoothing (order p, window W, right now $p = 3$ and $W = 13$) and compute aligned derivatives:

$$X = \mathrm{SG\_smooth}(R) \in \mathbb{R}^{n\times T'}, \qquad \dot{X} = \mathrm{SG\_derivative}(R) \in \mathbb{R}^{n\times T'}.$$

Edges of the SG filter are trimmed equally so X and $\dot{X}$ have the same length T'.

Per-channel normalisation. Let $\mu_i = \mathrm{mean}(X_i)$ and $\sigma_i = \mathrm{std}(X_i)$. We form z-scores

$$\tilde{X}_i = \frac{X_i - \mu_i}{\sigma_i}, \qquad \tilde{Y}_i = \frac{\dot{X}_i}{\sigma_i}.$$

Thus, states have unit scale per channel; derivatives are scaled by the same σ (no mean subtraction).

2 Local Taylor Neighbourhoods (Per-Window)

We create overlapping windows of length L with stride S. Right now $L = 1024$ and $S = 512$. For a window w we extract $\tilde{X}^{(w)} \in \mathbb{R}^{n\times L}$. We define the *Taylor centre* as the per-channel median

$$x_0^{(w)} = \mathrm{median}_t\big(\tilde{X}^{(w)}(:,t)\big).$$

We compute robust per-timestamp distances using a median/MAD scale per channel:

$$\mathrm{MAD}_i = \mathrm{median}_t\big|\tilde{X}_i^{(w)}(t) - \mathrm{median}_t(\tilde{X}_i^{(w)})\big|,$$

$$\sigma_i^{\mathrm{rob}} = 1.4826\,\mathrm{MAD}_i, \qquad Z_i(t) = \frac{\tilde{X}_i^{(w)}(t) - x_{0,i}^{(w)}}{\sigma_i^{\mathrm{rob}}},$$

$$d_{\mathrm{pc}}(t) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} Z_i(t)^2\right)^{1/2}.$$

We keep timestamps with $d_{\mathrm{pc}}(t) \le R_{\mathrm{target}}$. If too few/many remain, we take the best $K_{\min}$ or top $K_{\max}$ by proximity. A window is dropped only if it is globally wild *and* its kept subset remains too wide (95th percentile checks). The *same* kept timestamp indices are then applied to both state and derivative matrices. We centre the kept window by subtracting $x_0^{(w)}$.

3 SINDy Fitting (Per Window)

We fit a SINDy model with a polynomial library of degree $D_{\max}$ (typically 2) and STLSQ threshold λ. The regression is

$$\tilde{Y}^{(w)}(t) \approx \Theta\big(\tilde{X}^{(w)}(t) - x_0^{(w)}\big)\, \Xi^{(w)},$$

where Θ is the feature library (linear and cross terms for $D_{\max} = 2$) and $\Xi^{(w)}$ are the fitted coefficients (*targets* × *features*).

4 From Coefficients to Simplices (Per Window)

We work with *direct* scores extracted from $\Xi^{(w)}$, then optionally enforce closure inside the window.

Edge scores (linear terms). Let $A^{(w)} = |\Xi^{(w)}|$. For a linear feature x_j in the equation for dx_i/dt we take the directed score $A^{(w)}_{i\leftarrow j}$ and form a symmetric edge score

$$S_2^{(w)}(i,j) = \max\big\{A^{(w)}_{i\leftarrow j},\, A^{(w)}_{j\leftarrow i}\big\}, \qquad i \neq j.$$

Triangle scores (cross terms). For each target i and cross feature $x_j x_k$ with $j \neq k$ and $i \notin \{j,k\}$, we set

$$S_3^{(w)}(\{i,j,k\}) = \max\big(S_3^{(w)}(\{i,j,k\}),\, A^{(w)}_{i\leftarrow(j,k)}\big).$$

(We exclude degeneracies and any monomial that contains the target variable.)

Closure. Given a set of triangles $T^{(w)}$, we add all their faces: if $\{i,j,k\} \in T^{(w)}$ then $\{i,j\}, \{i,k\}, \{j,k\}$ are included as edges. (For $D_{\max} = 3$ we would similarly add faces of tetrahedra.)

5 Fixed Global Thresholds Across Windows

To enable temporal comparisons, we rebuild *every* window's complex using the same thresholds:

- **Edges.** We compute the *maximum spanning tree (MST) bottleneck* value $b^{(w)}$ for each window using $S_2^{(w)}$. The global edge threshold is

  $$\tau_2 = \min_w b^{(w)},$$

  which guarantees that each window's graph is connected at threshold τ_2.

- **Triangles (percentile regime).** Pool all triangle scores across windows and set

  $$\tau_3 = \mathrm{Quantile}_p\big(\cup_w \,\mathrm{values}(S_3^{(w)})\big), \qquad \text{e.g. } p = 97.5\%.$$

Given τ_2, τ_3, the final per-window complex is

$$E^{(w)} = \{\{i,j\} : S_2^{(w)}(i,j) \geq \tau_2\},$$
$$T^{(w)} = \{\{i,j,k\} : S_3^{(w)}(\{i,j,k\}) \geq \tau_3\},$$

followed by closure inside window w. These complexes are *unweighted* and *built only from that window's coefficients.* Cross-window unions are used *only* for global summaries.

6 Outputs

1. **Per-window edge/triangle lists** (closed complexes), plus counts.
2. **Global statistics:** average counts per window; unique hyperedges across time; node participation.
3. **CSV/JSON:** locality diagnostics, per-window counts, and a run summary (including τ_2, τ_3).

7 Quantifying Dynamics (Idea Sketch)

Let $\mathcal{S}_k$ be the set of all k-simplices observed across the entire run (e.g. all edges or all triangles). For a fixed simplex $s \in \mathcal{S}_k$ define a binary sequence across windows

$$b_s(w) = \begin{cases} 1, & s \in \text{complex of window } w, \\ 0, & \text{otherwise.} \end{cases}$$

Several complementary measures of *dynamism* can be derived:

(a) **Event rate / flip rate:** $r_s = \frac{1}{W-1} \sum_{w=1}^{W-1} \mathbf{1}[b_s(w) \neq b_s(w+1)]$.

(b) **Persistence:** $\pi_s = \frac{1}{W} \sum_w b_s(w)$ (fraction of windows in which s is present).

(c) **Temporal entropy:** $H_s = -p_s \log p_s - (1-p_s)\log(1-p_s)$ with $p_s = \pi_s$.

(d) **Ensemble diversity (optional):** compare sequences pairwise using Jaccard distance or Jensen–Shannon divergence between empirical 2-state Markov models; summarise via mean/median.

Global dynamism can be taken as the mean/quantiles of r_s over $s \in \mathcal{S}_k$, reported for $k = 1$ (edges) and $k = 2$ (triangles). This framework naturally generalises to higher-order simplices.

8 Design Choices and Rationale

- **Robust local neighbourhoods:** median/MAD ensures the Taylor expansion is taken around representative states; we keep enough near-centre samples ($K_{\min}$) to stabilise SINDy.
- **Per-window construction:** each window yields its complex *independently.* Closure is enforced inside the window.
- **Mapping rules:** edges come only from linear terms (symmetric max); triangles come from cross terms with distinct variables and no target participation (true 3-body structure).
- **Fixed global thresholds:** τ_2 ensures connectivity in every window. τ_3 (percentile regime) selects the top tail of triangle strengths consistently across time.

9 Key Hyperparameters ("Knobs")

SG smoothing:	window W, order p
Locality:	$R_{\text{target}}, K_{\min}, K_{\max}$, drop rules
SINDy:	degree $D_{\max}$, STLSQ threshold λ
Thresholds:	edge τ_2 (MST bottleneck min), triangle τ_3 (percentile p)

10 Some of the things to discuss

We have made several choices here and it's not clear if they are optimal for our purposes. some of these are

- It's not clear how to handle the monomials of the form $x_i.x_j$ while considering the channel i. The coeffecient gives us the interaction between $\{i, i, j\}$ which is a degenerate $2 - simplex$. Question is if this value is high while the value of interaction between the channel i and j (which is the coeffecient of the monomial x_j) is low, should we take this as an evidence of link between channel i and j or not.

- Edge scores and triangle scores: We can be conservative or liberal here while assigning the scores and this will play a role during thresholding, for instance instead of max one can do average, or median etc.

- Right now we are chossing thresholds globally so as to capture the dynamics and threshold for choosing the edge is simply the lowest value which esnures connectivity of all the graphs across the windows. One can tweak it by raising the value even further. A good threshold should capture the dynamics without being too noisy.

- The triangle (2-simplex) threshold is simply some percentile right now. Seems a bit random. Will need tuning, probably once we decide on quabtifying the dynamics.