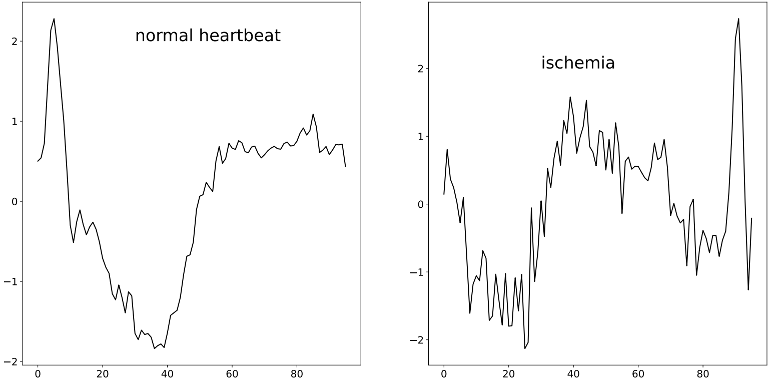
6.4. LIMESegment

Time series classification

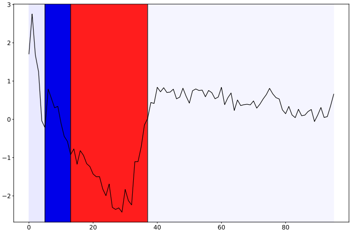
- **Time series:** ordered sequence of T observations
- **Example:**[44] ECG from one heartbeat, detect ischemia or not



[44] Olszewski, *Generalized feature extraction for structural pattern recognition in time-series data*, Carnegie Mellon, 2001

LIMESegment

- **Idea:**[45] adapt the LIME framework to time series
- similar high-level operation (differences in **bold**):
  1. **create interpretable features**
  2. **sample** n perturbed samples $x_1, \ldots, x_n$ from ξ
  3. **weight** the x_is
  4. train a local surrogate model
- **Output:** highlight important parts of the time-series



[45] Sivill, Flach, *LIMESegment: Meaningful, Realistic Time Series Explanations*, AISTATS, 2022

Step 1: interpretable features

- **Interpretable features:** homogeneous segments in the time series
- standard problem (usually called *change-point detection*[46])
- proposed methodology: `NNSegment`
- **Reminder:** empirical mean: let $A \in \mathbb{R}^\ell$,

$$\overline{A} := \frac{1}{\ell}\sum_{i=1}^{\ell} A_i \,.$$

- **Reminder:** empirical covariance / variance:

$$\widehat{\mathrm{Cor}}(A,B) := \frac{1}{\ell-1}\sum_{i=1}^{\ell-1}(A_i - \overline{A})(B_i - \overline{B})\,, \qquad \widehat{\mathrm{Var}}(A) := \frac{1}{\ell-1}\sum_{i=1}^{\ell}(A_i - \overline{A})^2\,.$$

[46] Truong, Oudre, Vayatis, *Selective review of offline change-point detection methods*, Signal Processing, 2020

Step 1: interpretable features

Definition-proposition: let A and B be two signals of length ℓ. We call *normalized cross-correlation* (*a.k.a.* sample correlation)

$$\psi(A,B) := \frac{\widehat{\mathrm{Cor}}(A,B)}{\sqrt{\widehat{\mathrm{Var}}(A)\widehat{\mathrm{Var}}(B)}} .$$

It holds that $\psi(A,B) \in [-1,1]$.

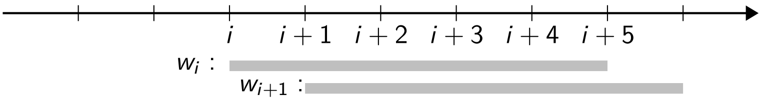
- **Intuition:** quantifies the linear relationship between A and B
- **Examples:**
  - if $B_i = \alpha A_i$, then $\widehat{\mathrm{Cor}}(A,B) = \alpha$
  - if A and B are "independent," then $\widehat{\mathrm{Cor}}(A,B) \approx 0$

Step 1: interpretable features

- back to `NNSegment`
- let w_s be a fixed window size, define

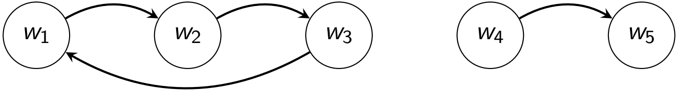
$$x_{a:b} := (x_a, x_{a+1}, \dots, x_b)^\top .$$

- for a given *window size* w_s, define $w_i := x_{i:(i+w_s)}$
- **Example:** indices corresponding to w_i with $w_s = 5$



Step 1: interpretable features

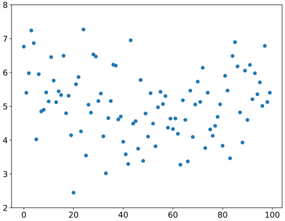
- **Global operating procedure:**
  1. compute all pairwise correlations between segments $\psi(s_1, s_2)$
  2. connect each segment to its *nearest neighbor*
  3. group adjacent segments together (nearest neighbor $=$ next segment)
- **Example:** (arrows denote nearest neighbor)

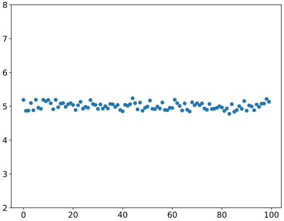


- in this example, we group w_1, w_2, and w_3 together

Step 1: interpretable features

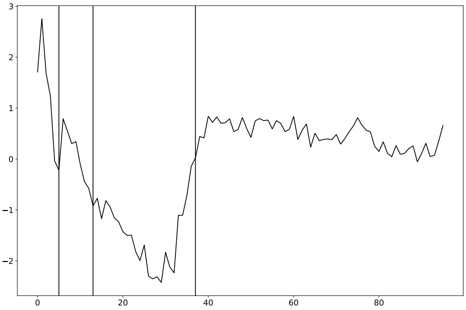
- **Further refinement:** look at difference in signal to noise ratio
$$\rho(w_i, w_j) := \left| \frac{\mu(w_i)}{\sigma(w_i)} - \frac{\mu(w_j)}{\sigma(w_j)} \right| ,$$
and then:
  - if $\rho(w_i, w_{i-w_s}) > \rho(w_i, w_{i+w_s})$, group i with $i + w_s$
  - if $\rho(w_i, w_{i-w_s}) < \rho(w_i, w_{i+w_s})$, group i with $i - w_s$
- stop doing this when we have reached the user-specified number of segments
- **Example:** left SNR ≈ 5, right SNR ≈ 50





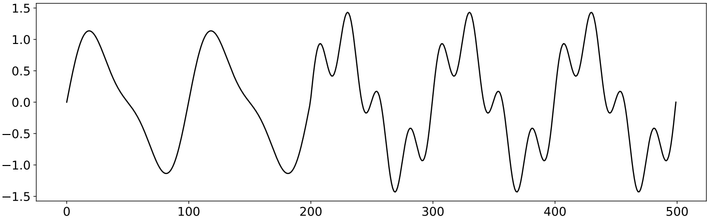
Step 1: interpretable features

- **Output:** segmented signal
- **Example:** here we obtain 4 segments, that is, 3 breakpoints



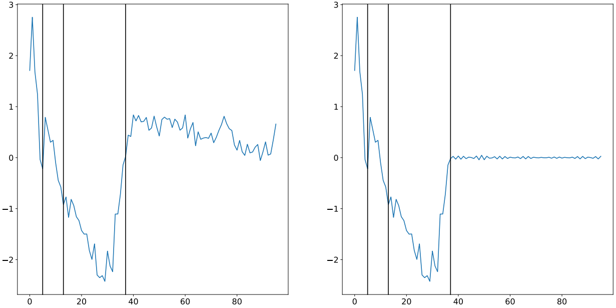
Step 2: perturbed examples

- **Idea:** identify background signal in the spectral domain
- **Discrete Short Time Frequency Transform (STFT):** $\rightarrow$ time-frequency domain
- **Example:** (local) spectrogram of superposition of sine waves



Step 2: perturbed samples

- identify a persistent frequency, map it back via inverse STFT
- **Example:** perturbing the last segment of the signal ($Z = (1, 1, 1, 0)^\top$)



Step 3: weights

- similar idea: exponential weights depending on a distance
- **Issue:** Euclidean distance between the z_i does not reflect distance between signals
- **Dynamic time warping (DTW):**[47] distance between signals taking alignment into account
- formally,

$$\mathrm{DTW}(x,x')^2 := \min_{\pi\in P(x,x')} \sum_{(i,i')\in\pi} d(x_i, x'_{i'}) \,,$$

  where π is an *admissible path*
- namely:
  - $\pi_1 = (1,1)$ (beginning of signals matched together);
  - $\pi_K = (S,T)$ (end of signals matched together);
  - writing π_k as (i_k, i'_k), both i and i' are non-decreasing.

[47] Bellman, Kalaba, *On adaptive control processes*, IRE Transactions on Automatic Control, 1959

Summary

- **Final steps:** surrogate model as before (ridge), coefficients given as importance
- **Main message:** a lot depends on the data-type and the kind of perturbation we want
- results depends a lot on the segmentation / sampling scheme
- no existing theoretical analysis
- many other methods[48]

[48] see Theissler et al., *Explainable AI for Time Series Classification: A review, taxonomy and research directions*, for an overview