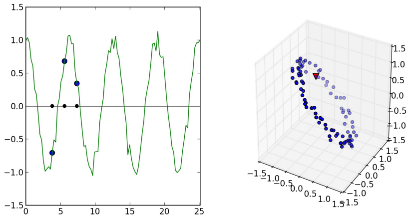
Topological Data Analysis for Time Series Analysis

Elizabeth Munch

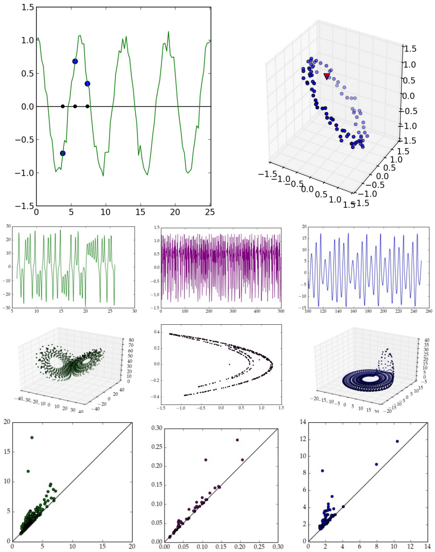
Michigan State University
::
Department of Computational Mathematics Science and Engineering
Department of Mathematics

June 6, 2018

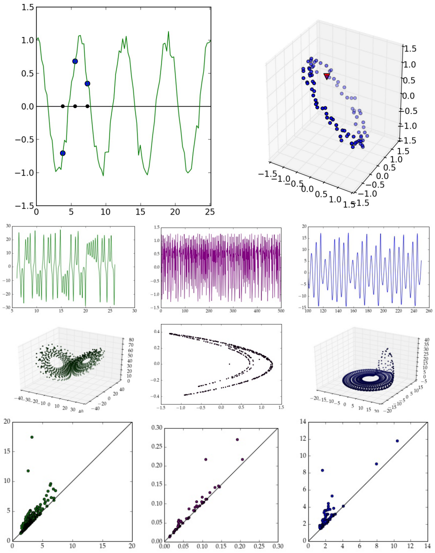
Time Series Data

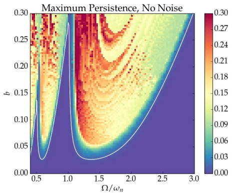


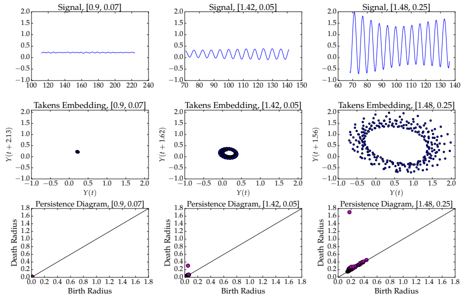
Time Series Data



Time Series Data

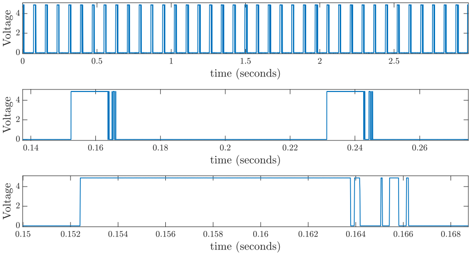


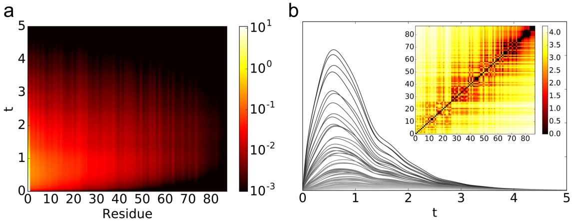




Outline

1. Pulse counting for Piecewise Constant Signals and RPM
2. Synchronization in Coupled Dynamical Systems

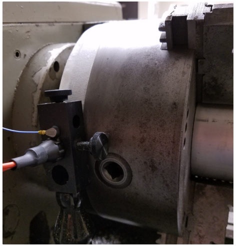


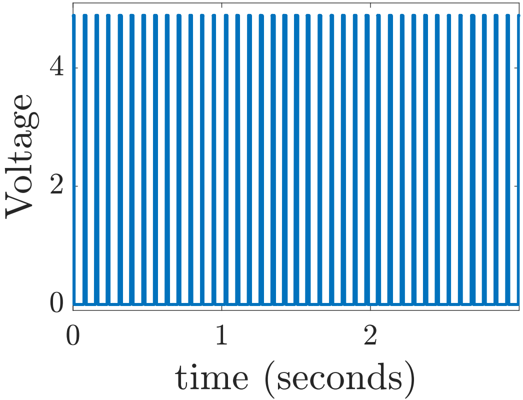


Section 1

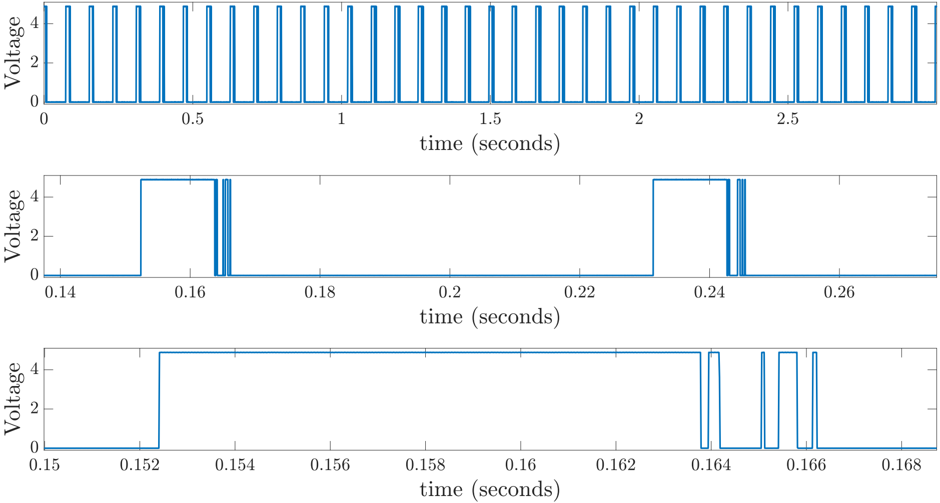
Pulse counting for Piecewise Constant Signals and RPM

Motivation: Spindle speed measurement using laser tachometer signals

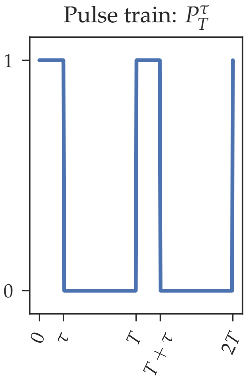




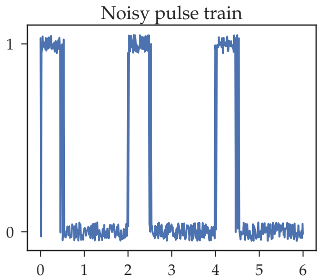
Digital Ringing



Noisy Signal Model (version 1)



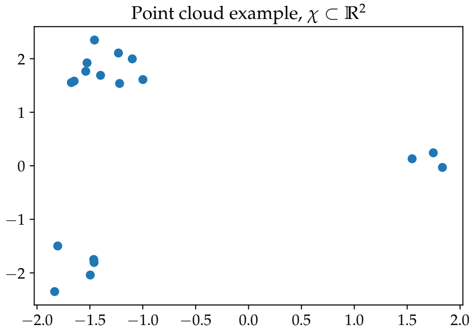
Duty cycle: τ
Pulse width: T

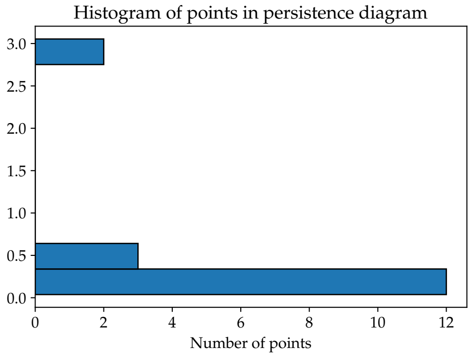


$$X(t) = P_T^\tau(t + \delta_x) + \delta_y$$

$\delta_x \sim \mathrm{unif}(-\alpha \cdot \tau, \alpha \cdot \tau)$
$\delta_y \sim \mathrm{unif}(-\beta, \beta)$ with
$\alpha \in [0, 1/2]$ and $\beta \in [0, 1]$.

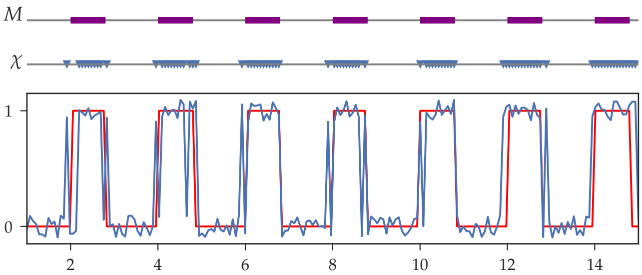
0-Dim Persistence Diagram as Histogram





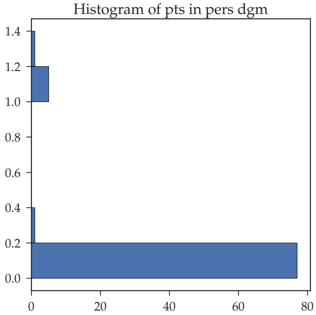
Thresholded persistence

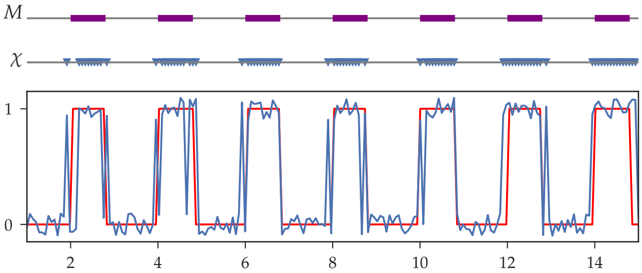
- Discrete
  - $X(0), X(t_1), \cdots, X(t_N)$
  - $\chi = \{t \mid X(t) \geq \frac{1}{2}\} \subset \mathbb{R}$
- Continuous:
  - $P_T^\tau : [0, t_N] \to \mathbb{R}$
  - $M = \{t \mid P_T^\tau(t) \geq \frac{1}{2}\}$



Thresholded persistence

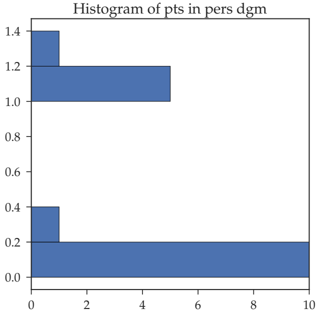
- Discrete
  - $X(0), X(t_1), \cdots, X(t_N)$
  - $\chi = \{t \mid X(t) \geq \frac{1}{2}\} \subset \mathbb{R}$
- Continuous:
  - $P_T^\tau : [0, t_N] \to \mathbb{R}$
  - $M = \{t \mid P_T^\tau(t) \geq \frac{1}{2}\}$

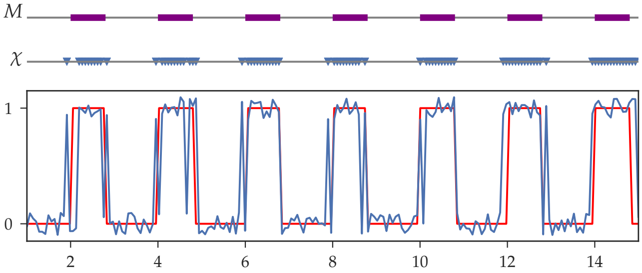




Thresholded persistence

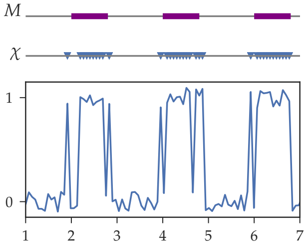
- Discrete
  - $X(0), X(t_1), \cdots, X(t_N)$
  - $\chi = \{t \mid X(t) \geq \frac{1}{2}\} \subset \mathbb{R}$
- Continuous:
  - $P_T^\tau : [0, t_N] \to \mathbb{R}$
  - $M = \{t \mid P_T^\tau(t) \geq \frac{1}{2}\}$

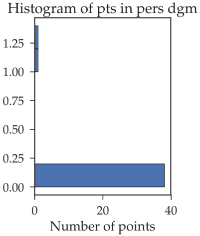




Algorithm

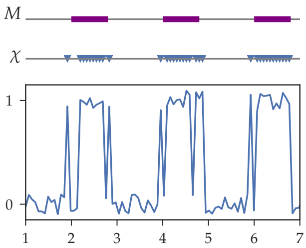
- Find $\chi = \{a_1 < \cdots < a_m\}$ and sort

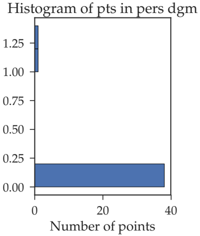




Algorithm

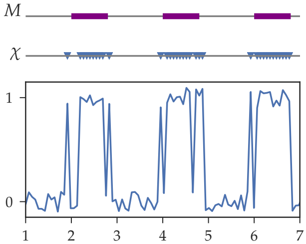
- Find $\chi = \{a_1 < \cdots < a_m\}$ and sort
- Get persistence diagram
  - $\{a_i - a_{i-1}\}$ because points are in $\mathbb{R}$
  - Sort and write as $\mathrm{dgm}(\chi) = \{d_1 < \cdots < d_\ell\}$

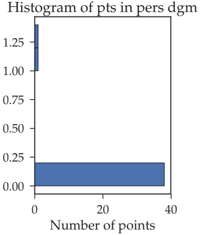




Algorithm

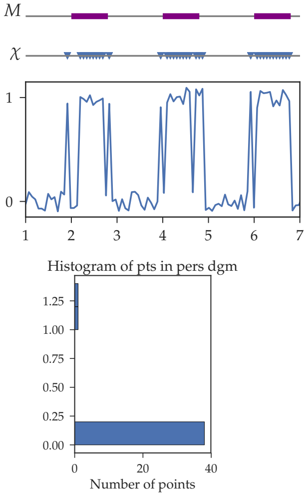
- Find $\chi = \{a_1 < \cdots < a_m\}$ and sort
- Get persistence diagram
  - $\{a_i - a_{i-1}\}$ because points are in $\mathbb{R}$
  - Sort and write as $\mathrm{dgm}(\chi) = \{d_1 < \cdots < d_\ell\}$
- Find largest split
  - Let $j = \mathrm{argmax}_k \{d_{k+1} - d_k\}$
  - Set $\mu \in (d_j, d_{j+1})$





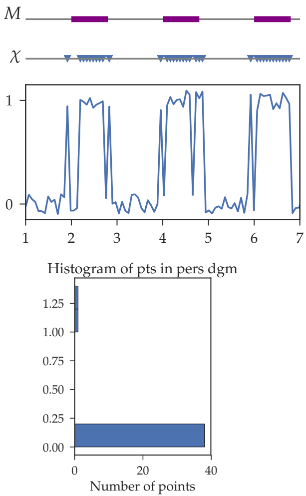
Algorithm

- Find $\chi = \{a_1 < \cdots < a_m\}$ and sort
- Get persistence diagram
  - $\{a_i - a_{i-1}\}$ because points are in $\mathbb{R}$
  - Sort and write as $\mathrm{dgm}(\chi) = \{d_1 < \cdots < d_\ell\}$
- Find largest split
  - Let $j = \mathrm{argmax}_k \{d_{k+1} - d_k\}$
  - Set $\mu \in (d_j, d_{j+1})$
- Count points in diagram above μ
  - $K = |\{d \in \mathrm{dgm}\chi \mid d > \mu\}|$



Algorithm

- Find $\chi = \{a_1 < \cdots < a_m\}$ and sort
- Get persistence diagram
  - $\{a_i - a_{i-1}\}$ because points are in $\mathbb{R}$
  - Sort and write as $\mathrm{dgm}(\chi) = \{d_1 < \cdots < d_\ell\}$
- Find largest split
  - Let $j = \mathrm{argmax}_k\{d_{k+1} - d_k\}$
  - Set $\mu \in (d_j, d_{j+1})$
- Count points in diagram above μ
  - $K = |\{d \in \mathrm{dgm}\chi \mid d > \mu\}|$
- Number of (partial) pulses seen is $K + 1$



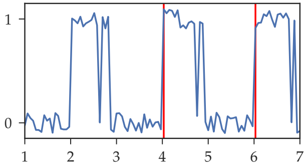
Theorem (Khasawneh, M 2018)

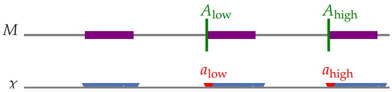
Given $\{X(t_i)\}_{i=1}^{N}$ *for evenly spaced* t_i, $A := t_0, B := t_N$, *with*

- $\alpha < \frac{1}{5}(\frac{T}{\tau} - 1)$, *Control horizontal noise*
- $\beta < .5$, *Control vertical noise*
- $t_i - t_{i-1} < \tau(1-2\alpha)/2$ *and* *See every pulse*
- $(t_N - t_0)/T > 3$. *See* ≥ 3 *pulses*

If K *is determined in algorithm, then there are* $K-1$ *pulses in the range*

$$A < A_{\mathrm{low}} := \left(\left\lceil \frac{A-\tau}{T} \right\rceil + 1 \right) T < \left\lfloor \frac{B}{T} \right\rfloor T =: A_{\mathrm{high}} \leq B.$$





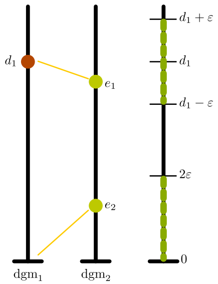
Theorem (Khasawneh, M 2018)

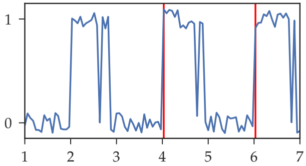
Given $\{X(t_i)\}_{i=1}^{N}$ *for evenly spaced* t_i, $A := t_0, B := t_N$, *with*

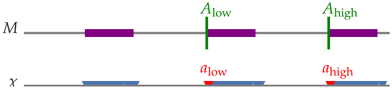
- $\alpha < \frac{1}{5}(\frac{T}{\tau} - 1)$, — *Control horizontal noise*
- $\beta < .5$, — *Control vertical noise*
- $t_i - t_{i-1} < \tau(1-2\alpha)/2$ *and* — *See every pulse*
- $(t_N - t_0)/T > 3.$ — *See ≥ 3 pulses*

If K *is determined in algorithm, then there are* $K-1$ *pulses in the range*

$$A < A_{\mathrm{low}} := \left(\left\lceil \frac{A-\tau}{T} \right\rceil + 1 \right) T < \left\lfloor \frac{B}{T} \right\rfloor T =: A_{\mathrm{high}} \leq B.$$



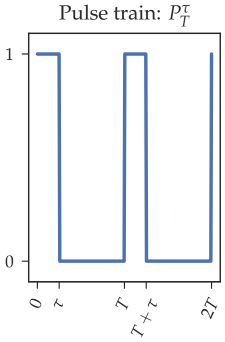




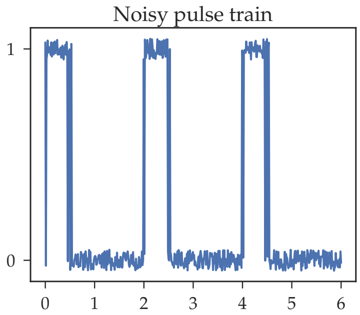
“Why don’t you just use Fourier?”

-Reviewer 2

Noisy Signal Model (version 2)

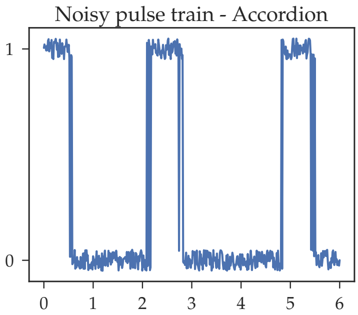


Duty cycle: τ
Pulse width: T



$$X(t) = P_T^\tau(t + \delta_x) + \delta_y$$

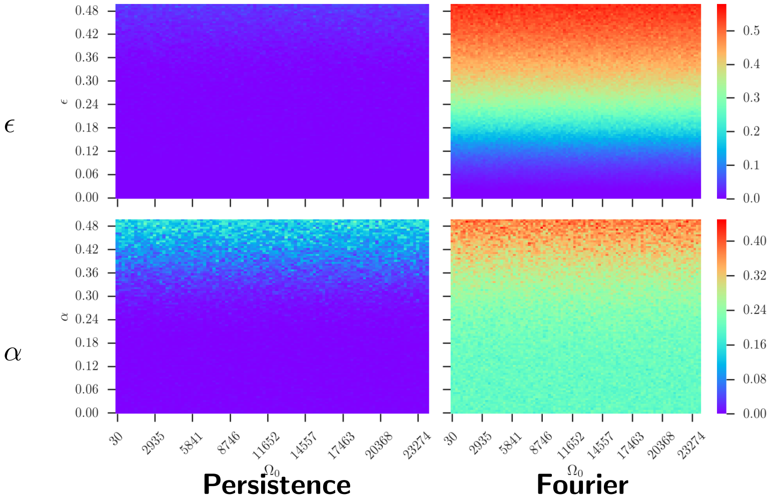
$\delta_x \sim \mathrm{unif}(-\alpha \cdot \tau, \alpha \cdot \tau)$
$\delta_y \sim \mathrm{unif}(-\beta, \beta)$ with
$\alpha \in [0, 1/2]$ and $\beta \in [0, 1]$.



$$X(s) = P_T^\tau(\varphi(s) + \delta_x) + \delta_y$$

Pulse length drawn from
$\mathrm{unif}((1-\varepsilon)T, (1+\varepsilon)T)$

Relative Error of RPM



case 1: $\tau = 0.05$, $\alpha = 0.1$, $\epsilon \in [0.02, 0.65]$

case 2: $\tau = 0.05$, $\epsilon = 0.25$, $\alpha \in [0, 0.5]$

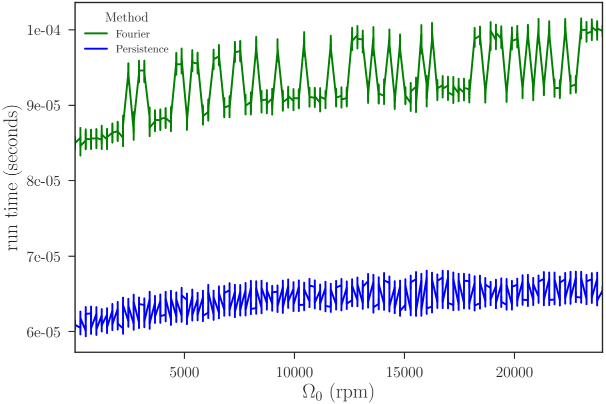
Accordion model:

$X(s) = P_T^\tau(\varphi(s) + \delta_x) + \delta_y$

ϵ: accordion parameter, higher values mean closer pulses

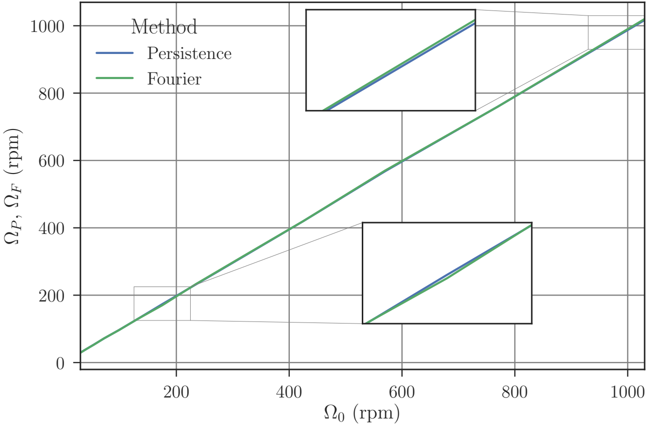
True step detection in piecewise constant signals

Average runtime for all trials (200 runs for each nominal RPM) as a function of the nominal RPM

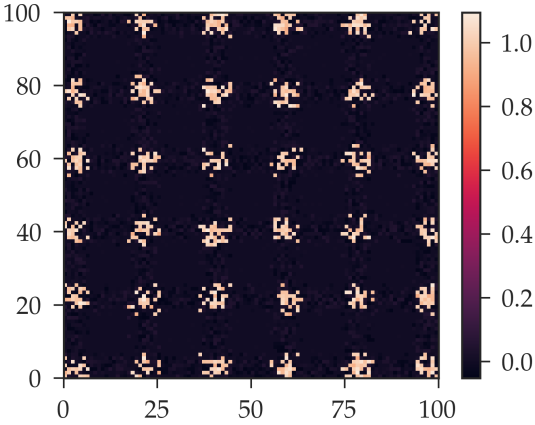


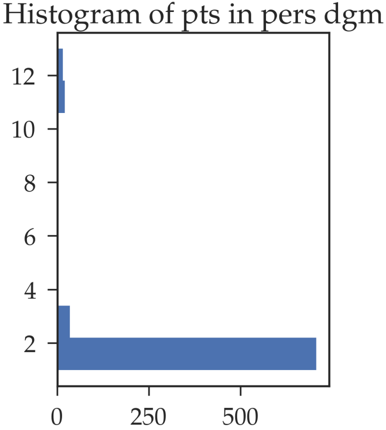
Theoretical worst case run-time for both algorithms: *O(n log(n))*

Experimental Data: RPM



Higher dimensional analogue of the persistence-based method for pulse counting





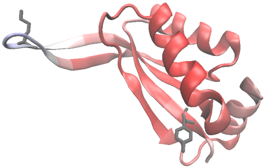
Section 2

Synchronization in Coupled Dynamical Systems

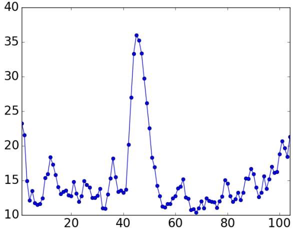
B-factor and protein

B-factor

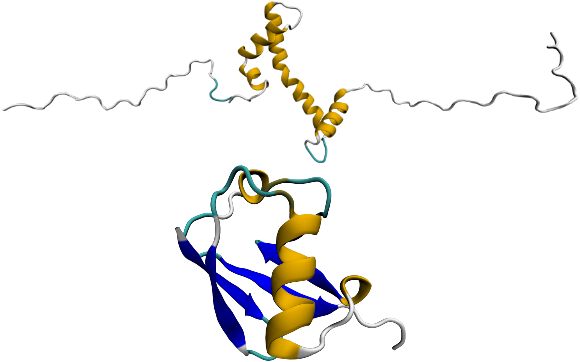
Quantitatively measures the relative thermal motion of each atom and reflects atomic flexibility.



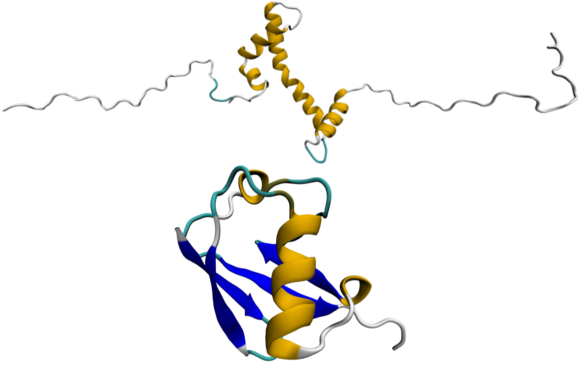
PDB:2NUH

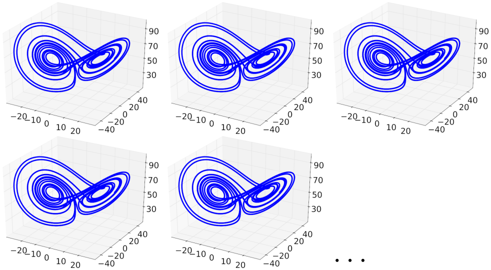


Setup: coupled dynamical system on a protein embedding



Setup: coupled dynamical system on a protein embedding



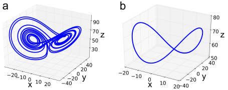


Mathematical forumlation

Lorenz Oscillators

$$\frac{d\mathbf{u}_i}{dt} = g(\mathbf{u}_i),\ i = 1, 2, \cdots, N$$

$$g(\mathbf{u}_i) = \begin{bmatrix} \delta(u_{i,2} - u_{i,1}) \\ u_{i,1}(\gamma - u_{i,3}) - u_{i,2} \\ u_{i,1}u_{i,2} - \beta u_{i,3} \end{bmatrix}$$



Coupling

$$d^{\mathrm{org}}(\mathbf{r}_i, \mathbf{r}_j) = \|\mathbf{r}_i - \mathbf{r}_j\|_2.$$

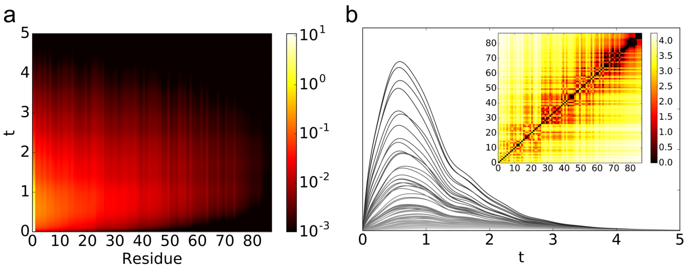
$$A_{ij} = \begin{cases} e^{-(d^{\mathrm{org}}(\mathbf{r}_i,\mathbf{r}_j)/\mu)^\kappa},\ i \neq j, \\ -\sum_{l \neq i} A_{il},\ i = j, \end{cases}$$

$$\frac{d\mathbf{u}}{dt} = \mathbf{G}(\mathbf{u}) + \epsilon(A \otimes \Gamma)\mathbf{u},$$

QUESTION

Given just the time series...
can we say something about the B-factors?

Perturbing an oscillator at residue i

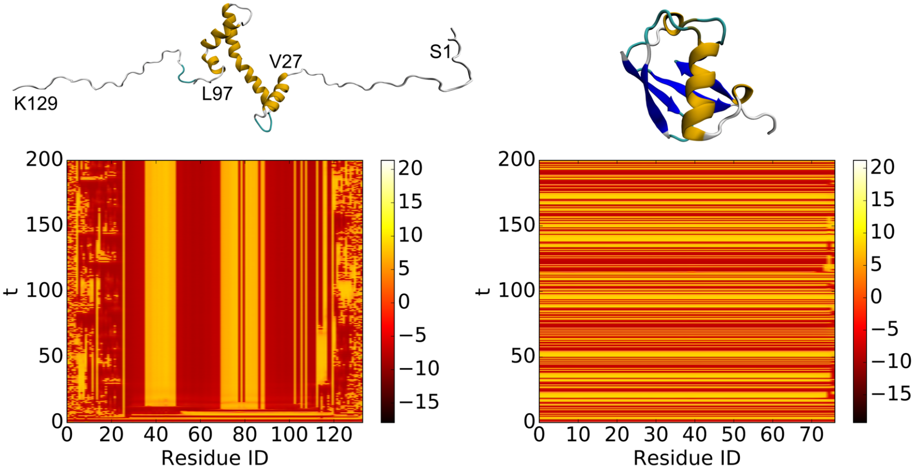


- N not yet synchronized oscillators $\{\mathbf{u}_1, \cdots, \mathbf{u}_N\}$
- N embedded points, $\{\mathbf{r}_1, \cdots, \mathbf{r}_N\} \subset \mathbb{R}^d$.
- Global synchronized state is a periodic orbit $\mathbf{s}(t)$ for $t \in [t_0, t_1]$ where $\mathbf{s}(t_0) = \mathbf{s}(t_1)$.

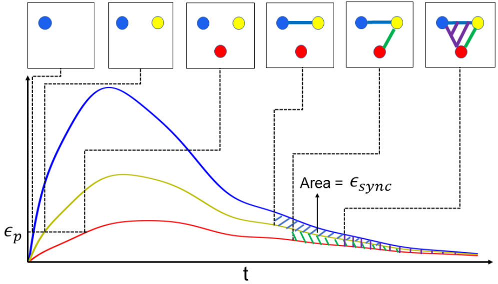
Transformation function

$$T(\mathbf{u}_i(t)) = \min_{t' \in [t_0, t_1]} \|\mathbf{u}_i(t) - \mathbf{s}(t')\|_2, \qquad \text{so } \widehat{\mathbf{s}}(t) := T(\mathbf{s}(t)) = 0$$

Protein Structure affects Synchronization



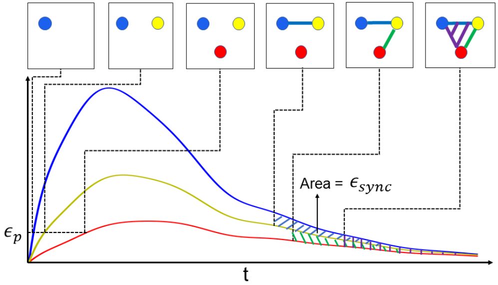
Constructing a filtration



- $\epsilon_p \geq 0$: far enough to be unsychronized
- $\epsilon_{\text{sync}} \geq 0$: close enough to be synchronized
- $\epsilon_d \geq 0$

$$V^i = \left\{ n_j \mid \max \lim_{t>0} \left\{ \min \lim_{t' \in [t_0, t_1]} \|\widehat{\mathbf{u}}_j^i(t) - \widehat{\mathbf{s}}(t')\|_2 \right\} \geq \epsilon_p \right\}$$

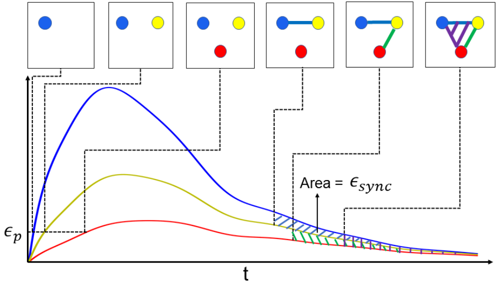
Constructing a filtration



- $\epsilon_p \geq 0$: far enough to be unsychronized
- $\epsilon_{\text{sync}} \geq 0$: close enough to be synchronized
- $\epsilon_d \geq 0$

$$t^i_{\text{sync}} = \min \left\{ t \mid \int_t^\infty \|\widehat{\mathbf{u}}^i_j(t') - \widehat{\mathbf{u}}^i_k(t')\|_2 \, dt' \leq \frac{\epsilon_{\text{sync}}}{2}, \forall j, k \right\}.$$

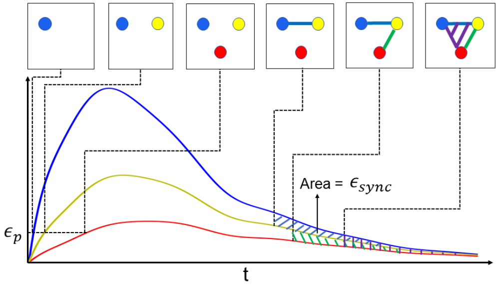
Constructing a filtration



- $\epsilon_p \geq 0$: far enough to be unsychronized
- $\epsilon_{\text{sync}} \geq 0$: close enough to be synchronized
- $\epsilon_d \geq 0$

$$f(n_j) = \min\left\{ \{t \mid \min_{t' \in [t_0, t_1]} \|\widehat{\mathbf{u}}_j^i(t) - \widehat{\mathbf{s}}(t')\|_2 \geq \epsilon_p\} \cup \{t^i_{\text{sync}}\} \right\}.$$

Constructing a filtration

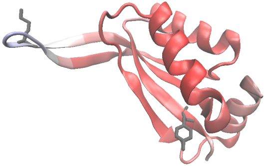


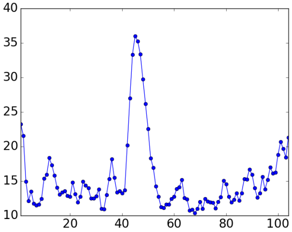
- $\epsilon_p \geq 0$: far enough to be unsychronized
- $\epsilon_{\text{sync}} \geq 0$: close enough to be synchronized
- $\epsilon_d \geq 0$

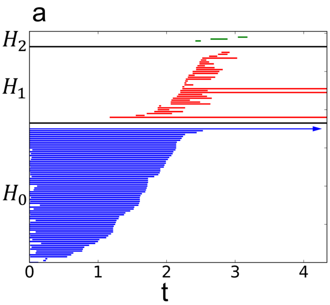
$$f(e_{jk}) = \begin{cases} \max\left\{ \min\{t \mid \int_t^\infty \|\widehat{\mathbf{u}}_j^i(t') - \widehat{\mathbf{u}}_k^i(t')\|_2 \, dt' \leq \epsilon_{\text{sync}}\}, f(n_j), f(n_k) \right\}, & \text{if } d_{jk}^{\text{org}} \leq \epsilon_d \\ t_{\text{sync}}^i, & \text{if } d_{jk}^{\text{org}} > \epsilon_d. \end{cases}$$

Results for two example proteins

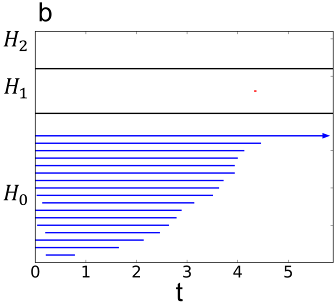
PDB:2NUH





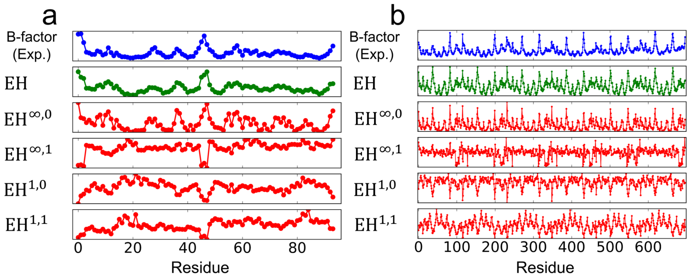


Residue 49



Residue 6

Featurization



Barcode

$B_i^k =$ k-dim barcode perturbing residue i

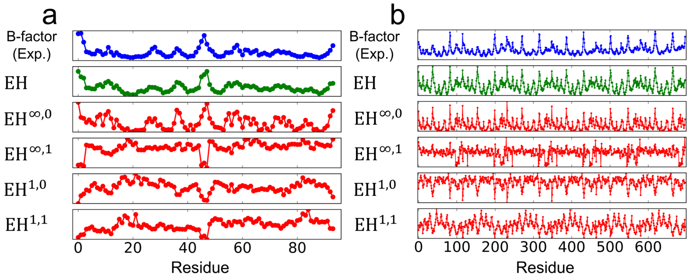
Feature

$$\mathrm{EH}_i^{p,k} = d_{W,p}(B_i^k, \emptyset),$$

Results: Predicting B-factor

Pearson correlation coefficient

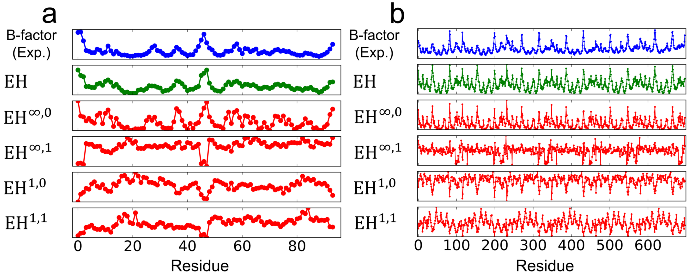
Method	R_P
$EH^{\infty,0}$	0.586
$EH^{\infty,1}$	-0.039
$EH^{\infty,2}$	-0.097
$EH^{1,0}$	-0.477
$EH^{1,1}$	-0.381
$EH^{1,2}$	-0.104
$EH^{2,0}$	0.188
$EH^{2,1}$	-0.258
$EH^{2,2}$	-0.100
EH	0.691



Results: Predicting B-factor

Pearson correlation coefficient

Method	R_P
$EH^{\infty,0}$	0.586
$EH^{\infty,1}$	-0.039
$EH^{\infty,2}$	-0.097
$EH^{1,0}$	-0.477
$EH^{1,1}$	-0.381
$EH^{1,2}$	-0.104
$EH^{2,0}$	0.188
$EH^{2,1}$	-0.258
$EH^{2,2}$	-0.100
EH	0.691

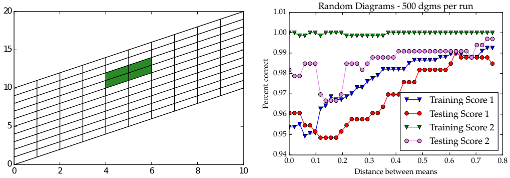


Comparison with state of the art

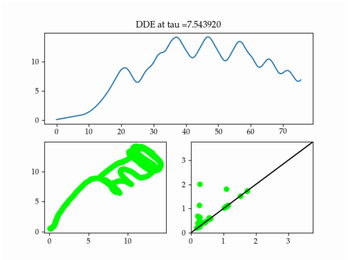
Method	R_P	Description
EH	0.691	Topological metrics
mFRI	0.670	Multiscale FRI
pfFRI	0.626	Parameter free FRI
GNM	0.565	Gaussian network model

Future work

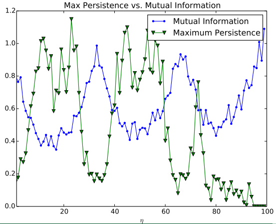
Extending the ML Theory
EM, Jose Perea



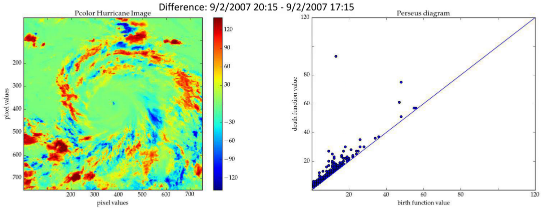
Chaos Characterization
Brian Bollen, Firas Khasawneh, EM



Parameter identification
Firas Khasawneh, EM, Chris Sukhu



Matrix-valued time series
Sarah Tymochko, EM, Kristen Corbosiero, Ryan Torn, Jason Dunion



Thank you!

Relevant papers

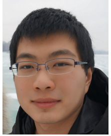
- FK, EM. *Topological Data Analysis for True Step Detection in Piecewise Constant Signals*. arXiv:1805.06403, 2018.
- ZC, EM, GW. *Evolutionary homology on coupled dynamical systems*. arXiv:1802.04677, 2018.
- FK, EM. *Chatter detection in turning using persistent homology*. MSSP, 2016.
- FK, EM, JP. *Chatter Classification in Turning Using Machine Learning and Topological Data Analysis*. arXiv:1804.02261, 2018.

Collaborators:


Firas Khasawneh


Jose Perea


Guo-wei Wei


Zixuan Cang

Teaspoon code available:
gitlab.msu.edu/TSAwithTDA/teaspoon

NSF

elizabethmunch.com
muncheli@egr.msu.edu

CMSE
Dept. of Computational Mathematics, Science, and Engineering

MICHIGAN STATE UNIVERSITY Founded 1855