Tokyo Research Laboratory

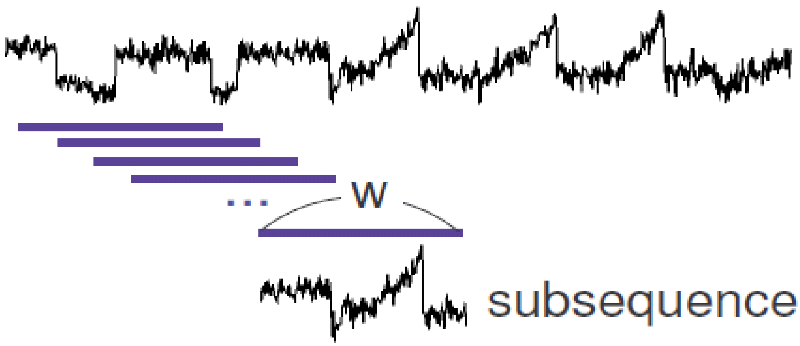
Why does subsequence time-series clustering produce sine waves?

IBM Tokyo Research Lab.
Tsuyoshi Idé

Contents

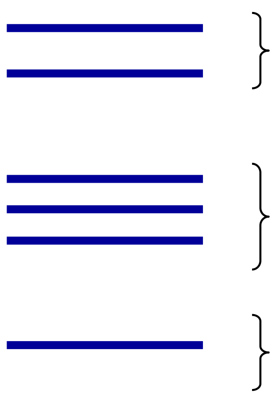
What is subsequence time-series clustering (STSC)?

What’s STSC: k-means clustering of subsequences generated from a time series. Cluster centers are patterns discovered.



(a) sliding window technique

- **subsequences generated by sliding window techniques**
- **subsequences are treated as independent data objects in k-means clustering**

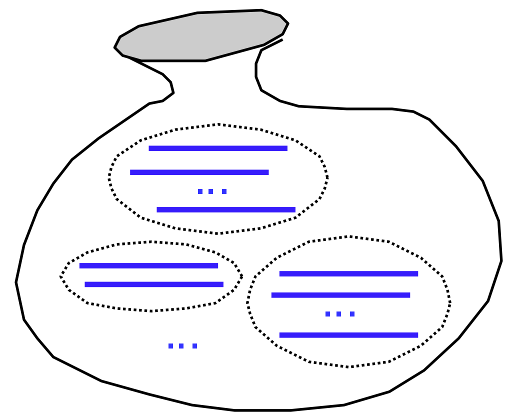


<u>Cluster centers (*centroids*)</u>

(the average of cluster members)

⇩

extracted pattern



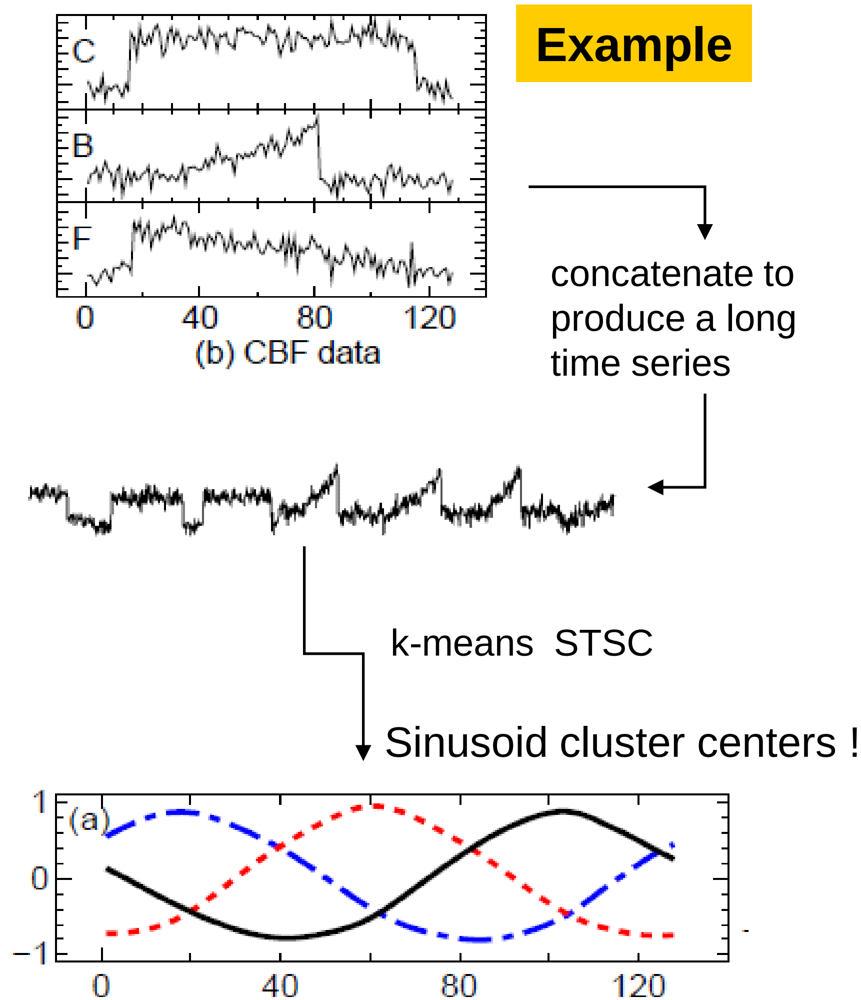
What's *sinusoid effect*: unexpectedly, cluster centers in STSC become sinusoids. The reason is unknown.

- **Shocking report**
  - Keogh-Lin-Truppel, "Clustering of time series subsequences is ***meaningless***", ICDM '03

- **k-means STSC *almost always* produces sinusoid cluster centers**
  - almost independent of the input time series
  - almost no relation to the original patterns

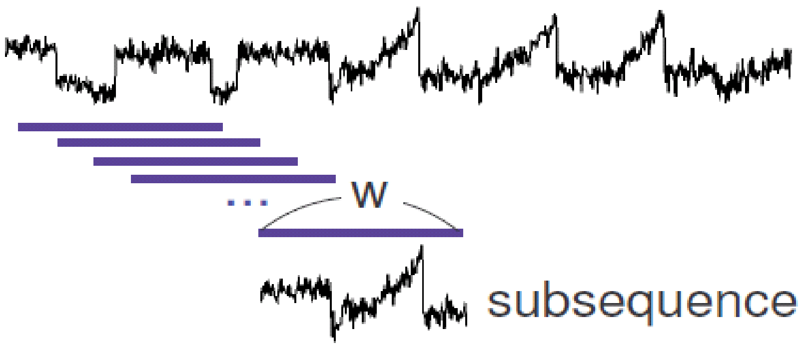
Explaining why is an open problem

We focus on explaining why.



Describing the dependence between subsequences

In reality, the subsequences are NOT independent at all.
We need to describe the dependence.



(a) sliding window technique

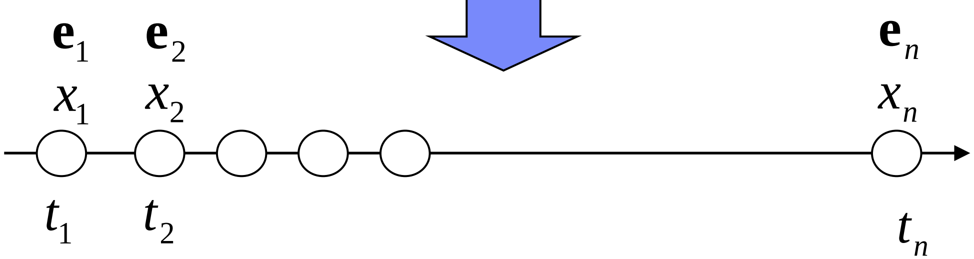
- subsequences generated by sliding window techniques
- subsequences are treated as **independent** data objects in clustering

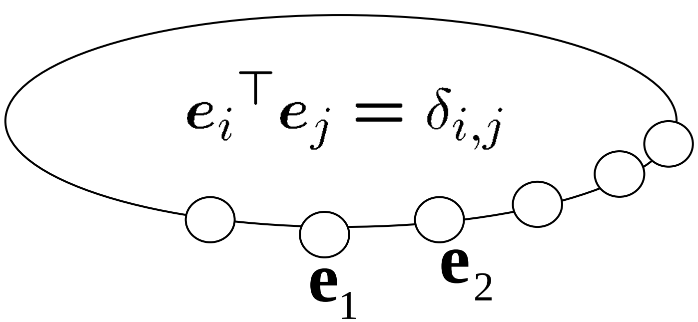
?

Let us study how the subsequences are dependent.

Theoretical model for time series:
Think of a time series as a “state” on a periodic ring.

$$\{x_t \mid t = 1,2,..,n\}$$





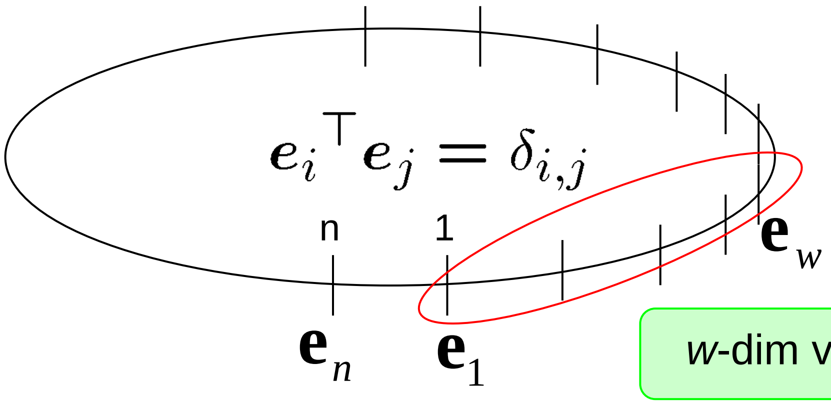
- Assign the value x_l on each lattice points (sites) l.
- Attach the orthonormal basis to the sites.
- Think of the time series as an n-D vector.

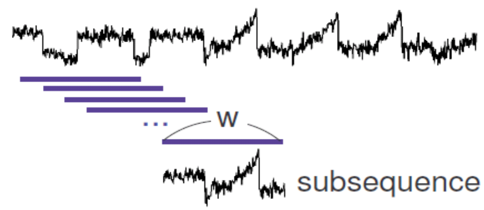
Whole time series

$$\boldsymbol{\Gamma} = \sum_{l=1}^{n} x_l \boldsymbol{e}_l$$

Artificially assume the periodic boundary condition (PBC)

Each subsequence s_p is concisely expressed using the translation operator τ





(a) sliding window technique

$$\Gamma = \sum_{l=1}^{n} x_l e_l$$

w-dim vector

w: window size

$$s_p \doteq \tau(-p)\Gamma$$

"Make *p* steps backward and take the sites from 1st thru w-th"

Definition of the translation operator

$$\tau(l) \equiv \sum_{l'=1}^{n} e_{l'+l} {e_{l'}}^T$$

Ex. shifts $\mathbf{e}_2$ with *l* steps $\tau(l)e_2 = \sum_{l'=1}^{n} e_{l'+l} {e_{l'}}^T e_2 = e_{2+l}$

Easer to analyze theoretically

Reducing *k*-means to eigen problem

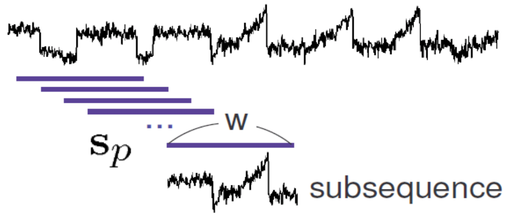
Simple but theoretically a little difficult to handle

Rewriting the objective of k-means using the indicator and the density matrix.

The objective function of k-means clustering

$$E = \sum_{j=1}^{k} \sum_{p \in \mathcal{C}_j} \left|\left| \mathbf{s}_p - \underbrace{\mathbf{m}^{(j)}}_{\text{centroid}} \right|\right|^2$$

objective to find $m^{(j)}$



$$= \text{const.} - \sum_{j=1}^{k} |\mathcal{C}_j| \, \mathbf{m}^{(j)\top} \mathbf{m}^{(j)}$$

Inserting the def of the centroid, and introducing an indicator $\mathbf{u}^{(j)}$ as

$$\mathbf{s}_p^{\mathrm{T}} \mathbf{u}^{(j)} = \begin{cases} 1/\sqrt{|C_j|} & \text{for } p \in C_j \\ 0 & \text{otherwise} \end{cases}$$

We finally get

objective to find $u^{(j)}$.

$$\boxed{E = \text{const.} - \sum_{j=1}^{k} ||\rho \boldsymbol{u}^{(j)}||^2}$$

$$\rho = \sum_{p=1}^{n} \boldsymbol{s}_p \boldsymbol{s}_p{}^T,$$

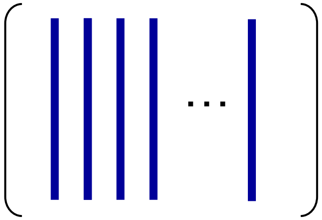
“density matrix”

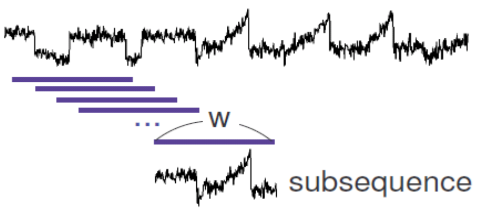
So what? Minimizing E is equivalent to eigen equation. Our goal is to solve it and to show the solution to be sinusoidal.

Minimizing E is equivalent to the eigen equation:

$$\rho \boldsymbol{u}^{(j)} = \lambda_j \boldsymbol{u}^{(j)} \ \text{ s.t. } \ \boldsymbol{u}^{(i)^T} \boldsymbol{u}^{(j)} = \delta_{i,j}$$

$$\rho = \mathrm{HH}^T, \qquad \mathrm{H} =$$





(a) sliding window technique

From the definition, it can be shown

eigenvector $\boldsymbol{u}^{(j)} \propto \boldsymbol{m}^{(j)}$ centroid

Let us study the sinusoid effect as ρ's eigen equation.

Deriving sine waves

By solving the eigen equation

Mathematical feature of ρ . The expression based on τ implies a translational symmetry. Fourier basis will simplify the problem.

By using $\boldsymbol{s}_p \doteq \tau(-p)\boldsymbol{\Gamma}$, the rho matrix can be written as

Summation of shifted ones

$$\rho = \sum_{p=1}^{n} \boldsymbol{s}_p \boldsymbol{s}_p^{\ T} \quad \Rightarrow \quad \sum_{p=1}^{n} \tau(l)^\top \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \tau(l)$$

This form suggests a (pseudo-) translational symmetry of the problem.

Translational invariant basis would be more natural

So, use the Fourier representation instead of the site representation

$$\boldsymbol{f}_q = \frac{1}{\sqrt{w}} \sum_{l=1}^{w} \mathrm{e}^{\mathrm{i} f_q (l - l_0)} \boldsymbol{e}_l$$

orthogonal transform

Window size

$$f_q = \frac{2\pi q}{w}, \quad q = -\frac{w}{2} + 1, ..., 0, 1, 2, ..., \frac{w}{2}$$

wave number

When represented in the Fourier basis, ρ is almost diagonal. Thus, the eigenstate is almost pure sinusoid.

If we take { $\boldsymbol{f}_q$ } as the basis, it follows (after straightforward calculations)

$$\rho \to n \begin{pmatrix} * & & & 0 \\ & * & & \\ & & * & \\ 0 & & & * \end{pmatrix} + \text{(off-diagonal)}$$

power of a Fourier component f_q

Will be small when a f_q is dominant.

Theorem
When a $|f_q|$ is dominant, the eigen state is well approximated by the sine waves with the wavelength of w/|q|, irrespective of the details of the input time series data.

Experiments

Let us see the correspondence between the two formulations using standard data sets.

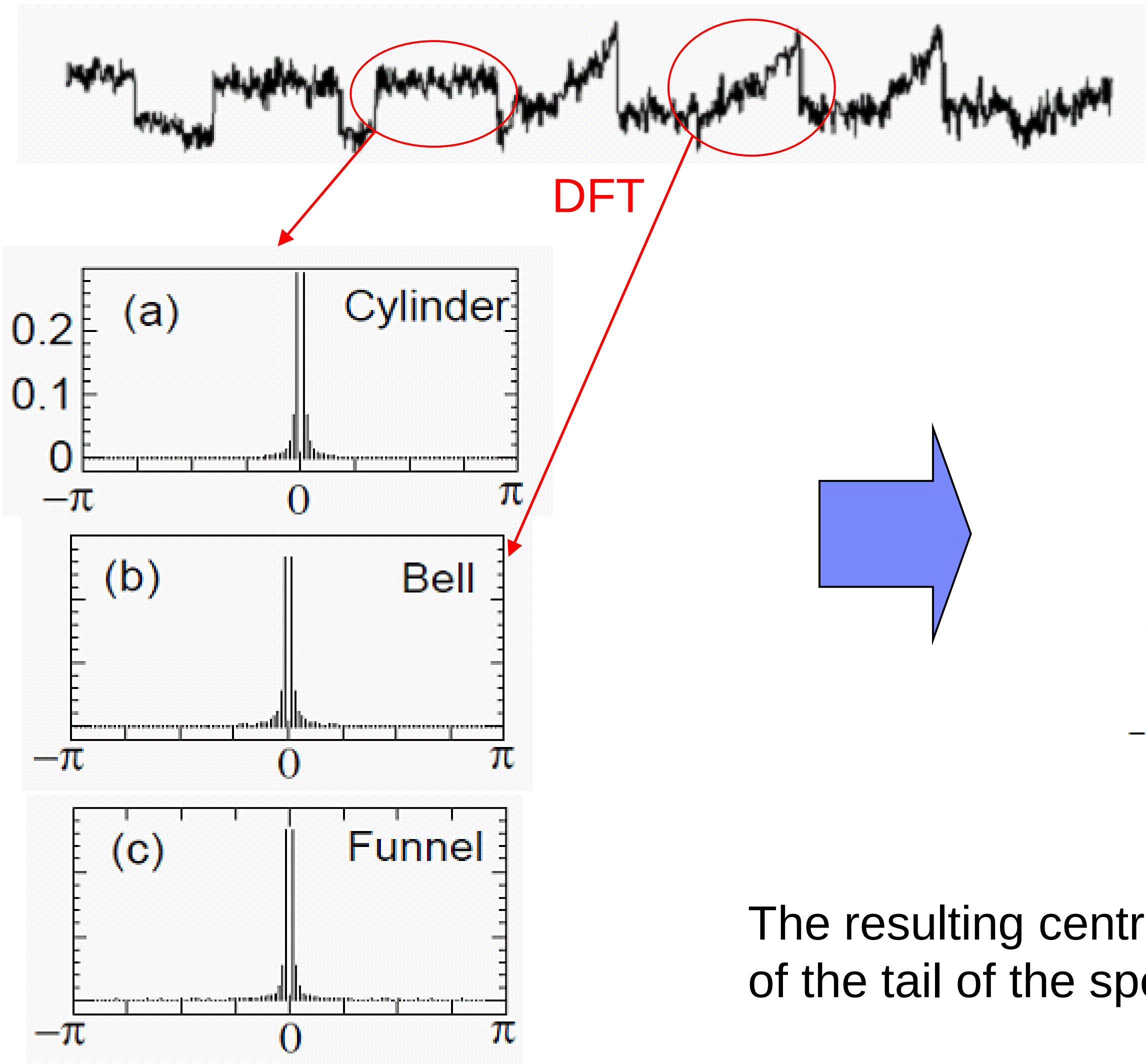
k-means STSC

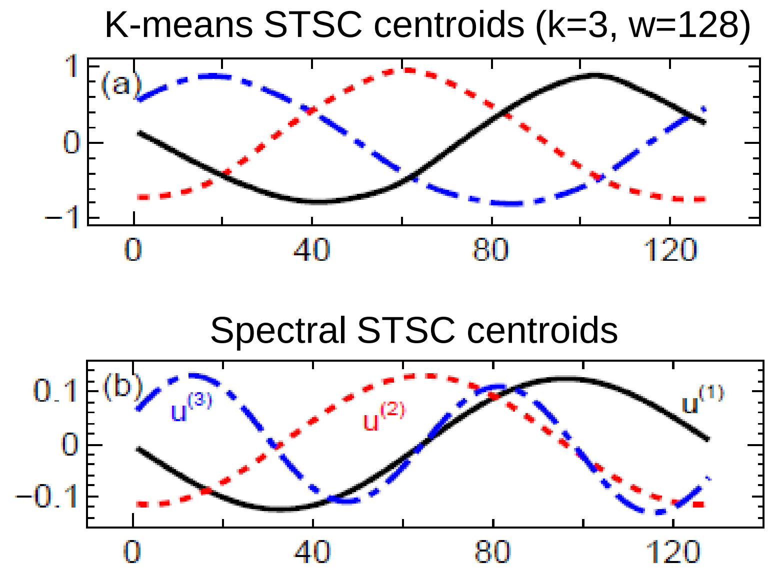
spectral STSC

k-means STSC

1. eigenvectors minimize the SoS objective
2. eigenvector → centroid
3. dominant F.c. governs the eigenvectors

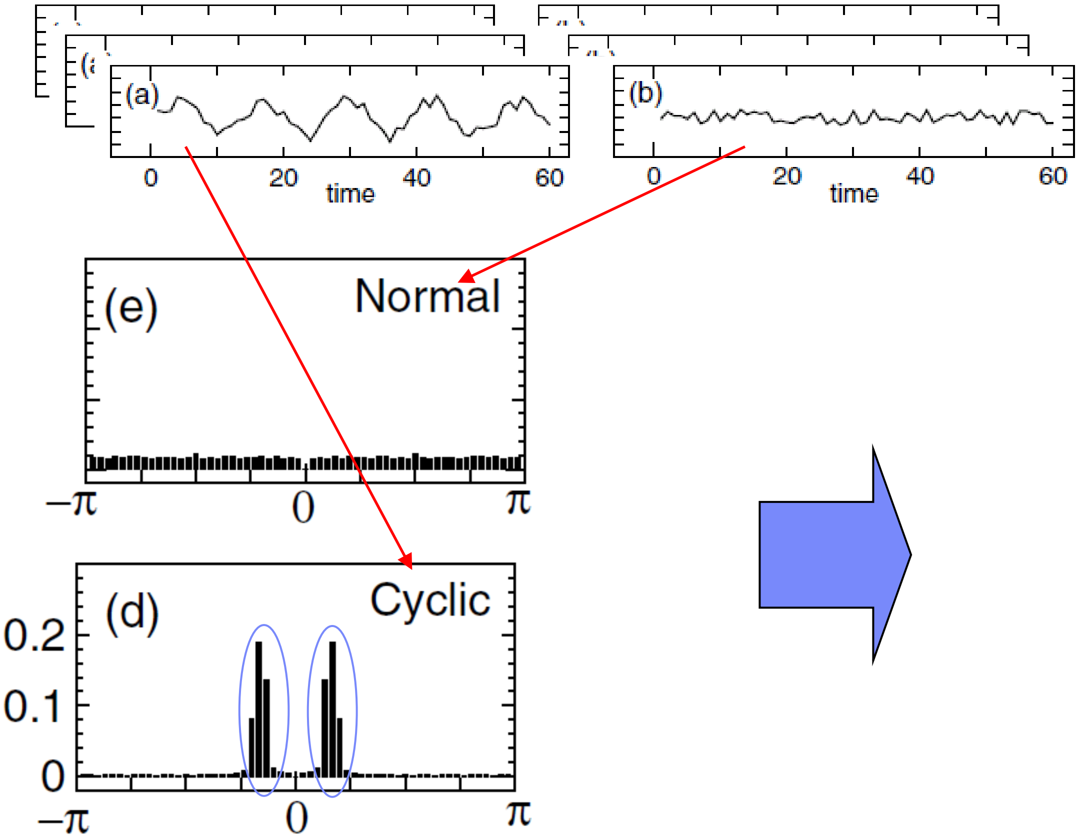
[1/2] For data with no particular periodicities, the power concentrates at the longest wavelength *w*. Only this peak does matter in the centroids.





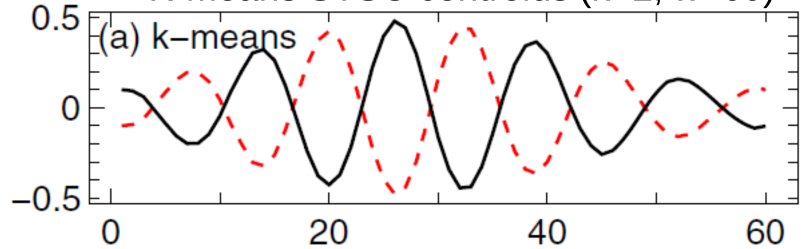
The resulting centroids are almost independent of the tail of the spectrum.

[2/2] STSC centroids become beat waves when a few neighboring |q| are dominant (k=2, w=60).

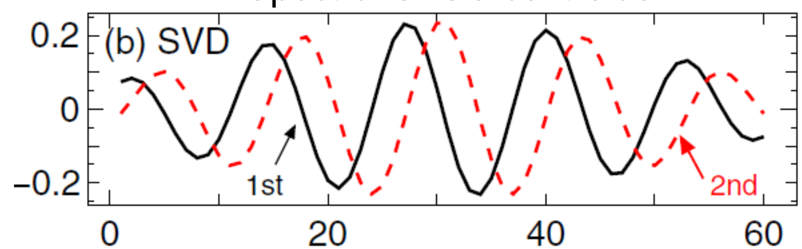


Concatenate 100 instances for each to make a long time series

K-means STSC centroids (k=2, w=60)



Spectral STSC centroids



|q| = 4, 5, 6 are dominant

Resulting sine waves exhibit *beat wave* by interference

Summary

Summary

- **The sinusoid effect is an important open problem in data mining.**
- **The pseudo-translational symmetry introduced by the sliding window technique is the origin of the sinusoid effect.**
- **In particular, if there is no particular periodicities within the window size, the clustering centers will be the sine waves of wavelength of *w*, irrespective of the details of the data.**

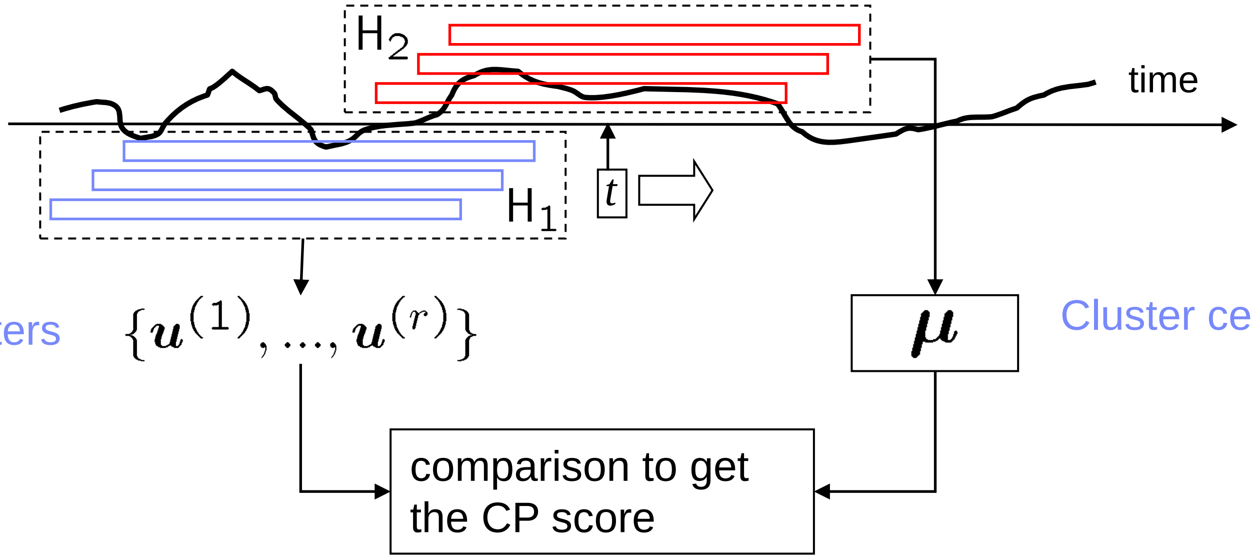
Thank you.

Appendix

STSC can produce useful results IF some of the conditions of the sinusoid effect are NOT satisfied.

For example, if STSC is done ***locally***…

A STSC-based change-point detection method (singular spectrum transformation [Ide, SDM'05])



Cluster centers in the past

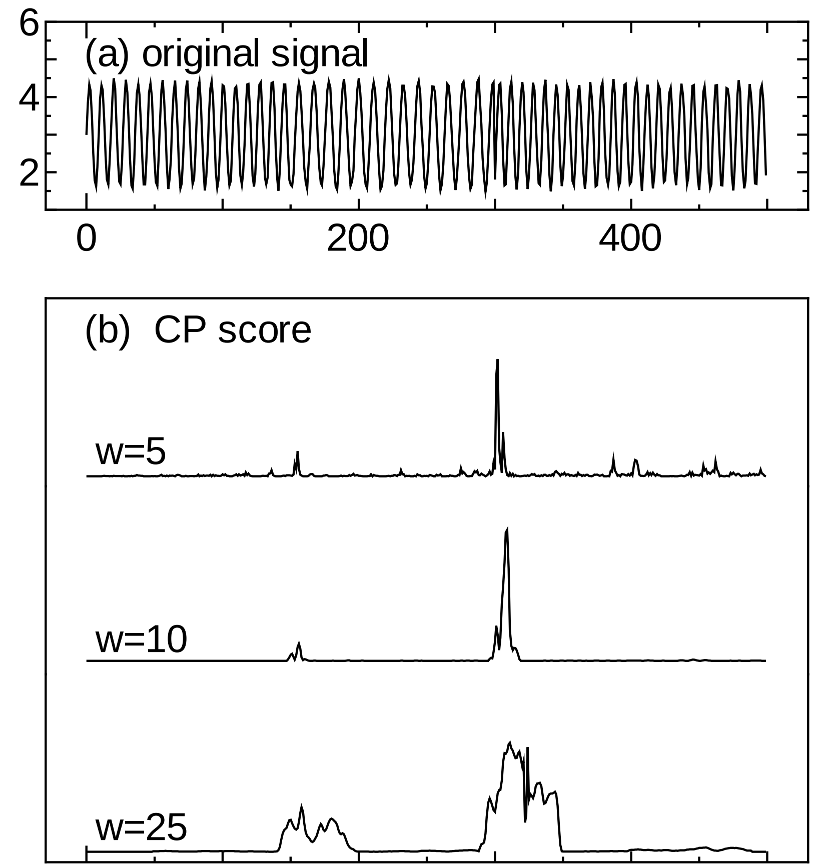
Cluster center at present

In this case, the locality of STSC leads to the loss of (pseudo) translational symmetry , resulting non-meaningless results.

SST can produce useful CP detection results, which do depend on the input signal.

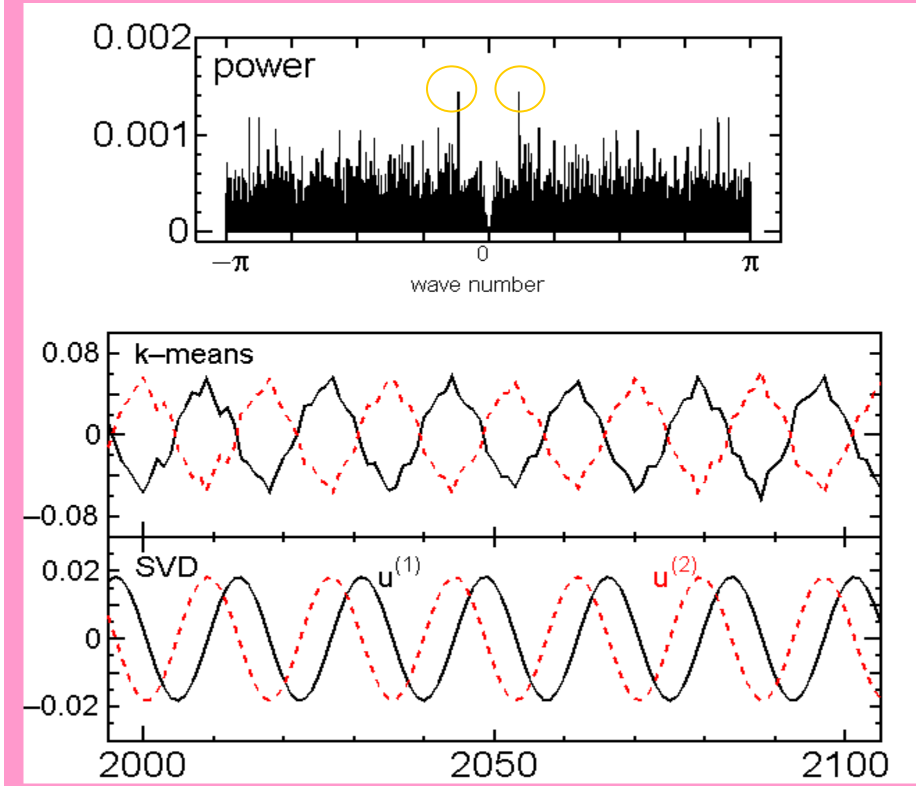
One general rule could be…

“Break the pseudo-translational symmetry”



Even for the random data, centroids become sinusoid when w = n (k=2, w=60 and 6000).

k=2, w=n=6000 (subseq = total)



Implication

The k-means STSC introduces a mathematical artifact.

It is so strong that the resulting centroids are dominated by it.