Why does Subsequence Time-Series Clustering Produce Sine Waves?

Tsuyoshi Idé

IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato, 242-8502 Kanagawa, Japan
goodidea@jp.ibm.com

Abstract. Data mining and machine leaning communities were surprised when Keogh et al. (2003) pointed out that the k-means cluster centers in subsequence time-series clustering become sinusoidal pseudo-patterns for almost all kinds of input time-series data. Understanding this mechanism is an important open problem in data mining. Our new theoretical approach (based on spectral clustering and translational symmetry) explains why the cluster centers of k-means naturally tend to form sinusoidal patterns.

1 Introduction

Subsequence time-series clustering (STSC) is one of the best-known pattern discovery techniques from time series data. In STSC, a time series data is represented as a set of subsequence vectors generated using the sliding window (SW) technique (see Fig. 1 (a)), and the generated subsequences are grouped typically using the k-means clustering technique. The cluster centers (the mean vector of the cluster members) are thought of as representative patterns of the time series.

STSC-based stream mining methods enjoyed popularity until a surprising fact was discovered in 2003 [8]: k-means STSC is "meaningless" as a pattern discovery technique in that the resultant cluster centers tend to form sinusoidal pseudo-patterns almost independent of the input time series.

For clarity, we reproduced the result of Ref. [8]. Figsure 2 (a) shows the k-means cluster centers calculated for the time series in Fig. 1 [1]. We set the number of clusters and the window size of SW to be $k = 3$ and $w = 128$, respectively. It is surprising that we have sinusoidal patterns in Fig. 2 (a), which are not similar to the original patterns in the data at all. Close inspection shows that the three sinusoids have the same wavelength of w, separated by a phase of $2\pi/3$.

So far, little effort has been made to theoretically pinpoint the origin of the sinusoidal pseudo-patterns, or the *sinusoid effect*. Empirical studies are substantially the only way to validate the attempts to improve STSC. It seems that the lack of theoretical understanding is causing a lack of progress in this area.

[1] A long time series (an example segment is shown in Fig. 1 (a)) was made by concatenating 90 random instances of the Cylinder, Bell, and Funnel (CBF) patterns, whose example instances are shown in Fig. 1 (b).

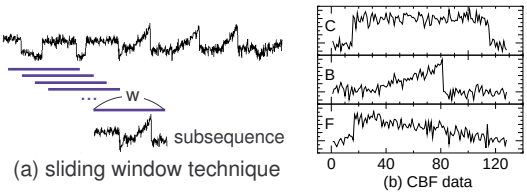

Fig. 1. (a) Sliding window technique and example segment of the concatenated CBF data. (b) Instances of the Cylinder(C)-Bell(B)-Funnel(F) data. There are 30 random instances for each.

This is a theoretical paper. We theoretically show that the SW-based k-means STSC introduces a mathematical artifact to the data, and, unexpectedly, that the artifact is so strong that the resulting cluster centers are dominated by it, irrespective of the details of the data. To the best of the author's knowledge, this is the first work that succeeds in theoretically explaining the sinusoid effect.

The layout of this paper is as follows. In Section 2, we summarize the sinusoid effects and point out a connection to spectral clustering. In Section 3, we present a new theoretical model for time series, which enables us to easily analyze symmetry properties hidden within the problem. In Section 4, we point out that k-means cluster centers can be found by directly solving an eigen equation. In Section 5, we explicitly show why the cluster centers in STSC become sinusoids. In Section 6, we validate our formulation using standard data sets. In the final section, we summarize the paper.

2 Sinusoid Effect in Spectral Clustering

Recently, spectral techniques have attracted great attention as a powerful method for clustering. Some authors [10, 9, 2, 3] has shown the theoretical connection between k-means and certain eigen problems. One interesting question here is whether or not the sinusoid effect is observed in spectral formulations of STSC. Experimentally, it seems that the answer is yes [6]. Specifically, if we think of subsequences generated from a time series as column vectors, and define a matrix H by putting the vectors as columns, the resulting left singular vectors of H will form sinusoids. We show in Fig. 2 (b) the top three left singular vectors calculated for the same concatenated CBF data. We see that the first ($\boldsymbol{u}^{(1)}$) and the second ($\boldsymbol{u}^{(2)}$) ones are sine waves with wavelength of w, showing clear similarities to Fig. 2 (a).

To summarize these observations in the CBF data,

Observation 1 *The cluster centers of k-means STSC are well approximated by sinusoids with a wavelength of w. While the additive phases are unpredictable, each sinusoid is separated by a phase of integer multiples of $2\pi/k$.*

Observation 2 *The left singular vectors of the subsequence matrix H are well approximated by sinusoids. A few top singular vectors have the wavelength of w.*

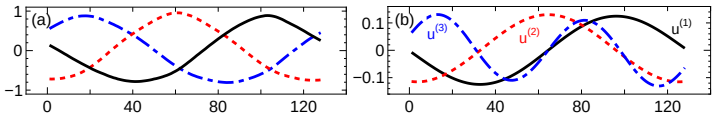

Fig. 2. (a) The k-means cluster centers ($k = 3, w = 128$). (b) The top three feature vectors by SVD ($w = 128$).

These observations suggest that the k-means and singular value decomposition (SVD) of H has a common mathematical structure, and the commonality is the origin of the sinusoid effect. Encouraged by this, we will elucidate the sinusoid effect (1) by reducing the k-means task to that of spectral clustering, and (2) by focusing on the translational symmetry of the problem. In (1), we will introduce a new formulation which directly seeks the cluster centers, instead of the standard formulation based on membership indicators.

3 Preliminaries

3.1 Lattice model for time series analysis

We define a time series $\mathit{\Gamma}$ as an ordered set of n real-valued variables $x_1, x_2, ..., x_n$. Given a $\mathit{\Gamma}$, a subsequence s_p of length $w \leq n$ is defined by $(x_p, x_{p+1}, ..., x_{p+w-1})$. A subsequence s_p can be viewed as a w-dimensional column vector $\boldsymbol{s}_p$. In STSC, $\boldsymbol{s}_p$s are thought of as independent vectorial data objects. We focus on SW-based STSC with unit step size and a fixed window size of w in this paper. The number of clusters is represented by k. All vectors are column vector hereafter.

Any time series $\mathit{\Gamma}$ can be represented as a vector $\boldsymbol{\mathit{\Gamma}}$ in an n-dimensional space. Consider a vector space $\mathcal{H}_0$ spanned by orthonormal bases $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_n\}$, and attach each base $\boldsymbol{e}_l$ to each time point l. By the orthonormality, $\boldsymbol{\mathit{\Gamma}}$ can be written as

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

with $x_l = \boldsymbol{e}_l^{\mathrm{T}} \boldsymbol{\mathit{\Gamma}}$. We call this expression the site-representation (SR) because we can think of our model as the one where each weight x_l is associated with each lattice point or *site* of a one-dimensional lattice having n lattice points.

3.2 Linear operators in $\mathcal{H}_0$

Let $\mathcal{L}$ be the set of linear operators which transforms a vector in $\mathcal{H}_0$ into another vector. We distinguish the operators by using $\hat{}$ hereafter. By definition, $\forall \hat{o} \in \mathcal{L}$ can be written as a matrix. In particular, it can be written with outer products of the bases in the SR as

$$\hat{o} = \sum_{l,l'=1}^{n} o_{l,l'} \boldsymbol{e}_l \boldsymbol{e}_{l'}^{\mathrm{T}}, \tag{2}$$

where the superscript T represents transpose.

The translation operator $\hat{\tau}(l)$

$$\hat{\tau}(l) \equiv \sum_{l'=1}^{n} \boldsymbol{e}_{l'+l} \boldsymbol{e}_{l'}^{\mathrm{T}} \tag{3}$$

is of particular importance. It is easy to verify $\hat{\tau}(l)\boldsymbol{e}_m = \boldsymbol{e}_{m+l}$ and $\boldsymbol{e}_m^{\mathrm{T}}\tau(l) = \boldsymbol{e}_{m-l}{}^{\mathrm{T}}$. The latter suggests

$$\hat{\tau}(l)^{\mathrm{T}} = \hat{\tau}(-l). \tag{4}$$

Hereafter, we assume the *periodic boundary condition* (PBC) to satisfy $\forall l, \boldsymbol{e}_{l+n} = \boldsymbol{e}_l$. As long as $n \gg 1$, the discrepancies due to this artificial condition will be negligible.

3.3 Discrete Fourier transformation

Consider a subspace $\mathcal{H}$ spanned by $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_w\} \subseteq \mathcal{H}_0$. Here we do not assume the periodicity of w in $\mathcal{H}$. For example, $\boldsymbol{e}_1 \neq \boldsymbol{e}_{1+w}$ unless $w = n$.

We define an orthogonal transformation from the site basis into the Fourier basis as

$$\boldsymbol{f}_q = \frac{1}{\sqrt{w}} \sum_{l=1}^{w} \mathrm{e}^{\mathrm{i}f_q(l-l_0)} \boldsymbol{e}_l \; ; \quad \boldsymbol{e}_l = \frac{1}{\sqrt{w}} \sum_{q \in \mathcal{D}_f} \mathrm{e}^{-\mathrm{i}f_q(l-l_0)} \boldsymbol{f}_q, \tag{5}$$

where l_0 is an arbitrary real number. For simplicity, we abuse the notation f_q to represent $2\pi q/w$, which we call the *wave number*. The subscript q runs over $\mathcal{D}_f = \{-\frac{w-1}{2}, ..., 0, 1, .., \frac{w-1}{2}\}$ when w is odd, and over $\{-\frac{w}{2}+1, ..., 0, 1, .., \frac{w}{2}\}$ when w is even. It is straightforward to show $\boldsymbol{f}_q^{\mathrm{T}} \boldsymbol{f}_{q'} = \delta_{q',q}$, and thus, $\{\boldsymbol{f}_q\}$ forms a complete set in $\mathcal{H}$.

For $\forall \boldsymbol{\gamma} \in \mathcal{H}$, the discrete Fourier transformation (DFT) is defined as

$$\boldsymbol{\gamma} = \sum_{q \in \mathcal{D}_f} \boldsymbol{f}_q \langle f_q | \gamma \rangle \; ; \quad \langle f_q | \gamma \rangle = \sum_{l=1}^{w} \langle f_q | e_l \rangle \langle e_l | \gamma \rangle, \tag{6}$$

where $\langle f_q | e_l \rangle = \frac{1}{\sqrt{w}} \mathrm{e}^{-\mathrm{i}f_q(l-l_0)}$, and we used the bracket notation to represent the inner product bewtween vectors ($\langle e_l | \gamma \rangle \equiv \boldsymbol{e}_l^{\mathrm{T}} \boldsymbol{\gamma}$, etc). We call the representation based on $\{\boldsymbol{f}_q\}$ the Fourier representation (FR). If γ is the expression of real-valued time series data, the weight on l must be real, so it follows that

$$\langle e_l | \gamma \rangle = \frac{1}{\sqrt{w}} \sum_{q \in \mathcal{D}_f} |\langle f_q | \gamma \rangle| \cos(f_q l + \phi), \tag{7}$$

where $\phi = -f_q l_0 + \arg\langle f_q | \gamma \rangle$.

4 Density Matrix Formulation of k-Means

4.1 Objective function of k-means

Consider a general k-means clustering task for a set of vectors $\{\boldsymbol{s}_p \in \mathcal{H} \mid p = 1, 2, ..., n\}$. It is well-known that the k-means algorithm attempts to minimize the sum-of-squared (SOS) error [4]:

$$E = \sum_{j=1}^{k} \sum_{p \in \mathcal{C}_j} \left|\left| \boldsymbol{s}_p - \boldsymbol{m}^{(j)} \right|\right|^2 = \sum_{p=1}^{n} \langle s_p | s_p \rangle - \sum_{j=1}^{k} |\mathcal{C}_j| \langle m^{(j)} | m^{(j)} \rangle, \tag{8}$$

where $\mathcal{C}_j$ and $|\mathcal{C}_j|$ represent the members of the j-th cluster and the number of members in the cluster, respectively. The centroid of $\mathcal{C}_j$ is denoted by $\boldsymbol{m}^{(j)}$.

The first term does not depend on the clustering. For the second term, E_2, by substituting the definition of the centroid $\boldsymbol{m}^{(j)} = \frac{1}{|\mathcal{C}_j|} \sum_{p \in \mathcal{C}_j} \boldsymbol{s}_p$, it becomes

$$E_2 = -\sum_{j=1}^{k} \frac{1}{|\mathcal{C}_j|} \sum_{p,r \in \mathcal{C}_j} \langle s_p | s_r \rangle. \tag{9}$$

To remove the restricted summation, we introduce an indicator vector $\boldsymbol{u}^{(j)} \in \mathcal{H}$, where $\langle s_p | u^{(j)} \rangle = 1/\sqrt{|\mathcal{C}_j|}$ for $\boldsymbol{s}_p \in \mathcal{C}_j$ and 0 otherwise, to have

$$E_2 = -\sum_{j=1}^{k} \sum_{p,r=1}^{n} \langle u^{(j)} | s_p \rangle \langle s_p | s_r \rangle \langle s_r | u^{(j)} \rangle.$$

Now introduce a linear operator $\hat{\rho}$ as

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

and call $\hat{\rho}$ the *density matrix*, following the statistical-mechanical terminology. Since the $\boldsymbol{s}_p$s are generated by the SW technique, we see

$$\hat{\rho} \doteq \sum_{l=1}^{n} \hat{\tau}(l)^{\mathrm{T}} \boldsymbol{\Gamma} \boldsymbol{\Gamma}^{\mathrm{T}} \hat{\tau}(l) \tag{10}$$

holds, where "$\doteq$" means "the left and the right sides have the same matrix elements when represented in $\mathcal{H}$ (not $\mathcal{H}_0$)".

Using $\hat{\rho}$, we get the final form of the objective function as

$$E_2 = -\sum_{j=1}^{k} \langle u^{(j)} | \hat{\rho}^2 | u^{(j)} \rangle, \tag{11}$$

where $\langle \cdot | \hat{o} | \cdot \rangle$ is defined as $\langle \cdot | \hat{o} \cdot \rangle$ for $\forall \hat{o} \in \mathcal{L}$. The k-means clustering task has now been reduced to seeking the solution $\{\boldsymbol{u}^{(j)}\}$ which minimizes E_2.

4.2 Connection to eigen problem

To this point, the vector $\boldsymbol{u}^{(j)}$ has been an artificially defined indicator to simplify the objective in Eq. (9). From the original definition, it is easy to see that $\{\boldsymbol{u}^{(j)}\}$ satisfy

$$\sum_{p=1}^{n} \langle u^{(i)}|s_p\rangle\langle s_p|u^{(j)}\rangle = \langle u^{(i)}|\hat{\rho}|u^{(j)}\rangle = \delta_{i,j}. \tag{12}$$

Now we relax the original binary restriction, and take this as the new restriction on the optimization problem, so that the k-means task is reduced to the generalized eigen problem which minimizes E_2 subject to Eq.(12). This eigen problem can be written as

$$\hat{\rho}\boldsymbol{u}^{(j)} = \lambda_j \boldsymbol{u}^{(j)} \ \text{ s.t. } \ \langle u^{(i)}|u^{(j)}\rangle = \delta_{i,j}, \tag{13}$$

where λ_j is the eigenvalue corresponding to the eigenstate $\boldsymbol{u}^{(j)}$ labeled in descending order of the eigenvalue. In the SR, $\langle e_l|\hat{\rho}|e_{l'}\rangle$ corresponds to the (l, l') element of $\mathsf{H}\mathsf{H}^{\mathrm{T}}$, where $\mathsf{H} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_n]$ (note that H has n columns by PBC). Thus, Eq. (13) can be written as

$$\mathsf{H}\mathsf{H}^{\mathrm{T}}\boldsymbol{u}^{(j)} = \lambda \boldsymbol{u}^{(j)}.$$

This equation also shows the $\boldsymbol{u}^{(j)}$s are the left singular vectors of H.

4.3 Cluster centers and eigenstates

Apart from the formal definition as the (relaxed) indicator, let us further consider the meaning of $\boldsymbol{u}^{(j)}$. Before the relaxation, the indicator satisfied

$$\boldsymbol{m}^{(j)} \equiv \frac{1}{|\mathcal{C}_j|}\sum_{p\in\mathcal{C}_j} \boldsymbol{s}_p = \frac{1}{\sqrt{|\mathcal{C}_j|}}\sum_{p=1}^{n} \boldsymbol{s}_p\langle s_p|u^{(j)}\rangle.$$

After the relaxation, $\boldsymbol{u}^{(j)}$ is the eigenstate of $\hat{\rho} = \sum_p \boldsymbol{s}_p\boldsymbol{s}_p^{\mathrm{T}}$. Thus, it follows that *the k-means cluster centers correspond to the eigenstates of $\hat{\rho}$*, or

$$\boldsymbol{m}^{(j)} \propto \boldsymbol{u}^{(j)}. \tag{14}$$

Note that our formulation directly seeks the cluster centers as the eigen vectors. This is in contrast to the standard spectral formulations [3].

Now, we summarize this section as Theorems:

Theorem 1 *The eigenstates of $\hat{\rho}$, which can be computed also as the left singular vectors of H, minimize the SOS objective.*

Theorem 2 *The eigenstates of $\hat{\rho}$ formally correspond to the k-means cluster centers.*

In spite of this, the correspondence between the k-means and our spectral formulation is not perfect. The major discrepancy comes from the fact that the eigenstates must be orthogonal to each other. The problem is that the cluster centers are not necessarily orthogonal in general. One reasonable expectation is that the top eigenstate $\boldsymbol{u}^{(1)}$ would be a good estimator representing the averaged direction of a few of the major k-means clusters. For the other eigenstates, the direction would be more or less influenced by the top one. [2] We will discuss this topic theoretically and experimentally later.

5 Fourier Representation of $\hat{\rho}$

5.1 The $w = n$ case

Let us consider the extreme case of $w = n$. In this case, $\mathcal{H}$ ($= \mathcal{H}_0$) can be thought of as periodic, so that the Fourier state $\boldsymbol{f}_q$ is the exact eigenstate of $\hat{\tau}(l)$. Explicitly,

$$\hat{\tau}(l)\boldsymbol{f}_q = \frac{1}{\sqrt{w}} \sum_{l'=1}^{n} \mathrm{e}^{\mathrm{i}f_q(l'-l_0)} \boldsymbol{e}_{l'+l} = \mathrm{e}^{-\mathrm{i}f_q l} \boldsymbol{f}_q. \tag{15}$$

Here we used the fact that $\mathrm{e}^{\mathrm{i}f_q n} = 1$ if $f_q = 2\pi q/n$.

Using Eqs. (10) and (15), we can calculate $\langle f_q|\hat{\rho}|f_{q'}\rangle$ as

$$\sum_{l=1}^{n} \langle f_q|\hat{\tau}(l)^{\mathrm{T}}|\Gamma\rangle\langle\Gamma|\hat{\tau}(l)|f_{q'}\rangle = \sum_{l=1}^{n} \langle f_q|\Gamma\rangle\langle\Gamma|f'_q\rangle \mathrm{e}^{\mathrm{i}(f_q - f_{q'})l} = n|\langle f_q|\Gamma\rangle|^2 \delta_{q,q'}, \tag{16}$$

which means the matrix representation of $\hat{\rho}$ is diagonal in FR. Thus, we conclude that *the Fourier state itself is the eigenstate of $\hat{\rho}$ completely independently of the input data*. Which f_q is chosen depends on the magnitude of $|\langle f_q|\Gamma\rangle|^2$, the *power* of the Fourier component. Note that the eigenstate must be a pure sinusoid even when the power spectrum does not have any dominant f_q. [3] When a q_1 was chosen, the resultant distribution is sinusoidal with the wave number f_{q_1} (see Eq. (7)). Thus, based on Theorems 1 and 2, the k-means cluster centers are expected to be approximated by the sinusoids apart from the orthogonality problem.

5.2 The $w < n$ case

For $w < n$, the $\boldsymbol{f}_q$s are not exactly the eigenstates of $\hat{\tau}(l)$, since $\mathcal{H}$ cannot be thought of as periodic. As a result, we have the matrix elements like

$$\langle f_q|\hat{\rho}|f_{q'}\rangle \approx n|\langle f_q|\Gamma\rangle|^2\delta_{q,q'} + \sum_{l=1}^{n} \mathrm{e}^{\mathrm{i}l\Delta_{q'q}} J_l(q,q'), \tag{17}$$

[2] The $k = 1$ case is special. The cluster center is the simple mean vector, and can be written as $|m\rangle = \sqrt{w}\bar{x}|f_0\rangle$, where $\bar{x}$ denotes the mean of Γ over the whole domain. This gives a constant distribution, having no relation with $\boldsymbol{u}^{(j)}$s.

[3] This is not the case when some of $|f_q|$s have exactly the same power. But it is unlikely in real-world time-series data under normal conditions.

instead of Eq. (16). It is straightforward to get the exact expression of $J_l(q, q')$ although we do not show it here. However, under normal conditions, we can assume that the first term is the leading term since $n \gg 1$ and phase cancellations are unavoidable in the second term. In particular, if the power spectrum has a single-peaked structure at $|f_q|$, which is the case in the CBF data (see the next section), the top eigenstate will be well approximated by $\boldsymbol{f}_{|q|}$, irrespective of the details of the spectrum. Again, Eq. (7) reads

$$\langle e_l | u \rangle \propto \cos(f_q l + \phi). \tag{18}$$

Since l_0 was arbitrary, the real number ϕ is also arbitrary. From this, we can naturally understand the unpredictability of the additive phase as stated in Observation 1. Now we get Theorem 3, which directly explains Observation 2:

Theorem 3 *When a $|q|$ is dominant, the singular vectors of H are well approximated by sinusoids with the wavelength of $w/|q|$, irrespective of the details of the input time series data.*

In addition, by considering Theorems 1 and 2, the k-means cluster centers will be sinusoidal except for the orthogonality problem. This is a mathematical explanation of Observation 1.

When the power spectrum is almost flat, the eigenvectors will be mixtures of many f_qs, so that the cluster centers will be far from pure sinusoids.

5.3 Optimizing the relative phases

If the data has a dominant q, the subsequences can be approximated as

$$\boldsymbol{s}_p = \sum_{q' \in \mathcal{D}_f} \boldsymbol{f}_{q'} \langle f_{q'} | s_p \rangle \approx \sum_{q' \in \mathcal{D}_f} \mathrm{e}^{\mathrm{i} f_{q'} p} \boldsymbol{f}_{q'} \langle f_{q'} | \Gamma \rangle \approx \boldsymbol{f}_q \langle f_q | \Gamma \rangle \mathrm{e}^{\mathrm{i} f_q p}. \tag{19}$$

Define $\boldsymbol{g}_{q,\phi} \in \mathcal{H}$ by $\langle e_l | g_{q,\phi} \rangle = \cos(f_q l + \phi)$. Equation (19) means that the k-means STSC is reduced to that for $\{\boldsymbol{g}_{q,\phi}\}$ with uniformly distributed ϕ.

Since $\{\boldsymbol{g}_{q,\phi}\}$ consists of sinusoids of f_q, the cluster centers must be sinusoids of f_q. Let the cluster centers be $\boldsymbol{g}_{q,\phi_j}$ $(j = 1, ..., k)$. The distribution of ϕ may be modeled as a continuous uniform distribution over $[0, 2\pi)$. The SOS objective is now written as

$$E(\phi_1, ..., \phi_k) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\phi \sum_{j=1}^{k} \theta_j(\phi) e(\phi, \phi_j), \tag{20}$$

where $1/(2\pi)$ represents the probability density of ϕ, and $e(\phi, \phi_j) \equiv ||\boldsymbol{g}_{q,\phi} - \boldsymbol{g}_{q,\phi_j}||^2 = 4 \sin^2 \frac{\phi - \phi_j}{2}$. The function $\theta_j(\phi)$ indicates cluster assignment, and takes 1 when the j-th cluster center is the closest to ϕ, 0 otherwise. For example, if we have $\phi_1 < \phi_2 < \phi_3$ when $k = 3$, $\theta_2(\phi)$ will be 1 for $\frac{\phi_1 + \phi_2}{2} \leq \phi < \frac{\phi_2 + \phi_3}{2}$ and 0 otherwise. Solving the minimization problem of E w.r.t. the phases is

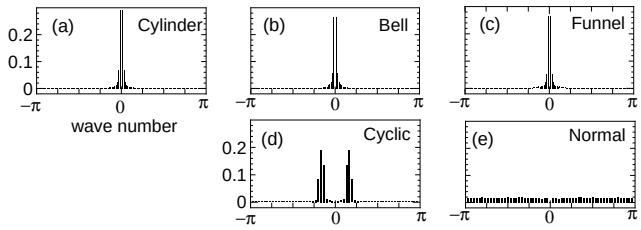

Fig. 3. Power spectra of each instance of the data.

straightforward but tedious. However, it is intuitively clear that the most balanced assignment is optimal. In fact, Eq. (20) is symmetric w.r.t. j. So, the solution should be also symmetric if it is the unique solution. Now we arrive at Theorem 4, which summarizes the theoretical proof of the phase issue in Observation 1:

Theorem 4 *If a Γ has a dominant f_q, the k-means STSC is reduced to that for uniformly distributed sinusoids of f_q. The optimal cluster centers are separated by a phase of integral multiples of $2\pi/k$.*

6 Experiments

6.1 Cylinder-Bell-Funnel data

The CBF data [7] includes three types of patterns of literal Cylinder, Bell, and Funnel shapes. We randomly generated 30 instances for each type (examples in Fig. 1 (b)) with a fixed length of 128 ($= w$) using the Matlab code provided by [7]. We also concatenated them in order after standardizing each one (zero mean and unit variance). We did 100 random restarts and chose the best one in the k-means calculation.

Figures 3 (a)-(c) show the power spectra of each instance as a function of the wave number. To handle the variation of the instances, we simply averaged the resultant spectra of all instances. We see that the most of the weight is concentrated on the $|q| = 1$ component in all of the cases. The f_0 component is naturally missing because of the standardization.

The results of k-means and SVD were shown in Fig. 2. The wavelength of w can be understood from the large $|q| = 1$ weight in Fig. 3 (a)-(c). Due to the orthogonality condition, the third singular vector necessarily has a wavelength of about $w/2$. This is an example of the difference between the two formulations in how the calculated cluster centers interact with each other. Apart from this, our formulation is completely consistent to the results.

6.2 Synthetic Control Chart data

The Synthetic Control Chart (SCC) data [7] consists of six types of 100 instances, each with 60 data values. Out of the six types, we focus on the Cyclic and

Normal types (Fig. 4), which have very different (averaged) power spectra from the CBF spectra, as shown in Fig. 3. We see that the weight is concentrated on the wavelengths of $\frac{w}{4}$, $\frac{w}{5}$, and $\frac{w}{6}$ in the Cyclic data ($w = 60$). In contrast, the distribution is almost flat for the Normal data, as expected for white noise.

We made a concatenated data set with 100 standardized Normal instances followed by 100 standardized Cyclic instances. Figures 5 (a) and (b) show the k-means cluster centers ($k = 2$, the best one among 100 random restarts) and the two highest singular vectors, respectively. We set $w = 60$. Since the $\boldsymbol{s}_p$s in the Normal part do not favor any particular direction, the clustering results seem to be dominated by the Cyclic part. In both figures, amplitude-modulated sinusoids with a periodicity of about $w/5$ are observed instead of pure sinusoids. The waves are separated by the phase intervals which can be naturally understood from Theorem 4 for (a) and from the orthogonality condition for (b).

The amplitude modulation can be understood as *beat* in physics. As shown in Fig. 3, the Cylinder part is dominated by the $f_{|4|}, f_{|5|}$ and $f_{|6|}$ components. Since SVD extracts the major direction of $\{\boldsymbol{s}_p\}$, the top singular vector $\boldsymbol{u}$ will be approximated as a linear combination of those components like

$$\langle e_l | u \rangle \approx \sum_{q=4}^{6} c_q \cos\left[f_q(l - l_0)\right].$$

Within the accuracy up to the order of $(2\pi/w)^2$, it reads

$$\langle e_l | u \rangle \propto \mathrm{e}^{-\frac{1}{2}\Delta_2{}^2(l-l_0)^2} \cos\left[(f_q - \Delta_1)(l - l_0)\right], \tag{21}$$

where $\Delta \equiv 2\pi/w$, and

$$\frac{\Delta_1}{\Delta} = \frac{c_4 - c_6}{c_4 + c_5 + c_6}, \quad \frac{\Delta_2}{\Delta} = \frac{\sqrt{4c_4c_6 + c_5(c_6 + c_4)}}{c_4 + c_5 + c_6}.$$

To derive this, we used Taylor expansion formulas such as ($\epsilon \ll 1$)

$$\ln\left(c_5 + c_6 \mathrm{e}^{\mathrm{i}\epsilon}\right) \approx \ln(c_5 + c_6) + \frac{\mathrm{i}\epsilon c_6}{c_5 + c_6} - \frac{\epsilon^2}{2}\frac{c_5 c_6}{(c_5 + c_6)^2}.$$

The line shape in Eq. (21) is known as beat in physics. If we set $c_q \propto |\langle f_q | \Gamma \rangle|$ (in the Cyclic part), we get $\Delta_1 = 0.1\Delta$ and $\Delta_2 = \Delta/1.3$ from Fig. 3 (d). This leads to a sine wave with wavelength $w/4.9$ modulated by a beat wavelength of $1.3w$. Except for the region where $\Delta(l - l_0) \simeq 1$, Equation (21) fairly explains Fig. 5.

It is interesting to see what happens when we have only the Normal data. As expected, the resulting cluster centers are far from sinusoids when $w = 60$ (not shown). However, STSC produces sinusoids when $w = n$ (=6000), despite the white noise nature of the data. Our theory clearly explains this counter-intuitive result. As discussed in Subsection 5.1, the top singular vector must be the pure sinusoid of the largest power. In this case, we have the largest power at $|f_q| = 0.358$ in Fig. 6 (a) (marked by the triangles). Thus, the wavelength must

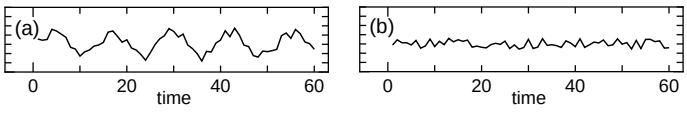

Fig. 4. Examples of (a) Cyclic and (b) Normal instances in the SCC data.

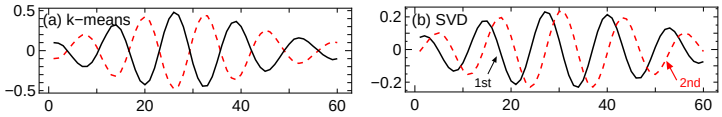

Fig. 5. (a) The k-means cluster centers and (b) the first and second singular vectors of H for the concatenated Normal-Cyclic data.

be $2\pi/|f_q| = 17.6$, which is completely consistent with Fig. 6 (c). In addition, we see that the singular vector is a good estimator of the k-means cluster center by comparing Figs. 6 (b) and (c).

While some authors attribute the sinusoidal patterns to simple smoothing effects caused by superposition of slightly shifted subsequences [1], such a discussion clearly fails to explain the origin of the sinusoidal curves for the Normal data, and that of the beat waves.

7 Concluding Remarks

We have performed theoretical analysis of the sinusoid effect in STSC. In particular, we pointed out that the k-means clustering task can be reduced to the eigen problem of the density matrix $\hat{\rho}$. Thanks to the translational symmetry of $\hat{\rho}$, the eigenstate can be approximated by a Fourier state if a single $|f_q|$ forms a conspicuous single peak in DFT. We also found that the k-means cluster centers produce beat waves (Fig. 5) when a few neighboring frequencies are dominant.

Mathematically, the sinusoid effect can be understood from the fact that the Fourier states are the irreducible representations of the translational group. In another paper [5], we used a point group for pattern discovery. This paper also can be seen as one of the first studies which introduce the concept of group into machine learning.

Our theory also provides a practical basis for attempts to make STSC meaningful. As long as the coherent superposition is used to define the cluster centers, sinusoid pseudo-patterns are more or less unavoidable. One possibility is to utilize incoherent superposition of subsequences. Medoid-based methods are perhaps the simplest way to use the incoherence, and are known to give better results than the simple STSC. Detailed discussion on this issue will appear elsewhere.

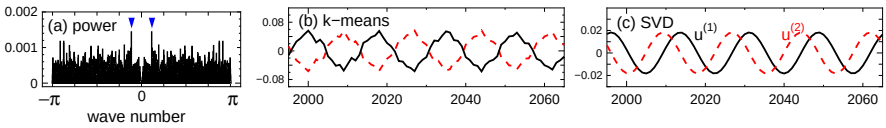

Fig. 6. The results for the concatenated Normal data with $w = n = 6000$. (a) The power spectrum, (b) a segment of the k-means cluster centers ($k = 2$), and (c) a segment of the top two left singular vectors.

Acknowledgement

The author thanks S. Jacobs for carefully reading the manuscript.

References

1. J. Chen. Making subsequence time series clustering meaningful. In *Proc. IEEE Intl. Conf. on Data Mining*, pages 114–121, 2005.
2. I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *Proc. Tenth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 551–556, 2004.
3. C. Ding and X. He. K-means clustering via principal component analysis. In *Proc. Int'l Conf. Machine Learning*, pages 225–232, 2004.
4. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, 2000.
5. T. Idé. Pairwise symmetry decomposition method for generalized covariance analysis. In *Proc. IEEE Intl. Conf. on Data Mining*, pages 657–660, 2005.
6. E. Keogh. Data mining and machine learning in time series databases. In *Tutorial Notes of the Tenth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*. 2004.
7. E. Keogh and T. Folias. The UCR time series data mining archive [`http://www.cs.ucr.edu/~eamonn/TSDMA/index.html`]. 2002.
8. E. Keogh, J. Lin, and W. Truppel. Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Proc. IEEE Intl. Conf. on Data Mining*, pages 115–122. IEEE, 2003.
9. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems, 14*, pages 849–856, 2001.
10. H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems, 14*, pages 1057–1064, 2001.