Translational symmetry in subsequence time-series clustering

Tsuyoshi Idé

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

Abstract. We treat the problem of subsequence time-series clustering (STSC) from a group-theoretical perspective. First, we show that the sliding window technique introduces a mathematical artifact to the problem, which we call the pseudo-translational symmetry. Second, we show that the resulting cluster centers are necessarily governed by irreducible representations of the translational group. As a result, the cluster centers necessarily forms sinusoids, almost irrespective of the input time-series data. To the best of the author's knowledge, this is the first work which demonstrates the interesting connection between STSC and group theory.

1 Introduction

Learning representative patterns from time series data is one of the most interesting tasks in data mining. Since the advent of a seminal work of Das et al. [1], subsequence time-series clustering (STSC) had enjoyed popularity as the simplest and the most reliable technique of stream mining. In STSC, time series data is represented as a set of subsequence vectors generated using a sliding window (see Fig. 1 (a)), and the generated subsequences are grouped using k-means clustering (Fig. 1 (b)). The cluster centers (the mean vectors of the cluster members) are thought of as representative patterns of the time series.

Currently, however, k-means STSC is considered to make little sense as a pattern discovery technique, since, as first pointed out by Keogh et al. [9], k-means STSC is "meaningless" in that the resultant cluster centers tend to form sinusoidal pseudo-patterns almost independent of the input time series. This *sinusoid effect* proved that even the simplest algorithms such as k-means STSC could be too dangerous to be used unless the mathematical structures are fully understood. We believe that the sinusoid effect raised a question to the general trend in the stream mining community that seemingly plausible analysis tends to be accepted without theoretical justifications.

In a previous paper [6], we theoretically studied the origin of the sinusoid effect. The original k-means STSC task was reduced to a spectral STSC task, and sinusoidal cluster centers were explicitly obtained by solving an eigen problem. In this paper, we discuss mathematical properties of STSC in more detail. In

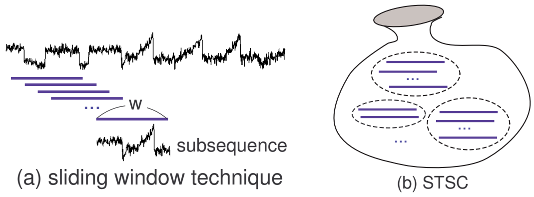

Fig. 1. (a) Sliding window technique to generate subsequences. (b) The generated subsequences are grouped as independent data items.

particular, we will point out that the cluster centers are inevitably governed by irreducible representations of the translational group, because of a hidden translational symmetry introduced by the sliding window technique. To the best of the author's knowledge, this is the first work that points out the interesting connection between STSC and group theory.

The layout of this paper is as follows: In Section 2, we reformulate STSC as the problem of linear algebra in a vector space, and introduce the notion of linear operators. In Section 3, we review the connection between k-means and spectral STSC. In Section 4, we introduce the concept of translational group, and explain its implications in spectral STSC. In Section 5, we derive the solution to spectral STSC from a group-theoretical perspective. In Section 6, we summarize the paper.

2 Lattice model for time series analysis

In this section, we introduce a lattice model for time series analysis, and show that this model provides us with a very handy way to express the subsequences of time series data.

2.1 Vector space in Dirac's notation

A common approach to express time-series data is to use a scalar function such as $x(t)$. However, this notation is not very effective in describing symmetry properties of the problem. We believe that this has made it difficult to pinpoint the origin of the sinusoid effect. Instead, we introduce a lattice model in Dirac's notation [12, 10] to represent time series data. While Dirac's notation is mathematically equivalent to the standard vector-matrix notation, it is much more powerful for describing linear operators, which play an essential role in this paper. In this subsection, we illustrate the notion of Dirac's notation, following [12].

Let $\mathcal{H}_0$ be a vector space spanned by n linearly independent bases $\{|1\rangle, |2\rangle, ..., |n\rangle\}$. By definition, any vector in $\mathcal{H}_0$ is represented as a linear combination of these bases. For example, a vector $|a\rangle \in \mathcal{H}_0$ may be expressed as

$$|a\rangle = \sum_{l=1}^{n} a_l |l\rangle,$$

where a_ls are constants (generally complex).

To introduce the metric into $\mathcal{H}_0$, we request that each of $|l\rangle$ has a unique counterpart in a dual space of $\mathcal{H}_0$. We denote the counterpart by $\langle l|$, and define that $c|a\rangle$ dual-corresponds to $c^*\langle l|$ where c is a complex constant and * denotes complex conjugate. Now, the inner product between vectors $|a\rangle, |b\rangle \in \mathcal{H}_0$ is defined as $\langle a|b\rangle \equiv \langle a| \cdot |b\rangle$, which is generally a complex number. Regarding the other choice $\langle b|a\rangle$, we assume that

$$\langle a|b\rangle = [\langle b|a\rangle]^* \tag{1}$$

holds as a premise. For example, the inner product between the above $|a\rangle$ and $|b\rangle = \sum_{l=1}^{n} b_l|l\rangle$ will be $\langle a|b\rangle = \sum_{l,l'} a_{l'}{}^* b_l \langle l'|l\rangle$, which is computable if $\langle l'|l\rangle$s are given.

As usual, we also request $\langle a|a\rangle \geq 0$ for any vector in $\mathcal{H}_0$. The notion of inner product allows us to define a normalized vector in the sense of unit norm. We assume that the bases $\{|l\rangle\}$ have been chosen to be orthonormal, i.e.,

$$\langle l|l'\rangle = \delta_{l,l'},$$

where $\delta_{l,l'}$ is Kronecker's delta.

2.2 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 from ordinary numbers by using ˆ hereafter. By definition, $\forall \hat{o} \in \mathcal{L}$ has an expression

$$\hat{o} = \sum_{l,l'=1}^{n} o_{l,l'} |l\rangle\langle l'|, \tag{2}$$

where $o_{l,l'} \in \mathbb{C}$ is called the (l, l') element of $\hat{o}$. Since $\{o_{l,l'}\}$ uniquely specifies $\hat{o}$ under a given orthonormal basis set, the $n \times n$ matrix $[o_{l,l'}]$ can be thought of as a matrix representation of $\hat{o}$. For $\forall |a\rangle \in \mathcal{H}_0$, we denote the dual element of $\hat{o}|a\rangle$ as $\langle a|\hat{o}^\dagger$, and call $\hat{o}^\dagger$ the Hermitian conjugate of $\hat{o}$. By definition, $\hat{o}^\dagger$ has the expression as $\hat{o}^\dagger = \sum_{l,l'=1}^{n} o_{l',l}{}^* |l\rangle\langle l'|$. A linear operator $\hat{o}$ such that $\hat{o}^\dagger = \hat{o}$ is called Hermitian. We see that this condition is equivalent to that the matrix representation is Hermitian.

As an example, consider a Hermitian operator

$$\hat{\vartheta}(w) \equiv \sum_{l'=1}^{w} |l'\rangle\langle l'|, \tag{3}$$

where $w \leq n$. To understand the nature of this operator, imagine an equi-interval one-dimensional lattice as shown in Fig. 2 (a), where each of the basis is attached to each site (lattice point). Since $\hat{\vartheta}(w)|l\rangle$ vanishes for $l > w$, and $\hat{\vartheta}(w)|l\rangle = |l\rangle$ otherwise, we see that $\hat{\vartheta}(w)$ works as the "cut-off operator". For example, if

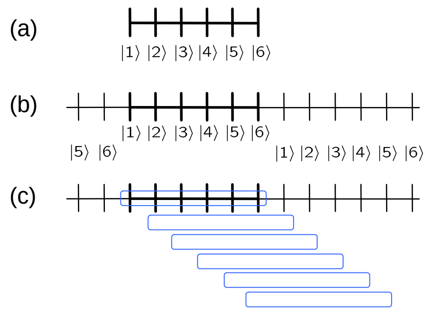

Fig. 2. (a) one-dimensional lattice with $n = 6$. (b) One-dimensional lattice under the periodic boundary condition. (c) Subsequences when $w = n = 6$.

$n = 6$ and $w = 3$, $\hat{\vartheta}(3)$ simply cuts off the portion which is not contained by the 3-dimensional lattice of $\{|1\rangle, |2\rangle, |3\rangle\}$, remaining the rest unchanged.

It is interesting to see how $\hat{\vartheta}(n)$ works. Clearly, this operator remains any vector in $\mathcal{H}_0$ the same. In other words, this is the *identity operator*. Explicitly, we define the identity operator in $\mathcal{H}_0$ as

$$\hat{1} \equiv \sum_{l'=1}^{n} |l'\rangle\langle l'|. \tag{4}$$

This operator is very useful when one wants to change the representation. For example, $\forall|a\rangle \in \mathcal{H}_0$ is equivalent to $\hat{1}|a\rangle$, so that $|a\rangle = \sum_{l=1}^{n} |l\rangle\langle l|a\rangle$ holds. Since $\langle l|a\rangle$s are just scalar, this equation gives the representation of $|a\rangle$ by $\{|1\rangle, ..., |n\rangle\}$.

2.3 Lattice model for time series data

Now, let us associate $\mathcal{H}_0$ with time-series data. Consider an equi-interval time series data with length n of $\{x_t \in \mathbb{R} \mid t = 1, 2, ..., n\}$. Since the data is assumed to be a collection of independent observations, it is equivalently expressed as a vector in $\mathcal{H}_0$:

$$|\Gamma\rangle = \sum_{l=1}^{n} x_l |l\rangle. \tag{5}$$

In Fig. 2 (a), this definition amounts to that each x_l is attached to the l-th site. We call this expression the site-representation of the time-series data. The coefficients x_l can be obtained by $x_l = \langle l|\Gamma\rangle$. By using $\hat{1}$, one can explicitly compute the squared norm of $|\Gamma\rangle$ as

$$\langle\Gamma|\Gamma\rangle = \langle\Gamma|\hat{1}\cdot\hat{1}|\Gamma\rangle = \sum_{l=1}^{n}\langle\Gamma|l\rangle\langle l|\sum_{l'=1}^{n}|l'\rangle\langle l'|\Gamma\rangle = \sum_{l=1}^{n}\langle\Gamma|l\rangle\langle l|\Gamma\rangle = \sum_{l=1}^{n}|x_l|^2,$$

which is just the squared sum for real valued time series data. We may simply denote this as $||\ |\Gamma\rangle\ ||^2$.

Hereafter, we impose the *periodic boundary condition* (PBC) on time-series data. As indicated in Fig. 2 (b), $\forall l, |l+n\rangle = |l\rangle$ holds under the PBC. As long as $n \gg 1$, the discrepancies due to this artificial condition will be negligible.

2.4 Translation operator on the lattice

As another instance of linear operators, let us focus on the translation operator $\hat{\tau}(l) \in \mathcal{L}$

$$\hat{\tau}(l) \equiv \sum_{l'=1}^{n} |l'+l\rangle\langle l'|. \tag{6}$$

The operator $\hat{\tau}(l)$ shifts the basis in the site-representation with l steps. To see this, for example, consider $\hat{\tau}(l)|2\rangle$. Thanks to the orthogonality, it follows

$$\hat{\tau}(l)|2\rangle = \sum_{l'=1}^{n} |l'+l\rangle\langle l'|2\rangle = \sum_{l'=1}^{n} |l'+l\rangle\delta_{l',2} = |2+l\rangle.$$

By Premise (1), it is straightforward to see $\hat{\tau}(l)^\dagger = \hat{\tau}(-l)$, i.e. $\hat{\tau}(l)$s are *unitary* operator, where, in general, $\hat{o} \in \mathcal{L}$ is said unitary if $\hat{o}\hat{o}^\dagger = \hat{o}^\dagger\hat{o} = \hat{1}$ holds.

The translation operator provides us with a handy way to express subsequences in STSC. If we use the expression of Eq. (5), the p-th subsequence with length w (i.e. *the window size* is w) is given by $\sum_{l=p+1}^{p+w} x_l|l\rangle$. While this subsequence should be viewed as a vector in $\mathcal{H}_0$ originally, it is clearly redundant in that only w dimensions are used out of the n dimensions. It is more reasonable to think of $|s_p\rangle$ as a vector in a subspace $\mathcal{H} \equiv \{|1\rangle, ..., |w\rangle\}$. Explicitly, we define $|s_p\rangle$ as

$$|s_p\rangle = \sum_{l=1}^{w} x_{l-p}|l\rangle = \hat{\vartheta}(w)\hat{\tau}(-p)|\Gamma\rangle \tag{7}$$

under the PBC.

Here we define the notion of *translational invariance* of operators:

Definition 1 (Translational invariance) *An operator $\hat{o} \in \mathcal{L}$ is said to be translationally invariant when*

$$\hat{o} = \hat{\tau}(l)^\dagger \hat{o}\hat{\tau}(l) \tag{8}$$

holds for $\forall l \in \{0, 1, .., n-1\}$.

The intuition behind this is that the matrix element of $\hat{o}$ between $\forall|a\rangle$ and $|b\rangle \in \mathcal{H}_0$ remains the same as that between $\hat{\tau}(l)|a\rangle$ and $\hat{\tau}(l)|b\rangle$. Since $\hat{\tau}(-l) = \hat{\tau}(l)^{-1}$ by definition, the invariance condition is equivalent to

$$\hat{o}\hat{\tau}(l) = \hat{\tau}(l)\hat{o}. \tag{9}$$

In other words, any operators invariant to translations must commute with $\hat{\tau}(l)$s.

3 Spectral clustering of subsequences

In this Section, we derive an eigen equation whose eigen vectors corresponds to the k-means cluster centers. It essentially follows the formulation in [6], but is the first treatment of spectral STSC with Dirac's notation.

As before, we use the whole space $\mathcal{H}_0 = \{|1\rangle, ..., |n\rangle\}$, and its subspace $\mathcal{H} = \{|1\rangle, ..., |w\rangle\}$ with $w \leq n$. Notice that we do *not* assume any periodicity in $\mathcal{H}$ unless $w = n$, despite the fact $\mathcal{H}_0$ is always periodic. The k-means STSC task is to group a set of vectors $\{|s_q\rangle \in \mathcal{H} \, |q = 1, 2, ..., n\}$, where the subsequences are thought of as vectors in $\in \mathcal{H}$

It is well-known that the k-means algorithm attempts to minimize the sum-of-squared (SOS) error [4]. In our notation, the SOS error is written as

$$E = \sum_{j=1}^{k} \sum_{p \in \mathcal{C}_j} \Big|\Big| \, |s_p\rangle - |m^{(j)}\rangle \, \Big|\Big|^2 = \sum_{p=1}^{n} \langle s_p | s_p \rangle - \sum_{j=1}^{k} \frac{1}{|\mathcal{C}_j|} \sum_{p,r \in \mathcal{C}_j} \langle s_p | s_r \rangle, \qquad (10)$$

where $\mathcal{C}_j$ and $|\mathcal{C}_j|$ represent the members of the j-th cluster and the number of members, respectively. The centroid of $\mathcal{C}_j$ is denoted by $|m^{(j)}\rangle$. To get the rightmost expression, we used the definition of the centroid $|m^{(j)}\rangle = \frac{1}{|\mathcal{C}_j|} \sum_{p \in \mathcal{C}_j} |s_p\rangle$.

Since the first term in the rightmost side does not depend on clustering, let us focus on the second term, which will be denoted by E_2. To remove the restricted summation, we introduce an indicator vector $|u^{(j)}\rangle \in \mathcal{H}$, where $\langle s_q | u^{(j)} \rangle = 1/\sqrt{|\mathcal{C}_j|}$ for $s_q \in \mathcal{C}_j$ and 0 otherwise, to have

$$E_2 = -\sum_{j=1}^{k} \sum_{p,r=1}^{n} \langle u^{(j)} | s_p \rangle \langle s_p | s_r \rangle \langle s_r | u^{(j)} \rangle = -\sum_{j=1}^{k} \langle u^{(j)} | \hat{\rho}^2 | u^{(j)} \rangle,$$

where we introduced a linear operator $\hat{\rho}$

$$\hat{\rho} = \sum_{p=1}^{n} |s_p\rangle \langle s_p| \qquad (11)$$

to get the rightmost expression. Note that, in contrast to Eq. (4), $\hat{\rho}$ is not the identity operator since $|s_p\rangle$s are not orthonormal.

The k-means clustering task has now been reduced to seeking the solution $\{|u^{(j)}\rangle\}$ which minimizes E_2. If we relax the original binary constraint on $|u^{(j)}\rangle$, and instead take

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

as the new restriction on the optimization problem, the k-means task now amounts to

$$\hat{\rho} | u^{(j)} \rangle = \lambda_j | u^{(j)} \rangle, \qquad (13)$$

where λ_j is the eigenvalue corresponding to the eigen vector $|u^{(j)}\rangle$. The site-representation allows us to solve this via standard matrix computations, i.e.,

$$\sum_{l'=1}^{w} \langle l|\hat{\rho}|l'\rangle\langle l'|u^{(j)}\rangle = \lambda_j \langle l|u^{(j)}\rangle.$$

Before the relaxation, the indicator vectors satisfied

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

After the relaxation, $|u^{(j)}\rangle$ is the eigen vector of $\hat{\rho}$. Thus, it follows that *the k-means cluster centers correspond to the eigenstates of* $\hat{\rho}$, or

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

This remarkable relation was first derived in [6]. As compared to previous work of spectral clustering [13, 11, 2, 3], our main contribution is that we introduced a new formulation which directly seeks the cluster centers, instead of the standard formulation based on membership indicators.

4 Group-theoretical properties of $\hat{\rho}$

In this section, we explain the basics of group theory. Specifically, we derive irreducible representations of the translational group, showing an interesting connection between Fourier components and the translational group.

4.1 Translational group

We have shown a theoretical connection between k-means STSC and spectral STSC. Since spectral STSC is expressed as the eigenvalue equation of $\hat{\rho}$, it is useful to study mathematical properties of $\hat{\rho}$.

Using the expression of Eq. (7), and arranging site indices accordingly, $\hat{\rho}$ can be written as

$$\hat{\rho} \doteq \sum_{l=1}^{n} \hat{\tau}(l)^{\dagger}|\Gamma\rangle\langle\Gamma|\hat{\tau}(l), \tag{15}$$

where we used a shorthand notation instead of using $\hat{\vartheta}(w)$, i.e. we define the symbol "$\doteq$" meaning "the left and the right sides have the same matrix elements when represented in $\mathcal{H}$ (not $\mathcal{H}_0$)".

In this expression of $\hat{\rho}$, it seems that a set of the translational operators

$$\mathcal{T}_n \equiv \{\hat{\tau}(0), \hat{\tau}(1), ..., \hat{\tau}(n-1)\}.$$

plays a key role. If we define $\hat{\tau}(n) = \hat{1}$, we see that $\mathcal{T}_n$ is closed in that any product between two of the operators remains in $\mathcal{T}_n$. For example, when $n = 6$ as shown in Fig. 2 (b), the operation of $\hat{\tau}(3)$ followed by $\hat{\tau}(4)$ must coincide with that of $\hat{\tau}(1)$ ($= \hat{\tau}(7)$) by definition. More exactly, for $\forall|a\rangle \in \mathcal{H}_0$, $\hat{\tau}(4)\hat{\tau}(3)|a\rangle = \hat{\tau}(1)|a\rangle$ must hold. In addition, first, for any integers $l, l', l'' \in \{0, 1, ..., n-1\}$,

$$\hat{\tau}(l)\hat{\tau}(l')\hat{\tau}(l'') = \hat{\tau}(l+l')\hat{\tau}(l'') = \hat{\tau}(l)\hat{\tau}(l'+l'')$$

is clearly satisfied. Second, $\mathcal{T}_n$ has the unit element of $\hat{\tau}(0) = \hat{1}$. Third, any of the elements in $\mathcal{T}_n$ has an inverse element. For example, $\hat{\tau}(2)$ is the inverse element of $\hat{\tau}(n-2)$, since $\hat{\tau}(2)\hat{\tau}(n-2) = \hat{\tau}(n-2)\hat{\tau}(2) = \hat{1}$.

These three properties are nothing but the axioms of *group*:

Definition 2 (Group) *A group $\mathcal{G}$ is a set of linear operators such that (1) any of three elements in $\mathcal{G}$ satisfy the associativity relation, (2) $\mathcal{G}$ includes the unit element, and (3) any of the elements in $\mathcal{G}$ has an inverse element in $\mathcal{G}$.*

Thus, $\mathcal{T}_n$ forms a group, which called the (one-dimensional) translational group.

A remarkable property of Eq. (15) is translational invariance of the right hand side (r.h.s.). Recall the definition of the invariance Eq. (8) and the PBC. Then, it follows

$$\hat{\tau}(l)^{\dagger}(\text{r.h.s.})\ \hat{\tau}(l) = \sum_{l'=1}^{n} \hat{\tau}(l+l')^{\dagger}|\Gamma\rangle\langle\Gamma|\hat{\tau}(l+l') = \sum_{l''=1}^{n} \hat{\tau}(l'')^{\dagger}|\Gamma\rangle\langle\Gamma|\hat{\tau}(l''), \quad (16)$$

showing the translational invariance of r.h.s. of Eq. (15).

Since the particular form of the r.h.s. in Eq. (15) is a direct consequence of the sliding window technique, we must say that this translational symmetry is just a mathematical artifact introduced by the sliding window technique. In this sense, we call the translational symmetry of the r.h.s. of Eq. (15) the *pseudo* translational symmetry.

4.2 Representation theory

In Subsection 2.2, we introduced the notion of matrix representation of linear operators in a way specific to $\mathcal{H}_0$. Here, we generalize the concept to groups:

Definition 3 (Representation) *A representation $\mathcal{D}$ of a group $\mathcal{G}$ is a set of $d \times d$ matrices, such that each element of $\mathcal{G}$ is associated with an element of $\mathcal{D}$, and $D(\hat{o}_i)D(\hat{o}_j) = D(\hat{o}_k)$ holds for any $\hat{o}_i, \hat{o}_j, \hat{o}_k \in \mathcal{G}$ satisfying $\hat{o}_i\hat{o}_j = \hat{o}_k$.*

In this definition, d is called the dimension of the representation. For example, each element of $\mathcal{T}_n$ can be expressed as an $n \times n$ matrix of $\langle l|\hat{\rho}|l'\rangle$, based on the basis of $\mathcal{H}_0$. The matrix representation of $\hat{\tau}(1)$ has ones for $l = l' + 1$, zeros otherwise.

As indicated by this example, a representation can be constructed by looking at how a group element operates on the basis of a vector space. Thus, it is natural to introduce the notion of representation spaces as

Definition 4 (Representation space) *A vector space is said invariant w.r.t. a group $\mathcal{G}$, if the space remains in the same space after the operation of $\forall \hat{o} \in \mathcal{G}$. If a subspace is invariant w.r.t. $\mathcal{G}$, then the space is said a representation space of $\mathcal{G}$.*

In the above example, $\mathcal{H}_0$ is a representation space of $\mathcal{T}_n$. Clearly, a representation space is not unique. For example, a different representation will be obtained if we use $2n$-dimensional space defined using a one-dimensional lattice having $2n$ sites. As expected from this example, there is generally no upper bound on the dimension of representation spaces.

One interesting question here is whether or not there exists a lower bound on the dimension of representation spaces. The answer is yes. It is known that, for a given group, there exist a certain number of "minimal" representation spaces, which are called the irreducible representation space of the group. Putting formally,

Definition 5 (Irreducible representation space) *If a representation space does not include any subspaces that are invariant w.r.t. $\mathcal{G}$, it is called an irreducible representation space.*

For example, while we have used the n-dimensional space to represent $\mathcal{T}_n$ so far, it can be shown that a vector $|f_0^n\rangle \equiv \frac{1}{\sqrt{n}} \sum_{l=1}^{n} |l\rangle$ spans a one-dimensional irreducible representation space of $\mathcal{T}_n$. In fact, it is easy to verify $\forall l$, $\hat{\tau}(l)|f_0^n\rangle = |f_0^n\rangle$. Clearly, this space is irreducible because it is one-dimensional. Since all the representation matrices are ones (1×1 identity matrix), this irreducible representation is also called the identity representation.

If representations of a group are turned to be irreducible, it is known that a strong theorem, which is a fundamental theorem in group theory and is also known as Schur's first lemma, holds. While Schur's first lemma is almost always expressed as mysterious-looking relations between representation matrices (for a proof, see, e.g. [7]), it essentially states the orthogonality of different irreducible representation spaces:

Theorem 1 (Schur) *For a given group $\mathcal{G}$, let $|q, m\rangle$, $|q', m'\rangle$ be bases in the representation spaces of irreducible representations labeled by q, q', respectively, where m or m' specifies a dimension of each representation space. For such a linear operator $\hat{o}$ that invariant w.r.t. any element of $\mathcal{G}$, $\langle q, m|\hat{o}|q', m'\rangle$ is zero if q and q' are different.*

Another fundamental theorem in group theory is one called Schur's second lemma. Combining the first and the second lemmas, one can prove a stronger relation (for proof, see [7]):

Theorem 2 (Selection rule) *In the same setting as Theorem 1, for such a unitary operator $\hat{o}$ that invariant w.r.t. any element of $\mathcal{G}$, the following relation holds:*

$$\langle q, m|\hat{o}|q', m'\rangle \propto \delta_{q,q'}\delta_{m,m'}.$$

This theorem is also known as the selection rule in quantum physics.

4.3 Irreducible representations of translational group

Since the identity operator is invariant w.r.t. any linear operator, we see from Theorem 2 that the bases of irreducible representation spaces are orthogonal to each other:

$$\langle q, m | q', m' \rangle \propto \delta_{q,q'} \delta_{m,m'}.$$

This orthogonal relation reminds us of subspace learning methods such as PCA, where mutually orthogonal directions are explored in terms of maximum variance. It is tempting to associate the irreducible representation spaces with extracted patterns. One important implication here is that group theory serves as a new tool for pattern learning, where each irreducible representation is thought of as an extracted pattern.

Let us find irreducible representations of $\mathcal{T}_n$. First, since $\hat{\tau}(l) \in \mathcal{T}_n$ is unitary, its matrix representation must be a unitary matrix. Second, irreducible representations of $\mathcal{T}_n$ must be all one-dimensional. To see this, suppose that an irreducible representation space is d_q-dimensional. By definition, for $\forall \hat{\tau}(l) \in \mathcal{T}_n$,

$$\hat{\tau}(l)|q, m\rangle = \sum_{m'=1}^{d_q} |q, m'\rangle \langle q, m' | \hat{\tau}(l) | q, m\rangle.$$

Since $\hat{\tau}(l)$ itself is translationally invariant, it follows from Theorem 2 that

$$\hat{\tau}(l)|q, m\rangle = |q, m\rangle \langle q, m | \hat{\tau}(l) | q, m\rangle.$$

This means that the original d_q-dimensional representation spaces has a subspace invariant w.r.t. $\mathcal{T}_n$, which contradicts to the assumption. Therefore, we conclude that each of the irreducible representations of $\mathcal{T}_n$ is one-dimensional. In particular, third, considering the first property, we see that

$$\hat{\tau}(l)|q\rangle = [\mathrm{e}^{\mathrm{i} f_q^n}]^l |q\rangle \tag{17}$$

holds for an irreducible representation q (we dropped the unnecessary index m), where

$$f_q^n = \frac{2\pi q}{n} \quad (q \text{ is an integer})$$

to satisfy the PBC. From these equations, we see that irreducible representation spaces are eigen spaces of $\hat{\tau(1)}$. As indicated by the exponential factor in Eq. (17), the eigen spaces are given by discrete Fourier transformation (DFT) in $\mathcal{H}_0$:

$$|f_q^n\rangle = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \mathrm{e}^{\mathrm{i} f_q (l - l_0)} |l\rangle, \tag{18}$$

where l_0 is a real constant, and the subscript q runs over $\mathcal{D}_f^n = \{-\frac{n-1}{2}, ..., 0, 1, .., \frac{n-1}{2}\}$ when n is odd, and $\{-\frac{n}{2} + 1, ..., 0, 1, .., \frac{n}{2}\}$ when n is even. It is easy to verify Eq. (17):

$$\hat{\tau}(l)|f_q^n\rangle = \frac{1}{\sqrt{n}} \sum_{l'=1}^{n} \mathrm{e}^{\mathrm{i} f_q^n (l' - l_0)} |l' + l\rangle = \mathrm{e}^{-\mathrm{i} l f_q^n} |f_q^n\rangle, \tag{19}$$

where we used $\mathrm{e}^{\mathrm{i} f_q^w n} = 1$.

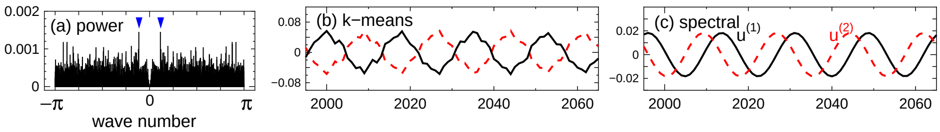

Fig. 3. The results for a white noise data with $w = n = 6000$. (a) The power spectrum. (b) A segment of the k-means cluster centers ($k = 2$), and (c) a segment of the top two eigen vectors of $\hat{\rho}$.

5 Solution to subsequence time-series clustering

5.1 The $w = n$ case

When $w = n$, i.e. the length of the subsequences is the same as the whole time series (see Fig. 2 (c)), the operator $\hat{\rho}$ has exact translational invariance. Thus, from the selection rule, it follows

$$\langle f_q^n | \hat{\rho} | f_{q'}^n \rangle \propto \delta_{q,q'}. \tag{20}$$

This means that the matrix representation of $\hat{\rho}$ is diagonal in the space spanned by $\{|f_q^n\rangle |\ q \in \mathcal{D}_f^n\}$. Thus, solving the eigen equation Eq. (13) for $|u^{(j)}\rangle$ is trivial, and the eigen vectors are nothing but $|f_q^n\rangle$s.

Using Eqs. (15) and (19), we can calculate a matrix element $\langle f_q^n | \hat{\rho} | f_q^n \rangle$ as

$$\langle f_q | \hat{\rho} | f_{q'} \rangle = \sum_{l=1}^{n} \langle f_q | \hat{\tau}(l)^\dagger | \Gamma \rangle \langle \Gamma | \hat{\tau}(l) | f_{q'} \rangle = \sum_{l=1}^{n} \langle f_q | \Gamma \rangle \langle \Gamma | f_q' \rangle = n |\langle f_q | \Gamma \rangle|^2. \tag{21}$$

Thus, the i-th top eigen vector occurs at a Fourier component having the i-th largest *power* $|\langle f_q | \Gamma \rangle|^2$. Since the power spectum becomes an even function for any real time series, the resulting cluster centers of spectral STSC are pure sinusoids. Note that each of the eigen vectors must be a pure sinusoid even when the power spectrum does not have any dominant f_q^n.

To validate this result, we performed k-means and spectral STSC for a white noise data having $n = 6000$ (the data used is Normal data in the Synthetic Control Chart data [8]). We did 100 random restarts and chose the best one in the k-means calculation. As shown in Fig. 3 (a), we have the largest power at $|f_q| = 0.358$ in (marked by the triangles). Thus, the wavelength must be $2\pi/|f_q| = 17.6$, which is completely consistent with Fig. 3 (c). In addition, we see that the eigen vector is a good estimator of the k-means cluster center by comparing Figs. 3 (b) and (c).

Note that the sinusoids are obtained as irreducible representations of $\mathcal{T}_n$. Thus, we conclude that the sinusoid effect is a direct consequence of the pseudo-translational symmetry introduced by the sliding window technique.

5.2 The $w < n$ case

Let us consider the general case of $w < n$. In this case, $\hat{\rho}$ does not have exact translational invariance, so we need to introduce DFT in $\mathcal{H}$ rather than $\mathcal{H}_0$:

$$|f_q^w\rangle = \frac{1}{\sqrt{w}} \sum_{l=1}^{w} \mathrm{e}^{\mathrm{i} f_q (l-l_0)} |l\rangle. \tag{22}$$

It is straightforward to show $\langle f_q^w | f_{q'}^w \rangle = \delta_{q',q}$, and thus, $\{|f_q^w\rangle|\ q \in \mathcal{D}_f^w\}$ forms the complete set in $\mathcal{H}$.

Consider a vector $\hat{\tau}(1)|f_q^w\rangle$. Using the fact $\mathrm{e}^{-\mathrm{i} f_q^w w} = 1$, it is easy to show

$$\hat{\tau}(1)|f_q^w\rangle = \mathrm{e}^{-\mathrm{i} f_q^w} |f_q^w\rangle + \frac{1}{\sqrt{w}} \mathrm{e}^{-\mathrm{i} f_q^w l_0} |B\rangle,$$

where $|B\rangle \equiv |w+1\rangle - |1\rangle$. By applying $\hat{\tau}(1)$ sequentially, we have

$$\hat{\tau}(l)|f_q^w\rangle = \mathrm{e}^{-\mathrm{i} f_q^w l} \left[|f_q\rangle + \frac{1}{\sqrt{w}} \sum_{l'=1}^{l} \mathrm{e}^{\mathrm{i} f_q^w (l'-l_0)} \hat{\tau}(l'-1)|B\rangle \right].$$

While we can omit the second term inside the bracket when both $w = n$ and the PBC hold, that is not the case here. Using this formula, we can calculate the matrix elements of $\hat{\rho}$ in the Fourier representation. The result will be [1]

$$\langle f_q^w | \hat{\rho} | f_{q'}^w \rangle \approx n |\langle f_q^w | \varGamma \rangle|^2 \delta_{q,q'} + \sum_{l=1}^{n} \mathrm{e}^{\mathrm{i} l \varDelta_{q'q}} J_l^w(q, q'). \tag{23}$$

It is straightforward to get the exact expression of $J_l^w(q, q')$ although we do not show it here. In the above equation, the first term is the same as Eq. (21). The second term is a perturbation term giving off-diagonal elements. 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 of a time-series data set has a single-peaked structure at a certain $|f_q^w|$, the top eigen vector will be well approximated by the $f_{|q|}^w$, irrespective of the details of the spectrum. On the other hand, when the power spectrum is almost flat, the eigenvectors will be mixtures of many f_q^ws, so that the cluster centers will be far from pure sinusoids. Therefore, even in this general case, we may claim that the sinusoid effect is a direct consequence of the pseudo-translational symmetry introduced by the sliding window technique.

To validate the theoretical result of Eq. (23), we performed experiments using Cylinder, Bell, and Funnel (CBF) data [8]. The CBF data includes three types of patterns literally having Cylinder, Bell, and Funnel shapes. We randomly generated 30 instances for each type with a fixed length of 128 ($= w$)

[1] Since $n \gg w$ under normal conditions, one may approximate as $\sum_{l=1}^{n} \mathrm{e}^{\mathrm{i} l \varDelta_{q',q}} \approx n\delta_{q,q'}$.

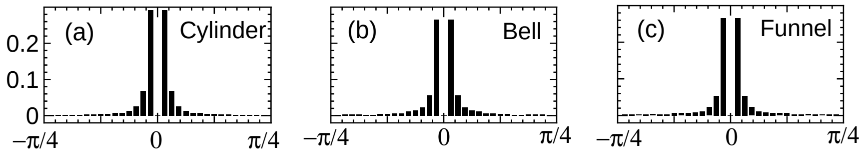

Fig. 4. Averaged power spectra of each instance of the CBF data. The horizontal axis represents f_q^w within $[-\frac{\pi}{4}, \frac{\pi}{4}]$ out of $(-\pi, \pi]$.

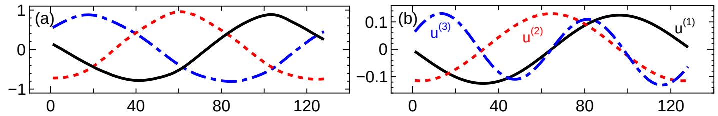

Fig. 5. (a) The k-means cluster centers ($k = 3, w = 128$). (b) The top three eigen vectors of the spectral STSC ($w = 128$).

using Matlab code provided by [8]. We concatenated the instances in order after standardizing each one (zero mean and unit variance). An example segment of the concatenated data was shown in Fig. 1 (a). Again, we did 100 random restarts and chose the best one in the k-means calculation.

Figures 4 (a)-(c) show the power spectra of each instance as a function of f_q^w. To handle the variation of the instances, we simply averaged the resultant spectra for all instances. We see that the most of the weight is concentrated on the $|q| = 1$ component (i.e. the wavelength of w) in all of the cases. The f_0 component is naturally missing because of the standardization.

The results of k-means and spectral STSC are shown in Fig. 5, where the sinusoid effect is clearly observed. The sinusoid of wavelength of w can be understood from the dominant $|q| = 1$ weight in Fig. 4 (a)-(c). Since the single-peaked structure in the power spectrum is naturally expected whenever there are no particular periodicities within the scale of window size, the sinusoid of wavelength w should be observed in a wide variety of the input data. This explains why the sinusoid effect is ubiquitous.

Due to the orthogonality condition, we see that the third singular vector necessarily has a wavelength of about $w/2$ in Fig. 5 (b). 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 with the results.

6 Concluding remarks

We have performed a group-theoretical analysis of the sinusoid effect in STSC. Based on a spectral formulation of STSC, we showed that the sinusoid effect is a group-theoretical consequence of the pseudo-translational symmetry introduced by the sliding window technique.

In Section 4, we claimed that finding irreducible representations can be viewed as pattern learning if the problem under consideration has a certain symmetry related to a group. In this sense, the sinusoid effect can be thought of as a result of (unintentional) pattern learning from time-series subsequences. In another paper [5], we introduced idea that certain correlation patterns are extracted as irreducible representation spaces of a point group. In this way, designing machine learning algorithms so that irreducible representations of a group can be effectively extracted would be interesting future direction of research.

References

1. G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, 1998.
2. I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 551–556, 2004.
3. C. Ding and X. He. K-means clustering via principal component analysis. In *Proc. Intl. 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. Data Mining*, pages 657–660, 2005.
6. T. Idé. Why does subsequence time-series clustering produce sine waves? In *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases, Lecture Notes in Artificial Intelligence*, volume 4213, pages 311–322, 2006.
7. T. Inui, Y. Tanabe, and Y. Onodera. *Group Theory and Its Applications in Physics, 2nd ed.* Springer-Verlag Telos, 1996.
8. E. Keogh and T. Folias. The UCR time series data mining archive [`http://www.cs.ucr.edu/~eamonn/TSDMA/index.html`]. 2002.
9. 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. Data Mining*, pages 115–122. IEEE, 2003.
10. R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proc. the 20th Intl. Conf. Machine Learning*, pages 361–368, 2003.
11. 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.
12. J. J. Sakurai. *Modern Quantum Mechanics.* Addison Wesley, 2nd Ed., 1994.
13. 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.