Noname manuscript No.
(will be inserted by the editor)

A Nested Two-Stage Clustering Method for Structured Temporal Sequence Data

Liang Wang · Vignesh Narayanan · Yao-Chi Yu · Yikyung Park · Jr-Shin Li

Received: date / Accepted: date

Abstract Mining patterns of temporal sequence data is an important problem across many disciplines. Under appropriate pre-processing procedures, a structured temporal sequence can be organized into a probability measure or a time-series representation, which grants a potential to reveal distinctive temporal pattern characteristics. In this paper, we propose a two-stage nested clustering method that integrates optimal transport and the dynamic time warping distances to learn the distributional and dynamic shape-based dissimilarity at the respective stage. The proposed clustering algorithm preserves both the distribution and shape patterns present in the data, which are critical for the datasets composed of structured temporal sequences. The effectiveness of the method is tested against existing Agglomerative and K-shape based clustering algorithms on Monte-Carlo simulated synthetic datasets, and the performance is compared through various cluster validation metrics. Furthermore, we apply the developed method to real-world datasets from three domains: temporal dietary records, online retail sales, and smart meter energy profiles. The expressiveness of the cluster and sub-cluster centroid patterns shows significant promise of our method for structured temporal sequence data mining.

This work was supported in part by the National Science Foundation under the awards ECCS-1509342 and CMMI-1763070, and by the NIH grant R01CA226937A1.

Liang Wang
Washington University in St. Louis
E-mail: liang.wang@wustl.edu

Vignesh Narayanan
Washington University in St. Louis
E-mail: vignesh.narayanan@wustl.edu

Yao-Chi Yu
Washington University in St. Louis
E-mail: y.yu@wustl.edu

Yikyung Park
Washington University School of Medicine
E-mail: yikyungpark@wustl.edu

Jr-Shin Li
Washington University in St. Louis
E-mail: jsli@wustl.edu

Keywords Clustering · Optimal Transport · Dynamic Time Warping · Structured Temporal Sequence

1 Introduction

Clustering has become a ubiquitous data mining tool employed to group a set of objects based on their similarities to potentially reveal any underlying structure in the given dataset. Owing to its application across diverse scientific domains such as engineering, medicine, and business applications, several clustering algorithms such as partitional clustering [**?**], hierarchical clustering [**?**] and density based clustering [**?**], etc, have been proposed in the literature. These clustering methods have been applied not only to organize static datasets but also to analyze dynamic datasets, such as the time-series data in which each sample consists of discrete data points in time order.

Designing clustering algorithms is challenging because the true dissimilarity measure (distance metric) for the underlying sample space is often impossible to know *a priori*. Oftentimes, there exist multiple dissimilarity measures with which a given dataset can be clustered and in general, the dissimilarity measure is tweaked to improve the clustering performance [**?**]. Especially, for time-series datasets, the data sequences are often distorted in some way and the dissimilarity measure need to satisfy a number of invariances such as scaling, translation, shift, occlusion and complexity invariances, to compare the sequences in a meaningful way [**?**]. In this context, several metric learning algorithms have been proposed to directly identify a suitable metric (with appropriate invariances) [**?**], which can then be used in the clustering algorithm. A brief summary of different categories of existing clustering algorithms and their properties along with related references are provided in Table. 1.

On the other hand, in some cases, the time-series data may inherently contain some structure. These can be local structures such as periodic/recurrent activity in Electroencephalogram (EEG), Electrocardiography (ECG) data, or global structures such as activities distributed over time as in temporal dietary records [**?**], retail product sales [**?**], and smart meter energy profile. Effective clustering of such datasets by explicitly accounting for the structural information can help organize the data efficiently and make informed decisions using the insights gained from the resulting clusters. For example, clusters from the daily dietary records can reveal healthy eating behavior in relation to disease outcome, clusters from retail product sales enable products bundled procurement, and clusters from electricity load measurements of smart home appliances can aid in utility contract design. These special time-series datasets, which consist of structural activities distributed over a fixed period of time besides local features, are referred in this paper as the structured temporal sequence data. In order to effectively cluster such structured temporal sequence data, developing a clustering algorithm that exploits the distribution structure in the data is critical. To the best of our knowledge, time-series clustering algorithms that explicitly incorporate such distribution structures with a time-series based dissimilarity measure are not reported and we aim to fill this gap.

To this end, in this paper, we propose a nested two-stage clustering approach that explicitly exploits the distribution structure (over a given time period) of the structured temporal sequence data. Specifically, we propose an optimal transport (OT) and dynamic time warping (DTW) distance based nested two-stage clustering method. The discrete distribution representation of the temporal sequence data, after normalization, fits well into the framework of the OT distance of two empirical discrete distributions. We propose to cluster the

structured temporal sequence data based on: a) the OT distance, which delineates the distributional similarity, and b) the DTW distance, which delineates the dynamic shape dissimilarity, between structured temporal sequences in a nested hierarchy. In doing so, both the distribution patterns and the shape patterns in data are preserved by the clustering algorithm. Additionally, we propose a variant of the OT based clustering by replacing the discrete distribution representation of the data with a continuous probability measure (referred as OTC), and present an efficient pre-processing and data representation strategy for all the cases (OT for discrete probability measure, OTC for continuous probability measure, and DTW for time-series representation). As a net result, with the proposed two-stage clustering approach and the pre-processing steps, the scale, translation, shift, distribution, and occlusion invariances are effectively captured without having to learn and identify a sophisticated distance metric.

To validate the efficacy of the proposed method, we present examples using Monte Carlo simulated synthetic datasets, and a comparative analysis with other algorithms such as DTW-DTW, MDTW-MDTW [**?**], OT-OT, OTC-DTW and Euclidean-Euclidean based K-means clustering. Further, the proposed algorithm is tested on real-world datasets such as the temporal dietary record, online retail data, and smart meter energy profile [**? ?**]. It is observed that the resulting cluster centroids provide relevant insights into the dietary energy, product sales and energy consumption patterns, conforming to each application's domain knowledge. Additionally, we include discussions on the ordering of the distance metrics and the other potential candidates (in place of OT and DTW distances) for the two stages of the proposed algorithm.

In summary, the contributions of the paper include the development of a nested two-stage clustering framework and the design of OT-DTW and OTC-DTW algorithms for structured temporal sequence data. We show that the proposed clustering method effectively captures the (macroscopic) distribution and (microscopic) shape patterns present in the considered structured temporal sequence data by using several examples with both synthetic and real-world datasets. We further demonstrate the broader applicability of the proposed algorithms in identifying synchronization clusters in oscillatory networks. The organization of the paper is as follows: in Section 2, we first provide a motivational example, and then review some of the related works on distance-based clustering using OT/OTC distance and DTW distance. In Section 3, the pre-processing step to achieve the discrete/continuous probability measure and time-series representation are introduced. Section 4 is devoted to introducing algorithm details of our two-stage clustering framework OT-DTW and OTC-DTW. All the experimental results with comparative analysis on synthetic data are presented in Section 5, and application to real-world datasets are presented in Section 6.

2 Motivation

To motivate the problem of clustering structured temporal sequence data, we present the following example. Three typical cases corresponding to the energy intake ratio by three individuals in a 24 hour period are recorded (see Fig. 1). The raw data shown in this figure for the three cases are typical examples of the temporal sequence data considered in this paper. In this example, Case I illustrates a three-meal dietary pattern with a gradual increase of energy portion from breakfast to dinner while Case II shows the same three-meal pattern, but with a shift of one hour for each meal. The temporal dietary pattern for Case III shares a similar energy distribution with patterns in Case I and Case II, but with more eating hours. The pairwise dissimilarity of these three cases under Euclidean distance, OT distance, and

Table 1: A brief summary of clustering models.

Clustering Approach	Models	Comments	References
Statistical	Bayesian non-parametric models	- Need uniform sampling	[?]
	ARMA	- High sampling frequency	[?]
	ARIMA	- Estimate correlations or transition probabilities on which the clusters are defined.	[?]
	Gaussian mixture models		[?]
	- Hidden Markov models		[?]
Dynamical system	Nonlinear finite impulse response models (NFIR)	- Choices of nonlinear mapping and regression vector are paramount	[?]
	Nonlinear auto regressive with exogenous input (NARX) models	- Input sequence selection is important	
	Nonlinear output error (NOE) models	- High sampling frequency	
	Nonlinear Box–Jenkins (NBJ) models	- Model identified can be used for prediction	
Feature	Structural characteristics	- Rely on extracting features based on the domain knowledge	[?]
	Sparsity-density entropy	- Dissimilarity measures are often customized	
		- Domain specific; hard to generalize	
Segmentation	Matrix profile	- Identify local abnormalities or repeated patterns	[?]
		- Requires a user-defined sliding window	

DTW distance are recorded in Table 2 (detailed explanation for calculating OT distance and DTW distance is provided in Section 4).

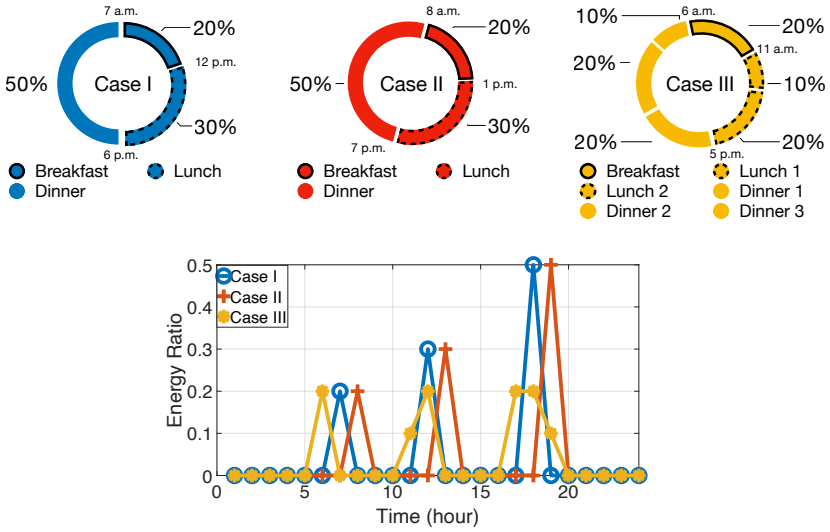


Fig. 1: (Top) Three donut charts illustrating three different eating patterns. The tilted donuts on Case II and Case III describe the shifts in the intake time for breakfast, lunch, and dinner compared to Case I. (Bottom) Three dietary temporal sequence data with x-axis representing hour of the day and y-axis representing energy intake ratio of each hour. The raw data (for all the three cases) represent the typical structured temporal sequence considered in this paper and this data posses a meaningful distributional structure, i.e., the energy intake distribution over a 24-hour cycle.

The insights gained from this example are as follows: The results recorded in Table 2 indicate that the Euclidean distance is sensitive to dynamic shifts yielding large distance values for all the three pairs (especially, between Cases I and II), and the OT distance well captures the distributional difference among these three samples, yielding relatively small values for all pairwise distances. However, the outputs from the OT suggests a smaller distance between Cases I and III than between Cases I and II, which creates ambiguity when the eating frequency (the number of total eating hours) is of interest. In addition, the outputs based on the DTW distance reveal no difference between Cases I and II. This is due to the (time warped) shape-invariant nature of the DTW distance, which also yields similar distances between Cases I and III and Cases II and III. In the context of the dietary data considered, these findings indicate that the DTW distance lacks of ability to distinguish the energy intake distributional difference between samples when the eating hours are not identical (see Case I and Case II).

Hence, the DTW distance is not equipped to detect high-level distributional dissimilarity in the data while the OT distance is incapable of detecting the dynamic shape similarity. Since the two key criteria to distinguish temporal dietary patterns are the energy distribution and the eating frequency, mining temporal distribution patterns using a single distance metric, evidently, does not provide desirable results, and the underlying reason for this difficulty is the unknown geometrical structure of the space from which the data is generated. Specifically, the dynamic shape and distributional differences are not simultaneously captured by the commonly used distance metrics such as OT and DTW. Therefore, explicitly accounting for the special distributional structure in the data along with its dynamic shape similarity while clustering is essential.

Table 2: Pairwise distances of the three cases in Fig. 1.

Pairwise Distance	Case I & II	Case I & III	Case II & III
Euclidean	0.8718	0.4899	0.6782
OT	0.0022	0.0014	0.0047
DTW	0	0.4	0.4

Motivated by this example, in this paper, we consider the structured temporal sequence data, and propose a nested two-stage clustering algorithm. The proposed algorithm can be categorized as an extension of the traditional hierarchical clustering algorithm [**?**]. However, in contrast to the traditional hierarchical algorithms (top-down or bottom-up), wherein the same distance metric persists at different stages, we propose to employ carefully designed distinct distance metrics in the two stages of our algorithm for distinct representations of the same structured temporal sequence data. We find that the difficulty in capturing the distribution and shape differences by a single metric, as shown in the aforementioned example, is mitigated with the proposed nested two-stage clustering algorithm.

Next, we develop the pre-processing steps to produce the discrete and continuous probability measure representations that are compatible with clustering using OT, and steps to produce the time-series representation for clustering using DTW.

3 Representations of Structured Temporal Sequences

In this section, we present the pre-processing steps required for the proposed clustering algorithm. We first discuss the details of two formulations of probability measure (both contin-

uous and discrete probability measures) to facilitate discrimination of temporal distribution pattern. Then we introduce the time-series representation and its importance for dynamic shape comparison with an appropriate time-series distance metric. In practice, the differences in the length of the time sequences in a given dataset and the differences in the scale of the data can be an impediment for clustering structured temporal sequence data. To conquer these difficulties, we design a pre-processing strategy including uniform interval-based aggregation and normalization with the sum as follows.

We consider the structured temporal sequence composed of real-valued data points indexed by time denoting their position in the sequence. Specifically, we call the raw data considered in this paper as the structured temporal sequence data. Such data are time-series data that possess a meaningful distributional structure. In contrast, when the distributional structure is not explicitly considered (as in Sec. 3.3), we call it a time-series. The main reason for this distinction is to highlight the fact that not all time-series data can have a valid distributional representation. Since we use OT distance as one of the distance metrics, for ease of exposition, we assume that the data points are re-scaled such that they are non-negative real numbers. For example, consider the i^{th} sample in the dataset. The temporal sequence, s_i, corresponding to this i^{th} sample is denoted as $s_i = \{(t_{i,1}, x_{i,1}), \ldots, (t_{i,\alpha}, x_{i,\alpha}), \ldots, (t_{i,M_i}, x_{i,M_i})\}$ where $t_{i,\alpha} \in [0, T_{max}]$ is the time of the α^{th} record, $x_{i,\alpha} \geq 0$ is the value of the α^{th} record, T_{max} is the upper limit of time for all samples, and M_i is the length of s_i. We further assume that the $x_{i,\alpha}$ values are summable. Most recorded discrete random event data, such as temporal dietary record, retail sales data, and smart meter energy profile [**? ?**], satisfy these properties.

For pre-processing, we assume that the entire time length, $[0, T_{max}]$, is partitioned into a uniform set of contiguous, disjoint n intervals (bins) $I = \{I_1, \ldots, I_n\}$, where $I_j = [\underline{I}_j, \bar{I}_j)$, $\underline{I}_1 = 0$ and $\bar{I}_n = T_{max}$. The i^{th} sample, s_i, is then transformed into a vector of tuples $\{(I_1, X_{i,1}), \ldots, (I_j, X_{i,j}), \ldots, (I_n, X_{i,n})\}$, where $X_{i,j} = \sum_{t_{i,\alpha} \in I_j} x_{i,\alpha}$. This aggregation can serve as a noise-reducing step when the data collection is in minuscule time scale, and it is unnecessary to study temporal patterns at such scale. Additionally, this uniform binning automatically equalizes the sequences with different number of events to the length n.

In practice, the bin size can be adjusted according to the requirement, such as temporal patterns at hourly scale, or daily scale, etc. To accommodate the temporal pattern's invariance to the scale of x, we further normalize each $X_{i,j}$ using $\sum_{j=1}^{n} X_{i,j}$ for each data sample and denote the normalized value as $p_{i,j} = \frac{X_{i,j}}{\sum_{j=1}^{n} X_{i,j}}$. So the processed i^{th} sample becomes $\{(I_1, p_{i,1}), \ldots, (I_j, p_{i,j}), \ldots, (I_n, p_{i,n})\}$. Viewpoints of the tuple $(I_j, p_{i,j})$ play an important role in designing a dissimilarity measure for an unsupervised clustering of structured temporal sequence task.

This motivates us to treat the processed sample as a discrete probability measure or a histogram of continuous probability measure, with the underlying space time-ordered. Alternatively, from a different perspective, the processed sequence of values $\{p_{i,j}\}$ is still a sequence of data points indexed by time, though the normalization process (in the pre-processing step) is different from the traditional z-score approach [**? **]. We introduce the details to construct these three representations in the following subsections.

3.1 Discrete Probability Measure Representation

For the i^{th} data sample, define $\mathbf{p}_i = [p_{i,1}, \ldots, p_{i,n}]^T$. The vector $\mathbf{p}_i$ denotes a probability vector as it satisfies the axiom of probability. Then, each $\mathbf{p}_i$ belongs to the probability simplex Σ_n

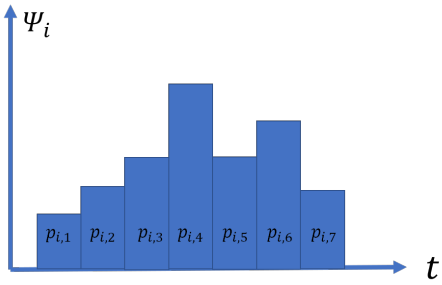


Fig. 2: An illustration plot of the histogram Ψ_i. $p_{i,j}$ represents the height of bin I_j for Ψ_i.

where

$$\Sigma_n := \left\{ \mathbf{q} = [q_1, \ldots, q_n]^T \in \mathbb{R}_+^n : \sum_{j=1}^{n} q_j = 1 \right\}.$$

To transform $\{(I_1, p_{i,1}), \ldots, (I_j, p_{i,j}), \ldots, (I_n, p_{i,n})\}$ to a discrete probability measure, we use a singleton, $\{T_j\}$, to represent the interval I_j such that $T_j = j$ for $j = \{1, \ldots, n\}$. Then for the discrete probability measure ν_i, $\nu_i(\{T_j\}) = p_{i,j}$. The singleton $\{T_j\}$ is called an atom of ν_i [**?**]. With this constructed discrete probability measure notation, the problem of calculating dissimilarity measure between two temporal sequences s_1 and s_2 is transformed to a problem of calculating a distance between two discrete probability measures ν_1 and ν_2.

3.2 Continuous Probability Measure Representation

In this section, we consider another feasible distribution representation of the temporal sequence, i.e., as a histogram of continuous probability measure which is assumed to be uniformly distributed within each interval I_j. In contrast to the traditional approach for constructing the histogram [**?**], we propose an alternative strategy, in which the histogram Ψ_i for s_i is directly constructed from $\{(I_1, p_{i,1}), \ldots, (I_j, p_{i,j}), \ldots, (I_n, p_{i,n})\}$, where $p_{i,j}$ becomes the height of bin I_j. The illustration is plotted in Figure 2.

Given the histogram Ψ_i, we can further define the continuous probability measure ϕ_i as follows. For each bin I_j, the distribution value $p_{i,j}$ can be treated as an integral from an uniform probability measure on I_j. So ϕ_i can represent a probability distribution on $[0, T_{max}]$ with uniform measure on each bin $I_j, j \in \{1, \ldots, n\}$. To numerically compute Φ_i, the empirical CDF (cumulative distribution function) of the continuous probability measure, ϕ_i, from Ψ_i, we start by calculating the cumulative weights of $p_{i,j}$ at each bin boundary. Let the cumulative weights $w_{i,j}$ at the right boundary of the bin I_j be calculated as

$$w_{i,j} = \sum_{l=1}^{j} p_{i,l} \quad \forall j = 1, ..., n. \tag{1}$$

Adopting the uniform distribution assumption within each I_j, $\Phi_i(t)$ becomes

$$\Phi_i(t) = w_{i,j-1} + \frac{t - \underline{I}_j}{\bar{I}_j - \underline{I}_j}(w_{i,j} - w_{i,j-1}), \ \forall t \in [\underline{I}_j, \bar{I}_j]. \tag{2}$$

Then the quantile function (inverse CDF) is a piecewise function defined as follows

$$\Phi_i^{-1}(w) = \underline{I}_j + \frac{w - w_{i,j-1}}{w_{i,j} - w_{i,j-1}}(\overline{I}_j - \underline{I}_j),\ \forall w \in [w_{i,j-1}, w_{i,j}]. \tag{3}$$

Given the assumption of uniform density over all pre-defined bins $I_j \in I$, the corresponding relationship between Φ_i and Ψ_i is one-to-one. In Section 4, the calculations for OT distance metric between Φ_{i_1} and Φ_{i_2} using $\Phi_{i_1}^{-1}(w)$ and $\Phi_{i_2}^{-1}(w)$ will be detailed.

3.3 Time-series Representation

For the comparison of time-series using DTW in the second stage, the order of values is more significant than the underlying metric space of time as in the OT based first stage. From Section 3.1, not only can $\mathbf{p}_i$ represent a probability vector, but it can also be treated as a time-series and the processed time-series representation of all structured temporal sequence samples have the same length n. The normalization in the pre-processing step automatically guarantees the time-series values are all at the same scale ($[0,1]$) and saves the additional z-score normalization step [**?**].

We devote the next section to design the appropriate dissimilarity measure for all the above three representations, and introduce the distance-based clustering algorithms.

4 The Nested Two-stage Clustering

Our proposed nested clustering framework is a novel approach to explicitly capture both the distributional difference and the dynamic shape difference between structured temporal sequences. As mentioned in Section 1, instead of customizing either the OT distance or the DTW distance to achieve an ideal dissimilarity measure, we choose to keep the simplicity of both distance metrics and tackle the structural temporal difference one at a time in a nested hierarchy. In this section, we first present the algorithms employed in our proposed nested two-stage clustering framework, and then, discussions on the choice of distance metrics, and their ordering in the two stages are presented.

4.1 Clustering of Discrete Probability Measure

4.1.1 Optimal Transport

Before presenting the algorithm used in our work, we briefly point out the original formulation of the problem of finding the optimal transport distance on discrete measures. The Kantorovich's formulation [**?**] of OT on discrete measures is a linear programming problem ($P0$):

$$(P0)\quad L_C(\nu_1,\nu_2) = \min_{\Pi \in U(\nu_1,\nu_2)} \langle \Pi, C\rangle = \min_{\Pi \in U(\nu_1,\nu_2)} \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} c_{j_1,j_2}\pi_{j_1,j_2}, \tag{4}$$

$$s.t. \sum_{j_2=1}^{n} \pi_{j_1,j_2} = p_{1,j_1}, \quad \sum_{j_1=1}^{n} \pi_{j_1,j_2} = p_{2,j_2}, \forall j_1, j_2 \in \{1,\ldots,n\},\ \ \pi_{j_1,j_2} \geq 0. \tag{5}$$

In this program, $U(\nu_1,\nu_2)$ is called the transport polytope defined by constraints (5), and $\Pi = \{\pi_{j_1,j_2}\}$, as a doubly stochastic transport matrix, consists values of a joint probability distribution with marginals $\mathbf{p}_1$ and $\mathbf{p}_2$. Here $\langle \Pi, C\rangle$ denotes the Hadamard product between Π and C, $C = \{c_{j_1,j_2}\}$ is the pairwise ground distance matrix and $c_{j_1,j_2} = ||T_{j_1} - T_{j_2}||_2^p$ is the p^{th} power of the Euclidean distance. Then, the p-Wasserstein distance is $W_p(\nu_1,\nu_2) := L_C(\nu_1,\nu_2)^{1/p}$, where $L_C(\nu_1,\nu_2)$ is the solution of $(P0)$ [**?**].

The formulation in $(P0)$ is a classical problem in linear programming and can be solved using the simplex method or the interior point method. However, the high computational complexity ($O(n^3\log(n))$) prohibits its application for large-scale problems. Therefore, the OT distance calculation in our approach is in line with the entropic regularized approximation, utilizing Sinkhorn's algorithm [**?**].

4.1.2 Entropic Regularized OT Distance

Let the discrete entropy of Π be defined as

$$H(\Pi) = -\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}[\pi_{j_1,j_2}\log(\pi_{j_1,j_2})]. \tag{6}$$

Then, the entropic regularized OT problem is given as

$$(P1) \quad L_C^{\varepsilon}(\nu_1,\nu_2) = \min_{\Pi\in U(\nu_1,\nu_2)} \langle \Pi, C\rangle - \varepsilon H(\Pi), \tag{7}$$

with the constraints (5) and $\varepsilon > 0$. The convergence of the solution from $(P1)$ towards the optimal solution of $(P0)$ has been shown in [**?**]. The regularized OT distance can then be calculated from equation (7) such that $W_2^2(\nu_1,\nu_2) \approx \langle \Pi^{\varepsilon *}, C\rangle - \varepsilon H(\Pi^{\varepsilon *})$. This entropic regularization brings significant numerical advantage for the calculation of OT distance between discrete measures due to the linear convergence rate of Sinkhorn's algorithm [**?**].

After computing the distance using OT, the samples are clustered using a variant of the K- means algorithm which is presented next.

4.1.3 Extension of Lloyd's Algorithm for OT Distance

The key to implementing a K-means type clustering algorithm is an iterative assignment and refinement procedure for all samples given initial cluster centroids. We denote N samples of processed discrete probability measure as $\{\nu_i : i = 1,...,N\}$. Then, the Lloyd's algorithm is a heuristic algorithm to minimize the following optimization objective

$$J = \sum_{k=1}^{K}\sum_{i:g(i)=k} W_2^2(\mu_k,\nu_i). \tag{8}$$

where g is a mapping from sample to cluster index, μ_k is the centroid of cluster k, and $W_2^2(\mu_k,\nu_i)$ is analogous to the within-cluster variance notation in the Euclidean K-means case. However, in the iterative refinement step, the centroid of the cluster is not a simple arithmetic mean of samples anymore. It turns out that when members of cluster k ($k \in \{1,2...,K\}$) are given, the calculation of the cluster centroid is essentially a Wasserstein barycenter (see Appendix for definition) calculation problem.

Note that [**?**] considers the general case with non-negative weights λ_i ($\sum_i \lambda_i = 1$) in front of each distance $W_2^2(\mu,\nu_i)$. We only consider uniform averaging as $\lambda_i = \frac{1}{N}$ here because

we do not assume any prior knowledge about the weights of different samples. For the numerical calculation of Wasserstein barycenter on the strictly positive probability simplex Σ_n (because of the entropic regularization term), we add a small positive shift ($\delta = 10^{-6}$) to all elements of the probability vector $\mathbf{p}_i$ and renormalize them to sum to 1. For practical clustering purposes, the effects of this small shift can be ignored.

4.2 Clustering of Time-series

Although DTW is not a true metric since it does not satisfy the triangle inequality, it is still the most popular time-series dissimilarity measure owing to its robustness to shift and dilation variance [**?** **?**]. Analogous to OT-means, DTW-means shares the same objective format as in (8) with $W(\cdot,\cdot)$ replaced by $E(\cdot,\cdot)$, such that

$$J = \sum_{k=1}^{K} \sum_{i:g(i)=k} E^2(\eta_k, \mathbf{p}_i). \tag{9}$$

where η_k is the DTW barycenter of cluster k (see Appendix for the definition of DTW barycenter). Then, η_k is calculated according to (14) using all samples $\{\mathbf{p}_i : g(i) = k\}$.

The change of distance metric from OT distance to DTW distance leads to a different solution approach for the centroid calculation. Since our goal in this paper is not to develop and compare DTW barycenter varieties, we adopt the classical DBA [**?**] approach in our experiments. Further discussions on the ordering and choice of algorithms in the two-stages are included at the end of this section.It is worth mentioning that the distributional characteristic at the first stage not only can be captured by the entropic regularized OT distance on discrete probability measures, but can also be studied using OT distance on continuous probability measures. Therefore, next, we present a computational approach for the clustering of the structured temporal sequence based on OT distance of continuous probability measures.

4.3 Clustering of Continuous Probability Measure

4.3.1 OT Distance for 1-D Continuous Measures

The main advantage of a 1-D continuous probability measure representation over a discrete probability measure representation boils down to a simpler expression for the 2-Wasserstein distance [**?**] using quantile functions. The 2-Wasserstein distance, D_2, for continuous measures is defined as

$$D_2(\Phi_1, \Phi_2) = \sqrt{\int_0^1 [\Phi_1^{-1}(w) - \Phi_2^{-1}(w)]^2 dw}. \tag{10}$$

This optimal transportation cost expression is true for any convex, nonnegative, symmetric ground cost function between continuous probability measures on the real line $\mathbb{R}$ (Theorem 2.18 of [**?**]). In this case, compared with the linear programming modeling for discrete measures, we get a simpler mathematical form as in (10) since all we need is the construction of the quantile function $\Phi_i^{-1}(w)$ for each sample i in the given dataset.

4.3.2 Clustering Under Continuous OT Distance

Similar to the case for discrete probability measures in Section IV.A, we consider a K-means clustering framework with 1-D continuous OT distance D_2. Given N samples of histogram representation $\{\Psi_i : i = 1,\ldots,N\}$ and the corresponding calculated $\{\Phi_i : i = 1,\ldots,N\}$, the minimization objective is

$$J = \sum_{k=1}^{K} \sum_{i:g(i)=k} D_2^2(\overline{\Phi}_k, \Phi_i), \tag{11}$$

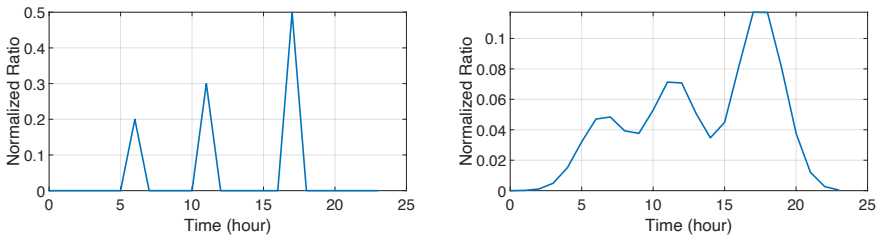
where the Wasserstein barycenter, $\overline{\Phi}_k$, of cluster k is defined as in Remark 1 in the Appendix.

This simple closed-form expression for Wasserstein barycenter of 1-D continuous measures greatly simplifies the theoretical analysis of barycenter properties [**?**], and the OT distance and Wasserstein barycenter of 1-D continuous measures are more succinct than that of discrete measures $(P0)$. However, whether this simple mathematical form also brings numerical accuracy or computational efficiency advantage over entropic regularized OT distance on discrete measures $(P1)$ is still in question. We further investigate the performance difference of these two approaches in the numerical results section.

4.4 Choice of Distance Metrics

While DTW distance is a popular choice to find the dynamic shape difference between time-series, there are many feasible distance metrics for probability measures, such as f-divergence based distance [**?**], Jensen-Shannon divergence [**?**], etc. However, these distances only measure the distribution difference corresponding to the same atom (for discrete probability measure) or the same bin (for continuous measure). This translates to high sensitivity to the choice of bin size and shift in the p_j values in time. In contrast, the optimal transport distance naturally compares distribution values including the cross-atom (T_{j_1} to $T_{j_2}, j_1 \neq j_2$) or cross-bin (I_{j_1} to $I_{j_2}, j_1 \neq j_2$) and yields intuitive geometrical meaning. So we considered the OT distance as the primary choice for calculating the distributional difference between temporal sequences for both discrete and continuous probability measure representations.

The proposed clustering framework holds a nested hierarchical structure, namely the second stage DTW-means is implemented within the cluster outputs of the first stage OT-means. The rationale for this ordering is the intrinsic higher hierarchy of the distribution invariance than the translation or shift invariance. To better visualize the difference of the two barycenters under OT and DTW distance, we include a plot of comparison of DTW barycenter and OT barycenter for Case I & II of Fig. 1 in Fig. 3. For the two samples with similar distribution value but a shift of one hour in the underlying support space, the DTW barycenter keeps the dynamic shape of the samples well (actually in this case, it is exactly equal to the sample 2), while the Wasserstein barycenter summarizes the distributional average of the two with three smoothed peaks of increasing height over time. The observations of Fig. 3 inspired the choice of distance metric for each hierarchy in the nested two-stage algorithm, which adopts OT-means or OTC-means to cluster the rough distributional structure at the first stage and then, utilizes DTW-means to cluster the finer shape structure such as the number of peaks, relative peak heights in the second stage. If instead, we switch the order of OT and DTW, and implement DTW at the first stage, the distribution invariance at the second stage would be interfered because DTW cluster outputs separate very similar



(a) (b)

Fig. 3: (a) DTW barycenter, which preserves the shape of the averaged samples. (b) OT barycenter, which disperses the peak value keeping only the rough shape.

distribution patterns with local shape difference into different clusters. Hence, characterization of distribution variance within these outputs become inconsistent. Empirically, we test this conjecture in the numerical experiments of synthetic dataset. Another practical merit of the current approach is that if we only care about temporal distributional pattern, the intermediate OT-means or OTC-means output can serve the needs. On the other hand, if we want a more delicate structural pattern analysis, the output of the second-stage DTW-means can provide the desired results.

In terms of the design choice of the clustering algorithms in both the stages, the main concern is scalability. Spectral clustering and hierarchical clustering methods require precomputing of the whole pairwise distance matrix, which is not scalable with the size of the data sample N. On the contrary, due to the developments of computational methods for the barycenter under OT and DTW geometry, we apply the K-means type clustering algorithm (Lloyd's algorithm, to be exact) with adapted OT or DTW distance, which significantly saves the computation time by calculating the distance from each sample to the cluster centroids only ($K \ll N$). The other benefit of a K-means type algorithm is the representation of the cluster using the cluster centroid which can further be used for nearest neighbor based classification. The cluster centroids out of the two stages of our algorithm can provide both visual discrimination and nearest neighbor based classification capabilities when additional samples are added after clustering.

5 Experimental Results

In this section, we evaluate the performance of our two-stage clustering method using two synthetic datasets with known class labels. We compare the clustering results obtained using the proposed algorithms with four additional distance-based K-means type clustering algorithms, which are OT-OT, DTW-DTW, MDTW-MDTW [**?**], and Euclidean-Euclidean. In these four algorithms, both of the two stages share the same distance metric as the name suggests. We also present a comparison based on the numerical performance of OT-DTW and OTC-DTW. In addition, we evaluate the performance of existing algorithms, including the popular agglomerative hierarchical cluster tree [**?**] with the aforementioned four distances, as well as a well-developed time-series clustering algorithm k-Shape with shape-based distance (SBD) and DTW-based distance [**?**]. All algorithms were implemented in MATLAB

2018a and tested on a desktop configured with Intel i7-5820k and 32GB of memory[1]. A comprehensive performance comparison is shown in the following subsections.

5.1 Synthetic Data Generation

In what follows, we describe how synthetic data were generated in both examples.

Example 1. The synthetic data for this example is generated through a Monte Carlo experiment, which has six known clusters each with 27 unique samples of length 24 (i.e., $K = 6$, $N_k = 27$, $n = 24$). Of the six clusters (see Fig. 4), three clusters are designed such that each of the 27 samples in them is chosen with 3 nonzero bins. Each bin index in the sequence is picked randomly from three sets $\{6,7,8\}$, $\{11,12,13\}$, and $\{17,18,19\}$ respectively, and the values are a permutation of $\{0.6,0.3,0.1\}$. The remaining three clusters are with samples chosen with six nonzero bins whose bin index are picked from the same three sets (two from each set), and whose values are a permutation of $\{0.3,0.3,0.15,0.15,0.05,0.05\}$. For example, for clusters (a), (c) and (e), we uniformly sample one bin index from each of the three sets and assign a distribution value from $\{0.6,0.3,0.1\}$ and the remaining 21 bins in the sequence are assigned zero. For clusters (b), (d) and (f), we uniformly sample two points within each set and assign, for both points, a value from $\{0.3,0.15,0.05\}$. Thus, the dataset is defined such that $T_j \in \{6,7,8,11,12,13,17,18,19\}$ and $p_{i,j} \in \{0.6,0.3,0.1,0.3,0.3,0.15,0.15,0.05,0.05\}$. Among the six clusters, the three pairs (clusters (a) and (b), clusters (c) and (d), and clusters (e) and (f)) share the same distribution pattern, and within each pair, samples in the second cluster is composed of twice the number of positive values as the first one. In this example, $\mathbf{p_i}$ for each sample is restricted to take values from $\{0.6,0.3,0.1\}$ or $\{0.3,0.3,0.15,0.15,0.05,0.05\}$, which is relaxed in the next example.

Example 2. The synthetic data generation process for Example 2 is similar to Example 1, but with different distributions for each of the six clusters. Here, for the six clusters of Example 2, we modify the samples in cluster (b), (d) and (f) of Example 1 to have three pairs of positive values sampled from $\{0.25,0.25,0.15,0.15,0.1,0.1\}$, and modify the distribution values in cluster (a) from $\{0.6,0.3,0.1\}$ to $\{0.7,0.2,0.1\}$ and the distribution values in cluster (c) changed from $\{0.1,0.3,0.6\}$ to $\{0.2,0.3,0.5\}$ (cluster (e) unchanged). This new synthetic dataset also has six known clusters with 27 samples each, and representatives of each cluster are shown Figure 5. For both examples, 100 runs of experiments with random initializations using the six different distance-based clustering algorithms are run and the results are summarized next.

5.2 Performance on Synthetic Data

With known class labels in the synthetic data, we evaluate different algorithms using popular external validation metrics, including the adjusted rand index (ARI) [**?**], the Mirkin index (MI) [**?**], the Hubert index (HI) [**?**], the Normalized Mutual Index (NMI) [**?**], the Jaccard index (Jaccard) [**?**], and the Fowlkes-Mallows index (FM) [**?**] as the performance quantification criteria for the comparison of different clustering methods. All of the clustering validation metrics have range $[0,1]$, and yield better performance when the value is closer to 1, except for the MI which yields better performance when the value is smaller.

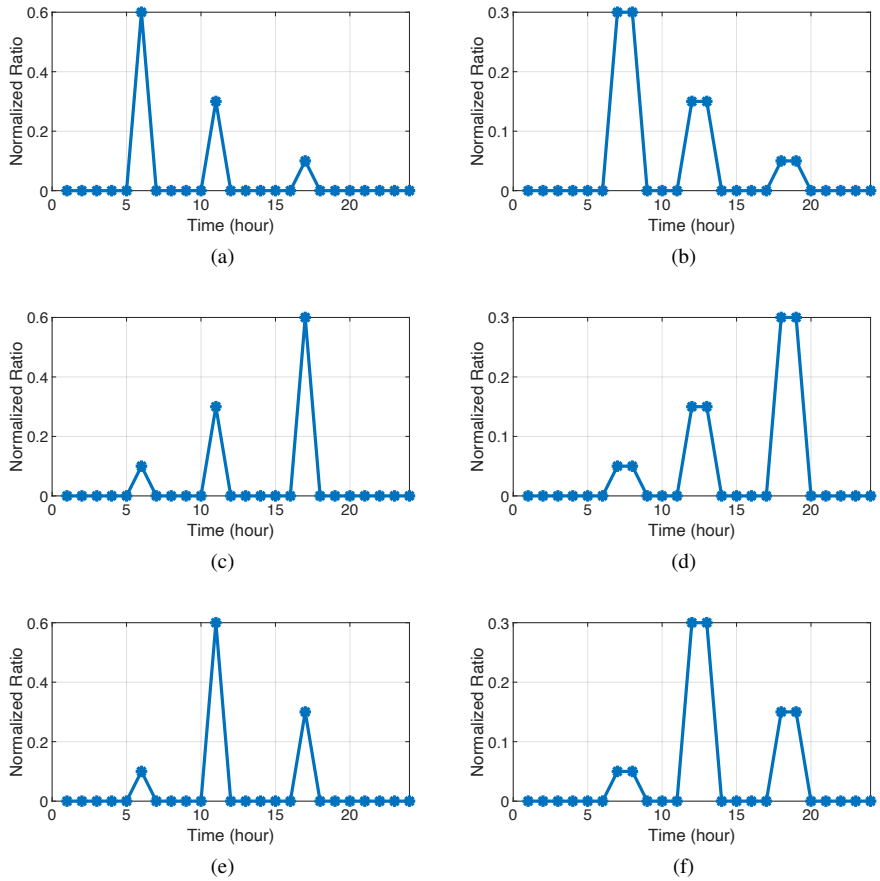


Fig. 4: The representative sample for the six clusters in Example 1. Each plot from (a) to (f) represents a randomly chosen sample from cluster (a) to (f). The remaining 26 samples within the same cluster only have time-shift difference with the representative. As the plots show, each pair in the same row denote the same distribution pattern, but the right plots always have double the number of positive values.

Across all experiments, we assign the number of clusters, $K_1 = 3$ in the first stage and $K_{2,1} = K_{2,2} = K_{2,3} = 2$ in the second stage, based on the prior knowledge of data. The OT-OT and DTW-DTW algorithm results are included, and for MDTW-MDTW, we implement the kernelized DTW-means [**?**] with DBA as the barycenter calculation method. The Euc-Euc is serving only as a baseline since Euclidean geometry is known not to be working well for these structured temporal sequences. The difference of OTC-DTW and OT-DTW exists only at the first stage and DTW-means is applied at the second stage for both methods. Also DTW-OT is tested on both examples to verify the importance of the order of two distance metrics.

[1] All the source codes have been made public on https://github.com/AML-wustl/OT-DTW

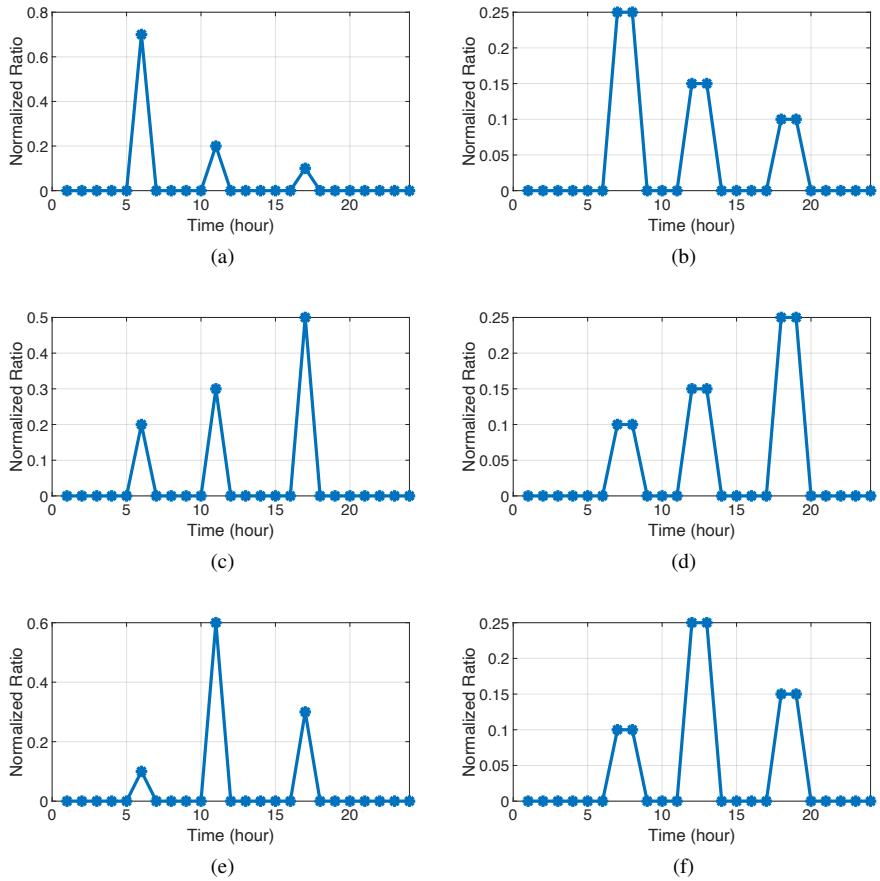


Fig. 5: The representative sample for the six clusters in Example 2. Each plot from (a) to (f) represents a randomly chosen sample from cluster (a) to (f). The remaining 26 samples within the same cluster only have time-shift difference with the representative. Compared with Fig. 4, each pair in the same row denote similar, but not the same distribution pattern anymore.

The mean values of the validation metrics out of 100 runs of experiments with random initializations using the seven different distance-based clustering algorithms for Example 1 and Example 2 are recorded in Table 3 and Table 4, respectively. In both tables, one-stage clustering algorithms generally show inferior performance. In two-stage algorithms, Euc-Euc yields the worst performance as expected, because it takes into account neither the distribution nor dynamic shape difference in the structured temporal sequence data. It is also observed that the MDTW-MDTW [**?**] results are not always better than the DTW-DTW, as the performance of MDTW depends on the choice of hyper-parameter (the additional penalty term corresponding to the time difference in the local distance). It is worth noting that although the OTC-DTW runtime is only about one third of OT-DTW runtime, the average performance is not as good as that of OT-DTW. A further investigation of the first stage

Table 3: (Example 1) mean values of six clustering validation metrics and runtime over 100 runs with different distance-based clustering methods and random initializations.

Clustering method (metrics) / Clustering validation metrics	k-means							Agglomerative				K-Shape	
	OT-OT	DTW -DTW	MDTW -MDTW	EUC -EUC	DTW -OT	OTC -DTW	OT -DTW	OT	DTW	MDTW	EUC	SBD	DTW
ARI	0.41	0.69	0.73	0.20	0.74	0.73	**1**	0.56	0.59	0.83	0.61	0.31	0.24
MI[a]	0.17	0.09	0.08	0.26	0.08	0.09	**0**	0.17	0.12	0.05	0.13	0.23	0.21
HI	0.67	0.82	0.83	0.48	0.84	0.83	**1**	0.67	0.75	0.90	0.74	0.53	0.58
NMI	0.62	0.81	0.84	0.37	0.87	0.86	**1**	0.76	0.77	0.92	0.81	0.54	0.41
Jaccard	0.34	0.60	0.63	0.21	0.65	0.65	**1**	0.48	0.50	0.75	0.52	0.29	0.23
FM	0.95	0.91	0.93	0.84	0.91	0.89	**1**	0.71	0.87	0.93	0.80	0.80	0.97
runtime(s)	1.65	0.99	1.15	0.01	1.19	0.50	1.33	7.02	0.62	0.66	0.01	0.26	1.76

[a] MI index shows better performance when the value is closer to 0.

Table 4: (Example 2) mean values of six clustering validation metrics and runtime over 100 runs with different distance-based clustering methods and random initializations.

Clustering method (metrics) / Clustering validation metrics	k-means							Agglomerative				K-Shape	
	OT-OT	DTW -DTW	MDTW -MDTW	EUC -EUC	DTW -OT	OTC -DTW	OT -DTW	OT	DTW	MDTW	EUC	SBD	DTW
ARI	0.65	0.72	0.69	0.08	0.73	0.68	**0.96**	0.46	0.70	0.84	0.39	0.34	0.19
MI[a]	0.01	0.09	0.10	0.36	0.09	0.11	**0.01**	0.22	0.09	0.04	0.24	0.23	0.32
HI	0.80	0.83	0.80	0.27	0.83	0.78	**0.98**	0.56	0.83	0.91	0.52	0.55	0.36
NMI	0.80	0.85	0.84	0.27	0.88	0.85	**0.98**	0.72	0.85	0.92	0.66	0.58	0.35
Jaccard	0.56	0.64	0.61	0.16	0.64	0.60	**0.96**	0.41	0.61	0.75	0.35	0.31	0.23
FM	0.94	0.89	0.88	0.84	0.88	0.83	**0.98**	0.66	0.90	0.95	0.68	0.78	0.71
runtime(s)	2.16	1.22	1.65	0.01	1.43	2.99	1.77	9.62	0.52	0.53	0.01	0.37	1.65

[a] MI index shows better performance when the value is closer to 0.

clustering output revealed that the OTC-means is highly sensitive with the random initialization of cluster centroids, and with good initialization in some of the 100 runs (which are not shown here) we can achieve comparable performance in that of OT-DTW. The inferior performance on DTW-OT, compared with OT-DTW and other one-stage algorithms, confirms our findings that the order of distance metric in this nested framework indeed matters. In summary, in both examples, OT-DTW yields the best performance validation metrics among all distance-based clustering methods and its performance is not vulnerable to the random initialization of cluster centroids.

Next, we discuss the choice of optimal K for both stages of our OT-DTW method when the data generation process is unknown because K is the key parameter in the K-means type clustering algorithm. Since we do not assume any prior knowledge of the number of clusters in the real-world dataset, no external truth for validation as in the synthetic data case is available. Therefore, the best K has to be determined with the help of cluster validity index, which measures the relative dispersion of clusters for a given K. Here, we choose two indices, the Davies-Bouldin (DB) index and the Calinski-Harabasz (CH) index to find an optimal choice of K at each stage. The smaller the value of DB index, the better the value of K. On the other hand, the larger the value of CH index, the better the value of K (see Appendix A for the definitions of the two indices).

In addition to choosing K, another parameter to choose in the clustering algorithm is the initialization of cluster centroids. In the simplest case, Lloyd's algorithm begins with

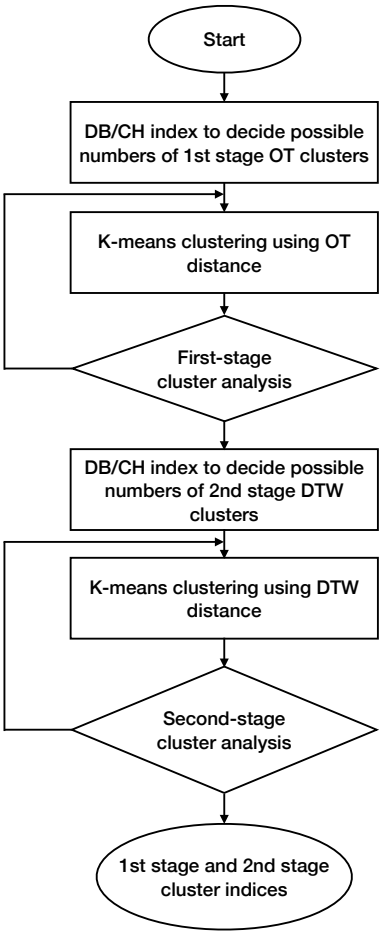


Fig. 6: A flowchart of the proposed OT-DTW algorithm.

K arbitrary centroids sampled uniformly from all samples. However, there is no guarantee that this random initialization would converge to a good local minimum. Therefore, here we choose the K-means++ [**?**] for our experiments, and extend all the sampling weights calculation using W_2 or D_2, instead of the original Euclidean distance due to the $O(\log K)$-optimal clustering guarantee in expectation for K-means++.

6 Application on Real-world Datasets

In this section, we demonstrate the performance of the proposed OT-DTW two-stage clustering method to retrieve the underlying patterns in structured temporal sequence dataset in three application domains: temporal dietary record, retail product sales, and smart meter energy profile [**? ?**]. All of our three applications follow the flowchart in Fig. 6.

6.1 Temporal Dietary Data

Temporal dietary pattern analysis has emerged as a multifaceted approach to examine the relationship between diet and the risk of chronic diseases. Compared with singular nutrient or dietary component approaches, it incorporates a more comprehensive picture of both the quantity and time of dietary intake. In the next sections, we first briefly describe the format of the temporal dietary record dataset and then demonstrate the OT-DTW clustering results compared with the other distance-based clustering algorithms used in the synthetic data case.

6.1.1 Data Description

The temporal dietary record data consists of dietary intake events in a 24-hour period from 1021 participants. It is collected from an "Interactive Diet and Activity Tracking in AARP (IDATA)" study [**?**], in which participants who were older than 50 years old and living in Pittsburgh, PA reported their diet activity using ASA24 (Automated Self-Administered 24-hour recall). The data records were anonymized due to privacy protection reasons. For each participant, the number of recalls varied from 4 to 6 over the course of a year. Nevertheless, to focus on the illustration of our method, we ignore the modeling for the averaging of multiple recalls and only use his/her first recall as the representative pattern. All dietary intake events were reported with an accuracy of minutes, and we aggregate them in hours using the pre-processing step. The validity is owing to the number of eating events for each person ranging from 1 to 9, and discussing the temporal dietary pattern in the time-scale of minutes is unnecessary in practice.

6.1.2 Empirical Performance of First-stage OT-means

First, we varied the K value from 2 to 10 and calculated both the DB index and CH index (see Figure 18 in the Appendix), and $K = 4$ and $K = 8$ seem to be good candidates for the number of clusters. The cluster centroids of first-stage OT-means for $K = 4$ are plotted in

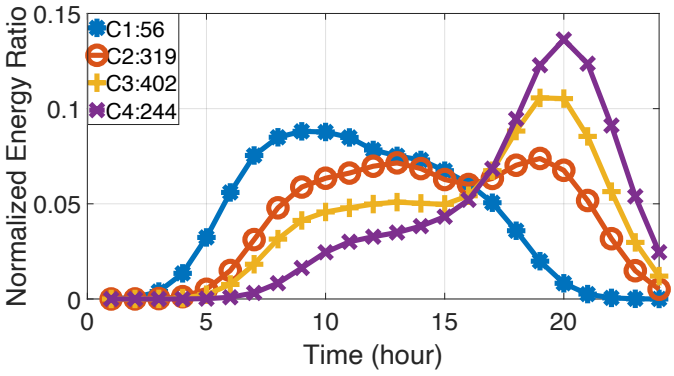


Fig. 7: First-stage OT-means cluster centroids on temporal dietary dataset. The number in the legend represent the number of samples in each cluster C1-C4.

Figure 7. The different clusters are labeled C1-C4 with different colors and the cluster size is

shown in the legend. In the following, we further discuss these centroids' possible suggested dietary distribution patterns, which serve as the representative of the corresponding clusters. Cluster C3 and C4, as the largest two clusters, represent the dinner-dominated type of eating pattern, which match US population's common eating habit with a large dinner. Compared with C3, C4 has a higher energy ratio in the night and also a higher chance to skip the breakfast in the morning. Nevertheless, cluster C2 leans towards eating patterns with intake events happening equally as likely during the day time. Cluster C1, as the smallest cluster, represents a rare pattern that energy intake distribute mainly in periods before afternoon. Because this dataset consists mostly of people older than 60 years old, one possible reason for the emergence of this cluster could be some early risers with aging who start their day early. A reminder is that OT-means centroids here only tell a rough distributional average and dots in Figure 7 do not directly correspond to dietary intake events. These distributional clusters will serve as a pre-conditioning mask for the second-stage DTW-means clustering and the more delicate temporal dietary patterns should be determined from the second-stage clustering outputs.

6.1.3 Empirical Performance of Second-stage DTW-means

In this subsection, we further apply DTW-means clustering to cluster outputs from the first-stage OT-means clustering. The K value in the second-stage DTW-clustering is also determined with the assistance of DB and CH index, as in Table 5 (in Appendix C). If the optimal choice of K are not consistent between the two indices, we choose the smaller value (all the chosen K are in bold font in Table 5 in Appendix C). In Figure 8, the sub-cluster centroids of each of the four cluster outputs are plotted, with the sub-cluster sample number shown in the legend. The shape of sub-cluster centroids match with our preliminary observations of DTW barycenter of two samples in Figure 3. As illustrated in the previous subsection, the OT barycenter only tells the rough distribution of probability mass across time bins, and sub-clusters within the same cluster could exhibit different time and frequency of eating behavior. For example for the smallest cluster C1, the five sub-cluster centroids show totally different eating time and eating frequency. The 1st sub-cluster centroid shows one dominating meal in the noon, while the 2nd sub-cluster centroid shows a three meal pattern with the main energy intake in the lunch. However, it is clear that they are distinctive from subclusters of the other three clusters because the majority of their energy is distributed in the morning. The other three larger clusters can also be analyzed similarly using the corresponding sub-cluster centroids for comparison of subtler eating patterns in terms of eating frequency, eating time, etc. Overall, the aggregated view of sub-cluster centroids within the same cluster all match well with the cluster centroids in Figure 7.

To better visualize the two-stage cluster outputs for practical use, we plot in Figure 9 using piechart and barchart to visualize the sample size distribution of clustering results, and in Figure 10 using the mean energy ratio weighted time to visualize an averaged property of clusters. Figure 9 reveals a rare energy distribution pattern, and it turns out that C1 is a cluster corresponding to a rare pattern with majority of energy intake in the morning. And for the smallest subcluster C25 in C2, when we check the centroids plot in Figure 8b, the centroid pattern has only one major meal in the noon, which again is a rare pattern revealed by the clusters. These results (Figure 9) demonstrate the potential of the proposed clustering approach to help detect abnormality in temporal dietary patterns. In Figure 10, utilizing the property of discrete probability representation, we can calculate the energy ratio weighted average time for each sample, in other words, the expected value for the concentration of energy, and the average expected time for each cluster. The clear difference of energy ratio

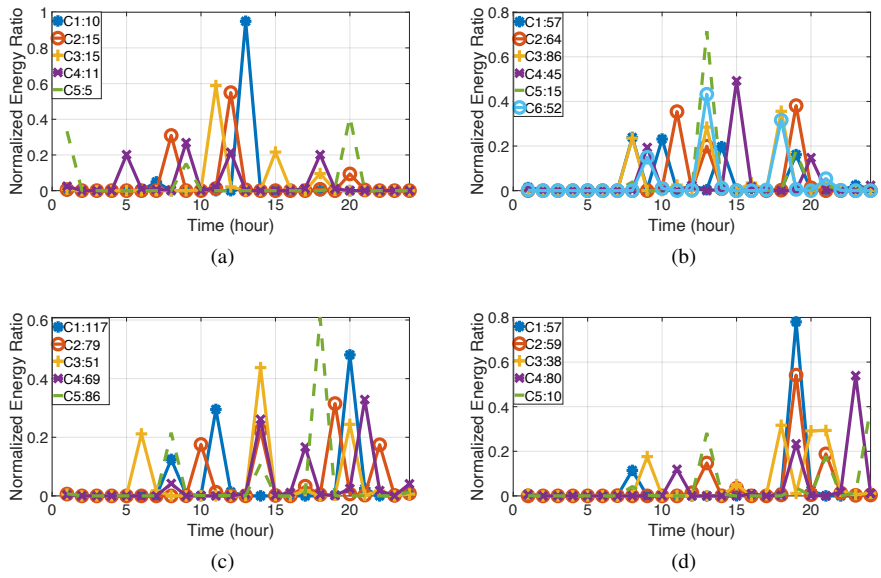


Fig. 8: Second-stage DTW-means cluster centroids for OT-DTW on temporal dietary dataset. The four subplots (a)-(d) correspond to the four clusters C1-C4 in Figure 7.

weighted average time across first-stage clusters confirms the distributional discrimination power of our method.

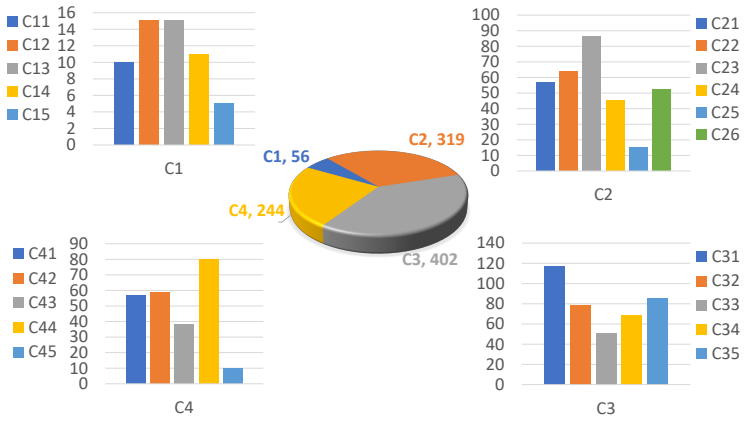


Fig. 9: The two-stage OT-DTW cluster output chart for Temporal Dietary Dataset. The piechart shows the sample size of the four cluster outputs of the first-stage. The four bar charts show the subcluster output sample size distribution of the second-stage.

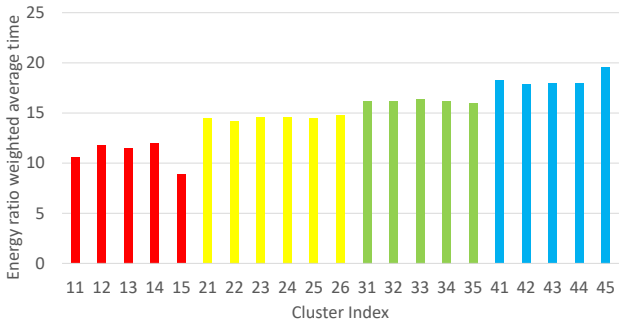


Fig. 10: The plot shows the mean energy ratio weighted average time for each OT-DTW cluster of Temporal Dietary Dataset. And x-axis is the cluster index, denoted in digits, namely '11' stands for subcluster 1 within cluster 1. Distinct first-stage clusters are colored in distinct colors.

6.1.4 Results From Other Methods

We implement OT-OT and DTW-DTW to plot the centroids as a comparison. In Figure 11, given the four cluster outputs from 1st-stage OT-means, we further apply OT-means with the optimal choice of K to get subclusters. We observe many overlaps owing to distribution similarities from these subcluster centroids, especially in (c) and (d). This demonstrates that the objective to distinguish dynamic shape difference cannot be achieved by hierarchically increasing the number of clusters under OT distance.

In Figure 12 and Figure 13, we plot the first stage and second stage cluster outputs of DTW-DTW, respectively. Although the four cluster centroids in Figure 12 seem different in terms of distribution, but compared with Figure 7, those samples with dominating energy distribution in the morning or the noon are missing. If instead, we try to interpret first stage results as dynamic shape difference only, subclusters in each of them as shown in Figure 13 cannot clearly tell the distribution difference. Above all, DTW-DTW cannot achieve the desired simultaneous discrimination of structural distribution and dynamic shape difference.

6.2 Online Retail Data

This online retail data is from the UCI Machine Learning Repository [**?**] and contains all the transactions occuring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retailer. The company mainly sells unique all-occasion gifts to wholesalers. We organize the sales quantity by StockCode (Product code) and aggregate it in month to get a sales quantity sequence of length 12 for each product (We only consider 12 entire months and ignore the sales in December of 2011). After cleaning the products with negative values for sales quantity, we finally have a monthly sales dataset for 3202 unique products. The goal is then to apply our OT-DTW algorithm to cluster these products to discover interesting sales pattern, and further optimize the bundled procurement decision.

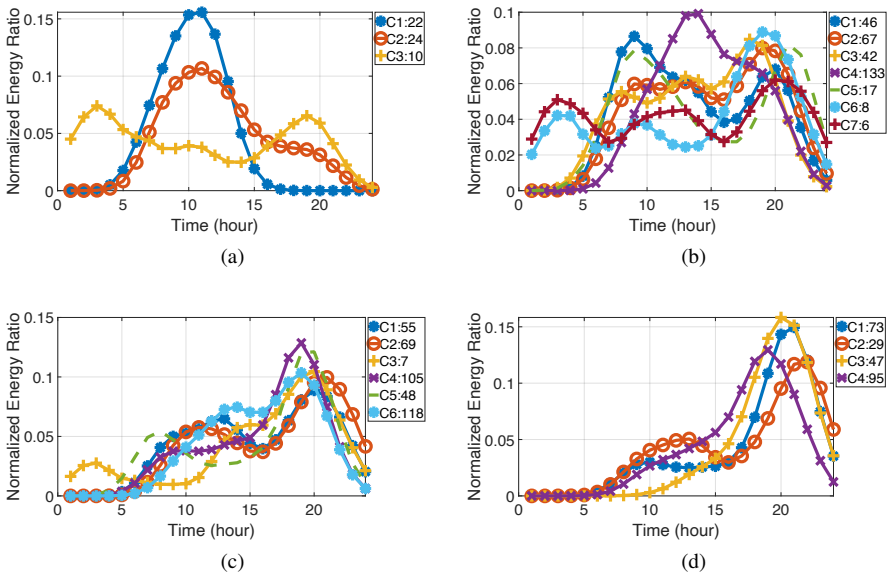


Fig. 11: Second-stage OT-means cluster centroids for OT-OT on temporal dietary dataset. The four subplots (a)-(d) correspond to the four clusters C1-C4 in Figure 7

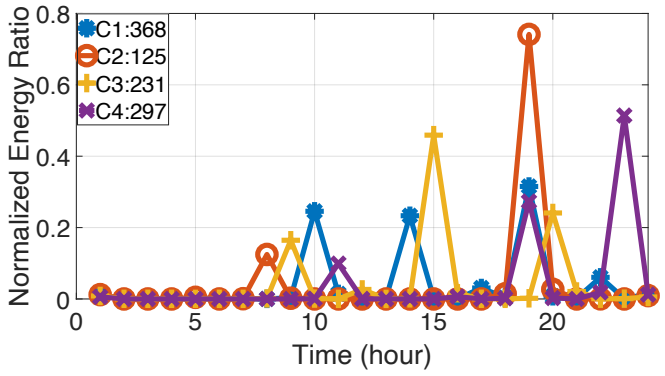


Fig. 12: First-stage DTW-means cluster centroids for DTW-DTW on temporal dietary dataset. The number in the legend represent the number of samples in each cluster.

6.2.1 Empirical Performance of First-stage OT-means

Here again, we varied the K value from 2 to 10 and calculated both the DB index and CH index first, and found that $K = 6$ is a good candidate for the number of clusters (see Figure 19 in the Appendix). We plot the cluster centroids of the first-stage OT-means for $K = 6$ in Figure 14. The different clusters are labeled C1-C6 with different colors and the cluster size is shown in the legend. The difference across products' seasonal sales pattern can be observed from the plot, where C1 and C4 tend to represent product clusters with

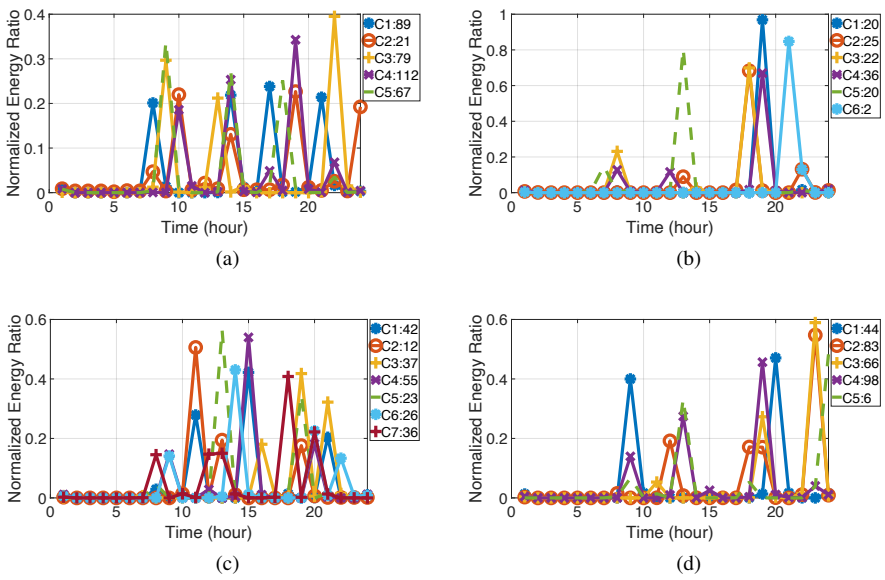


Fig. 13: Second-stage DTW-means cluster centroids for DTW-DTW on temporal dietary dataset. The four subplots (a)-(d) correspond to the four cluster C1-C4 in Figure 12.

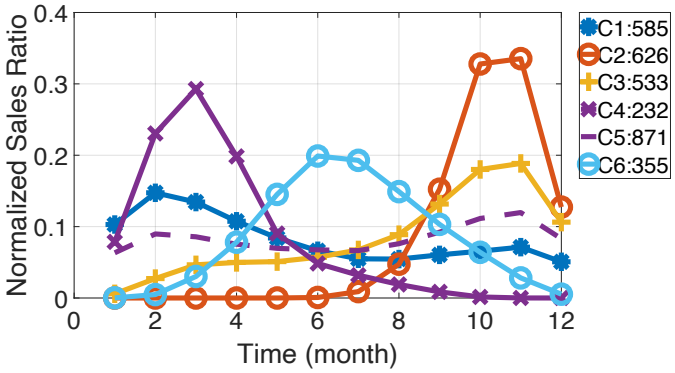


Fig. 14: First-stage OT-means cluster centroids on online retail dataset. The number in the legend represent the number of samples in each cluster. The six cluster centroids show six distinctive sales patterns over 12 months of the year.

more sales in the spring, C2 and C3 tend to represent product clusters with more sales in the winter, C6 is the summer major sales product cluster, and finally C5's sales is the most stable throughout the year. A further sample exploration confirms these product clusters' composition. For example, C6 contains "paper pocket travelling fan" and "sandlewood fan" and C2 contains "first class holiday purse" and "set 10 cards christmas tree". Analogous to the case for temporal dietary data, we investigate details of the dynamic shape difference from the time-series perspective in the following second-stage DTW means.

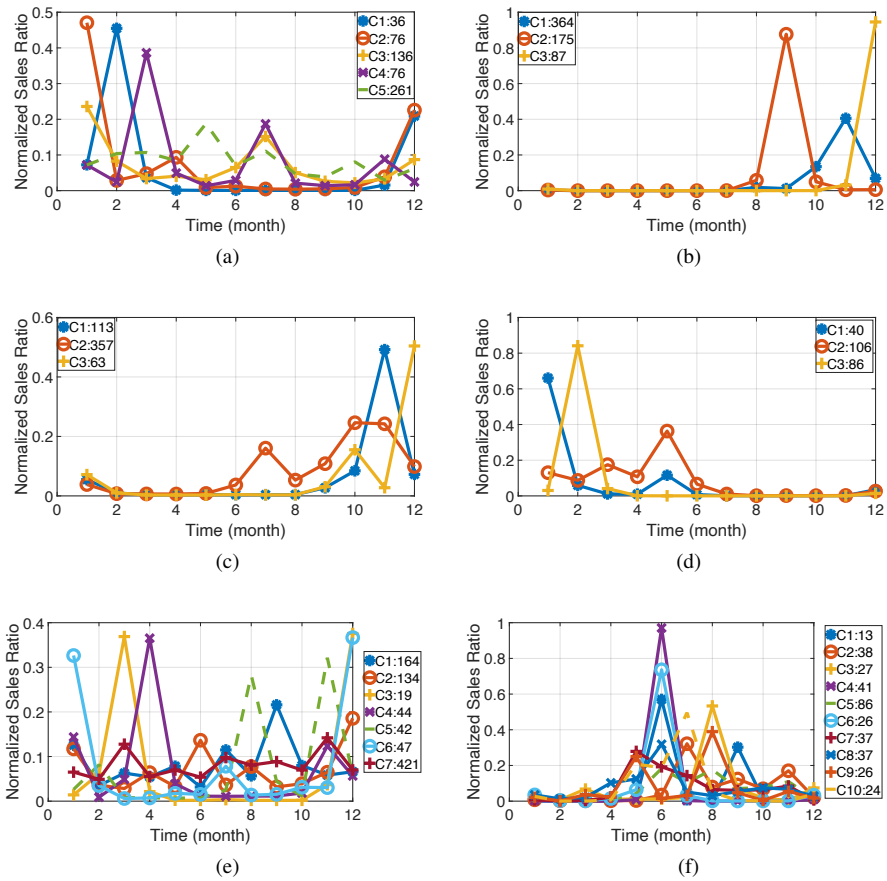


Fig. 15: Second-stage DTW-means cluster centroids for OT-DTW on online retail dataset. The six subplots (a)-(f) correspond to the six clusters C1-C6 in Figure 14.

6.2.2 Empirical Performance of Second-stage DTW-means

In this subsection, we further apply DTW-means clustering to cluster outputs from the first-stage clustering. The K value in the second-stage DTW-clustering is also determined with the assistance of DB and CH index, as in Table 6 (see Appendix C). For cluster 6, we further check K values from 11 to 15 and $K = 10$ is indeed the local minimum for DB_6. In Figure 15, the sub-cluster centroids of each of the four cluster outputs are plotted in color, with the subcluster sample number shown in the legend. The subcluster centroids's dynamic shape difference is reflected in the peak number and relative height, etc. Nevertheless, the aggregated view of subcluster centroids within the same cluster all match well with the cluster centroids in Figure 14. Similar to the results on temporal dietary dataset, the visualization of the two-stage clustering results for online retail dataset is recorded in Figure 21, and the comparison of the mean energy ratio weighted average time in Figure 20 (included in the Appendix).

6.3 Smart Meter Energy Consumption Data

This smart meter energy consumption data is collected from London Households that took part in the UK Power Networks led Low Carbon London project between November 2011 and February 2014[2]. All the readings were taken at half hourly intervals and the customers in the trial were recruited as a balanced sample representing the Greater London population. We processed the data according to the pre-processing step introduced in Section 3 and obtain a sample of 5084 households with a temporal sequence of length 48. The cluster number K at both stages are also determined using the DB and CH index as in the previous examples. The plots of the first stage OT-means cluster centroids are presented in Figure 16, and the plots of the second stage DTW-means subcluster centroids are recorded in Figure 17. The resulting cluster results on smart meter energy consumption dataset in Figure 23, and the comparison of the mean energy ratio weighted average time in Figure 22 validates that the proposed algorithm preserves distributional and shape patterns, and reveals valuable insights on the energy consumption patterns (see Appendix).

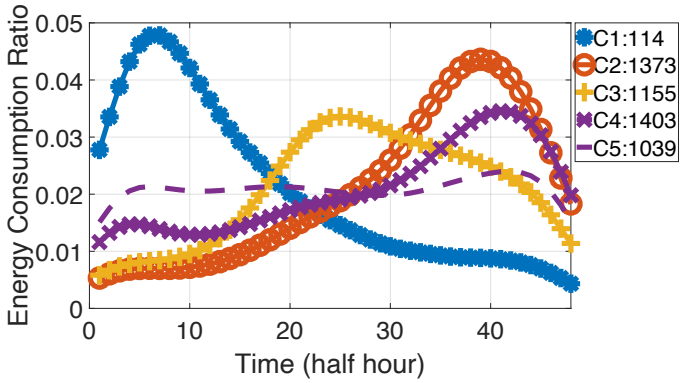


Fig. 16: First-stage OT-means cluster centroids on smart meter energy consumption dataset. The number in the legend represents the number of samples in each cluster.

6.4 Discussions

The results on the three real-world datasets all successfully reveal the hierarchical distributional and dynamic shape-based cluster of patterns in each application. Depending on the goal of the data mining exploration, the revealed clusters can be used for outlier detection, or reveal patterns valuable for crucial decision making process.

6.4.1 Limitation and Extension

The proposed algorithm has some limitations at its present form, and extensions can be made in the furture work to accomodate these constraints. One is concerned with the ne-

[2] The data is publicly available at https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

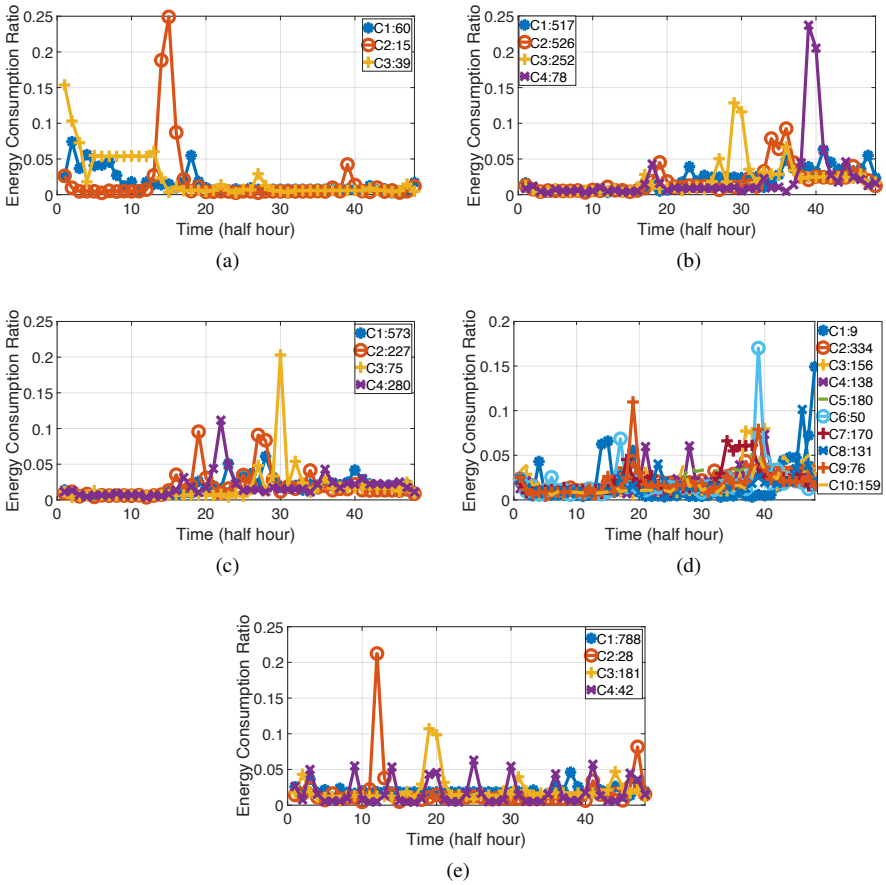


Fig. 17: Second-stage DTW-means cluster centroids for OT-DTW on smart meter energy consumption dataset. The five subplots (a)-(e) correspond to the five clusters C1-C5 in Figure

glect of absolute input values before normalization of the temporal patterns learned using our method. For example, in the temporal dietary data case, two persons with the same energy intake ratio of 60% at 6pm could yield absolute intake of 500 calories and 2000 calories, respectively. The daily energy intake standard is known to be correlated with the person's body weight, food preference, and health condition, etc. Thus, one way to further extend our method is to add a parametric model component to the dissimilarity measure incorporating all these factors. Another restriction of our method is the inability to deal with multi-dimensional structured temporal sequences. For example, in the smart meter energy profile data, each household may have multiple days of observations of energy usage. How to find a valid way to combine these multiple observations of the same household, and then define the dissimilarity measure is a viable future extension.

6.4.2 Broader Applications

Apart from the temporal pattern discovery in the applications discussed in Section 6, the proposed clustering algorithm appears to possess some desirable properties which would extend its use in the synchronization detection application in an oscillator network [**?**] (see Appendix B). Traditionally, this problem requires pre-processing of the data by peak-finding or Hilbert transform (to extract phase information from the measured data), and further, clustering according to the phase difference [**?**]. Our method can save the expensive phase processing step, and directly work with the raw recordings to achieve very similar results with the phase difference based approach (see Appendix B for preliminary results). We conjecture that the distributional difference and the dynamic shape difference in the time domain have some intrinsic correlation with the phase synchronization and plan to pursue this direction further.

7 Conclusions

Mining patterns from temporal sequence data is an important data mining problem with broad applications. In this paper, we develop a nested OT-DTW distance-based clustering method for one type of structured temporal sequence data. The first-stage OT-means clustering captures the macroscopic temporal distributional pattern, while the second-stage DTW-means clustering produces subtle subclusters according to dynamic shape difference. At the same time, we also investigate the data representation differences for OT-based clustering algorithm from discrete probability measure to histogram of continuous probability measure. The experiment results on the synthetic data confirm the clear performance edge of OT-DTW over other popular distance-based clustering methods. Above all, the OT-DTW clustering results on the three real-world datasets, temporal dietary record, online retail data, and smart meter energy profile all demonstrate distinct, satisfying cluster and subcluster centroids and are easily accessible to further outlier detection or characteristics analysis.

For future work, we plan to further extend the current nested clustering framework with two distinct distance metrics to a general number of n stages and with different combinations of distances for specific unsupervised data exploration use. We believe this nested clustering structure with unique distance metric at each stage has more potential in data mining problems than the current shown structured temporal sequence.

Appendix

Remark 1 The Wasserstein barycenter $\overline{\Phi}_k$ of n_k continuous distributions $\{\Phi_1, \ldots, \Phi_{n_k}\}$ of cluster k under the objective of Definition (13) satisfies

$$\overline{\Phi}_k^{-1}(w) = \frac{1}{n_k} \sum_{i:g(i)=k} \Phi_i^{-1}(w), \forall\, w \in [0,1]. \tag{12}$$

Remark 2 (Wasserstein Barycenter, [**?**]) A Wasserstein barycenter of N measures $\{\nu_i : i = 1, ..., N\}$ in $\mathbb{P} \subset P(\Omega)$ is a minimizer of f over $\mathbb{P}$, where

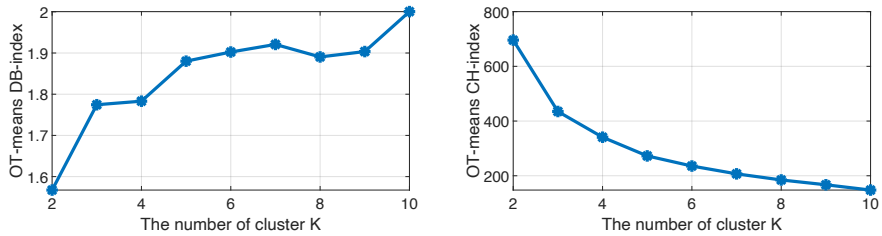
$$\mu^* := \arg\min_{\mu} f(\mu) = \arg\min_{\mu} \sum_{i=1}^{N} \lambda_i W_2^2(\mu, \nu_i). \tag{13}$$

Remark 3 (DTW Barycenter) A DTW barycenter of N time-series $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ in a space $\mathbb{E}$ induced by DTW metric is a minimizer of the sum of squared distance to the set P, where

$$\eta^* := \arg\min_{\eta} \frac{1}{N} \sum_{i=1}^{N} E^2(\eta, \mathbf{p}_i). \tag{14}$$

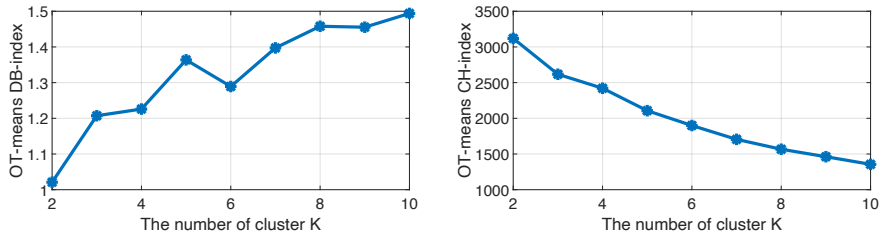
A. Results

Based on the definition of DB and CH indices, we seek to find the local minimum of DB index and the local maximum of CH index. From Figure 18b, the CH index strictly decreases with increasing K and there is no clear kink point towards plateau, which provides little information for the optimal choice of K. From Figure 18a, due to the relative smaller DB index and clearer separation of cluster centroids, we set $K = 4$ in the current experiment (Example of temporal dietary dataset). From Figure 19b, the CH index also strictly decreases with increasing K and provides little information for the optimal choice of K. But from Figure 19a, $K = 6$ becomes a good candidate for the number of clusters since the DB index achieves local minimum then.



(a) DB index (b) CH index

Fig. 18: Cluster validity index DB and CH for experiments with K ranging from 2 to 10, to determine the optimal choice of K for the first stage OT-means on temporal dietary dataset.



(a) DB index (b) CH index

Fig. 19: Cluster validity index DB and CH for experiments with K ranging from 2 to 10, to determine the optimal choice of K for the first stage OT-means on Online Retail Dataset.

K	2	3	4	5	6	7	8	9	10
DB_1	0.91	1.11	1.09	**0.84**	0.94	0.97	0.95	0.92	0.86
CH_1	82.1	51.0	41.6	**47.0**	39.6	39.3	37.0	36.1	36.0
DB_2	1.20	1.55	1.74	1.58	**1.49**	1.42	1.49	1.43	1.42
CH_2	177	142	111	99.3	**103**	98.6	91.6	86.4	84.0
DB_3	1.39	1.46	1.53	**1.43**	1.33	1.40	1.31	1.42	1.44
CH_3	198	163	135	**138**	119	105	98.0	89.5	85.8
DB_4	1.29	1.19	1.29	**1.14**	1.23	1.16	1.14	1.27	1.32
CH_4	133	152	99.0	**132**	107	92.1	85.9	84.7	81.8

Table 5: *DB* and *CH* index values of the second-stage DTW-means clustering for varying *K* values. The subscript of DB and CH is the OT-means cluster number 1 to 4.

K	2	3	4	5	6	7	8	9	10
DB_1	1.02	1.19	1.39	**1.36**	1.56	1.55	1.49	1.60	1.57
CH_1	404	350	284	**240**	214	196	183	163	155
DB_2	1.01	**0.64**	0.84	1.07	1.02	0.99	1.17	1.15	1.20
CH_2	414	**1245**	973	968	944	898	829	819	785
DB_3	1.16	**1.12**	1.48	1.18	1.40	1.29	1.55	1.56	1.55
CH_3	278	**260**	213	248	213	204	166	157	148
DB_4	1.25	**0.81**	1.04	1.07	1.11	1.41	1.16	1.26	1.12
CH_4	121	**219**	212	204	181	162	162	140	151
DB_5	1.26	1.30	1.55	1.70	1.51	**1.50**	1.60	1.55	1.63
CH_5	402	374	284	256	242	**220**	201	193	184
DB_6	0.74	1.01	1.29	1.36	1.48	1.52	1.49	1.45	**1.41**
CH_6	496	420	323	295	257	227	207	193	**180**

Table 6: *DB* and *CH* index values of the second-stage DTW-means clustering for varying *K* values. The subscript is the OT-means cluster number 1 to 6.

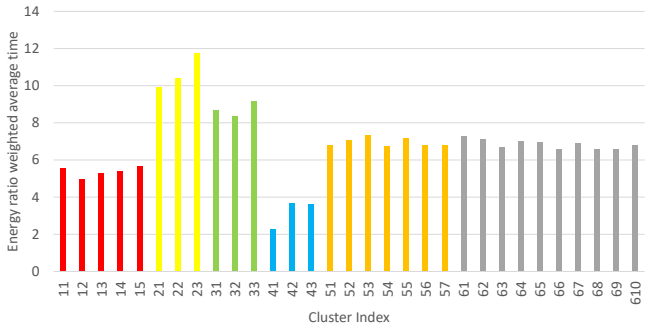


Fig. 20: The plot shows the mean energy ratio weighted average time for each cluster of Online Retail Dataset. Distinct first-stage clusters are colored in distinct colors.

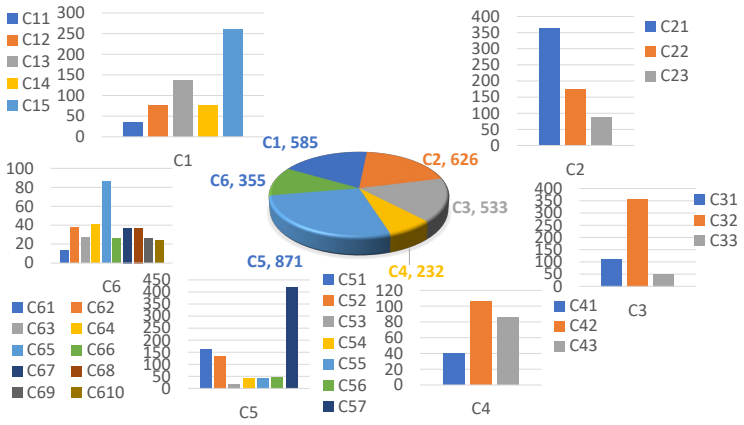


Fig. 21: The two-stage OT-DTW cluster output chart for Online Retail Dataset. The piechart in the middle shows the sample size of the six cluster outputs of the first-stage. The six bar charts show the subcluster output sample size distribution of the second-stage.

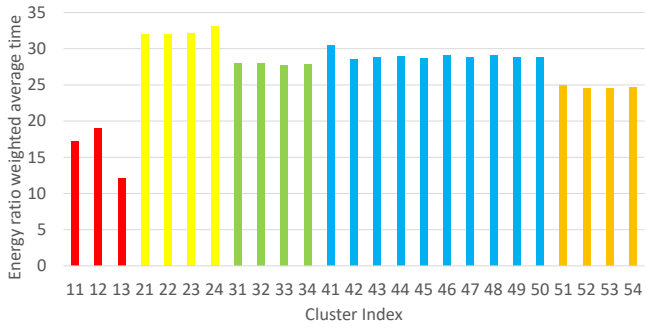


Fig. 22: The plot shows the mean energy ratio weighted average time for each cluster of Smart Meter Energy Consumption Dataset. Distinct first-stage clusters are colored in distinct colors.

B. Applications

Apart from the temporal pattern discovery in the applications discussed in Section 6, the proposed clustering algorithm appears to posses some desirable properties which would extend its use in synchronization detection application in an oscillator network [**?**]. The synchronization detection problem is defined as follows: in an oscillator network, each oscillator can be treated as a node in the network, and the coupling between oscillators are the edges. Each oscillator's dynamics consists of two parts- its own intrinsic dynamics, and the coupling functions from other oscillators. The network starts from an arbitrary initial condition, and evolves over time (according to the oscillator dynamical equations). Given the time-series measurement corresponding to the output of each oscillator, we aim to determine which

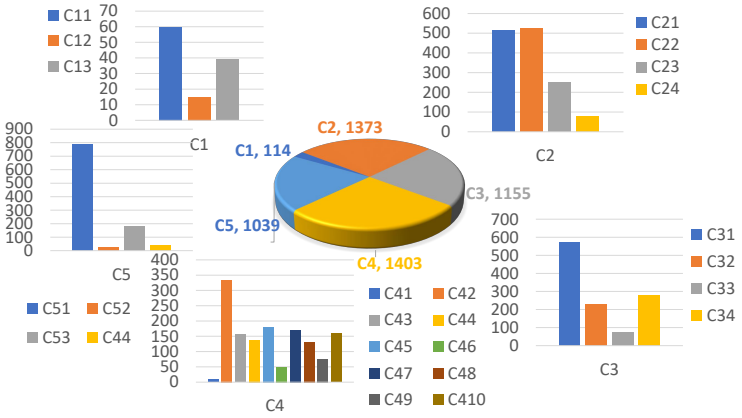


Fig. 23: The two-stage OT-DTW cluster output chart for Smart Meter Energy Consumption Dataset. The piechart in the middle shows the sample size of the five cluster outputs of the first-stage. The five bar charts show the subcluster output sample size distribution of the second-stage.

of the oscillators (nodes) are phase synchronized. Traditionally, this problem requires preprocessing of the data by peak-finding or Hilbert transform (to extract phase information from the measured data), and further, clustering according to the oscillator phase model [**?**]. Our method saves the expensive phase processing step, and can directly work with the recordings. For example, Figure 24 is an illustration of a synthetic oscillator network with 15 oscillators and cluster results from our OT-DTW method. The colored nodes in the left network plot provide the synchronization clusters based on phase difference calculation. On the right is our two-stage cluster outputs, and except oscillator 14, our cluster results match very well with the phase based synchronization clusters (our results also separate oscillator 7, 12, and 13 into a separate cluster from oscillator 2, 3, 6, and 8). This leads to our conjecture that the distributional difference and the dynamic shape difference in the time domain have some intrinsic correlation with the phase synchronization and we plan to pursue this direction in a future study.

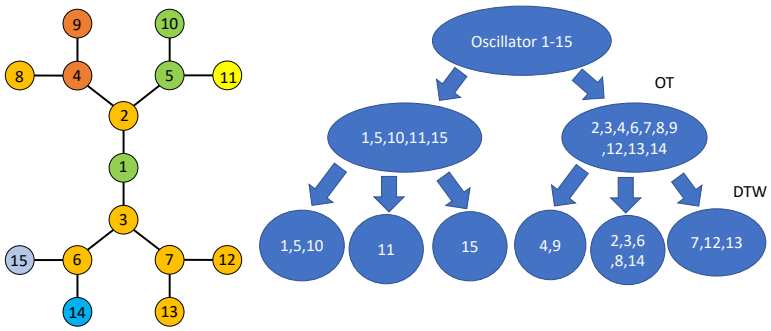


Fig. 24: The left is a 15 oscillator network with shown connection topology, and phase difference based synchronization clusters are colored in different colors. The right is an illustration of the cluster result using out OT-DTW method.