



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이학석사 학위논문

Step Data Clustering via
Thick Pen Transformation

굵은 펜 변환을 이용한 걸음 수 자료 군집화

2021년 2월

서울대학교 대학원

통계학과

김민지

Step Data Clustering via Thick Pen Transformation

지도교수 오 회 석

이 논문을 이학석사학위논문으로 제출함

2020년 10월

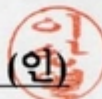
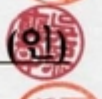
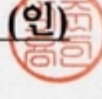
서울대학교 대학원

통계학과

김 민 지

김민지의 석사학위논문을 인준함

2021년 1월

위 원 장	<u>이 상 열</u>	(인) 
부 위 원 장	<u>오 회 석</u>	(인) 
위 원	<u>PARK JUN YONG</u>	(인) 

Abstract

Minji Kim

The Department of Statistics

The Graduate School

Seoul National University

This thesis studies clustering time-series data by suggesting a new similarity measure and an optimization algorithm. To illustrate, we propose a new time-series clustering method based on the Thick Pen Transformation (TPT) of Fryzlewicz and Oh (2011), whose basic idea is to draw along the data with a pen of given thicknesses. The main contribution of this research is that we suggest a new similarity measure for time-series data based on the overlap or gap between the two thick lines after transformation and smoothing. This method of applying TPT to measure the association exploits the strengths of the transformation; it is a multi-scale visualization technique that can be defined to provide some information on neighborhood values' temporal trends. Moreover, we further suggest an efficient iterative clustering optimization algorithm appropriate for the proposed measure. Our main motivation is to cluster a large number of physical step count data obtained from a wearable device. Moreover, a numerical simulation is performed to compare our method to some existing methods, suggesting that the proposed scheme can be adapted to more general cases.

Keywords: Time-series data, clustering, multi-scale method, dynamic time warping, thick pen transform

Student Number: 2019-28751

Contents

1	Introduction	1
2	Background	4
2.0.1	Thick Pen Transform	4
2.0.2	Thick Pen Measure of Association	5
3	Methodology	8
3.0.1	Transformation and smoothing	8
3.0.2	Similarity Measure	9
3.0.3	Optimization Problem for Clustering	12
3.0.4	Clustering Algorithm	14
4	Real data Analysis	16
5	Simulation study	20
6	Conclusions	24
	References	25

List of Tables

3.1	Hierarchical clustering results	10
4.1	Summary of clustering results	17
5.1	Means (standard deviations) of the correct classification rate (CCR) for each method	23

List of Figures

1.1	Three different step count data	2
2.1	(top) Round pen bound with thickness 30, (bottom) round pen bound with thickness 60.	6
3.1	(a) Original step data. (b) smoothed data by simple moving average with window size 5, and (c)-(f) smooth the data and apply the TPT: (c) square pen with thickness 30; (d) square pen with thickness 100; (e) ensemble square pen with thickness 30; (f) ensemble square pen with thickness 100	9
3.2	Six groups of synthetic data with different trends: (a) normal, (b) cyclic, (c) increasing, (d) decreasing, (e) upward shift, (f) downward shift.	10
3.3	Hierarchical clustering dendrogram for the TPMA result . . .	11
3.4	Two data in group (e) matched using the TPMA by the DTW algorithm	11
3.5	(a) Visualization of the overlapping areas between two data, colored by blue and red respectively. (b) $TPMA_0$ values of (a) . . .	11
4.1	Clustering results by using the TPT with $\tau = 30$	18
4.2	Clustering results by using the TPT with $\tau = 100$	18
4.3	Map of the distribution of individuals included in each group . . .	19
5.1	Four groups of sinusoidal data with different variabilities. . . .	21

5.2	Four groups of block data with different patterns.	21
5.3	Three groups of block data with different amount and patterns.	22
5.4	Different optimization settings used for the comparison. . . .	23

Chapter 1

Introduction

Clustering is an unsupervised classification problem where data objects with similar features are grouped together. In this thesis, our main focus is to cluster time-series data, which is intrinsically high-dimensional, and values tend to co-vary and thus are dependent on their neighborhoods. To cluster such high-dimensional data, much work has been done on suggesting new data representation method ([1]), or distance measure ([2],[3]). Moreover, functional data clustering approach which assumes that curves can be represented by a set belonging to an infinite dimensional space can be applied to time-series data ([4], [5], [6]). Numerical works in the literature are motivated by such data as gene expression data ([7],[8]), bike sharing systems data [9], power load supply data ([10]), and so on. Another example is Lim *et al.*'s ([11]) functional clustering of accelerometer data after transforming input variables based on the rank-based transform and the thick-pen transform, which is highly related to our motivating example.

Our main concern is to suggest new time-series similarity measure. Choosing an adequate distance measure is a controversial and important matter in time-series clustering domain. Euclidean distance and Dynamic Time Warping (DTW) are the most common methods for similarity measure in the time-series clustering ([12]). However, Euclidean distance considers each component

as a part of a long vector of independent values, which fails to take into account temporal trends and similarities in shape in time-series data sets. While dynamic time warping allows a non-linear mapping between two temporal sequences and provides a way to measure the similarity of sequences of different lengths, it is unsuitable when intending to reflect the time gap. Also, since the DTW matching can be applied to various distance measures or cost matrices, we note that proposing a new component-wise measure can employ the DTW algorithm as well.

The motivation of this study is to cluster a large set of step count data measured every minute from a wearable device. Each data is 1440-dimensional count data per day per individual recorded from 00:00 a.m. to 11:59 p.m. We aim to identify different step patterns of 19604 days over 79 people. Figure 1.1 shows three example plots of step data. As we can see, the data possess unique and interesting characteristics: it is high-dimensional, zero-inflated, and steps tend to occur discontinuously, that is, there are numerous moments when people take a break between each step. Our goal is to present a computationally efficient clustering method that can identify different trends of movements regarding their amounts and patterns.

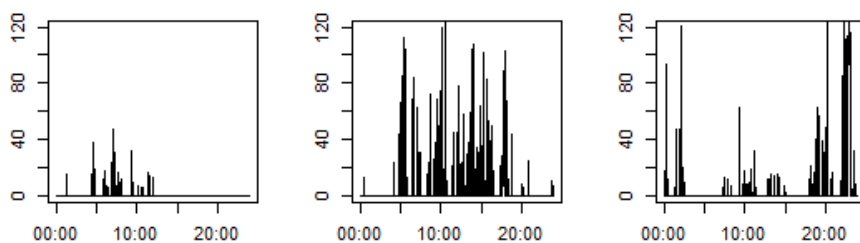


Figure 1.1: Three different step count data

In this study, we propose a novel time-series clustering scheme after transformation and smoothing, inspired by the Thick Pen Transform (TPT) of Fryzlewicz and Oh ([13]). TPT is a novel way of viewing time series at multiple scales using a range of pens with various thicknesses. The effectiveness

of applying TPT to measure the association lies in its flexibility. To be specific, the shape of a pen can vary; and, we define a new shape based on the mean of bounds of simple square pens so that the boundaries could encompass time-series trends of neighboring data points. The thickness of a pen can vary, enabling us to explore the multi-scale nature of TPT with larger thicknesses of pens bringing out coarser scale features of the data, which may diminish noise effects. We propose a new similarity measure for time-series clustering based on the overlap of the areas under the upper boundaries after applying thick pen transformation and smoothing. One of the main characteristics of the measure is that it is defined for between components of vectors, so that we can apply the DTW algorithm to find an optimal path between two vectors. Finally, we show that applying the k -medians algorithm for the logarithms of upper boundaries is appropriate for the clustering optimization problem for the suggested similarity measure.

The rest of this thesis is organized as follows. Chapter 2 introduces Thick Pen Transform and Thick Pen Measure of Association of Fryzlewicz and Oh (2011) as a background concept needed for our method. Chapter 3 suggests our clustering scheme, including data representation methods, a new similarity measure, and an appropriate clustering algorithm. Chapter 4 deals with a real data analysis of accelerometer data, and chapter 5 presents results on simulation data to compare our results with other methods. Chapter 6 concludes the thesis with some remarks.

Chapter 2

Background

2.0.1 Thick Pen Transform

The TPT of Fryzlewicz and Oh (2011) is based on the idea of drawing along the time series data points with a pen with its own shape and thickness. Let $\mathcal{T} = \{\tau_i : i = 1, \dots, |\mathcal{T}|\}$ denote the set of thickness parameters. The formal definition of the thick pen transform $\mathbf{TP}_{\mathcal{T}}(X_t)$ of a real valued univariate process $(X_t)_{t=1}^n$ is the following sequence of pairs of boundaries,

$$\mathbf{TP}_{\mathcal{T}}(X_t) = \{(L_t^{\tau_i}, U_t^{\tau_i})\}_{i=1, \dots, |\mathcal{T}|},$$

where $L_t^{\tau_i}$ and $U_t^{\tau_i}$ respectively represent the lower and the upper boundary of the area covered by a pen of thickness τ_i at time t .

The TPT plays three important roles in reflecting the time series data characteristics. To be specific, different shapes of a pen can be defined to manage how the transformed values are affected by the temporal trends of neighborhood values. For example, Fryzlewicz and Oh proposed the square and the round pen as follows.

(a) *Square pen* :

$$U_t^{\tau} = \max\{X_{t-\frac{\tau}{2}}, \dots, X_{t+\frac{\tau}{2}}\} + \frac{\tau}{2}\gamma$$

$$L_t^\tau = \min\{X_{t-\frac{\tau}{2}}, \dots, X_{t+\frac{\tau}{2}}\} - \frac{\tau}{2}\gamma$$

(b) *Round pen* :

$$U_t^\tau = \max_{k \in [-\frac{\tau}{2}, \frac{\tau}{2}] \cap \mathbb{Z}} \{X_{t+k} + \gamma\sqrt{\tau^2 \setminus 4 - k^2}\}$$

$$L_t^\tau = \min_{k \in [-\frac{\tau}{2}, \frac{\tau}{2}] \cap \mathbb{Z}} \{X_{t+k} - \gamma\sqrt{\tau^2 \setminus 4 - k^2}\}$$

In the above definition, \mathbb{Z} denotes the set of integers and γ is the scaling factor defined for adjusting the difference between the thickness of the pen and the variability of the data. Second, it has a multi-scale nature of viewing data at a different distance according to the thickness of a pen. To be specific, applying large τ values corresponds to zoom out and see trends of the data in a coarse way, while small τ values sensitively catch original features. Finally, the transformation is visually intuitive and informative. Figure 2.1 shows round pen boundaries with thickness 30 and 60 applied to a step count data. Different pen shapes are further addressed in the next chapter, shown in Figure 3.1.

2.0.2 Thick Pen Measure of Association

Fryzlewicz and Oh (2011) also proposed a way to measure the association between two time-series data based on the TPT. Let $L_t^\tau(Z)$ and $U_t^\tau(Z)$ be the lower and upper boundary for a generic process Z at time $t \in T$ and thickness τ . The thick pen measure of association (TPMA) between X and Y is defined as

$$\rho_t^\tau(X, Y) = \frac{\min\{U_t^\tau(X), U_t^\tau(Y)\} - \max\{L_t^\tau(X), L_t^\tau(Y)\}}{\max\{U_t^\tau(X), U_t^\tau(Y)\} - \min\{L_t^\tau(X), L_t^\tau(Y)\}}.$$

Some remarks can be made here.

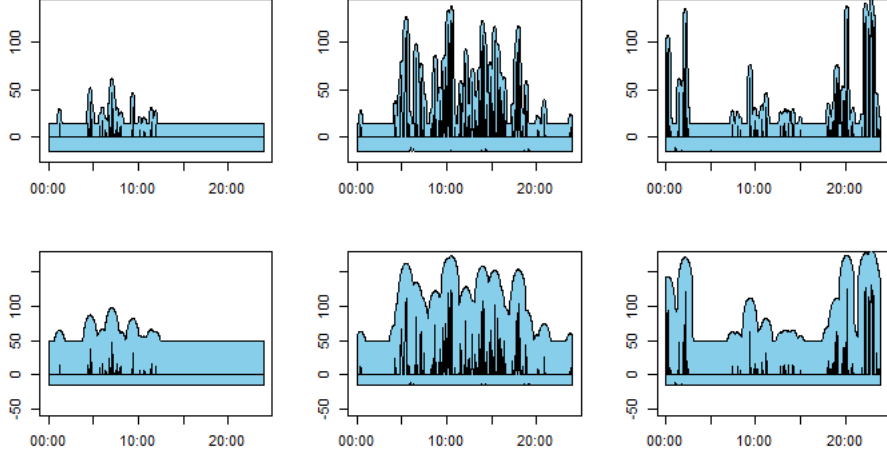


Figure 2.1: (top) Round pen bound with thickness 30, (bottom) round pen bound with thickness 60.

- The measure satisfies that $\rho_t^\tau(X, Y) \in (-1, 1]$. Note that $\rho_t^\tau(X, Y) > 0$ holds when there exists an overlap between the two thick boundaries, while $\rho_t^\tau(X, Y) < 0$ with a gap between them. This idea of measuring time-series dependence based on the overlap or gap of pen areas is intuitively perceived when we visualize the transform.
- The association is defined at each time t , preserving the dimension of the original process, $n = |T|$. This allows various applications to be attempted. For example, we can define a summary measure between two series in a various way; an overall mean, i.e. $\bar{\rho}_{1,n}^\tau(X, Y) = \frac{1}{n} \sum_{t=1}^n \rho_t^\tau(X, Y)$, mean of the first half, i.e. $\bar{\rho}_{1,n/2}^\tau(X, Y) = \frac{2}{n} \sum_{t=1}^{n/2} \rho_t^\tau(X, Y)$, or product, i.e. $\frac{1}{n} \prod_{t=1}^n \rho_t^\tau(X, Y)$, can be used as summarized similarity measures.
- In addition, the similarity measure is computed coordinate-wise so that the dynamic time warping algorithm is applicable on the TPMA to find an optimal match between two long vectors. This characteristic is also addressed in the next chapter, see Figure 3.4.

In this study, we work on defining a new similarity measure for clustering time-series data based on an application of the TPMA.

Chapter 3

Methodology

3.0.1 Transformation and smoothing

When we decide extra smoothness is advantageous, we first apply simple moving average to the data with a chosen window size. When applied to step count data, this smoothing process weakens the effect of a short momentary step generated between consecutive zeros. It can be observed in Figure 3.1 by comparing (a) and (b). Then, we apply TPT to get transformed pairs of boundaries. As previously stated, through this transformation, we can employ useful features of the TPT that is multi-scale and visually enlightening, embracing time-series local dependence structure. In this study, we define a variation of the square pen to get a smoothed version of thick pen boundaries. To illustrate, we define “Ensemble square pen” as ensemble means of bounds of simple square pens with different starting points.

The definition is as follows. Suppose that we have a real-valued uni-variate process $(X_t)_{t=1}^n$. Let $\mathcal{T} = \{\tau_i : i = 1, \dots, |\mathcal{T}|\}$ be the set of thickness parameters, γ be the scaling factor and τ be the thickness of a pen.

(a) *Ensemble square pen* :

$$U_t^\tau = \frac{1}{\tau + 1} \sum_{i=0}^{\tau} \max\{X_{t-i}, \dots, X_{t+\tau-i}\} + \frac{\tau}{2}\gamma$$

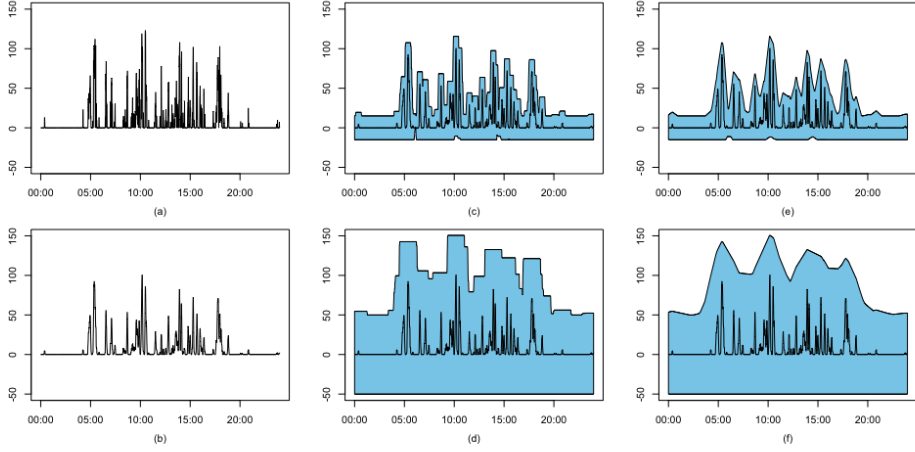


Figure 3.1: (a) Original step data. (b) smoothed data by simple moving average with window size 5, and (c)-(f) smooth the data and apply the TPT: (c) square pen with thickness 30; (d) square pen with thickness 100; (e) ensemble square pen with thickness 30; (f) ensemble square pen with thickness 100

$$L_t^\tau = \frac{1}{\tau + 1} \sum_{i=0}^{\tau} \min\{X_{t-i}, \dots, X_{t+\tau-i}\} - \frac{\tau}{2}\gamma$$

Figure 3.1 shows examples of transformed and smoothed results. In the chapter 4, we use the TPT with the ensemble square pen with thicknesses 30 and 100 to cluster step data and compare the results.

3.0.2 Similarity Measure

To check if the TPMA in the previous chapter can be used as a similarity measure for a time-series clustering problem, we performed a hierarchical clustering using the measure to cluster synthetic data. Figure 3.2 shows six groups of different trends of synthetic data used for the experiment. Each group has five elements with different trends of (a) normal, (b) cyclic, (c) increasing, (d) decreasing, (e) upward shift, and (f) downward shift. Since we aim to differentiate the overall trends, it does not matter if shifts occur at a different timing. Thus, we use the dynamic time warping (DTW) algorithm based on the TPMA similarity measure to find an optimal match between two

data ([2]). Figure 3.4 shows an example of an optimal match between two data in group (e) obtained by the DTW algorithm. Table 3.1 shows the hierarchical clustering result using the TPMA with DTW, together with results using the Euclidean L_2 distance and the Euclidean distance with DTW. The average-linkage criteria is considered here. As we can see from Figure 3.3, using the TPMA as a similarity measure not only correctly identifies all clusters, but also groups Group (a) and (b), (c) and (e), (d) and (f) together when we tend to cluster them into three groups.

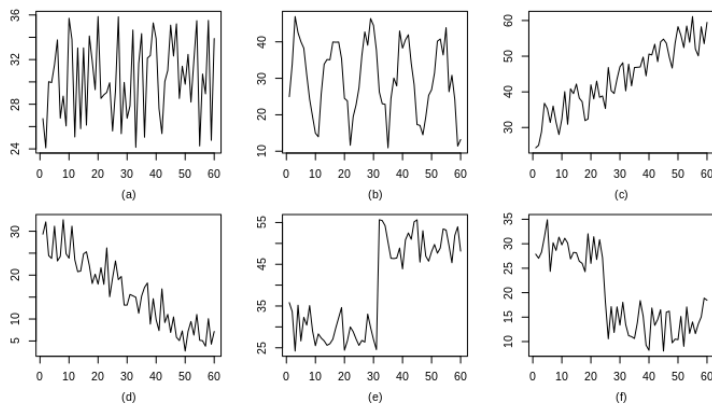


Figure 3.2: Six groups of synthetic data with different trends: (a) normal, (b) cyclic, (c) increasing, (d) decreasing, (e) upward shift, (f) downward shift.

Group	1	2	3	4	5	6
DTW + TPMA	5	5	5	5	5	5
DTW + Euclidean	5	4	1	8	10	2
Euclidean	5	3	1	1	10	10

Table 3.1: Hierarchical clustering results

Dynamic Time Warping has been shown to be powerful and computationally efficient time-series measure widely-used in recent studies ([14]). Therefore, the availability of applying the DTW to the TPMA is indeed a powerful potential for the measure. However, there are computational limitations to proceed clustering large-scale data set using the TPMA as a simi-

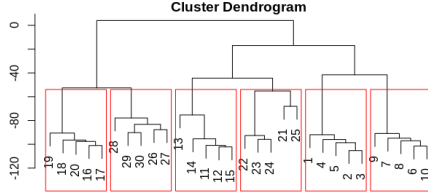


Figure 3.3: Hierarchical clustering dendrogram for the TPMA result

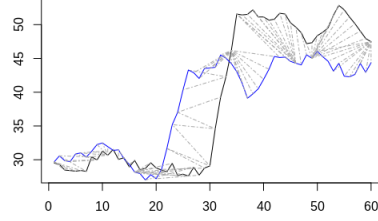


Figure 3.4: Two data in group (e) matched using the TPMA by the DTW algorithm

larity measure, since iterative algorithms such as the k -means algorithm are hardly guaranteed to converge to the local optimum. Instead, the pairwise distance/similarity matrix should be computed to proceed PAM clustering or other search methods. To solve this problem and employ an efficient clustering algorithm, we consider a special form of the TPMA in this thesis.

As we can see from Figure 2.1 and Figure 3.1, lower bounds of step data barely fluctuate. So it is natural to try setting lower bounds to zero when applying TPMA. In essence, measuring a similarity between two data by applying TPMA with zero lower-bounds corresponds to measuring the ratio of the overlapping areas under the upper boundaries obtained by the thick pen transformation at each time t . In Figure 3.5, we visualize an example of two step data with zero lower bounds and their TPMA measure values.

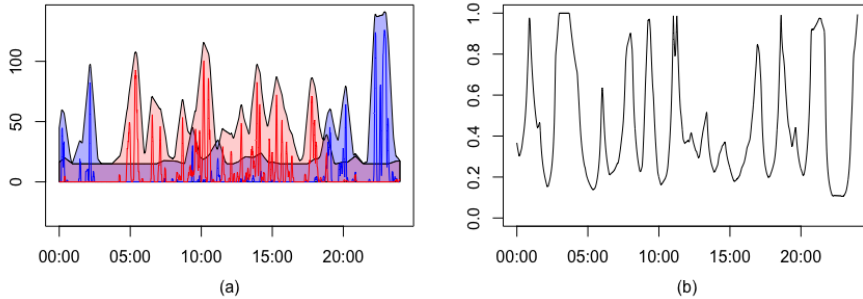


Figure 3.5: (a) Visualization of the overlapping areas between two data, colored by blue and red respectively. (b) $TPMA_0$ values of (a)

We propose to use the TPMA measure after setting the lower bound of the pen to 0, denoted by TPMA_0 , as a new similarity measure for the clustering problem.

$$(\rho_0)_t^\tau(X, Y) = \frac{\min\{U_t^\tau(X), U_t^\tau(Y)\}}{\max\{U_t^\tau(X), U_t^\tau(Y)\}}$$

In practice, first calculate upper bounds of the data, $U = \{U_1, \dots, U_N\}$, where $U_i = (u_i(t))_{t=1}^{1440} = (U_t(X_i))_{t=1}^{1440}$, and rewrite the similarity measure by means of those upper boundaries. We suppress the dependence on thickness τ to simplify the notation.

$$\begin{aligned} \eta(u_i(t), u_j(t)) &= \rho_0(X_i(t), X_j(t)) \\ &= \frac{\min\{u_i(t), u_j(t)\}}{\max\{u_i(t), u_j(t)\}} \end{aligned}$$

When using the measure $\eta(x, y) = \frac{\min(x, y)}{\max(x, y)}$, the center of the two real values is their geometric mean. In other words, $\eta(x, a) = \eta(a, y)$ holds when $a = \sqrt{xy}$. In the next chapter, we use some features of this function to obtain a simple and appropriate clustering algorithm.

3.0.3 Optimization Problem for Clustering

In this section, we view the clustering time-series data as an optimization problem. The goal is to determine K optimal partitions of a set of observations $X = \{X_1, \dots, X_N\}$. Let $P = \{P_1, \dots, P_K\}$ be the set of K partitions of the data which satisfies that $\bigcup_{c=1}^K P_c = X$ and $P_i \cap P_j = \emptyset$ for $i \neq j$. Suppose that each data X_i belongs to a domain set E . Then, find a set of cluster prototypes $M = \{m_1, \dots, m_K : m_c \in E, c = 1, \dots, K\}$. Given a distance function d , we define the clustering problem as minimizing the cost function

$$W(P, M) = \sum_{c=1}^K \sum_{x \in P_c} d(x, m_c).$$

An iterative algorithm proceeds the optimization process in two steps:

Update P : Given a set of cluster prototypes M , update P with

$$P_c = \{x_i : \operatorname{argmin}_{m \in M} d(x_i, m) = m_c, i = 1, \dots, N\} \text{ for each } c \in \{1, \dots, K\}.$$

Update M : Given a partition P , update M with

$$m_c = \operatorname{argmin}_{m \in E} \sum_{x \in P_c} d(x, m) \text{ for each } c \in \{1, \dots, K\}.$$

Note that the cost function decreases for each iteration step. The well-known k -means algorithm deals with L_2 distance, which leads to the mean of each components as a cluster prototype when $E = \mathbb{R}^n, n \in \mathbb{N}$. Also, L_1 distance function uses medians as cluster prototypes, leading to the k -medians algorithm.

Going back to our similarity measure, note that

$$\log \{\eta(u_i(t), u_j(t))\} = \log \frac{\min\{u_i(t), u_j(t)\}}{\max\{u_i(t), u_j(t)\}} = -|\log \frac{u_i(t)}{u_j(t)}|$$

holds for each time t . For $i \in \{1, \dots, N\}$ and given partition P , let $c_i = c$ such that $c \in \{1, \dots, K\}$ and $X_i \in P_c$. Assume that $\mu = \{\mu_c : 1 \leq c \leq K\}$ be a set of cluster representatives. Then, according to the following analogue, maximizing the product of TPMA₀'s for each time t and element i is equivalent to minimizing the sum of L_1 distance with respect to the logarithms of upper boundaries.

$$\begin{aligned} & \underset{P, \mu}{\text{maximize}} \prod_{t=1}^T \prod_{i=1}^N \eta^\tau(u_i(t), \mu_{(c_i)}(t)) \\ \iff & \underset{P, \mu}{\text{maximize}} \sum_{t=1}^T \sum_{i=1}^N \log \{\eta^\tau(u_i(t), \mu_{(c_i)}(t))\} \end{aligned}$$

$$\iff \underset{P, \mu}{\text{minimize}} \sum_{t=1}^T \sum_{i=1}^N |\log u_i(t) - \log \mu_{(c_i)}(t)|$$

In other words, we define the cost function to be minimized given a partition as follows.

$$W(P, \mu) = \sum_{t=1}^T \sum_{i=1}^N |\log u_i(t) - \log \mu_{(c_i)}(t)|$$

Since it is the L_1 optimization problem with respect to the logarithms of upper bounds, applying the k -medians algorithm to $\{LU_i : LU_i = (\log u_i(t)), 1 \leq i \leq N\}$ guarantees monotone decrease in the cost function. The algorithm depends on the initialization, thus we repeat k -medians algorithm several times and get the final cluster with the minimal cost function as the final cluster.

We have a remark on this idea of log transformation of upper boundaries. $0 < \eta(u_i, u_j) \leq 1$ holds with $\tau > 0$, so it is problematic when $\eta(u_i, u_j)$ goes near zero and the logarithm diverges toward the negative infinity. This is why the thick pen transformation matters in our setting. Under the zero-bounded setting, data are non-negative and a thickness of a pen guarantees the minimal value of upper boundaries sufficiently bigger than zero. Thus, we can assume that there exists δ such that $\eta(u_i, u_j) = \frac{\min\{u_i, u_j\}}{\max\{u_i, u_j\}} > \delta > 0$ for any $i, j \in \{1, \dots, N\}$ as long as the upper boundaries of the transformed data are bounded above. For instance, our step data with thickness 30 satisfies $-3.5 \leq \log \eta(u_i, u_j) \leq 0$.

3.0.4 Clustering Algorithm

Now we present the whole clustering scheme with the clustering objective of maximizing products of TPMA_0 's between each observation and its cluster prototype.

Step 1: Smooth and Transform the data via moving average and the thick pen transformation to obtain the upper boundaries $\{u_1, \dots, u_N\}$.

Step 2: Randomly initialize the cluster.

Step 3: For each cluster, obtain the cluster prototype as

$$m_c := \log \mu_c = \text{med}\{\log u_1^{(c)}, \dots, \log u_{n_c}^{(c)}\}, c \in \{1, 2, \dots, K\}$$

Step 4: Assign every curve to the cluster with the minimal L_1 distance between the logarithm of the upper bounds of the curve and cluster prototypes.

Step 5: Iterate Step 3 - Step 4 until no more curves are regrouped.

Step 6: Repeat Step 2 - Step 5 for sufficiently many times and get the final cluster with the minimum cost function.

To perform the analysis, observe the given data and choose a proper smoothing window size, a thickness τ , and a shape of a pen appropriate for the data and clustering objective. For instance, applying a thicker pen tends to see the data at a distance, focusing on the big trends, while narrower thickness values tend to catch the pattern sensitively. To determine the number of clusters K , we use the gap statistics of Tibshirani *et al.* ([15]).

Chapter 4

Real data Analysis

We perform clustering of step data after smoothing and transforming the data with the ensemble square pen. For this, we consider two different parameter settings as follows.

- A. Smoothing window size 3, pen thickness $\tau = 30$, and scaling parameter $\gamma = 0.2$.
- B. Window size 5, pen thickness $\tau = 100$, and scaling parameter $\gamma = 0.2$.

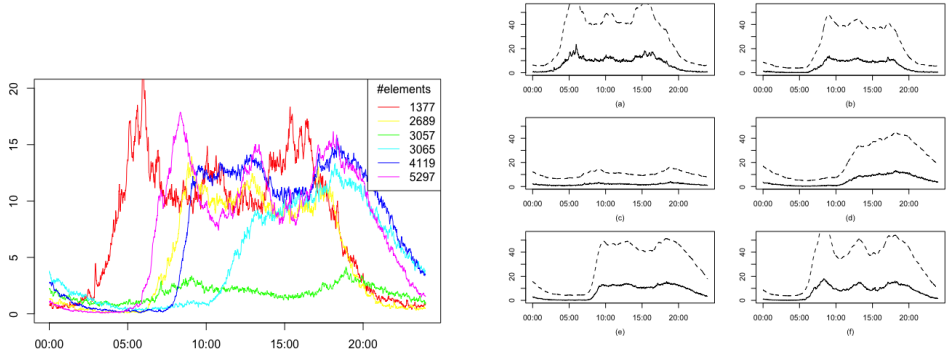
The main difference between the settings is in τ values. As we can see from Figure 3.1, with a large pen thickness, the pen boundary values tend to less fluctuate than smaller thicknesses, keeping boundaries large when there exists a movement in the nearby time. In other words, in the case with large τ values, some changes in the amount of walking do not affect the upper boundaries much, and the boundaries instead mainly reflect the existence of the movement.

Figure 4.1 and 4.2 plot the mean curves of the original step data and the pen means for each resulted cluster respectively using the setting A and B. We chose $K = 6$ for the both cases. Note that since each cluster consists of a large number of step data where zero steps occur repeatedly, the mean curves might tend to regress downward than the original trend that each group represents.

Cluster ID			1	2	3	4	5	6
Number of Days	thickness	30	1377	2689	3057	3065	4119	5297
		100	1649	3315	2298	3074	4225	5043
Mean Step Count	thickness	30	11600	7433	2683	7303	10665	10925
		100	11112	5904	1833	7392	12444	10101
Weekend (%)	thickness	30	9.9	29.7	38.9	47.1	33.3	9.1
		100	11.3	44.3	39.8	48.9	23.5	7.1

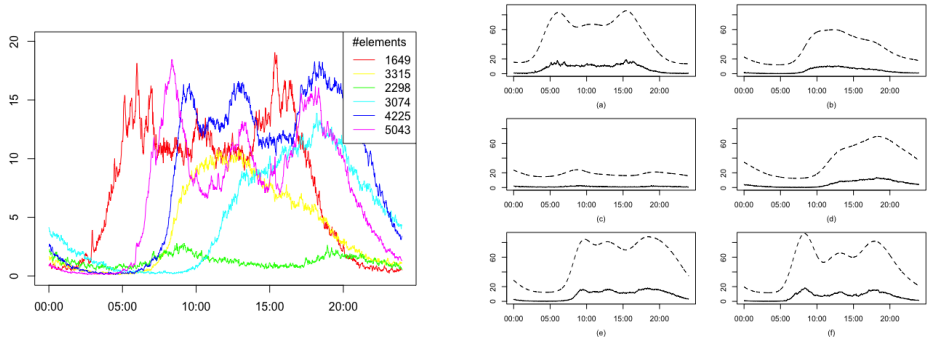
Table 4.1: Summary of clustering results

Table 4.1 has information about the resulted cluster size the mean step counts, and the percentage of weekend days. Figure 4.3 further maps the distribution of individuals included in each group. In both cases, group 1 (red) represents days with early wake-up, early sleep, and a lot of walks, where only a few people, who may be early birds, are included according to Figure 4.3. Group 2 (yellow) represents days with late rising and less walks, which relatively differs in the shape a lot between the $\tau = 30$ and $\tau = 100$ results. Group 3 (green) consists of the laziest days with the smallest total steps while group 4 (sky-blue) keeps late hours. Both groups have a large weekend proportion, where group 4 shows the largest percentage of weekend days among six groups. Finally, group 5 (blue) and 6 (pink) both show large amount of mean step counts, with different average wake-up times. The distribution of individuals between group 5 and 6 is quite different, which might represents two groups of people sharing different office hours or morning routines.



(a) Mean curves of step data for each cluster (b) Mean curves of step count data (—) and the pen means (---) in each group

Figure 4.1: Clustering results by using the TPT with $\tau = 30$



(a) Mean curves of step data for each cluster (b) Mean curves of step count data (—) and the pen means (---) in each group

Figure 4.2: Clustering results by using the TPT with $\tau = 100$

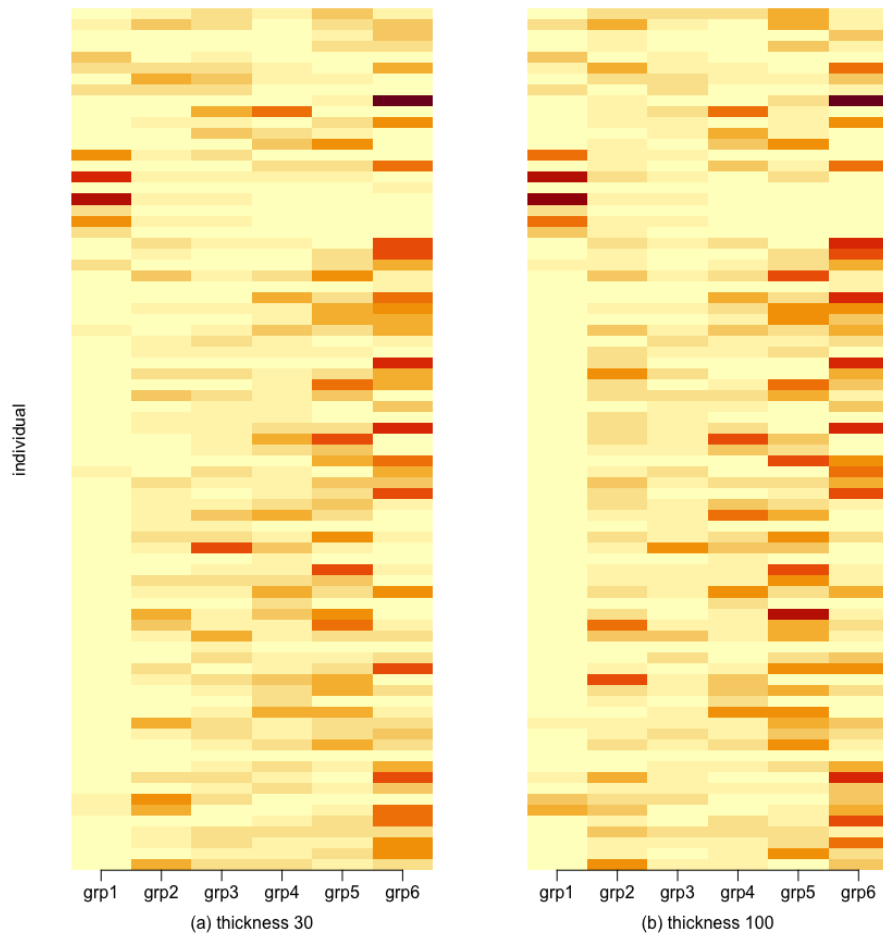


Figure 4.3: Map of the distribution of individuals included in each group

Chapter 5

Simulation study

In this chapter, we generate synthetic data to compare the proposed clustering scheme with existing methods. The list of simulation sets used are as follows. Figure 5.1, 5.2, and 5.3 represent each data from different groups of respective simulation set.

- *Sinusoidal data with different variability*: we generate random sinusoidal curves with several variances defined as $Y(t) = |\sin(\frac{5(t-1)}{T}) + \epsilon(t)|$, $t \in \{1, \dots, T\}$, where $T = 1024$ and $\epsilon \sim N(0, \sigma^2)$ with $\sigma \in \{0.01, 0.1, 0.7, 1\}$. We consider clustering 200 random curves into 4 groups, 50 samples generated for each σ value.
- *Block data with different patterns*: we generate random block curves with different patterns defined as $Y(t) = |\sum_{j=1}^5 h_j \{1 + \text{sgn}((t-1)/T - \xi_j)\}/2 + \epsilon(t)|$, $t \in \{1, \dots, T\}$, where $T = 512$ and $\epsilon \sim N(0, 3^2)$. h_j satisfies $|h_j| \sim U(0, 20)$, $h_1, h_3 < 0$, $h_2, h_4 > 0$, and $\sum_{j=1}^5 h_j = 0$, whose values are related to the height of each vertical jump. ξ_j is a randomly real number chosen from a specified interval, which determines where the jump occurs. We set four different intervals as $(0, \frac{2}{5})$, $(\frac{1}{5}, \frac{3}{5})$, $(\frac{2}{5}, \frac{4}{5})$ and $(\frac{3}{5}, 1)$, where possible jumps can occur. Here we generate 50 samples for each interval and cluster the samples into 4 groups.

- *Block data with different amount and patterns:* we generate three different groups of random block curves with different amount and patterns. Data are generated in a similar way like above. Group 1 data have three jumps in $(0, \frac{1}{5})$ with $|h_j| \sim U(0, 30)$ and five jumps jumps in $(\frac{2}{5}, \frac{3}{5})$ with $|h_j| \sim U(0, 20)$, with errors having $\sigma = 5$. Group 2 data have five jumps in $(0, \frac{2}{5})$ with $|h_j| \sim U(0, 10)$ and three jumps jumps in $(\frac{2}{5}, \frac{4}{5})$ with $|h_j| \sim U(0, 5)$, with errors having $\sigma = 3$. Finally, Group 3 data have four jumps in $(\frac{1}{5}, \frac{2}{5})$ with $|h_j| \sim U(0, 20)$, three jumps in $(\frac{2}{5}, \frac{4}{5})$ with $|h_j| \sim U(0, 15)$ and three jumps jumps in $(\frac{4}{5}, 1)$ with $|h_j| \sim U(0, 20)$, with errors having $\sigma = 5$. 50 samples are generated per each group and groups show different amount and patterns of jumps.

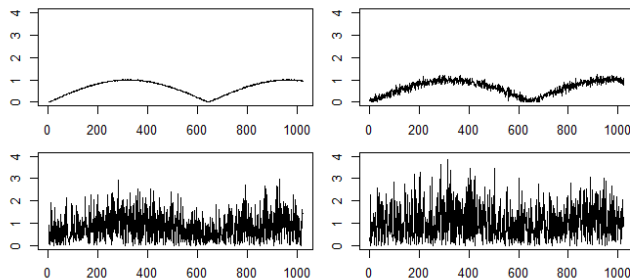


Figure 5.1: Four groups of sinusoidal data with different variabilities.

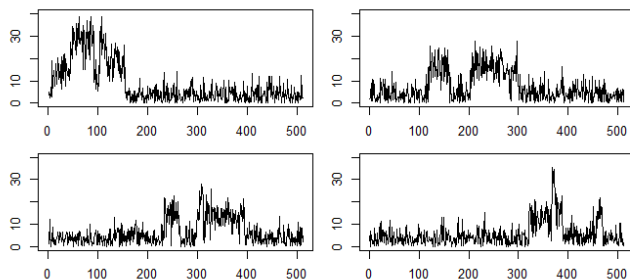


Figure 5.2: Four groups of block data with different patterns.

We compare the proposed method with different optimization settings and

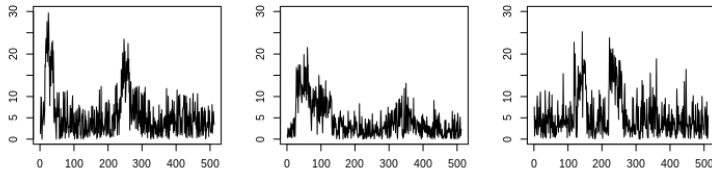


Figure 5.3: Three groups of block data with different amount and patterns.

some functional clustering methods. Figure 5.4 shows different optimization schemes considered in this study. Our method using TPMA_0 as a similarity measure is depicted on the top, where we apply logarithms to the thick pen upper boundaries and then apply the k -medians algorithm. In the other settings, we use k -medians or k -means algorithm respectively for L_1 or L_2 optimization. For these simulation sets, we smooth the original data with window size 3 and then apply the ensemble square pen with thickness 30 to get the transformed data. For the five settings depicted in Figure 5.4, we repeat algorithms $N = 20$ times and choose the cluster result with the minimum cost. For functional clustering methods, we use (a) funFEM: functional clustering using discriminative functional mixture model by Bouveron, Come and Jacques (2014, [9]), and (b) funHDDC: clustering functional data based on modeling each group within a functional subspace by Bouveyron and Jacques ([16]).

The simulation results are in the Table 5.1 based on the correct classification rate (CCR) criteria defined as

$$\text{CCR} = \frac{\text{the number of correctly classified curves}}{\text{total number of curves}}.$$

The average CCR rates and their standard deviations after simulated 100 times using the seven different methods are listed in Table 5.1. We observe several remarks: (a) Methods in Figure 5.4 have standard deviations close to zero. This means that repeating clustering algorithm $N = 20$ times within those methods yields stable outcomes. (b) Overall, the proposed method, TPMA_0

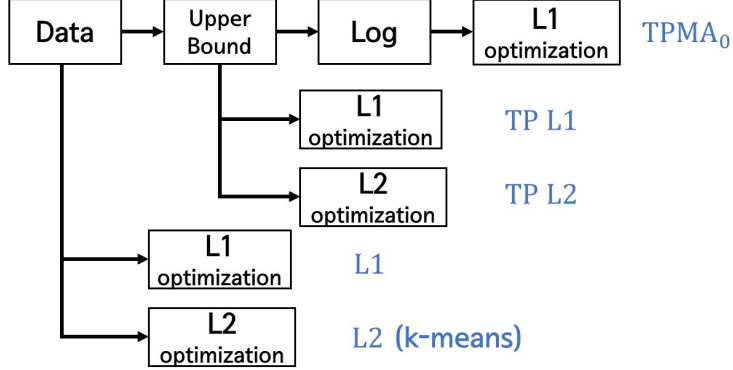


Figure 5.4: Different optimization settings used for the comparison.

Signal	Results for the following methods:						
	TPMA ₀	TP L1	TP L2	L1	L2	funFEM	funHDDC
Sinusoidal	1.00 (0)	1.00 (0)	0.68 (0.09)	0.64 (0.09)	0.71 (0.05)	0.92 (0.14)	0.76 (0.16)
Block (pattern)	0.94 (0)	0.92 (0)	0.88 (0.01)	0.77 (0.02)	0.78 (0.01)	0.83 (0.11)	0.83 (0.08)
Block (pattern and amount)	1.00 (0)	0.99 (0)	0.97 (0)	0.74 (0.01)	0.80 (0.01)	0.91 (0)	0.91 (0.03)

Table 5.1: Means (standard deviations) of the correct classification rate (CCR) for each method

outperforms other methods in our simulated data, suggesting that the measure might be generally applied to cluster non-negative count data. (c) However, applying L_1 optimization to upper bounds worked as well as the proposed method, which implies that taking logarithms to the upper boundaries is not a critical choice for the performance. We might skip that step when it is not appropriate to apply log transform to the data.

Chapter 6

Conclusions

As shown by the hierarchical clustering example of the synthetic data, the TPMA of Fryzlewicz and Oh is a good time-series similarity measure, which can be further applied to clustering problems. However, measuring the similarity of all pairs to proceed PAM clustering is computationally disadvantageous. In this study, we have proposed a simple and effective algorithm applicable to the new similarity measure, TPMA_0 , which is a special form of the TPMA. We examine that the proposed method can be applied in general for time series data distributed on the same side along the axis, whose similarities are measurable in the form of a proportion of overlapping areas. Indeed, we have shown that measuring time-series similarity using the TPMA_0 and optimizing their total products is equivalent to the Log L_1 optimization of the transformed upper boundaries.

The proposed measure has its strength in the use of the novel thick pen transformation, which is visually inspiring multi-scale method, representing time-series dependence structure. Moreover, since the measure is computed coordinate-wise, we can also employ the dynamic time warping algorithm, one of the most widely-used and effective time-series matching algorithm. Indeed, these properties endow the measure a great potential for various applications. For the further study, since the TPMA can be summarized in various ways and

different choices of parameters yield different similarity distribution, comprehensive approaches such as ensemble clustering or fuzzy clustering can be further applied for the measure. In general, we expect that the TPMA, together with its special form called TPMA_0 , might become an overarching similarity measure assessing the time-series association with its intuitive structure and great flexibility.

Bibliography

- [1] R. Agrawal, C. Faloutsos, and A. Swami, “Efficient similarity search in sequence databases,” vol. 730, pp. 69–84, 01 1993.
- [2] D. Lemire, “Faster retrieval with a two-pass dynamic-time-warping lower bound,” *Pattern Recognition*, vol. 42, no. 9, pp. 2169 – 2180, 2009.
- [3] J. Rodgers and A. Nicewander, “Thirteen ways to look at the correlation coefficient,” *American Statistician - AMER STATIST*, vol. 42, pp. 59–66, 02 1988.
- [4] J. Jacques and C. Preda, “Functional data clustering: A survey,” *Advances in Data Analysis and Classification*, vol. 8, pp. 231–255, 09 2013.
- [5] C. Abraham, P. Cornillon, E. Matzner-Løber, and N. Molinari, “Unsupervised curve clustering using b-splines,” *Scandinavian Journal of Statistics*, vol. 30, pp. 581 – 595, 09 2003.
- [6] J.-M. Chiou and P.-L. Li, “Functional clustering and identifying substructures of longitudinal data,” *Journal of the Royal Statistical Society Series B*, vol. 69, pp. 679–699, 09 2007.
- [7] N. Serban and L. Wasserman, “Cats: Clustering after transformation and smoothing,” *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 990–999, 2005.

- [8] C. Möller-Levet, F. Klawonn, K.-H. Cho, and O. Wolkenhauer, “Fuzzy clustering of short time-series and unevenly distributed sampling points,” vol. 2810, pp. 330–340, 08 2003.
- [9] C. Bouveyron, E. Côme, and J. Jacques, “The discriminative functional mixture model for a comparative analysis of bike sharing systems,” *Ann. Appl. Stat.*, vol. 9, pp. 1726–1760, 12 2015.
- [10] A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi, “Clustering functional data using wavelet,” *International Journal of Wavelets Multiresolution and Information Processing*, vol. 11, p. 1350003, 01 2013.
- [11] Y. Lim, H. Oh, and Y. Cheung, “Functional clustering of accelerometer data via transformed input variables,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 68, 09 2018.
- [12] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, “Time-series clustering – a decade review,” *Information Systems*, vol. 53, pp. 16 – 38, 2015.
- [13] P. Fryzlewicz and H.-S. Oh, “Thick pen transformation for time series,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 4, pp. 499–529, 2011.
- [14] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, M. B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, “Searching and mining trillions of time series subsequences under dynamic time warping,” vol. 2012, 08 2012.
- [15] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *Journal of the Royal Statistical Society Series B*, vol. 63, pp. 411–423, 02 2001.
- [16] A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze, and P. Martin, “Clustering multivariate functional data in group-specific functional subspaces,” 07 2018.

국문초록

본 학위 논문에서는 시계열 자료의 군집화 방법을 탐색하며 새로운 유사도 척도와 최적화 알고리즘을 제시한다. 보다 구체적으로 Fryzlewicz and Oh 가 2011년 제시한 굵은 펜 변환법에 기반하여 새로운 시계열 자료 군집화 방식을 제안한다. 이 변환법의 기본 발상은 두께가 있는 굵은 펜을 이용하여 자료를 따라 그리는 것이다. 우리 연구의 주된 기여는 변환과 평활화 작업을 진행한 후 얻어지는 두꺼운 선들 끼리의 겹침 혹은 격차를 기반으로 두 시계열 자료의 유사성을 정의하는 것에 있다. 굵은 펜 변환의 강점은 이것이 다중 척도의 성질을 띄는 시각화 기법이고, 인접한 값들의 추세를 반영하여 변환이 정의될 수 있다는 것에 있다. 따라서 해당 변환을 이용해 연관성을 측정하는 우리의 방식은 앞서 언급한 강점들을 적극 활용하게 된다. 나아가 우리는 제안한 유사도 척도를 이용하여 군집화를 최적화 문제로 정의하고, 이를 해결하는 효율적인 반복 알고리즘을 제시한다. 본 연구는 기기를 통해 측정된 대용량의 걸음 수 자료를 군집화하여 비슷한 걸음 양상끼리 분류하는 것을 목표로 시작되었다. 나아가 합성 자료에 대해 군집화를 진행하며 기존의 다른 방법들과 성능을 비교하고, 우리의 방법이 다른 자료에도 일반적으로 적용 가능하다는 것을 확인한다.

주요어: 시계열 자료, 군집화, 다중 척도 방법, 동적 시간 워핑, 굵은 펜 변환
학 번: 2019-28751