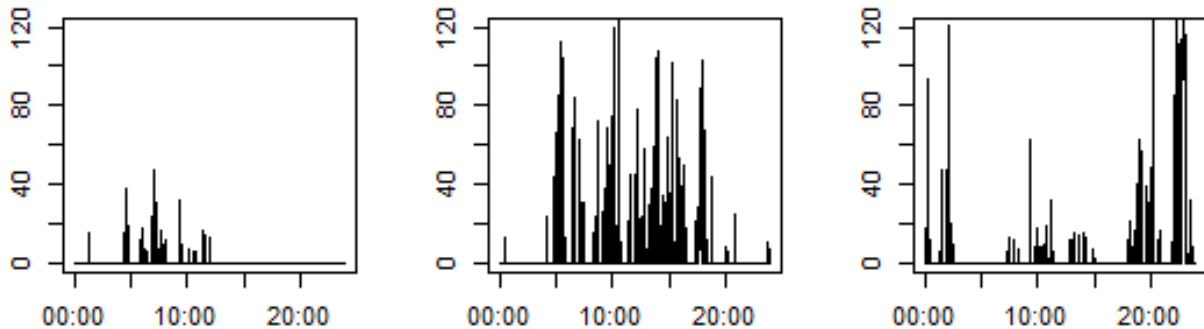# Clustering Step Data via Thick pen Transformation

Minji Kim

UNC Chapel Hill STOR

March 28, 2022
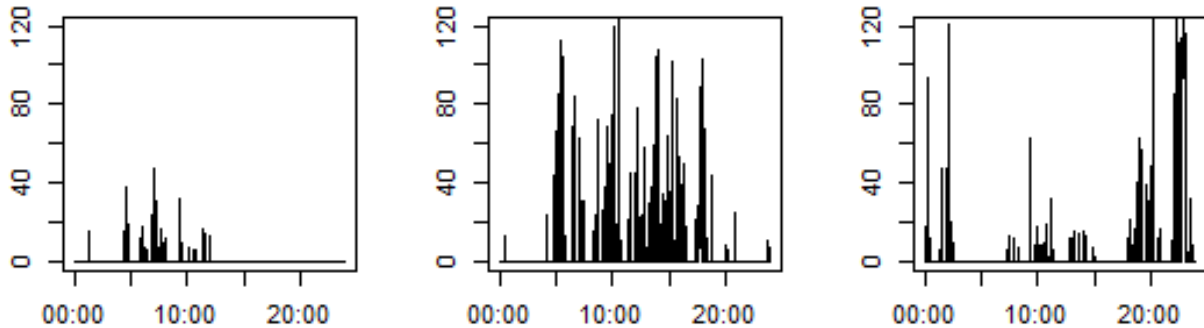
Data Observation

Figure: Three different step count data

- ▶ The motivation of this study is to cluster a large set of step count data measured every minute from a wearable device.
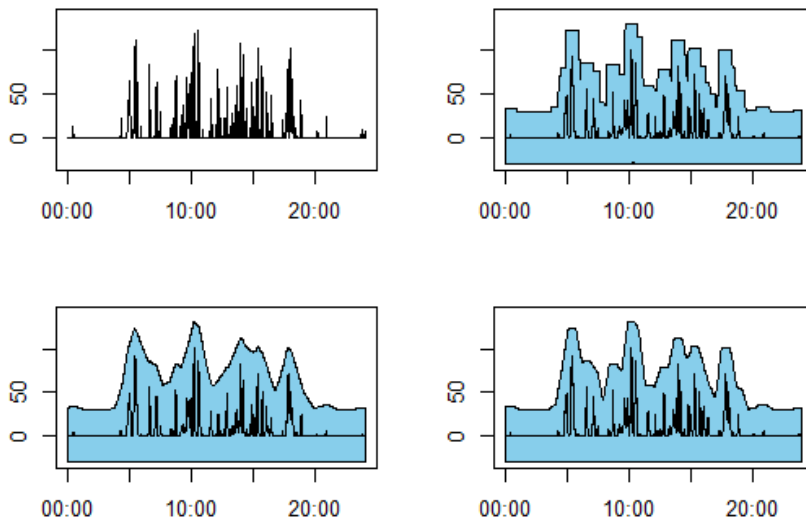
Data Observation

Figure: Three different step count data

- ▶ High-Dimensional and discrete
- ▶ Zero-inflated with numerous moments people taking a break between each step.
- ▶ 19604 days over 79 people item

Motivation

Figure: Thick Pen Transform with various thickness and shape

- ▶ Multi-scale Visualization of Time Series data
- ▶ Draw along the points with a pen with its own shape and thickness

Motivation

Definition (Ensemble Square pen)

Let $\mathcal{T} = \{\tau_i : i = 1, \ldots, |\mathcal{T}|\}$ be the set of thickness parameters. For each $\tau_i \in \mathcal{T}$, scaling factor γ, let $U_t^{\mathcal{T}_i}$ and $L_t^{\mathcal{T}_i}$ be the upper and lower boundary of the area that is covered by a pen of thickness τ_i while connecting the points $(t, X_i)_{t=1}^n$.

▶ *Ensemble Square pen* :

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

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

Thick Pen Measure of Association

Definition (Thick Pen Measure of Association/TPMA)

To measure the overlap between the areas that are created by the thick pen transforms of X and Y, define TPMA between X and Y at time t, thickness τ as

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

This measure describes how the two time series appear to *covary* when seen from the distance corresponding to thickness τ. A natural averaged version of $\rho_t^\tau(X, Y)$ is
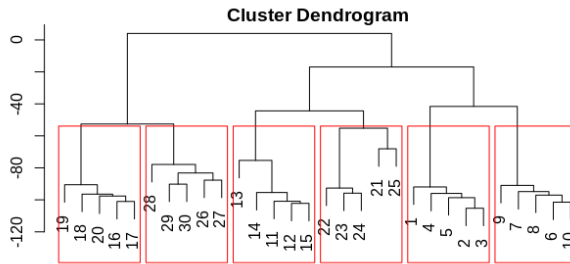
$$\bar{\rho}_{1,n}^\tau = \frac{1}{n} \sum_{t=1}^{n} \rho_t^\tau(X, Y)$$
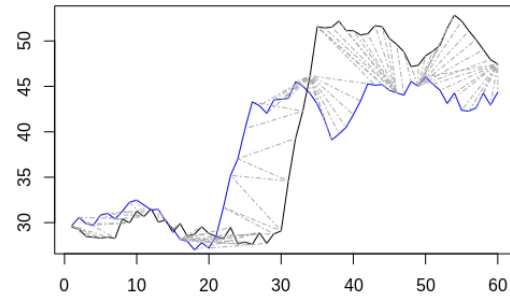
Similarity Measure
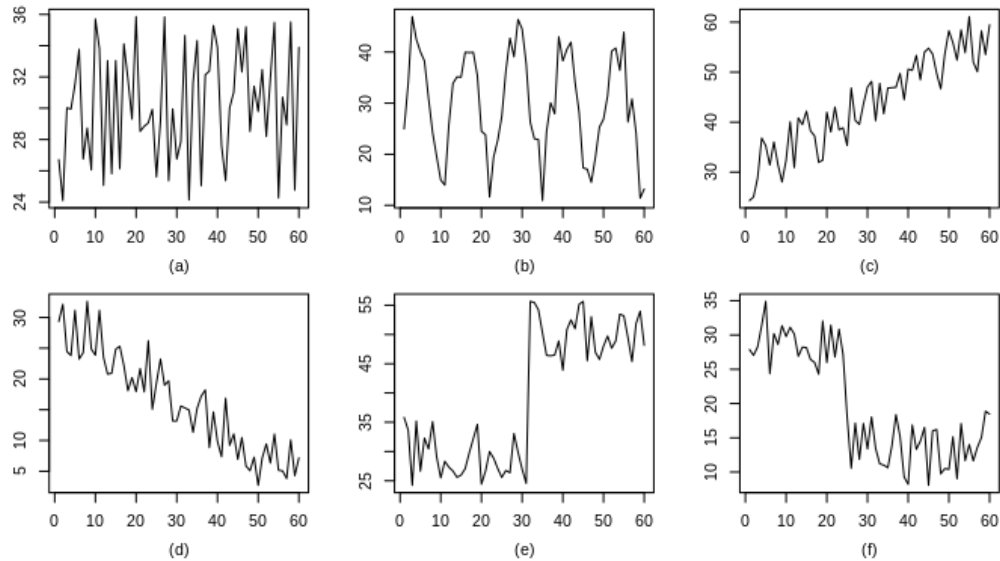
Figure: Six groups of synthetic data with different trends: (a) normal, (b) cyclic, (c) increasing, (d) decreasing, (e) upward shift, (f) downward shift.

Similarity Measure

Figure: Hierarchical clustering dendrogram for the TPMA result

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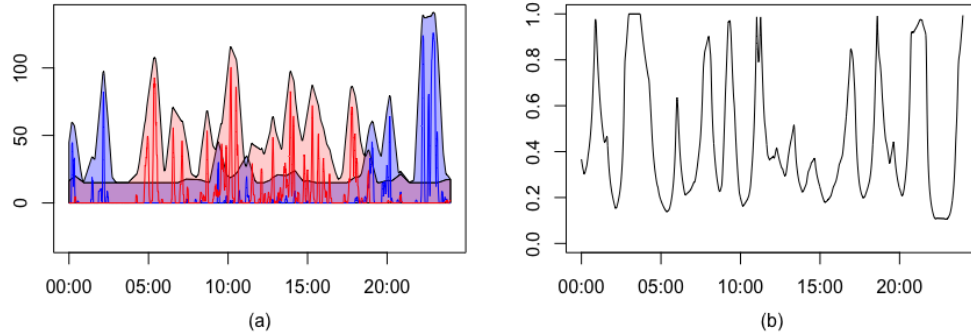
Similarity Measure

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: Hierarchical clustering results

▶ 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.

Similarity Measure

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

▶ Setting the lower bound of the pen to 0, we propose TPMA_0 as a new similarity measure to deal with large-scale data clustering.

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

Clustering 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.

$$\underset{P,\ \mu}{\text{maximize}} \prod_{t=1}^{T} \prod_{i=1}^{N} \eta^\tau (u_i(t), \mu_{(c_i)}(t))$$

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

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

▶ 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.

Clustering Algorithm

With partition P and cluster prototypes M, we view clustering as an optimization problem minimizing the following cost function,

$$W(P, M) = \sum_{c=1}^{K} \sum_{x \in P_c} d(x, m_c) = \sum_{t=1}^{T} \sum_{i=1}^{N} |\log u_i(t) - \log \mu_{(c_i)}(t))|.$$

An iterative algorithm proceeds in two steps:

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

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

Update M: Given a partition P, update M with

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

Clustering Algorithm

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

Step 2: Randomly initialize the cluster.

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

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

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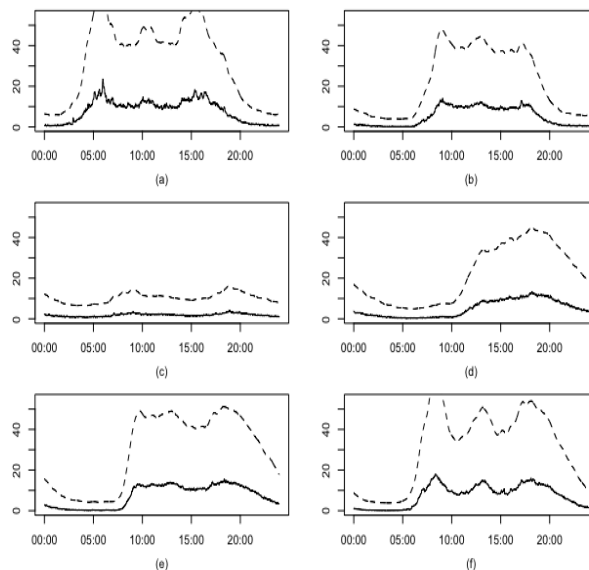
Step 5: Iterate Step 3 - Step 4 until no more curves are regrouped.

Step 6: Repeat Step 2 - Step 5 for sufficiently many times and get the final cluster with the minimum cost function.
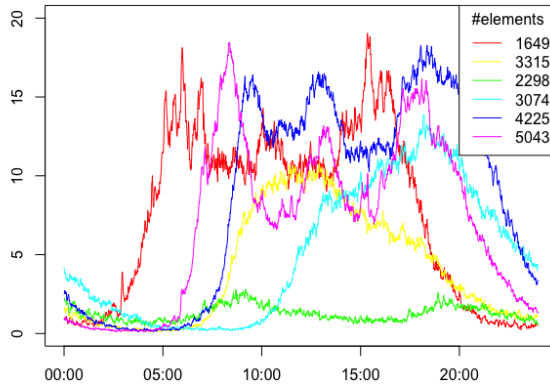
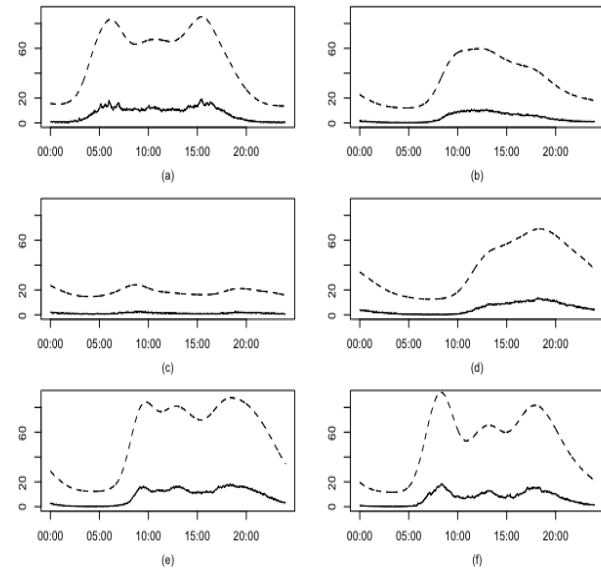# Real Data Analysis
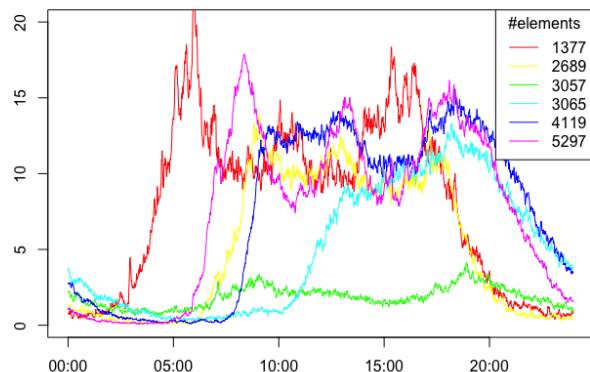
(a) Mean curves of step data for each cluster

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

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

Real Data Analysis

(a) Mean curves of step data for each cluster

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

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

Real Data Analysis

▶ 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 10.

▶ 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.

▶ 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 represent two groups of people sharing different office hours or morning routines.
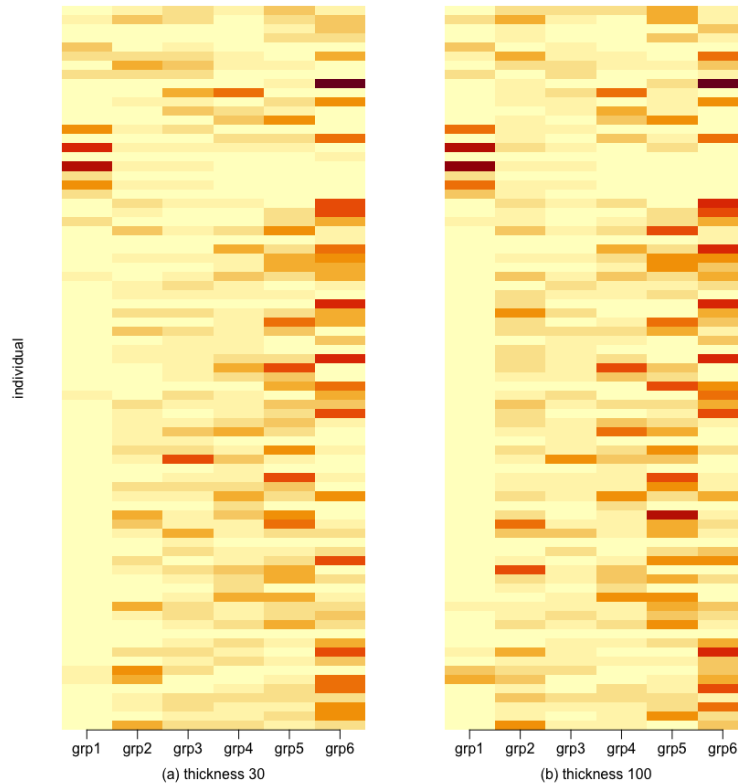
Real Data Analysis

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: Summary of clustering results

Real Data Analysis


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

Simulation Study

▶ We generate three different types of synthetic data to compare the proposed clustering scheme with existing methods: *Sinusoidal data with different variability, Block data with different patterns, and Block data with different amount and patterns*.

▶ Different optimization schemes are considered based on $k-$medians or $k-$means algorithm respectively for L_1 or L_2 optimization.

▶ For the five iterative optimization settings, 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, [1]), and (b) funHDDC: clustering functional data based on modeling each group within a functional subspace by Bouveyron and Jacques ([2]).
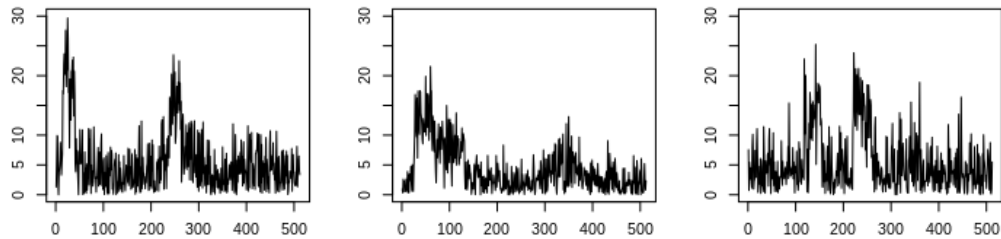
Simulation Study
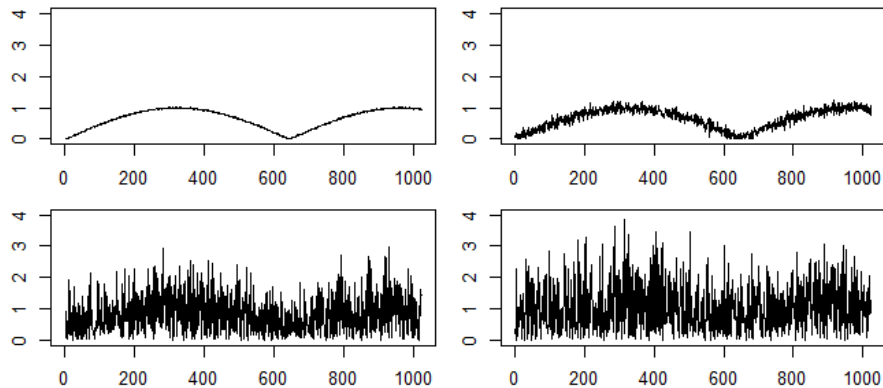
Figure: Four groups of sinusoidal data with different variabilities.
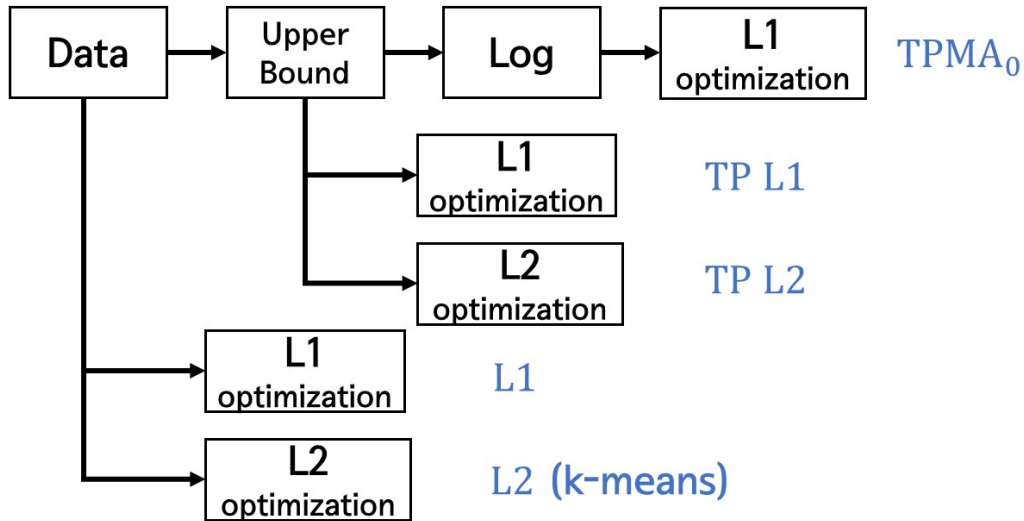
Simulation Study
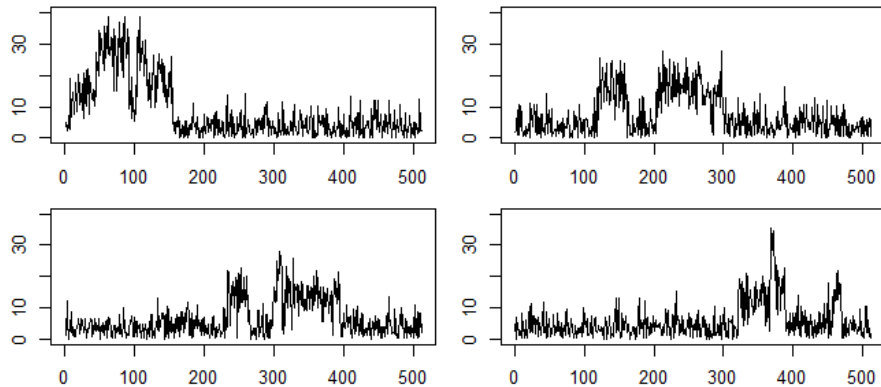
Figure: Four groups of block data with different patterns.

Simulation Study

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

Simulation Study

Figure: Different optimization settings used for the comparison.

Simulation Study

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

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

▶ Overall, the proposed method, TPMA_0 outperforms other methods in our simulated data, suggesting that the measure might be generally applied to cluster non-negative count data.

▶ For the five iterative optimization settings, we repeat algorithms $N = 20$ times and choose the cluster result with the minimum cost.

▶ 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.

Simulation Study

Signal	Results for the following methods:						
	$TPMA_0$	TP L1	TP L2	L1	L2	funFEM	funHDDC
Sinusoidal	**1.00**	**1.00**	0.68	0.64	0.71	0.92	0.76
	(0)	**(0)**	(0.09)	(0.09)	(0.05)	(0.14)	(0.16)
Block	**0.94**	0.92	0.88	0.77	0.78	0.83	0.83
(pattern)	**(0)**	(0)	(0.01)	(0.02)	(0.01)	(0.11)	(0.08)
Block	**1.00**	0.99	0.97	0.74	0.80	0.91	0.91
(pattern and amount)	**(0)**	(0)	(0)	(0.01)	(0.01)	(0)	(0.03)

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

Conclusion

- ▶ TPMA 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.
- ▶ To overcome the computation issue, 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.

References I

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.

A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze, and P. Martin, "Clustering multivariate functional data in group-specific functional subspaces," 07 2018.

S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – a decade review," *Information Systems*, vol. 53, pp. 16 – 38, 2015.

T. Furtuna, "Dynamic programming algorithms in speech recognition," *Informatica Economica Journal*, vol. XII, 01 2008.

References II

📄 J. Jacques and C. Preda, "Functional data clustering: A survey," *Advances in Data Analysis and Classification*, vol. 8, pp. 231–255, 09 2013.

📄 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.

📄 D. Lemire, "Faster retrieval with a two-pass dynamic-time-warping lower bound," *Pattern Recognition*, vol. 42, no. 9, pp. 2169 – 2180, 2009.

📄 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.

Thank you