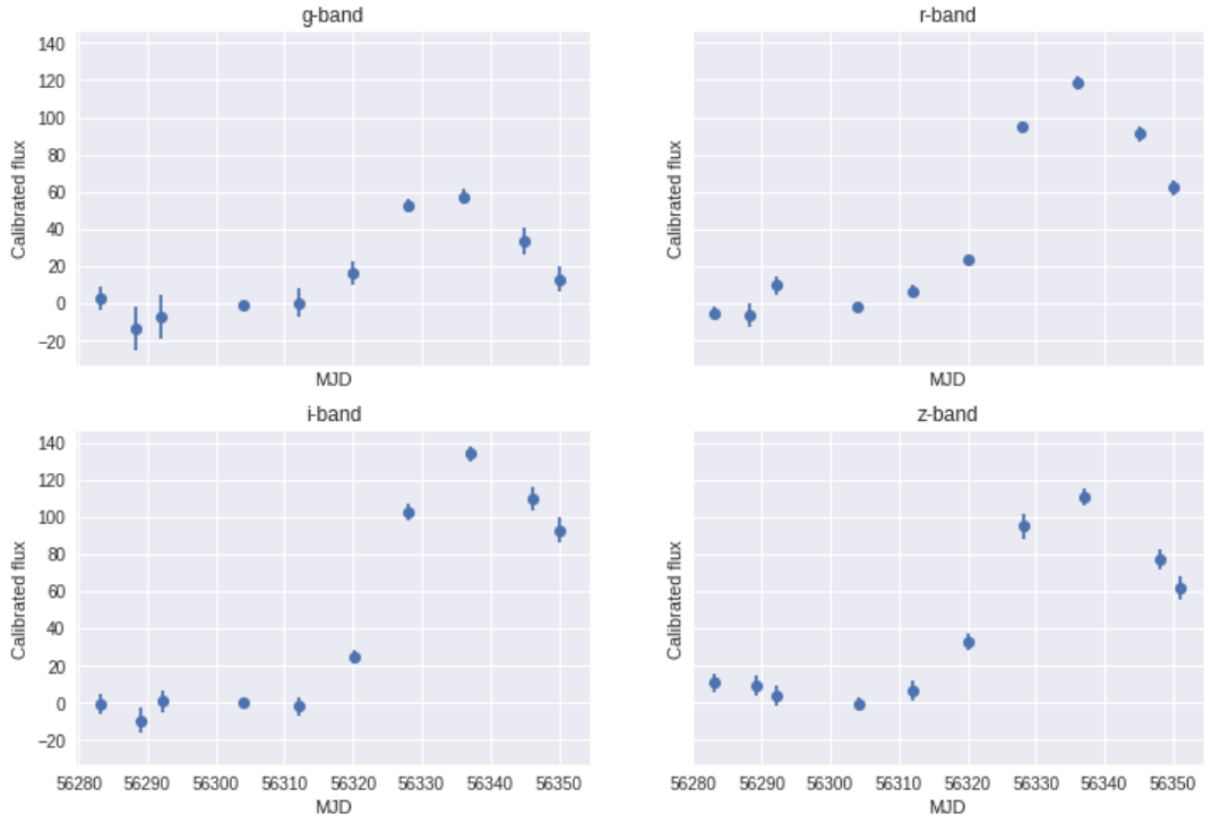
Clustering Time Series using Unsupervised-Shapelets

JESIN ZAKARIA ABDULLAH MUEEN EAMONN KEOGH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF CALIFORNIA, RIVERSIDE

Motivations



Motivations

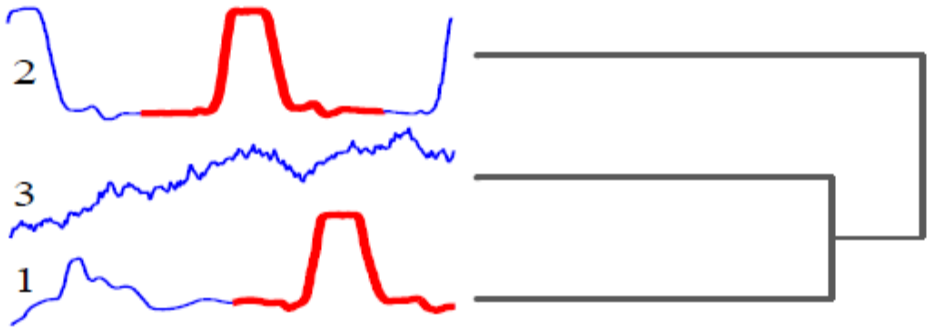


Figure 1. A Euclidean distance clustering of two exemplars from the "raw" Gun-Point dataset, together with a random walk sequence. The hundreds of papers that have used the Gun-Point dataset have only considered the human edited version, corresponding to *just* the red/bold data.

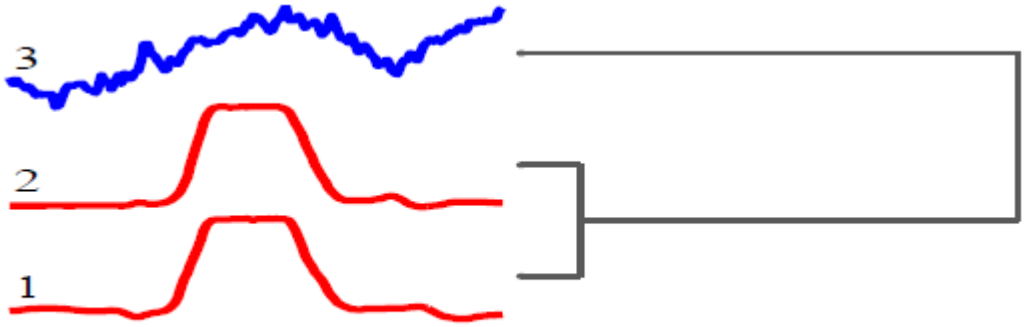


Figure 2. Clustering Gun-Point after ignoring some data.

Problem

How can we discover shapelets from a dataset without having any knowledge of the class labels?

Definitions and Background

Definition 1: *Time Series*, a time series $T = T1, T2, \ldots, Tn$ is an ordered set of real values. The total number of real values is equal to the length of the time series. A dataset $\boldsymbol{D} = \{T1, T2, \ldots TN\}$ is a collection of N such time series.

Definition 2: *Subsequence*, a subsequence Si,l, where $1 \leq l \leq n$ and $1 \leq i \leq n$, is a set of l continuous real values from a time series, T, that starts at position i.

Definitions and Background

Definition 3: The *Subsequence distance* between a subsequence *S* of length *m* and a time series *T* of length *n* is the distance between *S* and the subsequence of *T* that has minimum distance. We denote it as *sdist*(*S*, *T*).

Definition 4: An *unsupervised-shapelet S'* is a subsequence of a time series *T* for which the *sdists* between S' and the time series from a group *DA* are much smaller than the *sdists* between and rest of the time series *DB* in the dataset ***D*** .

Definitions and Background

Definition 5: A *Distance map* contains the *sdists* between each of the u-shapelets and all the time series in the dataset. If we have m u-shapelets for a dataset of N time series, the size of the distance map is $[N \times m]$ where each column is a distance vector of a u-shapelet.

A Discrete Analogue of U-Shapelet

San *Jose; Earth* ***Day****;* ***San*** *Francisco; Memorial* ***Day****; Fink Nottle; Labor* ***Day****; Bingo Little.*

	San Jose	***Earth Day***	***San Francisco***	***Memorial Day***	***Fink Nottle***	***Labor Day***	***Bingo Little***
San	0	2	0	2	2	2	2
Day	2	0	2	0	3	0	3

Algorithm - intuition

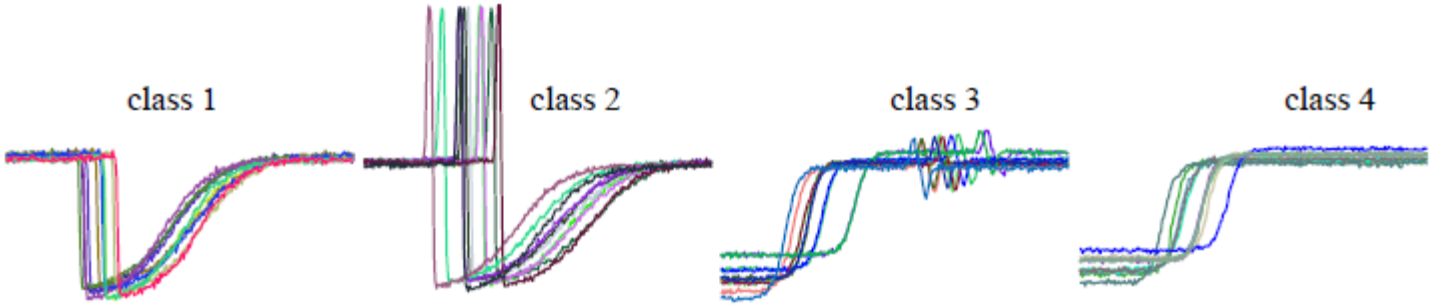
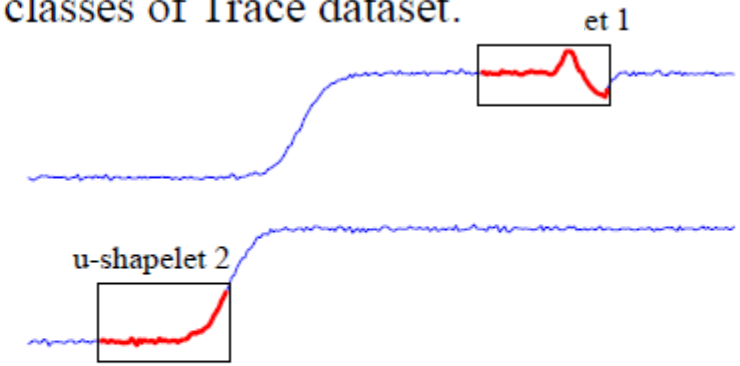


Figure 3. Sample time series from four classes of Trace dataset.



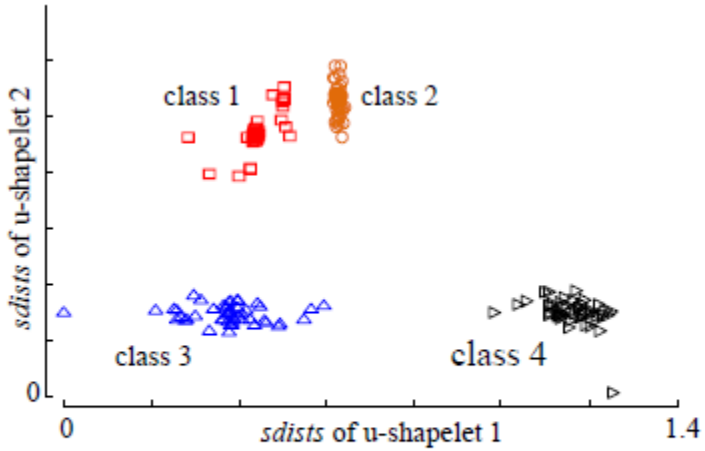


Figure 4. (*left*) two *u-shapelets* (marked with red) used for clustering *Trace* dataset. (*right*) a plot of *distance map* of the *u-shapelets*.

Algorithm - A Formal Description

$$gap = \mu_B - \sigma_B - (\mu_A + \sigma_A)$$

Abraham ***Lincoln*** *lived here for many years*. (English)
She is looking for Ibrahim. (Arabic)
You can find Abrahan in that house. (Portuguese)
Michael is singing a song for her. (English)
She bought a gift for Michaël. (Dutch)
She can teach Michales chess (Hebrew)

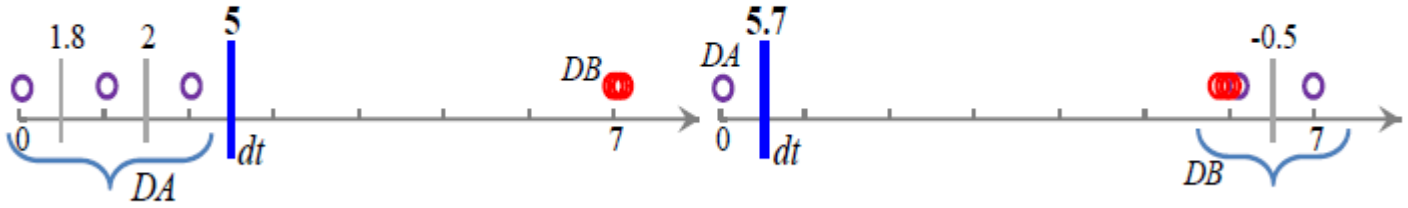


Figure 5. Orderline for (*left*) "*Abraham*", (*right*) "*Lincoln*".

Hamming distance for Abraham
[0, 2, 1, 7, 7, 7]

Algorithm - A Formal Description

$$\Theta = mean\left(sdist\left(\acute{S}, D_A\right)\right) + std\left(sdist\left(\acute{S}, D_A\right)\right)$$

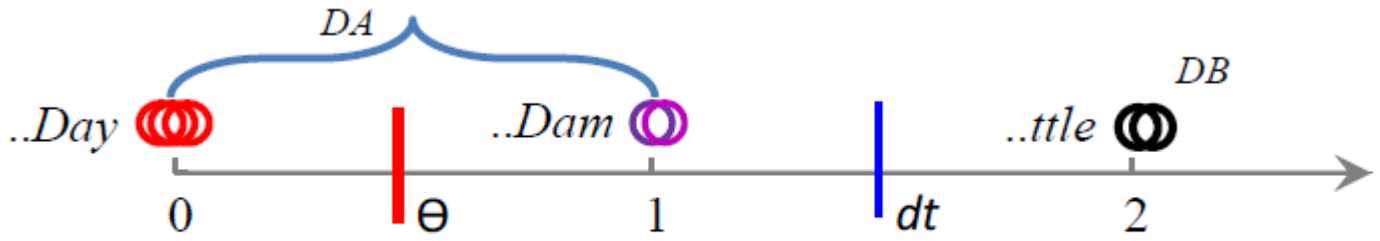


Figure 6. Orderline for "Day". Θ is shown with red/thick line and *dt* is shown with blue/thin line.

Algorithm - A Formal Description

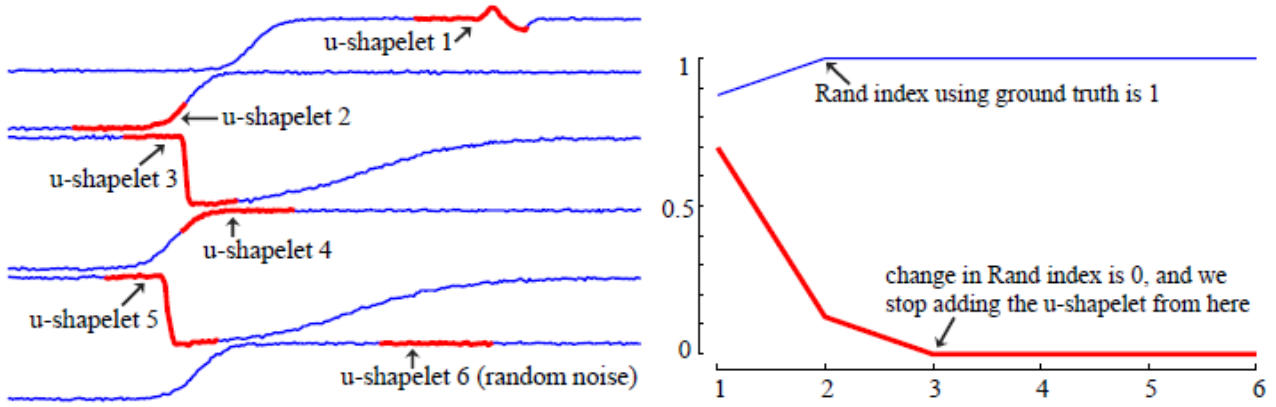


Figure 7. (*left*) The six u-shaplets returned by our algorithm on the Trace dataset. (*right*) The *CRI* (red/bold) predicts the best number of u-shapelets to use is two. By peeking at the ground truth labels (blue/light) we can see that the choice does produce a *perfect* clustering.

EXPERIMENTAL EVALUATION

TABLE VI. COMPARISON TO RIVAL METOHDS

Dataset (# of class)	Rand index			Number of u-shapelets used
	Extracted Features [33]	u-Shapelets	Time Series ED	
Trace (4)	0.74	**1**	0.75	**2**
Syn-Control (6)	0.85	**0.94**	0.87	5
Gun Point (2)	0.49	**0.74**	0.49	1
ECG (3)	0.4	**0.7**	*not-defined*	1
Population (2)	0.8	**0.9**	0.5	1
Temperature (2)	0.8	0.9	**1**	1
Income (2)	0.5	0.5	0.5	1

Drawback

The computation of orderline is time consumming.

Scalable Clustering of Time Series with U-Shapelets

LIUDMILA ULANOVA NURJAHAN BEGUM EAMONN KEOGH
UNIVERSITY OF CALIFORNIA, RIVERSIDE

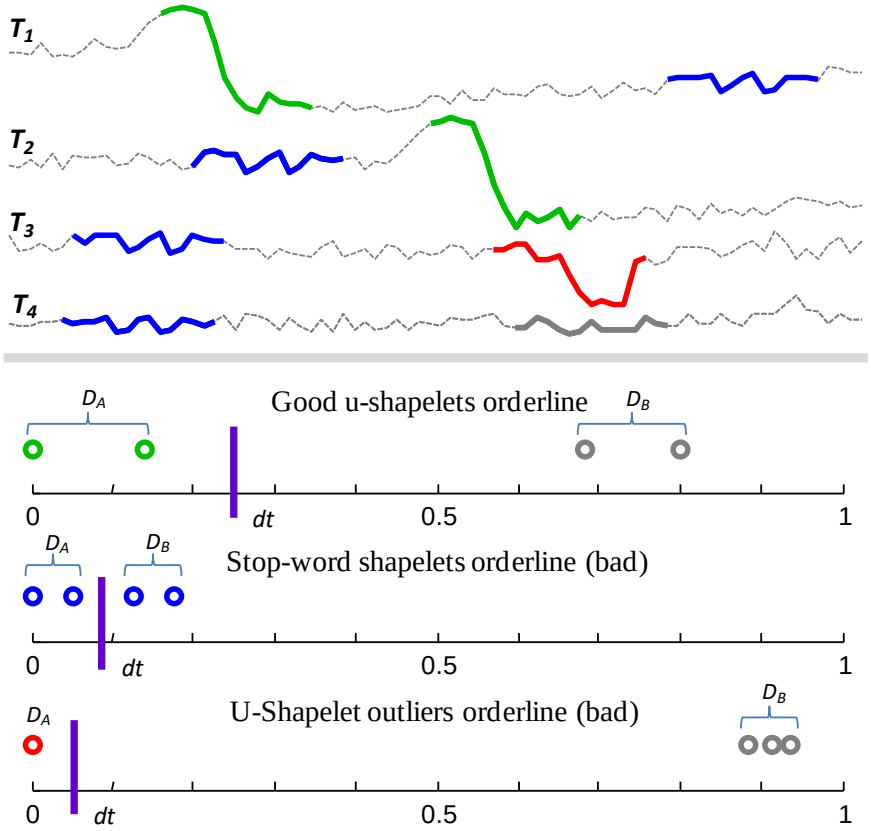


Figure 8. *top*) Good u-shapelet candidates are presented as green subsequences in time series T_1 and T_2; bad u-shapelets are shown in red in time series T_3 (close to only one subsequence in *one* time series) and blue (all time series in the dataset have similar subsequences). *bottom*) Orderlines for the three different types of u-shapelets (best viewed in color)

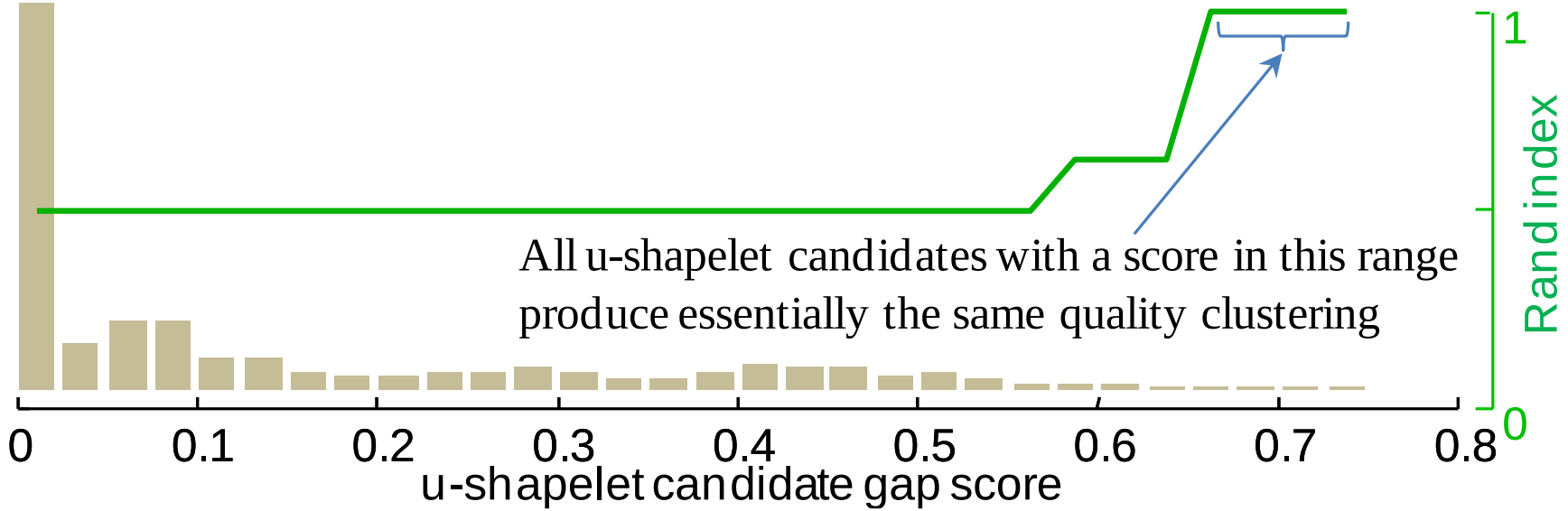


Figure 5. *gray*) The distribution of all u-shapelet scores computed during a brute force search. *green*) The minimum Rand index of these u-shapelets. Once the u-shapelet score is greater than about 0.65, it can achieve the same Rand index as the best u-shapelet

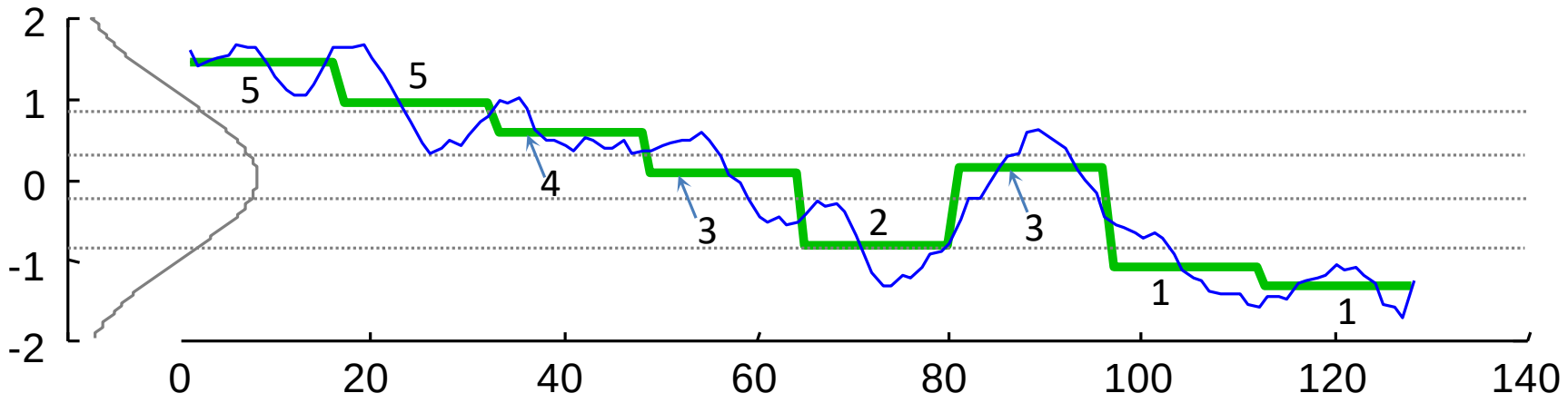


Figure 6. Representation of time series T (blue) in PAA (green/bold) converted into a SAX word, $\text{SAX}(T)_{5,8} = \{5,5,4,3,2,3,1,1\}$, with cardinality 5 and dimensionality 8

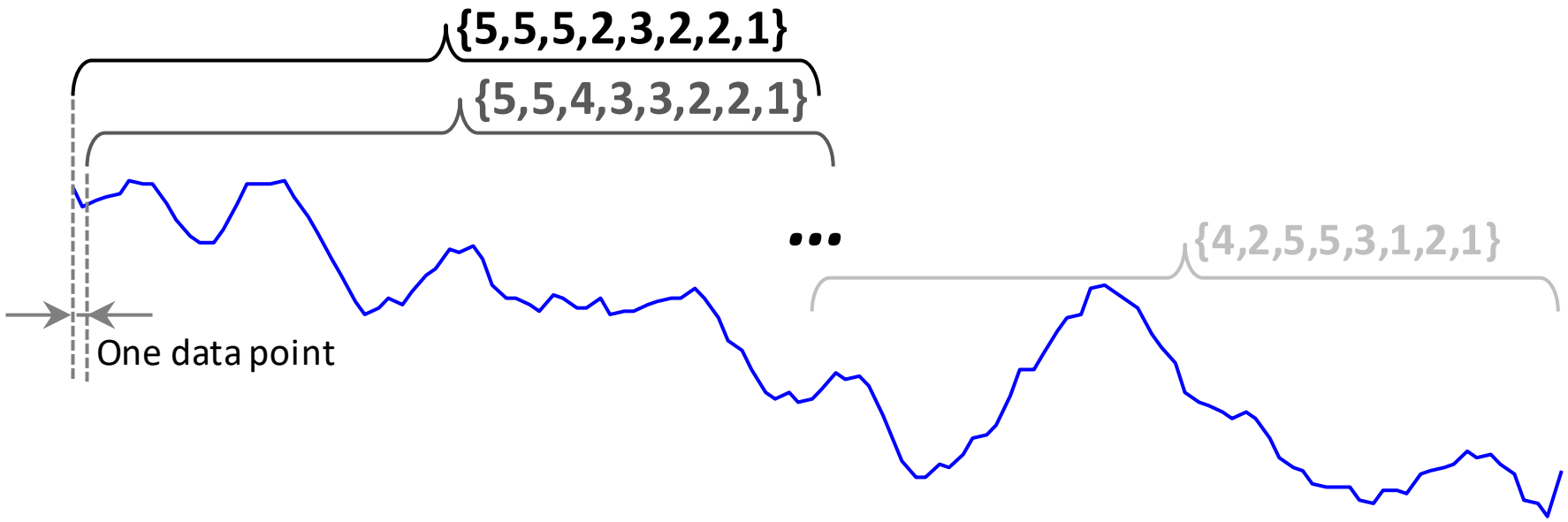


Figure 7. Time series *T* converted into a set of SAX words, {5,5,5,2,3,2,2,1}, {5,5,4,3,3,2,2,1}, ..., {4,2,5,5,3,1,2,1}, using a sliding window of length 64

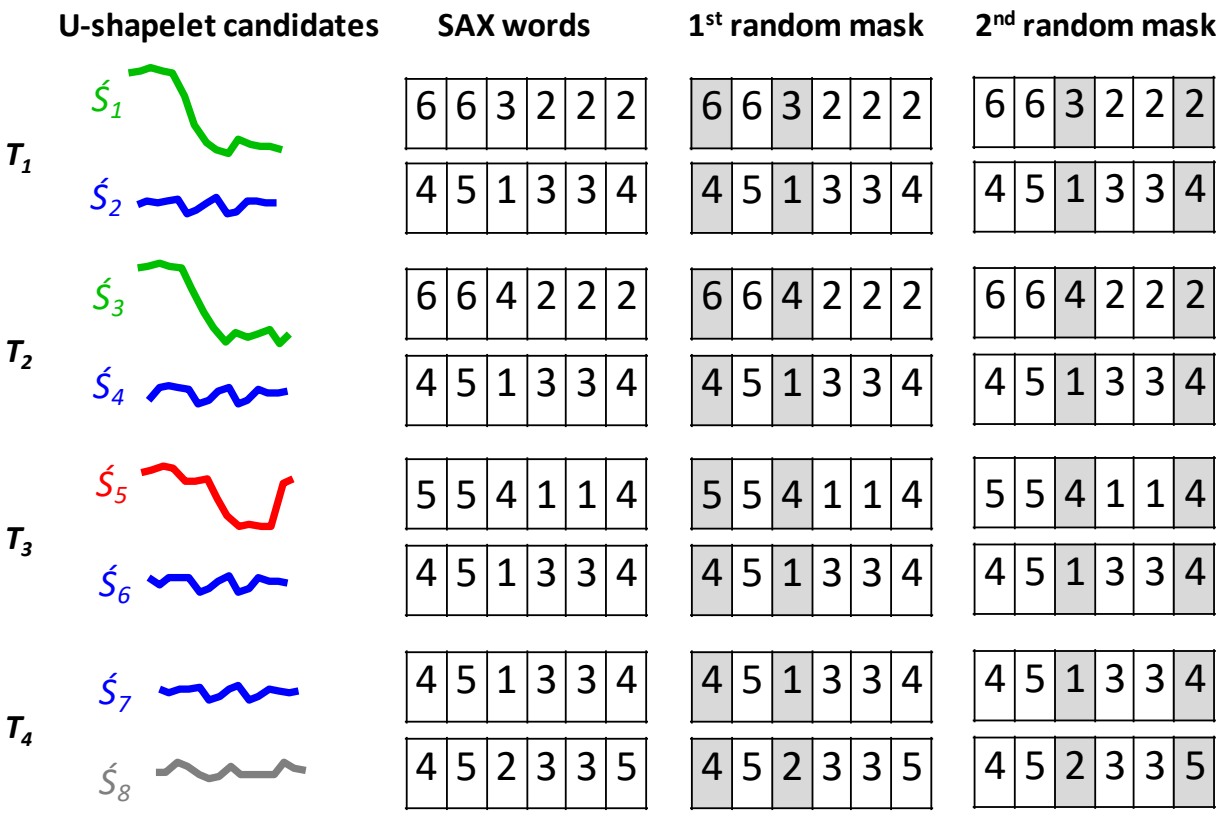


Figure 8. Several randomly chosen u-shapelet candidates, their original SAX representation (c = 6 and d = 6) and SAX words after two rounds of random masking of 2 symbols

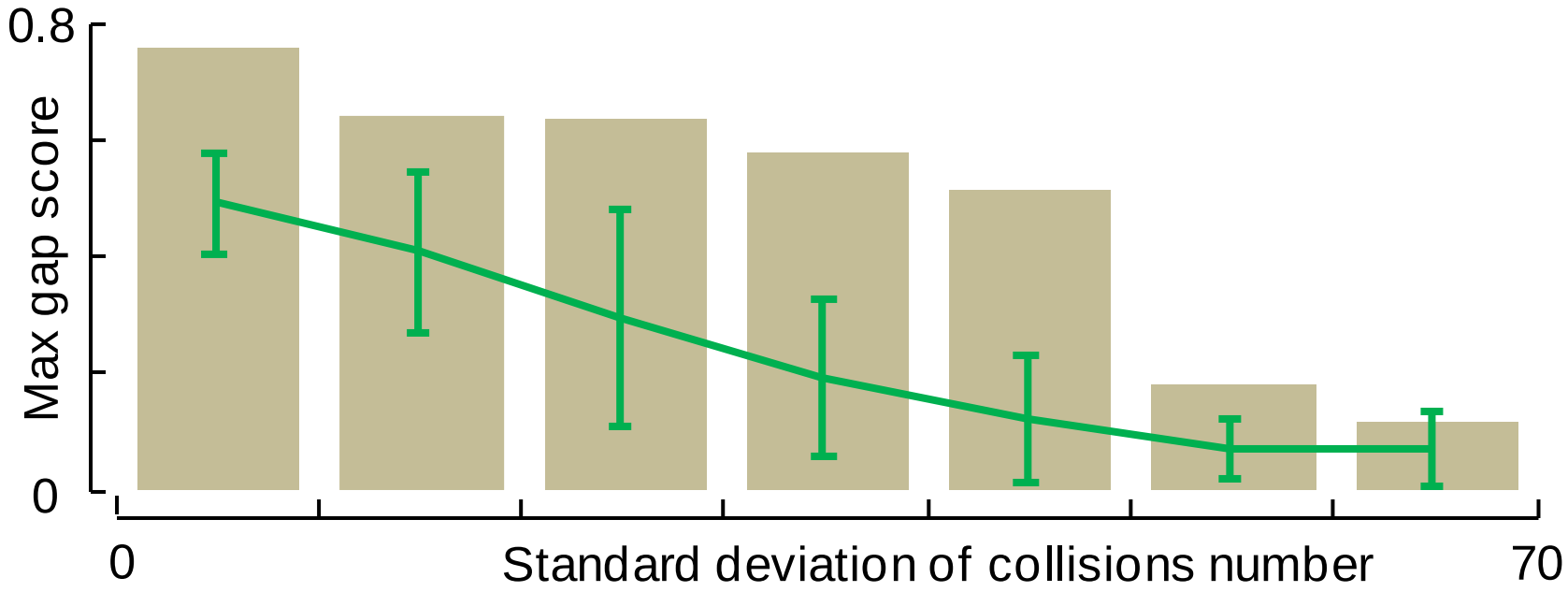


Figure 9. *gray*) Distribution of maximum values of gap score per interval. *green*) Mean and standard deviation of gap score values per interval

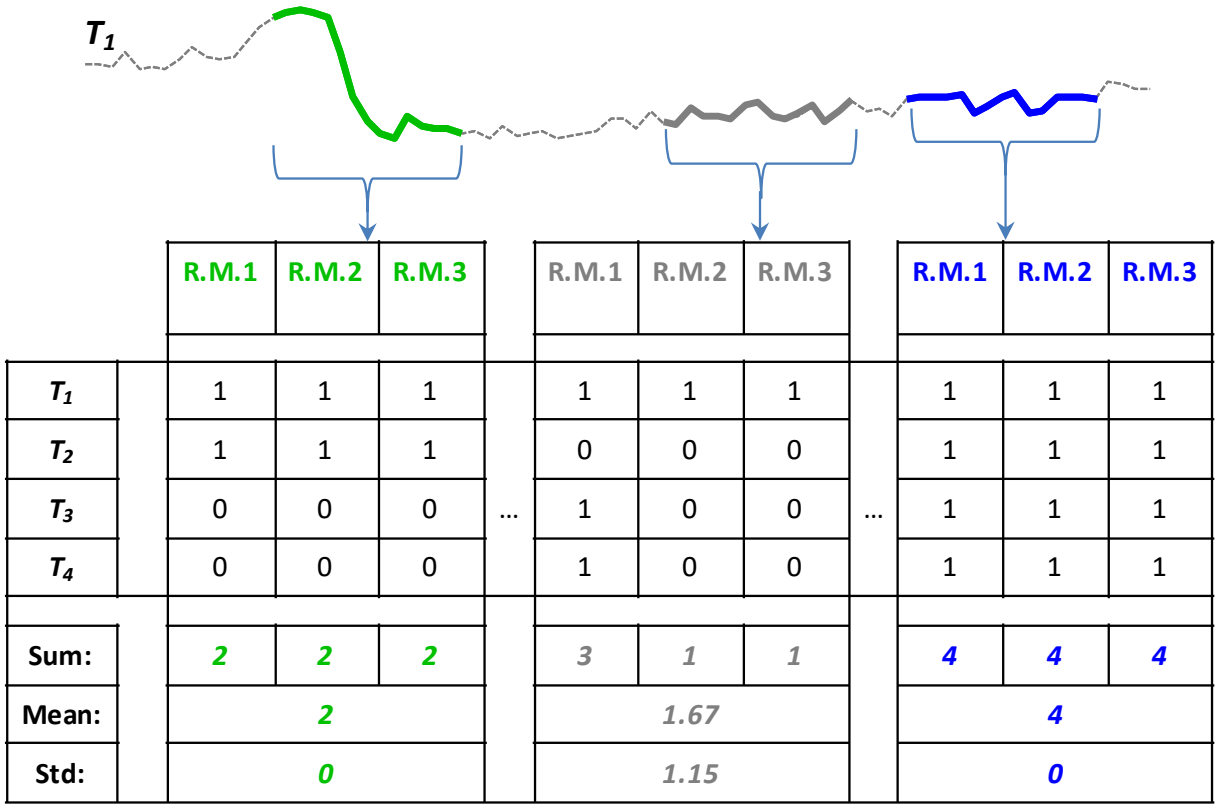


Figure 10. For each u-shapelet candidate we count how many time series have a subsequence that shares the masked SAX signature with this candidate (number of collisions). U-shapelet candidates having a low variability of the number of collisions are very likely to be better candidates

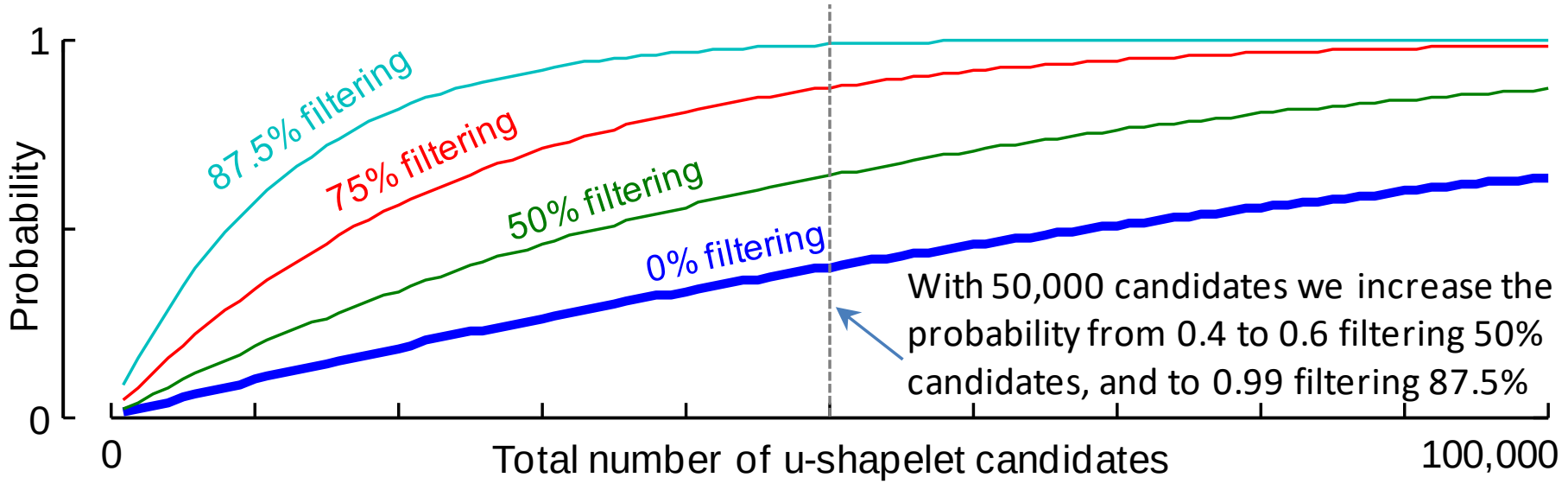


Figure 11. Probability of finding a “good enough” u-shapelet after searching 1% of candidates

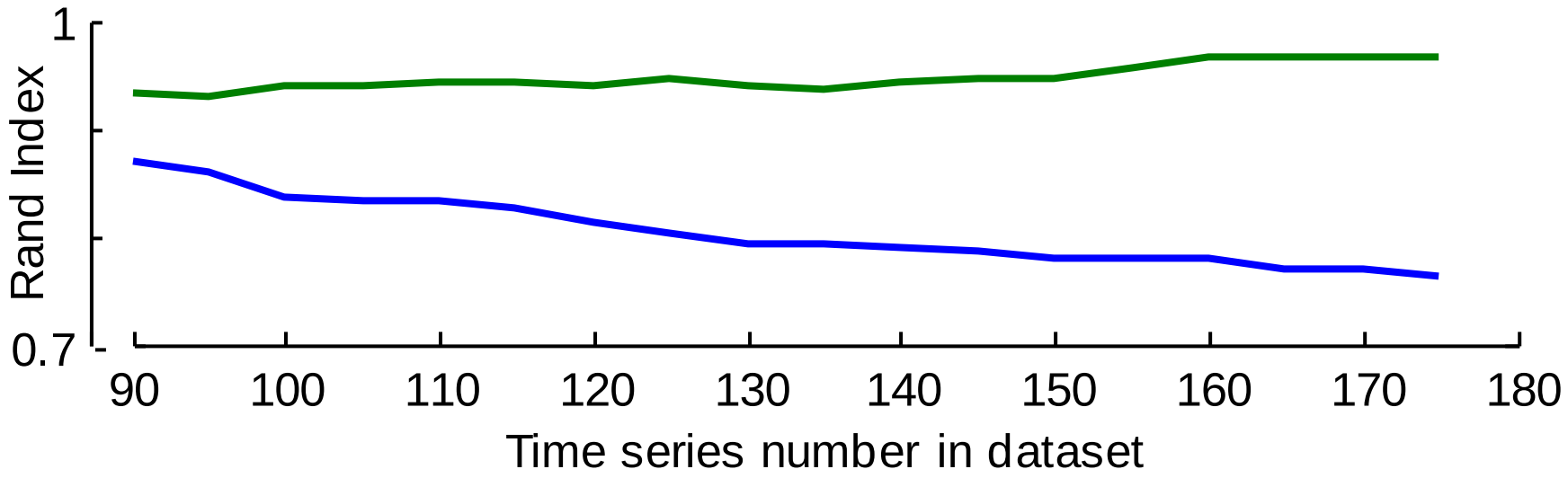


Figure 13. Rand index for k-means clustering (blue) and clustering with u-shapelets (green). The addition of spurious data does not hurt the quality of clustering with u-shapelets (averaged by 10 random runs)