An Invariance-guided Stability Criterion for Time Series Clustering Validation

ICPR 2020

Florent Forest, Alex Mourer, Mustapha Lebbah, Hanane Azzag, Jérôme Lacaille

f@florentfo.rest
http://florentfo.rest
FlorentF9

January 11-15, 2021

Table of contents

Time series clustering: challenges and algorithms

Time series: Type of data naturally organized as sequences. Functional data varying along one dimension (curve), often time but not necessarily.

Examples: sensor measurements, biological data, economic data…

Time series: Type of data naturally organized as sequences. Functional data varying along one dimension (curve), often time but not necessarily.

Examples: sensor measurements, biological data, economic data...

Clustering: Finding groups called clusters such that elements sharing the same cluster are similar, and elements belonging to different clusters are dissimilar.

Time series clustering: challenges and algorithms

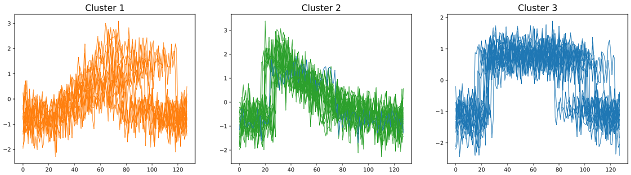
Time series: Type of data naturally organized as sequences. Functional data varying along one dimension (curve), often time but not necessarily.

Examples: sensor measurements, biological data, economic data...

Clustering: Finding groups called clusters such that elements sharing the same cluster are similar, and elements belonging to different clusters are dissimilar.

Challenges

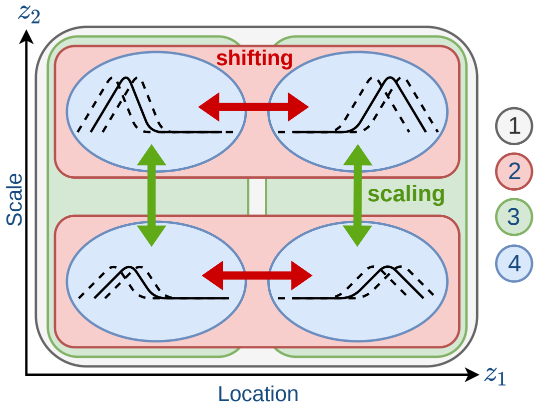
- High dimensionality
- Temporal correlation
- Invariance to transformations
- Varying lengths



Time series clustering: challenges and algorithms

Time series invariances and similarity measures (see [Giusti and Batista, 2013])

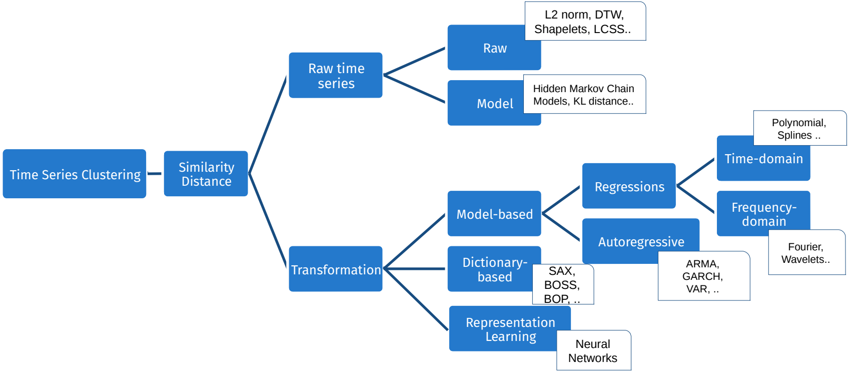
- **Scale, offset** $\rightarrow$ normalization, correlation-based similarities (shape-based) [Paparrizos and Gravano, 2015] ...
- **Shifting** $\rightarrow$ find optimal shifting between 2 series
- **Warping** (speed & delay) or **uniform temporal scaling** (speed) $\rightarrow$ Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978]
- **Occlusion** $\rightarrow$ subsequences, shapelets...
- **Complexity, noise** $\rightarrow$ smoothing, complexity-invariant distance [Batista et al., 2011] ...


z_2
shifting
scaling
Scale
Location
z_1
1
2
3
4

A bit of taxonomy (see [Warren Liao, 2005, Aghabozorgi et al., 2015])

Tasks: Whole time series, Subsequence clustering, Time point/Segmentation

Methods:



Time series clustering: challenges and algorithms

This work focuses on **whole raw time series** clustering, experiments with 2 algorithms: *K*-**medoids (PAM)** [Kaufman and Rousseeuw, 1990, Ng and Han, 1994] with EUC/COR/DTW and *K*-**shape** [Paparrizos and Gravano, 2015].

Time series clustering: challenges and algorithms

This work focuses on **whole raw time series** clustering, experiments with 2 algorithms: K**-medoids (PAM)** [Kaufman and Rousseeuw, 1990, Ng and Han, 1994] with EUC/COR/DTW and K**-shape** [Paparrizos and Gravano, 2015].

$$\mathrm{EUC}(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||_2 = \sqrt{\sum_{t=1}^{T}(x_t - y_t)^2} \qquad \mathrm{DTW}(\mathbf{x},\mathbf{y}) = \min\sqrt{\sum_{i=1}^{P} w_i}$$

$$\mathrm{COR}(\mathbf{x},\mathbf{y}) = 1 - \mathrm{NCC}_0(\mathbf{x},\mathbf{y}) \qquad \mathrm{SBD}(\mathbf{x},\mathbf{y}) = 1 - \max_{w} \mathrm{NCC}_w(\mathbf{x},\mathbf{y})$$

Time series clustering: challenges and algorithms

This work focuses on **whole raw time series** clustering, experiments with 2 algorithms: *K*-**medoids (PAM)** [Kaufman and Rousseeuw, 1990, Ng and Han, 1994] with EUC/COR/DTW and *K*-**shape** [Paparrizos and Gravano, 2015].

$$\mathrm{EUC}(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||_2 = \sqrt{\sum_{t=1}^{T}(x_t - y_t)^2} \qquad \mathrm{DTW}(\mathbf{x},\mathbf{y}) = \min \sqrt{\sum_{i=1}^{P} w_i}$$

$$\mathrm{COR}(\mathbf{x},\mathbf{y}) = 1 - \mathrm{NCC}_0(\mathbf{x},\mathbf{y}) \qquad \mathrm{SBD}(\mathbf{x},\mathbf{y}) = 1 - \max_{w} \mathrm{NCC}_w(\mathbf{x},\mathbf{y})$$

Method/Invariance	Scale	Shift	Warping
K-medoids + EUC	✗	✗	✗
K-medoids + COR	✓	✗	✗
K-medoids + DTW	✗	✓	✓
K-shape	✓	✓	✗

Model selection in time series clustering

Model selection in clustering

Clustering validation

”Evaluating results of cluster analysis in a *quantitative* and *objective* fashion” [Roth et al., 2002], in order to select the *right* number of clusters in a data set, or to tune any hyperparameter of a clustering algorithm.

Clustering validation

”Evaluating results of cluster analysis in a *quantitative* and *objective* fashion” [Roth et al., 2002], in order to select the *right* number of clusters in a data set, or to tune any hyperparameter of a clustering algorithm.

No universally admitted loss function or ground-truth as in supervised ML
$\rightarrow$ challenging problem! [von Luxburg et al., 2012, Ben-David, 2018]

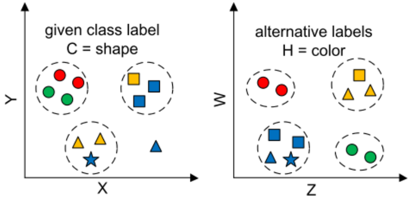


Figure 1: Toy example on alternative clusters. [Färber et al., 2010]

Model selection in clustering

No labels → **Internal clustering validity indices** (CVIs) (see [Arbelaitz et al., 2013])

- Indices based on within-cluster/between-cluster distances (compactness VS separateness): Davies-Bouldin, Silhouette...strong priors on the geometry!
- Model-based: likelihood criteria (AIC, BIC, ICL...)
- Statistical robustness: cluster stability analysis

Not well studied in TS clustering!

Clustering stability

Stability principle

A clustering algorithm applied with the same parameters to perturbed versions of a data set should find the same structure and obtain similar results. ”to be meaningful, a clustering must be both good and the only good clustering of the data, *up to small perturbations*” [Meilă, 2018]

1. Generate several samples from the data distribution (resampling, perturbation).
2. Run the clustering algorithm on each sample.
3. Measure similarities between the obtained partitions.
4. Aggregate these similarities into a stability score.
5. (optional: normalization step.)

See [Von Luxburg, 2009] for a review.

Stadion: a criterion based on a stability trade-off

Definition in [Mourer et al., 2020]

A clustering is a partitioning of data into groups so that the partition is stable, and within each cluster, there exists no stable partition.

Stadion: a criterion based on a stability trade-off

Definition in [Mourer et al., 2020]

A clustering is a partitioning of data into groups so that the partition is stable, and within each cluster, there exists no stable partition.

- Between-cluster stability $\mathrm{Stab}_{\mathrm{B}}(\mathcal{C}_K)$: How much does the partition change when adding uniform or Gaussian noise?
- Within-cluster stability $\mathrm{Stab}_{\mathrm{W}}(\mathcal{C}_K, \mathbf{\Omega})$: Are there any stable partitions within any of the clusters?

Stadion: a criterion based on a stability trade-off

Definition in [Mourer et al., 2020]

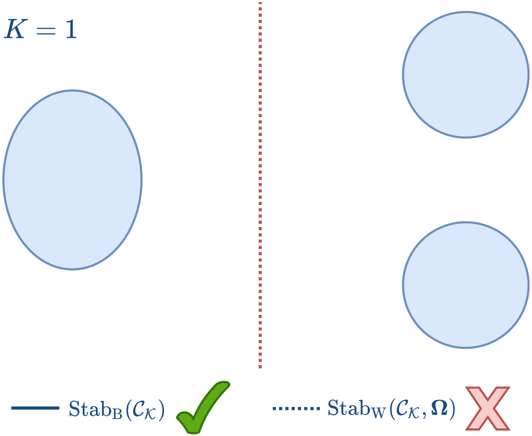
A clustering is a partitioning of data into groups so that the partition is stable, and within each cluster, there exists no stable partition.

- Between-cluster stability $\mathrm{Stab_B}(\mathcal{C}_K)$: How much does the partition change when adding uniform or Gaussian noise?
- Within-cluster stability $\mathrm{Stab_W}(\mathcal{C}_K, \mathbf{\Omega})$: Are there any stable partitions within any of the clusters?

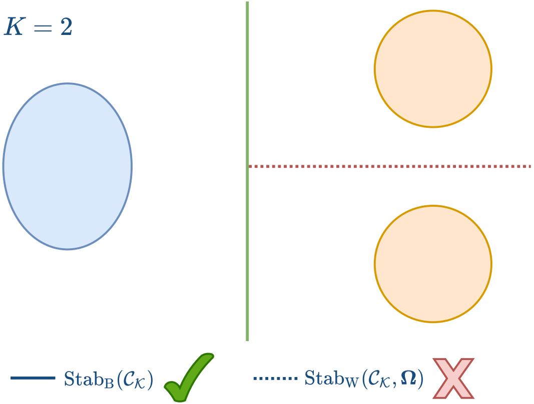
Stability **di**fference criter**ion**:

$$\mathrm{Stadion}(\mathcal{C}_K, \mathbf{\Omega}) := \mathrm{Stab_B}(\mathcal{C}_K) - \mathrm{Stab_W}(\mathcal{C}_K, \mathbf{\Omega})$$

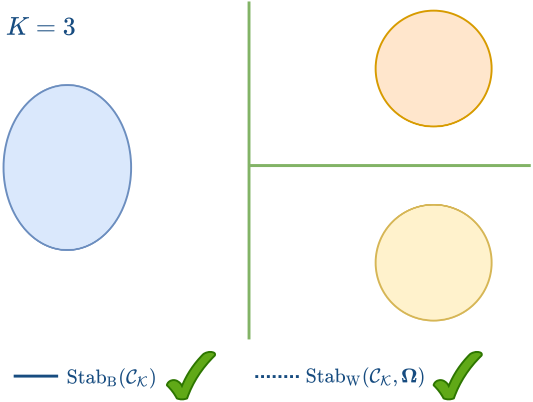
Stadion: a criterion based on a stability trade-off



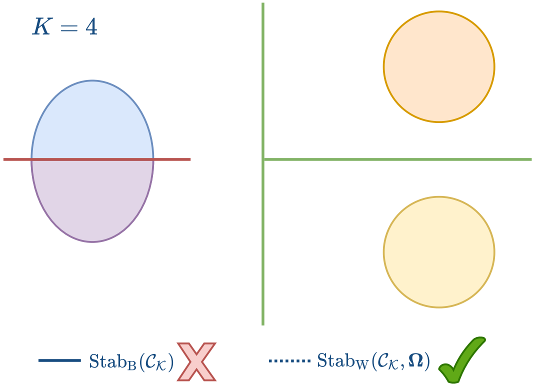
Stadion: a criterion based on a stability trade-off



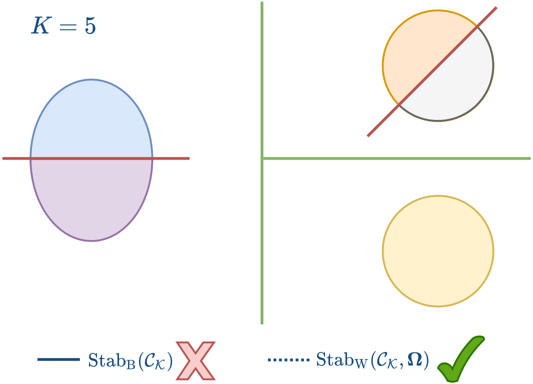
Stadion: a criterion based on a stability trade-off



Stadion: a criterion based on a stability trade-off



Stadion: a criterion based on a stability trade-off



Invariance-guided time series clustering validation

Let invariances guide the perturbation process

Prerequisite: Prior knowledge of invariances of the data (domain knowledge).

- Additive uniform/Gaussian noise is not adapted to time series (won't hit the cluster boundaries).

Let invariances guide the perturbation process

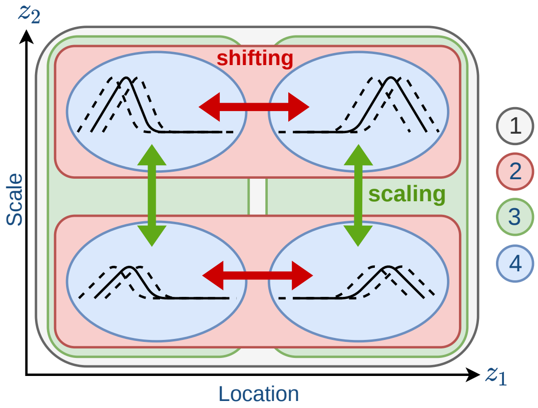
Prerequisite: Prior knowledge of invariances of the data (domain knowledge).

- Additive uniform/Gaussian noise is not adapted to time series (won't hit the cluster boundaries).
- Domain-specific perturbation process (first approach in finance [Marti et al., 2016]).

Let invariances guide the perturbation process

Prerequisite: Prior knowledge of invariances of the data (domain knowledge).

- Additive uniform/Gaussian noise is not adapted to time series (won't hit the cluster boundaries).
- Domain-specific perturbation process (first approach in finance [Marti et al., 2016]).
- **Idea:**
  - Leverage data invariances to guide the perturbation process.
  - Perturbing latent factors of variation.
  - Finding structures that are resilient to perturbation.


z_2
shifting
scaling
Scale
1
2
3
4
z_1
Location

Experiments

Selecting the K

How many clusters are there in a data set?

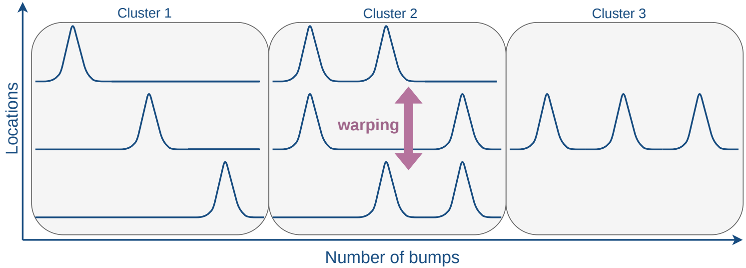
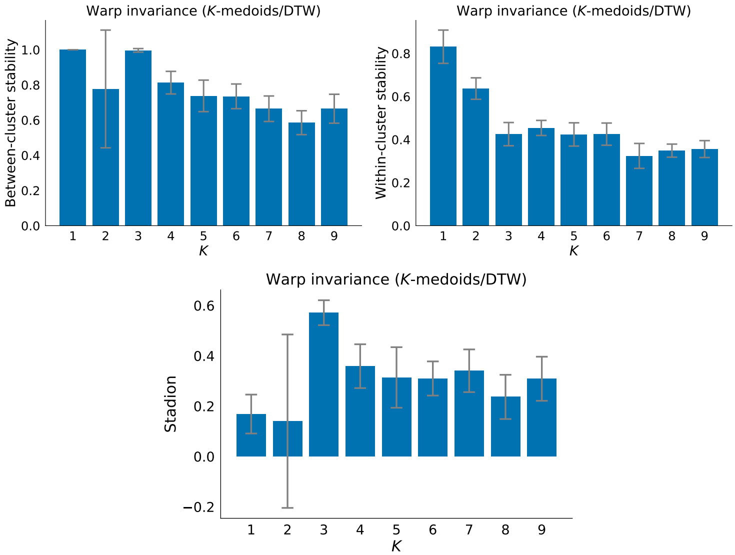


Figure 2: Toy data set with 1, 2 or 3 bumps ar random locations.

- Perturbation: random warping
- Algorithm: K-medoids + DTW

Selecting the K



Selecting the K

How many clusters are there in my data set?

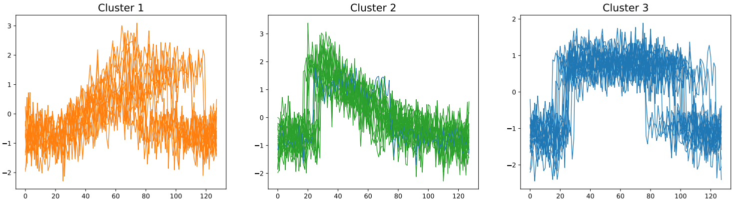
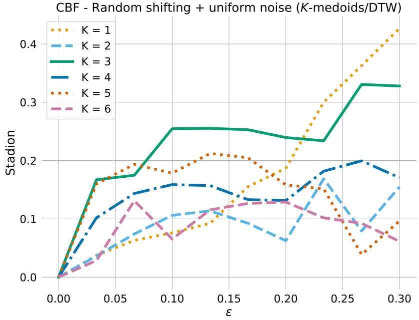
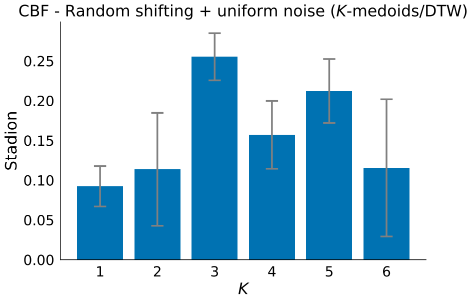


Figure 3: CBF data set.

- Perturbation: random shifting + uniform noise
- Algorithm: K-medoids + DTW

Selecting the K





Thank you for watching, feel free to read the paper for more details!

Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015).
Time-series clustering - A decade review.
Information Systems, 53:16–38.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013).
An extensive comparative study of cluster validity indices.
Pattern Recognition, 46(1):243–256.

Batista, G. E., Wang, X., and Keogh, E. (2011).
A Complexity-Invariant Distance Measure for Time Series.
In *SIAM International Conference on Data Mining (SDM)*, pages 699–710.

References ii

Ben-David, S. (2018).
Clustering - What both theoreticians and practitioners are doing wrong.
AAAI Conference on Artificial Intelligence, pages 7962–7964.

Färber, I., Günnemann, S., Kriegel, H.-P., Kröger, P., Müller, E., Schubert, E., Seidl, T., and Zimek, A. (2010).
On Using Class-Labels in Evaluation of Clusterings.
KDD International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust), page 9.

Giusti, R. and Batista, G. E. (2013).
An empirical comparison of dissimilarity measures for time series classification.
In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 82–88.

Kaufman, L. and Rousseeuw, P. J. (1990).
Finding Groups in Data: An Introduction to Cluster Analysis.
John Wiley & Sons.

Marti, G., Very, P., Donnat, P., and Nielsen, F. (2016).
A proposal of a methodological framework with experimental guidelines to investigate clustering stability on financial time series.
In *International Conference on Machine Learning and Applications (ICMLA)*, pages 32–37.

Meilă, M. (2018).
How to tell when a clustering is (approximately) correct using convex relaxations.
In *NeurIPS*, number 1, pages 7407–7418.

References v

Mourer, A., Forest, F., Lebbah, M., Azzag, H., and Lacaille, J. (2020).
Selecting the Number of Clusters K with a Stability Trade-off: an Internal Validation Criterion.

Ng, R. T. and Han, J. (1994).
Efficient and Effective Clustering Data Mining Methods for Spatial Data Mining.
In *International Conference on Very Large Data Bases (VLDB)*, pages 144–155.

Paparrizos, J. and Gravano, L. (2015).
k-Shape: Efficient and Accurate Clustering of Time Series.
ACM SIGMOD, pages 1855–1870.

Roth, V., Lange, T., Braun, M., and Buhmann, J. (2002).
A Resampling Approach to Cluster Validation.
Compstat, pages 123–128.

Sakoe, H. and Chiba, S. (1978).
Dynamic Programming Algorithm Optimization for Spoken Word Recognition.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1):43–49.

Von Luxburg, U. (2009).
Clustering stability: An overview.
Foundations and Trends in Machine Learning, 2(3):129–168.

von Luxburg, U., Williamson, R. C., and Guyon, I. (2012).
Clustering: Science or Art?
JMLR: Workshop and Conference Proceedings, 27:6579.

Warren Liao, T. (2005).
Clustering of time series data - A survey.
Pattern Recognition, 38(11):1857–1874.