An Invariance-guided Stability Criterion for Time Series Clustering Validation

Time series clustering [1,2]

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... and many others!

Invariances

- Scale, offset
- Shifting
- Warping
- Occlusion
- Complexity, noise...

Algorithms

Method/Invariance	Scale	Shift	Warping
K-medoids + EUC	✗	✗	✗
K-medoids + COR	✓	✗	✗
K-medoids + DTW [3]	✗	✓	✓
K-shape [4]	✓	✓	✗

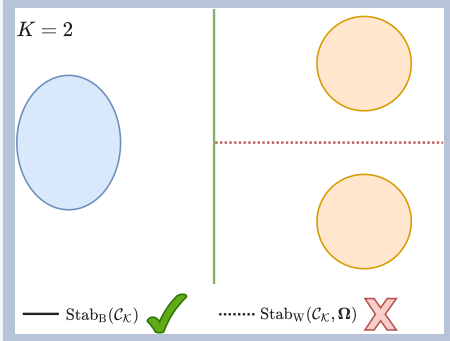
Cluster stability analysis for model selection [6,7]

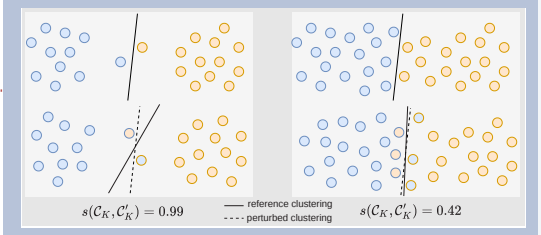
Model selection: Evaluating results of cluster analysis in a *quantitative* and *objective* fashion, in order to select the *right* number of clusters in a data set, or to tune any hyper-parameter of a clustering algorithm. **Internal clustering validity indices** [5] incorporate **strong priors on cluster geometry**.

Cluster stability analysis: a model-agnostic principle [6]

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.

Stadion: trade off between- and within-cluster stability [7]





github.com/FlorentF9/skstab

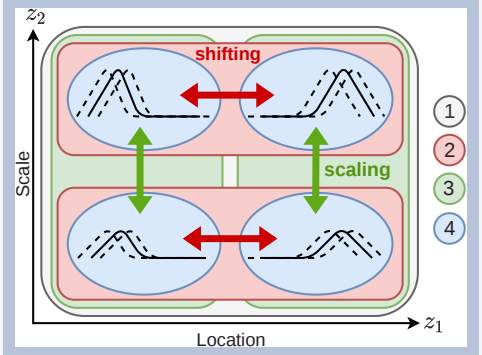
Invariance-guided stability-based model selection

Principle

Use prior knowledge on data invariances to guide the perturbation process.
E.g. shift-invariant and noisy data → random shifting and noise perturbation.

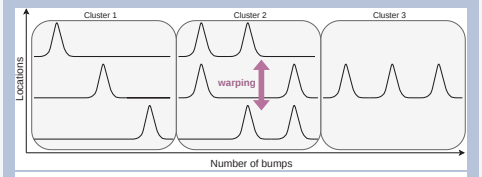
Perturbing latent factors of variation to discover resilient structures in the data.

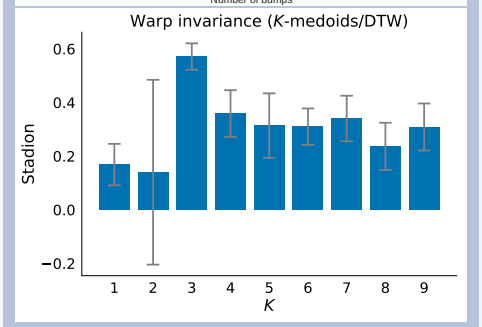
Data invariances in clustering

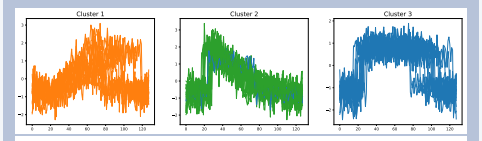


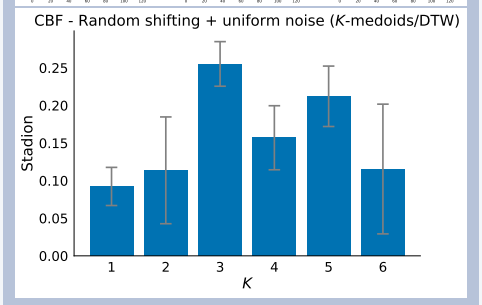
Invariances determine the cluster structure

Selecting K: a few results









References

[1] Warren Liao (2005). Clustering of time series data - A survey. *Pattern Recognition*.
[2] Aghabozorgi, Seyed Shirkhorshidi & Ying Wah (2015). Time-series clustering - A decade review. *Information Systems*.
[3] Sakoe & Chiba (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing.
[4] Paparrizos & Gravano (2015). k-Shape: Efficient and Accurate Clustering of Time Series. *ACM SIGMOD*.
[5] Arbelaitz, Gurrutxaga, Muguerza, Pérez & Perona (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*.
[6] Von Luxburg (2009). Clustering stability: An overview. *Foundations and Trends in Machine Learning*.
[7] Mourer, Forest, Lebbah, Azzag & Lacaille (2020). Selecting the Number of Clusters K with a Stability Trade-off: an Internal Validation Criterion. https://arxiv.org/abs/2006.08530

Contact

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

Florent Forest[1,3], Alex Mourer[2,3], Mustapha Lebbah[1], Hanane Azzag[1], Jérôme Lacaille[3]
1 - LIPN, Université Sorbonne Paris Nord 2 - SAMM, Université Panthéon Sorbonne 3 - Safran Aircraft Engines