Time Series Clustering

data preparation

- data import
- data cleaning
- data exclusions
- data transformations

1. variable selection

field- and concept specific

acculturation specific:
– affect
– behavior
– cognition
– desire

2. feature engineering

A. feature extraction

psychological TS:
– diversit
  – multivariate
– (in)stability
  – mean of sum squared differences
  – mean absolute change
– central tendency
  – mean
  – median
– variance
  – variance / standard deviation
  – median absolute deviation
– inertia
  – autocorrelation (lag-1)
  – pereodicity/seasonality (day, week)
– linear trend
  – OLS regression slope
– nonlinear trend
  – GAM spline est. degrees of freedom
    (line wigglieness)

B. dimension reduction

– linear
  – Principal Component Analysis (PCA)
  – Factor Analysis (FA)
  – Linear Discriminant Analysis (LDA)
– nonlinear
  – Kernel PCA
  – t-distributed Stochastic Neighbor
    Embedding (t-SNE),
  – Multidimensional Scaling (MDS)
  – Isometric mapping (Isomap)

OR

B. feature selection

– filter
  – univariate
  – multivariate
– wrapper
  – sequential
  – bio-inspired
  – iterative
– hybrid
– (embedded)

3. clustering

centroid-based
(e.g., k-means)

– non-probabilistic
– no assumptions
– fast
– sensitive to local minima

distribution-based
(e.g., gaussian mixture model [GMM])

– probabilistic
– (gaussian) distributions assumed
– more flexible shape vanilla centroid

density-based
(e.g., dbscan)

– no shape assumption
– outlier detection

hierarchical-based
(e.g., ward hierarchical clustering)

– divisive (top-down) or
  agglomerative (bottom-up)
– no pre-specified number of clusters
– easy visual inspection (dendogram)
– best for small number of cases
– no reversal of assignments

hybrid
(e.g, GMM with kmeans centroids as init.)

– helps avoid individual shortcomings
  (e.g., local minima avoidance)