Model-Based Clustering of Longitudinal Data

Marco Zanotti

University Milano-Bicocca

Contents

1. Model-based Clustering
2. GMM with Cholesky-Decomposed Covariance Structure
3. Applications
4. Conclusions

1. Model-based Clustering

Model-based Clustering

Model-based clustering is a method for clustering data through the imposition of a mixture modelling framework. A Gaussian Mixture Models is most frequently used and its density can be expressed in the form

$$f(x) = \sum_{g=1}^{G} \pi_g \phi(x|\mu_g, \Sigma_g)$$

where π_g is the probability of membership of group g, and $\phi(x|\mu_g, \Sigma_g)$ is the density of a multivariate Gaussian distribution with mean μ_g and covariance matrix Σ_g.

Model-based Clustering

Many authors exploited an eigenvalue decomposition of the group covariance matrices to to give a wide range of parsimonious covariance structures and their contributions culminated in the so-called "MClust" family of models.

These consist of 10 mixture models arising from the imposition of different constraints upon the group covariance matrix $\Sigma_g = \lambda_g H_g A_g H_g'$, where λ_g is a constant, H_g is a matrix of eigenvectors of Σ_g, and A_g is a diagonal matrix with entries proportional to the eigenvalues of Σ_g.

Model-based Clustering

In the classical approach each alternative covariance structure corresponds to a member of the family of mixture models.

picture

with explanation

Longitudinal Data Problem

Although classical model-based clustering extends into many application areas, none of these models have a covariance structure designed for the analysis of longitudinal data.

Since, longitudinal data arise when measurements are taken on each subject at a number of points in time, modelling this data requires special considerations. In particular, the correlation between different measurements in time on each subject must be taken into account.

Hence, a covariance structure that explicitly accounts for the relationship between measurements at different time points is necessary.

2. GMM with Cholesky-Decomposed Covariance Structure

Cholesky Decomposition

The covariance matrix Σ can be decomposed using the relation

$$T\Sigma T' = D$$

or equivalently

$$\Sigma^{-1} = T' D^{-1} T$$

where T is a unique lower triangular matrix with diagonal elements 1, and D is a unique diagonal matrix with strictly positive diagonal.

This relation is known as the (modified) Cholesky decomposition.

Cholesky Decomposition

The values of T and D have interpretations as generalized autoregressive parameters and innovation variances, so that the linear predictor of Y_t based on $Y_{t-1}, ..., Y_1$ can be written as

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1}(-\phi_{ts})(Y_s - \mu_s) + \sqrt{d_t}\epsilon_t$$

where $\epsilon_t \sim N(0,1)$, the ϕ_{ts} are the sub-diagonal elements of T and d_t are the diagonal elements of D.

GMM

Introduce a family of mixture models with a covariance structure specifically designed to analyse time series data.

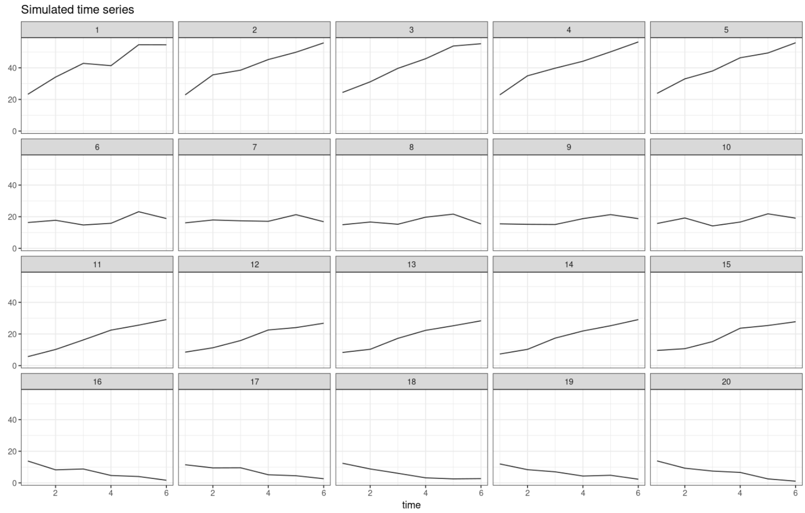
3. Applications

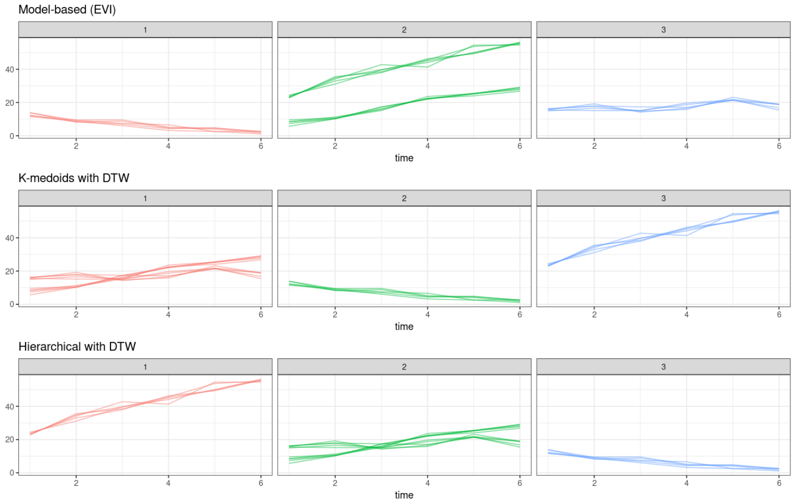
Datasets

I have applied the discussed approach on three different datasets:

- a simulated dataset of 20 time series generated through multivariate Gaussian distributions with different means and variance-covariance structures
- the rat body weight dataset of 16 rats time series with known groups (diets)
- a real dataset of 85 time series with five different underlying behaviours

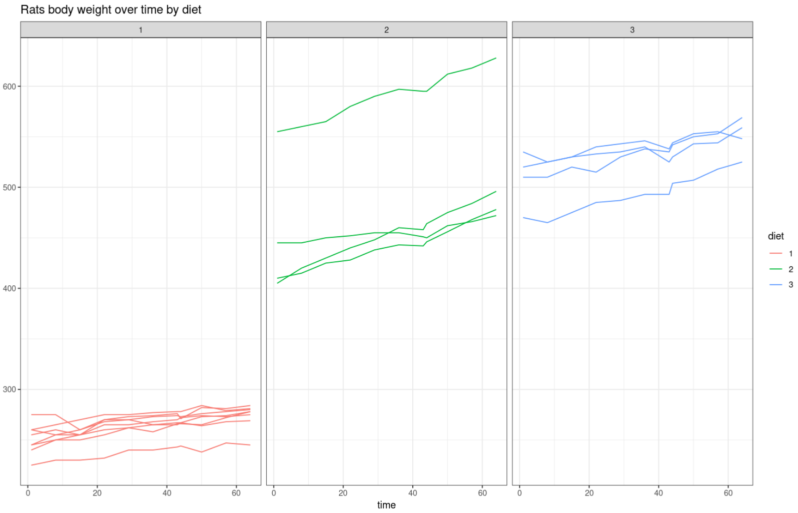
Methods of measuring distance in time series have been grouped based on approach that is shape based, feature based and model based. DTW distance is classified as shape based. This type takes into account the overall shape and matches time series based on that aspect.


Simulated time series
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
20
40
2
4
6
time

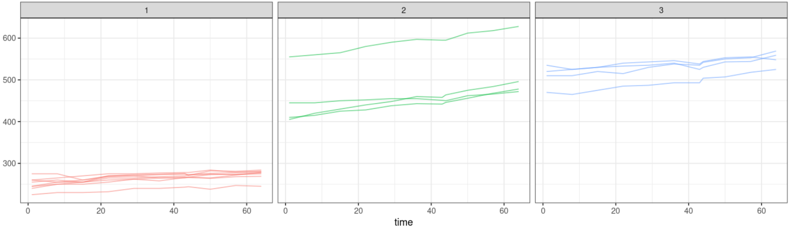

Model-based (EVI)

K-medoids with DTW

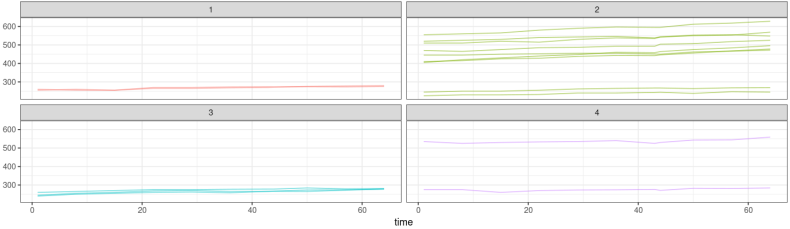
Hierarchical with DTW

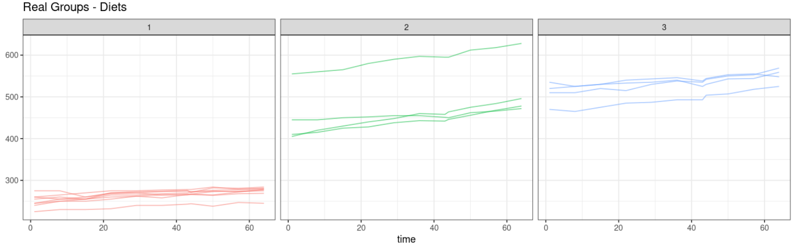

Rats body weight over time by diet
1
2
3
600
500
400
300
0
20
40
60
time
diet
1
2
3

Real Groups - Diets

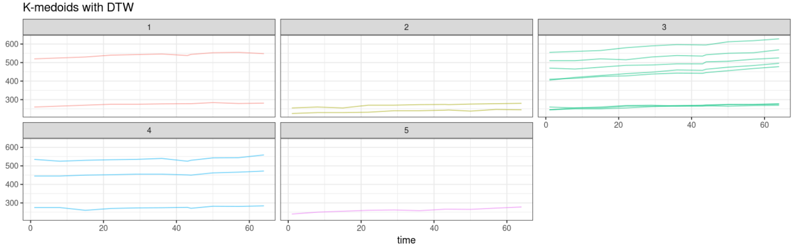


Model-based (EVI)

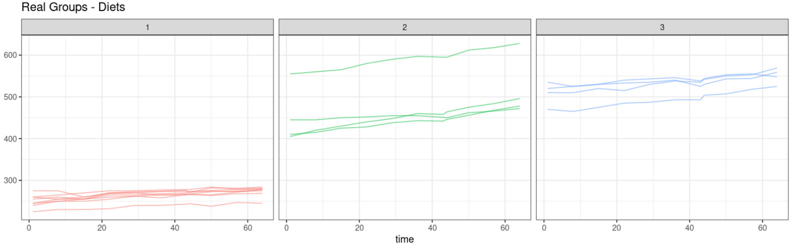


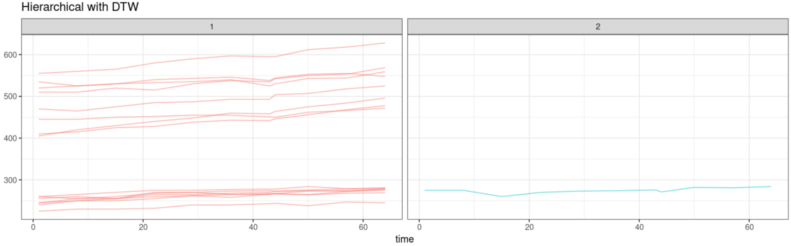


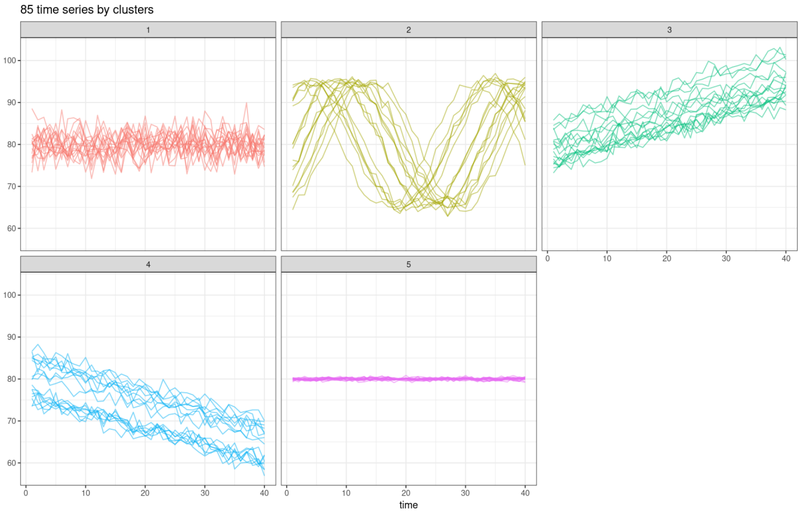
Real Groups - Diets

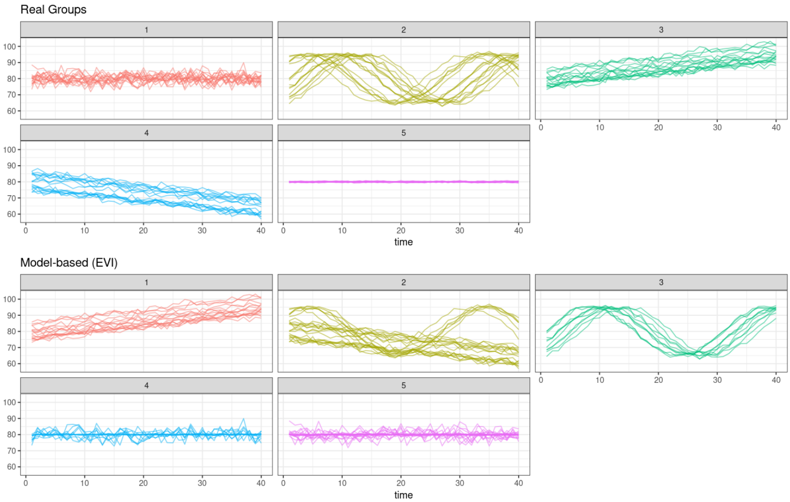


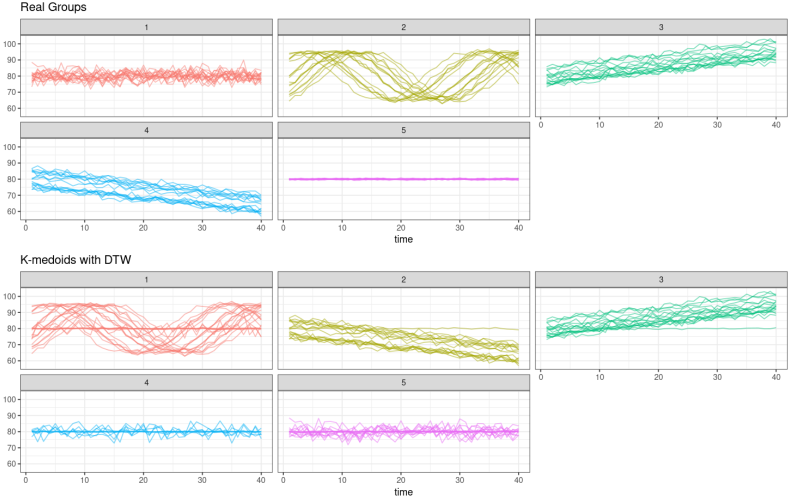
K-medoids with DTW


Real Groups - Diets
1
2
3
600
500
400
300
0
20
40
60
time

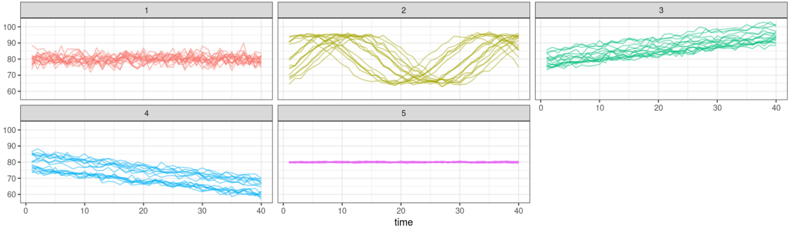

Hierarchical with DTW
1
2
600
500
400
300
0
20
40
60
time


85 time series by clusters
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time

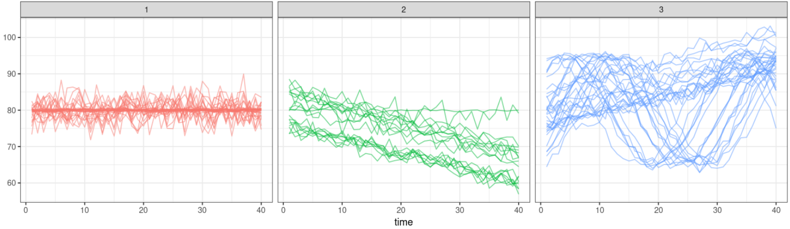

Real Groups
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time
Model-based (EVI)
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time


Real Groups
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time
K-medoids with DTW
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time

Real Groups



Hierarchical with DTW



4. Conclusions

Conclusions

Bibliografy

Paul, D. McNicholas & T. Brendan Murphy (2010), 'Model-based clustering of longitudinal data', The Canadian Journal of Statistics, No.1, 153–168.

Thank you!