Model-Based Clustering of Longitudinal Data

Marco Zanotti

University Milano-Bicocca

Contents

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 several mixture models arising from the imposition of different constraints upon the group covariance matrix

$$\Sigma_g = \lambda_g V_g D_g V_g'$$

- λ_g is a constant controlling the **volume**
- V_g is a matrix of eigenvectors of Σ_g controlling the **orientation**
- D_g is a diagonal matrix controlling the **shape**.

Model-based Clustering

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

Model	Σ_g	Distribution	Model	Σ_g	Distribution
EII	$\lambda \mathbf{I}$	Spherical	EVE	$\lambda \mathbf{V}\mathbf{D}_g\mathbf{V}'$	Ellipsoidal
VII	$\lambda_g \mathbf{I}$	Spherical	VEE	$\lambda_g \mathbf{V}\mathbf{D}\mathbf{V}'$	Ellipsoidal
EEI	$\lambda \mathbf{D}$	Diagonal	VVE	$\lambda_g \mathbf{V}\mathbf{D}_g\mathbf{V}'$	Ellipsoidal
VEI	$\lambda_g \mathbf{D}$	Diagonal	EEV	$\lambda \mathbf{V}_g\mathbf{D}\mathbf{V}_g'$	Ellipsoidal
EVI	$\lambda \mathbf{D}_g$	Diagonal	VEV	$\lambda_g \mathbf{V}_g\mathbf{D}\mathbf{V}_g'$	Ellipsoidal
VVI	$\lambda_g \mathbf{D}_g$	Diagonal	EVV	$\lambda \mathbf{V}_g\mathbf{D}_g\mathbf{V}_g'$	Ellipsoidal
EEE	$\lambda \mathbf{V}\mathbf{D}\mathbf{V}'$	Ellipsoidal	VVV	$\lambda_g \mathbf{V}_g\mathbf{D}_g\mathbf{V}_g'$	Ellipsoidal

The parameters λ_g, V_g and D_g can be constrained to be equal or variable across the clusters in different ways, obtaining a family of 14 possible models.

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 $Y_{t-1}, \ldots, Y_1$ 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.

It is possible to introduced a family of mixture models exploiting this covariance structure to analyse longitudinal data.

GMM with Cholesky-decomposed Covariance Structure

Assuming a Gaussian mixture model with a modified Cholesky-decomposed covariance structure for each mixture component, the density of an observation x_i in group g is given by

$$f(x_i|\mu_g, T_g, D_g) = \frac{1}{\sqrt{(2\pi)^p|D_g|}}\, exp\left\{-\frac{1}{2}(x_i - \mu_g)'T'D^{-1}T(x_i - \mu_g)\right\}$$

where T_g is the $p \times p$ lower triangular matrix and D_g is the $p \times p$ diagonal matrix that follow from the modified Cholesky decomposition of Σ_g.

GMM with Cholesky-decomposed Covariance Structure

There are three different constraint that can be imposed:

- Constraining T_g to be equal across groups suggests that the autoregressive structure is the same for all groups
- Constraining D_g to be equal across groups suggests that the variability at each time point is the same for all groups
- Imposing the isotropic constraint $D_g = \delta_g I_p$ suggests that the variability is the same at all time point.

Model	T_g	D_g	D_g
EEA	Equal	Equal	Anisotropic
VVA	Variable	Variable	Anisotropic
VEA	Variable	Equal	Anisotropic
EVA	Equal	Variable	Anisotropic
VVI	Variable	Variable	Isotropic
VEI	Variable	Equal	Isotropic
EVI	Equal	Variable	Isotropic
EEI	Equal	Equal	Isotropic

Model Estimation

The models are estimated using the **Expectation-Maximization** (EM) algorithm.

The missing data are taken to be the group membership labels (z) and the complete-data likelihood for the mixture model is given by

$$\mathcal{L}_c(\pi_g, \mu_g, T_g, D_g) = \prod_{i=1}^{n} \prod_{g=1}^{G} (\pi_g f(x_i | \mu_g, T_g, D_g))^{z_{ig}}$$

where $z_{ig} = 1$ if observation i is in group g and $z_{ig} = 0$ otherwise.

Model Estimation

The expected value of the complete-data log-likelihood is given by

$$Q(\pi_g, \mu_g, T_g, D_g) = \sum_{g=1}^{G} n_g \; log(\pi_g) - \frac{np}{2} \; log(\pi) + \\ - \sum_{g=1}^{G} \frac{n_g}{2} \; log(|D_g|) - \sum_{g=1}^{G} \frac{n_g}{2} \; tr(T_g S_g T_g' D_g^{-1})$$

where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}, \quad S_g = (1/n) \sum_{i=1}^{n} \hat{z}_{ig}(x_i - \mu_g)(x_i - \mu_g)'$,

and z_{ig} have been replaced by their expected values $\hat{z}_{ig}$.

Model Estimation

The parameter estimates are derived by maximizing Q. For T_g and D_g the estimation depends also upon the constraints imposed by the model (Table 2).

The **Aitken acceleration** is used to provide an asymptotic estimate of the log-likelihood at each iteration and this estimate is used to determine the convergence of the EM algorithm.

The **Bayesian information criterion** (BIC) is used to select the best member among this family of Gaussian mixture models since it gives consistent estimates of the number of components in a mixture model.

Constraining Sub-Diagonals

In many applications the relative magnitude of elements of the estimated $\hat{T}_g$ matrix is very small (almost 0). This leads to the general notion of **constraining various sub-diagonals** of T_g to 0, introducing to a more parsimonious class of models.

This constrained covariance structure has the effect of removing any autocorrelation structure over large time lags. That is, T_g constrained to contain zeros below the d^{th} sub-diagonal implies an order d autoregressive structure.

Models where all sub-diagonal elements are 0 are equivalent to the diagonal "MClust" models, hence do not exploit any longitudinal data covariance structure.

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 covariance matrices
- an **experimental** dataset measuring the rat body weight of 16 rats over time for 3 different diets
- a **real** dataset of 85 time series with five different underlying behaviours

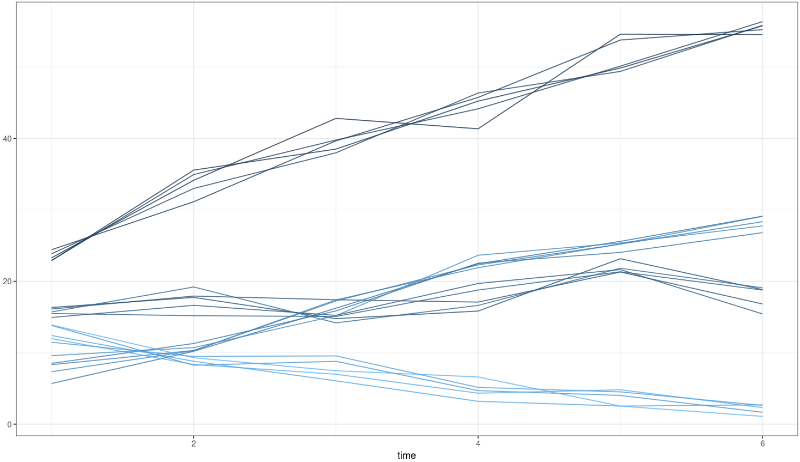
Comparison of Results

There exists different methods for clustering time series data and these have been grouped based on approach that is **shape based**, **feature based** and **model based**.

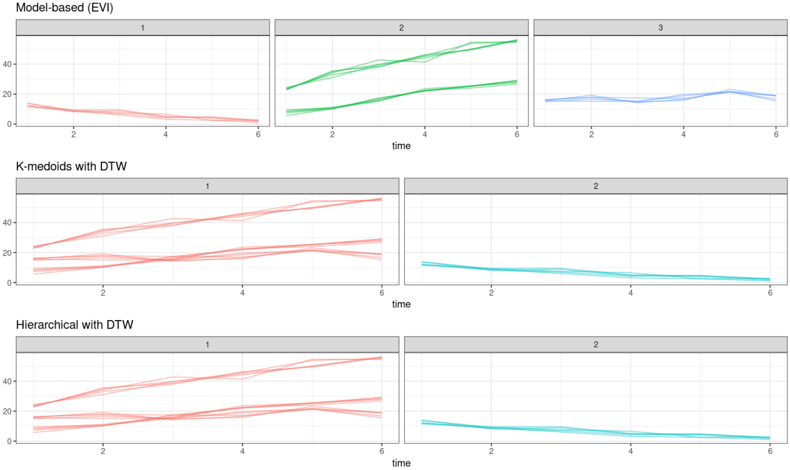
I compare the results obtained from the model based approach proposed with a shape based approach that uses the **Dynamic Time Warp** (DTW) distance, to measure the similarity between time series, coupled with both **K-medoids** and **Hierarchical** clustering algorithms.

The choice of the optimal number of clusters is based on the **BIC** for the model based and on the **Dunn Index** for the others.

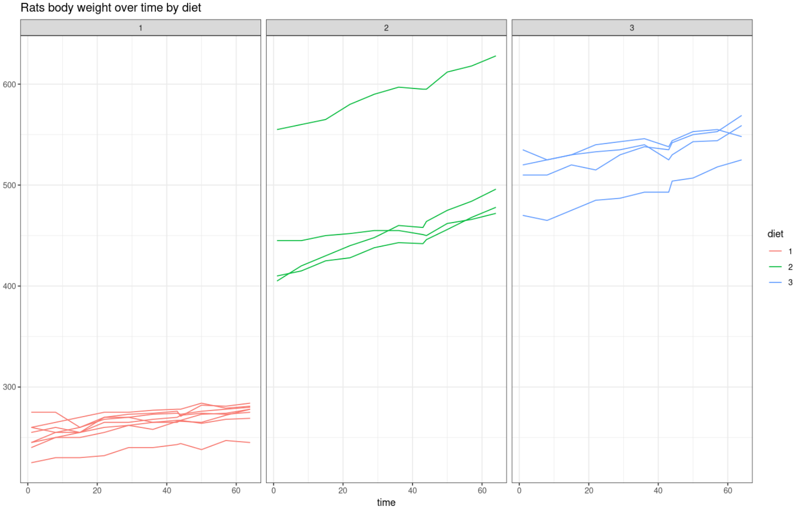
Simulated Data



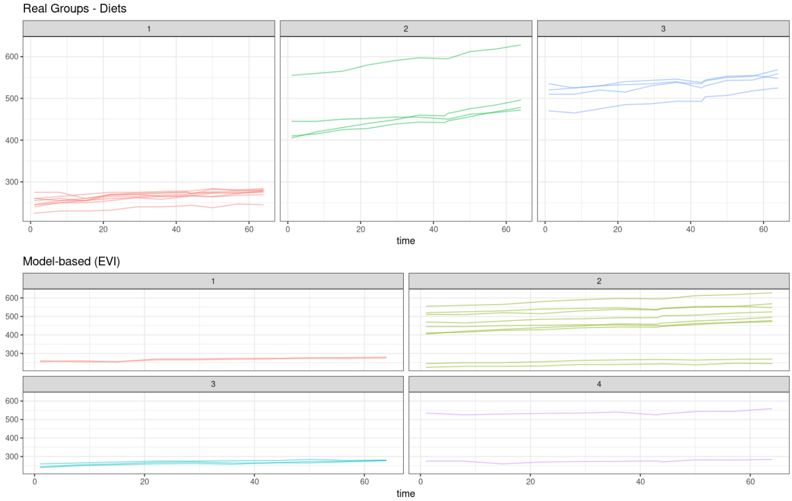
Simulated Data



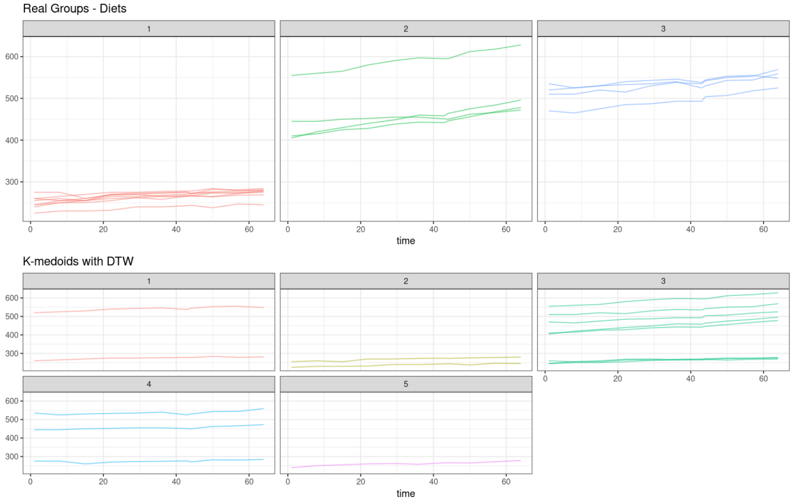
Experimental Data



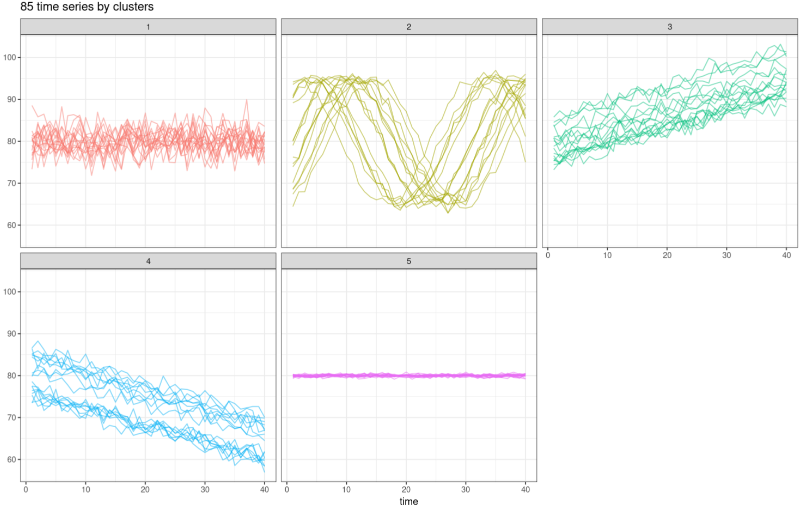
Experimental Data



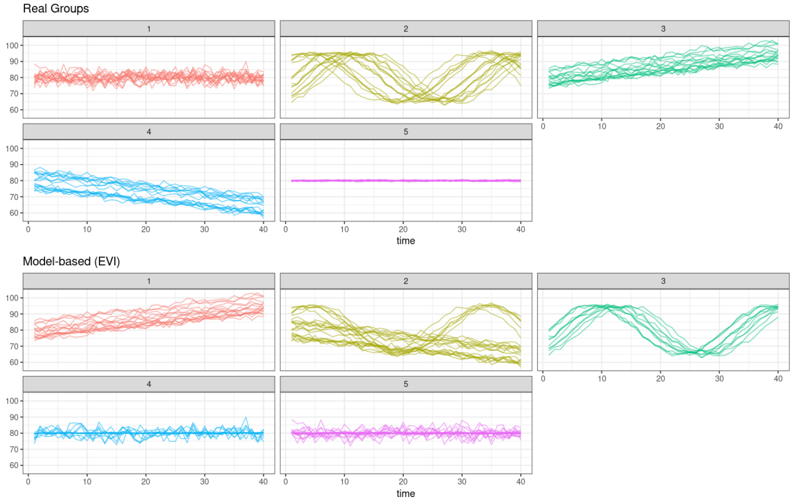
Experimental Data



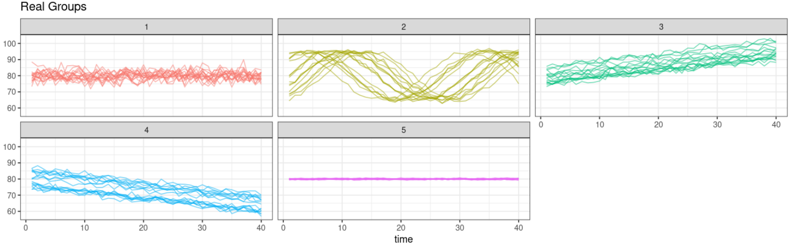
Real Data

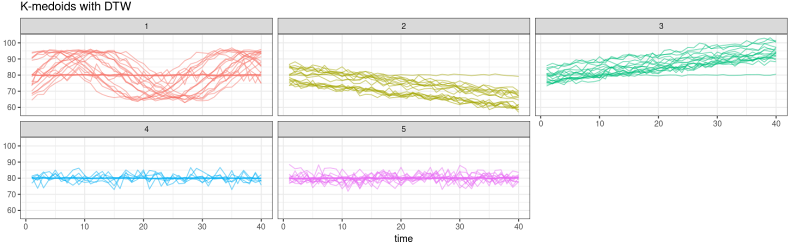

85 time series by clusters
1
2
3
4
5
100
90
80
70
60
0
10
20
30
40
time

Real Data



Real Data


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

4. Conclusions

Conclusions

A model-based clustering method, using Gaussian mixture models, for the analysis of longitudinal data is introduced.

This family of mixture models follows the classical approach, so that each member of the family has different constraints imposed on the modified Cholesky-decomposed covariance structure.

Eight member of this family are available but more parsimonious models can be obtained by constraining certain sub-diagonals of the autoregressive matrix T_g to be 0.

The empirical results on the tested datasets do not show a very high performance.

Bibliografy

McNICHOLAS, Paul D., and T. Brendan MURPHY. "Model-Based Clustering of Longitudinal Data." The Canadian Journal of Statistics / La Revue Canadienne de Statistique, vol. 38, no. 1, 2010, pp. 153–68. JSTOR

Thank you!