Model-Based Clustering of Time Series in Group-Specific Functional Subspaces

Dani Chu

Simon Fraser University

STAT 843, Feb 11th 2019

Introduction

About the Paper

- Paper by Charles Bouveyron & Julien Jacques
- Published in 2011
- Published in the international, journal Advances in Data Analysis and Classification (ADAC)
- Cited 22
- Link to paper

- Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300

National Football League (NFL) Example

Figure: Example of using model based clustering for multivariate functional data to identify similar shaped routes in the NFL. (Open in Adobe Acrobat)

Table of Contents

Clustering

Clustering

- Identify groups of homogeneous data without prior knowledge of these groups
- Examples: k-means, hierarchical classification, model-based clustering
- Clustering time series data is difficult since the data lives in an infinite dimensional space

Time Series Clustering

- Notion of pdfs do not exist for functional data
- Transform the data first to a finite dimensional problem
- Clustering time series data is difficult since the data lives in an infinite dimensional space
    - Discretizing the time interval
    - Decomposing into basis of functions
    - FPCA

Time Series Clustering Issues

- Descretization of the observed curves leads to high-dimensional data sets (sometimes $n < p$).
- In these cases, model-based clustering suffer from numerical problems.
- Can assume that high-dimensional data live in group-specific subspaces

Functional Latent Mixture Model

Plan

- Identify K homogeneous clusters among the data
- Introduce a family of mixture models designed for multivariate functional data

Reconstructing the Functional Form

- Let $\{x_1, \ldots, x_n\}$ be independent realizations of a L_2-continuous stochastic process $X = \{X(t)\},\ t \in [0, T]$
- We observe discrete observations $x_{i,j} = x_i(t_{i,j})$ at a finite set of ordered times $\{t_{i,j} :\ j = 1, \ldots, m_i\}$
- Must reconstruct the functional form of the data from the discrete observations
- Assume curves belong to a finite dimensional space spanned by a basis of functions

Basis Expansion

- Consider a basis $\{\psi_1, \ldots, \psi_p\}$
- $\gamma = (\gamma_1(X), \ldots, \gamma_p(X))$ a random vector in $\mathbb{R}^p$
- p is assumed to be fixed and known
- Assume that the stochastic process X admits the basis expansion

$$X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t)$$

Estimation of Basis Expansion

- Each observed curve is $x_i(t) = \sum_{j=1}^{p} \gamma_{i,j}\psi_j(t)$
- Estimated by least square smoothing.
- A latent mixture model is proposed for modelling the coefficient vectors $\{\gamma_1, \dots, \gamma_n\} \in \mathbb{R}$ of the observed curves
- $\gamma_i = (\gamma_{i,1}, \dots, \gamma_{i,p})$, for $i = 1, \dots, n$

Assumptions

- Consider a set of n_k observed curves described by their coefficient vectors $\{\gamma_1, \ldots, \gamma_{n_k}\} \in \mathbb{R}$
- Assume the gammas are independent realizations of a random vector $\Gamma \in \mathbb{R}^p$
- Actual stochastic process associated with the kth cluster can be described in a low-dimensional functional latent subspace $\mathbb{E}_k[0, T]$ of $L_2[0, T]$ with dimension $d_k \leq p$
- Let $\mathbb{E}_k[0, T]$ be spanned by the first d_k elements of a group-specific basis of functions $\{\phi_{kj}\}_{j=1,\ldots,d_k}$ in $L_2[0, T]$

Assumptions

- The group-specific basis is obtained from $\{\psi\}_{j=1,\ldots p}$ by a linear transformation $\phi_{kj} = \sum_{l=1}^{p} q_{k,jl}\psi_l$
- With an orthogonal $p \times p$ matrix $Q_k = (q_{k,jl})$
- Split into two parts $[U_k, V_k]$
- U_k of size $p \times d_k$ with $U_k^t U_k = I_{d_k}$
- V_k of size $p \times (p - d_k)$ with $V_k^t V_k = I_{p-d_k}$
- $U_k^t V_k = 0$

Latent Expansion Coefficients

- $\lambda_1, \ldots, \lambda_{n_k}$ is the latent expansion coefficients of the curves in the group-specific basis $\{\phi_{kj}\}_{j=1,\ldots,d_k}$
- Assumed to be independent realizations of a latent random vector $\Lambda \in \mathbb{R}^{d_k}$
- The relationship between $\{\phi_{kj}\}_{j=1,\ldots,d_k}$ and $\{\psi_j\}_{j=1,\ldots,p}$ suggest s that the random vectors Γ and Λ are linked through the following linear transformation for the kth group

$$\Gamma = U_k \Lambda + \epsilon$$

- where $\epsilon \in \mathbb{R}^p$ is an an independent and random noise term

Distributional Assumptions

- Λ assumed to be distributed according to a multivariate Gaussian density, $\Lambda \sim \mathcal{N}(m_k, S_k)$
- For the kth group we have m_k the mean and $S_k = \text{diag}(a_{k1}, \ldots, a_{kd_k})$ the covariance matrix
- ϵ distributed according to a multivariate Gaussian density $\epsilon \sim \mathcal{N}(0, \Xi_k)$
- We then get that for the kth cluster $\Gamma \sim \mathcal{N}(\mu_k, \Sigma_k)$
  - $\mu_k = U_k m_k$
  - $\Sigma_k = U_k S_k U_k^t + \Xi_k$

Variances

- Assume that the noise covaraiance matrix Ξ_k is such that $\Delta_k = \text{cov}(Q_k^t \Gamma) = Q_k^t \Sigma_k Q_k$ has the following form
    - A diagonal matrix
    - $a_{k1} \dots a_{kd_k}$ along the first d_k entries
    - $b_k \dots b_k$ along the remaining $(p - d_k)$ entries.
    - $a_{kj} > b_j$ for $j = 1, \dots, d_k$
- Practically the variance of the actual data of the kth group is modeled by $a_{k1} \dots a_{kd_k}$
- b_k models the variance of the noise
- The dimension d_k is the intrinsic dimension of the latent subspace of the kth group.

The Problem Set Up

- Consider a set of n observed curves $\{x_1 \ldots, x_n\}$, where $x_i = \{x_i(t)\}_{t \in [0,T]} (1 \leq i \leq n)$
- We want to cluster the curves into K homogeneous groups
- Assume there exists a latent random variable $Z = (Z_1, \ldots, Z_K) \in \{0,1\}^K$ indicating the group membership of X
- Z_k is equal to 1 if X belongs to the kth group and 0 otherwise
- Want to predict the value of $z_i = (z_{i1}, \ldots, z_{iK})$ of Z for each observed curve x_i

Summary of Model

- Each x_i is a sample path of X admitting a basis expansion summarised by the coefficient vector γ_i whose distribution is now a mixture of Gaussians with density

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; \mu_k, \Sigma_k)$$

- where ϕ is the standard Gaussian density function, $\mu_k = U_k m_k$ $\Sigma_k = Q_k \Delta_k Q - k^t$ and $pi_k = P(Z_k = 1)$ is the prior probability of the kth group
- This mixture model will will be referred to as $\text{FLM}_{[a_{kj} b_k Q_k d_k]}$ model

Constraints on the Full Model

- Can constrain the model parameters within or between groups
- Constrain the first d_k diagonal elements of Δ_k to be equal within each class ($a_{k1} = \ldots = a_{kd_k}$)
  - Assumes each matrix Δ_k only has two different eigenvalues a_k and b_k.
- Can fix the parameters b_k to be common across classes.
  - Assume that the behavior of the error components outside the class specific subspaces is common
  - modelling the noise outside the latent subspace by b

Submodels

For

- $\rho = Kp + K - 1$ the number of parameters needed to estimate the means and proportions
- $\tau = \sum_{k=1}^{K} d_k[p - (d_k + 1)/2]$ the number of parameters neeeded to estimate the orientation matrices Q_k and $D = \sum_{k=1}^{K} d_k$

Here are a selection of submodels and the number of parameters required to estimate the model

FLM Model	Number of Parameters
$[a_{kj} b_k Q_k d_k]$	$\rho + \tau + 2K + D$
$[a_{kj} b Q_k d_k]$	$\rho + \tau + K + D + 1$
$[a_k b_k Q_k d_k]$	$\rho + \tau + 3K$
$[a b_k Q_k d_k]$	$\rho + \tau + 2K + 1$
$[a_k b Q_k d_k]$	$\rho + \tau + 2K + 1$
$[a b Q_k d_k]$	$\rho + \tau + K + 2$

Maximum Likelihood (EM)

funHDDC Algorithm

- In model-based clustering, the estimation of model parameters is traditionally done by maximizing the likelihood through the EM algorithm
- Iterative alogirhtm consists in maximizing the complete likelihood rather than directly maximizing the likelihood which is an intractable problem with incomplete data

Log-likelihood

Given the coefficient vectors $\gamma_1, \dots, \gamma_n$ of the observed curves $x_1, \dots, x_n$ the complete log-likelihood of the data under the FLM model has the following form

$$l_c(\theta; \gamma_1, \dots, \gamma_n, z_1 m \dots, z_n) = -\frac{1}{2} \sum_{k=1}^{K} \eta_k \Bigg[\sum_{j=1}^{d_k} \left(\log(a_{kj}) + \frac{q_{kj}^t C_k q_{kj}}{a_{kj}} \right) + \left(\log(b_k) + \frac{q_{kj}^t C_k q_{kj}}{b_k} \right) - 2\log(\pi_k) \Bigg] + \xi$$

where $\theta = (\pi_k, \mu_k, a_{kj}, b_k, q_{kj})$ for $1 \leq j \leq d_k$ and $1 \leq k \leq K$, q_{kj} is the jth column of Q_k, $C_k = \frac{1}{\eta_k} \sum_{i=1}^{n} z_{ik}(\gamma - \mu_k)^t(\gamma_i - \mu_k)$, $\eta_k = \sum_{i=1}^{n} z_{ik}$ and ξ is a term not depending on the parameter θ.

E step I

- At iteration q, compute the expectation of the compelte log-likelihood conditionally on the current value of the parameter $\theta^{(q-1)}$
- So just need to commpute $t_{ik}^{(q)} = e[Z_{ik}|\gamma_i, \theta^{(q-1)}]$
- For the $\text{FLM}_{[a_k b_k Q_k d_k]}$ model, the posterior probability $t_{ik}^{(q)}$ can be computed for iteration q.

$$t_{ik}^{(q)} = 1 \Big/ \sum_{l=1}^{K} \exp\left(H_k^{(q-1)}(\gamma_i) - H_l^{(q-1)}(\gamma_i) \right)$$

E step II

With $H_k^{(q-1)}(\gamma)$ defined for $\gamma \in \mathbb{R}^p$ as:

$$
\begin{aligned}
H_k^{(q-1)}(\gamma) =& ||\mu_k^{(q-1)} - P_k(\gamma)||^2_{D_k} + \\
& \frac{1}{b_k^{(q-1)}||\gamma - P_k(\gamma)||^2} + \sum_{j=1}^{d_k} \log\left(a_{kj}^{(q-1)}\right) + \\
& (p - d_k)\log\left(b_k^{(q-1)}\right) - 2\log\left(\pi_k^{((q-1)}\right)
\end{aligned}
$$

where $||.||^2_{\mathcal{D}_k}$ is a norm on the latent space $\mathbb{E}_k$ defined by $||y||^2_{\mathcal{D}_k} = y^t\mathcal{D}_k y$, $\mathcal{D}_k = \widetilde{Q}\Delta_k^{-1}\widetilde{Q}^t$ and $\widetilde{Q}$ is a $p \times p$ matrix containing the d_k vectors of U_k completed by zeros such as $\widetilde{Q} = [U_k, 0_{p-d}]$, P_k is the projection operator on the latent space $\mathbb{E}_k$ defined by $P_k(\gamma) = U_k U_k^t(\gamma - \mu_k) + \mu_k$

E step III

- $H_k(\gamma)$ is mainly based on two distances
  1. The distance between the projection of γ on $\mathbb{E}_k$ and the current mean of the kth group
  2. The distance between the observation and the subspace $\mathbb{E}_k$
- The classification function favors the assignment of a new observation to the class for which it is close to the subspace and for which its projection on the class subspace is close to the mean of the class
- The variance terms a_k and b_k balance the importance of both distances
- If the data is noisy (b_k is large) it is natural to balance the distance $||\gamma - P_k(\gamma)||^2$ by $1/b_k$

M step I (Mixture Proportions and Means

- The next step is to estimate the model parameters by maximizing the expectation of the complete likelihood conditionally on the posterior probabilities $t_{ik}^{(q)}$ computed in the E step.
- Mixture proportions and means are updated as usual by:

$$\pi_k^{(q)} = \frac{n_k^{(q)}}{n}, \ \mu_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \gamma_i$$

- where $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$

M step II (a_{kj}, b_k, q_{kj})

- Introduce $C_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \left(\gamma_i - \mu_k^{(q)}\right)^t \left(\gamma_i - \mu_k^{(q)}\right)$, the sample covariance matrix of group k
- Introduce $W = (w_{jk})_{q \leq j, k \leq p} = \int_0^T \psi_j(t)dt$, the matrix of inner products between the basis functions
- the first d_k columns of Q_k are updated by the eigenvectors associated with the largest eigenvalues of $W^{\frac{1}{2}} C_k^{(q)} W^{\frac{1}{2}}$
- Variance parameters $a_{kj}, j = 1, \ldots, d_k$ are updated by the d_k largest eigenvalues $W^{\frac{1}{2}} C_k^{(q)} W^{\frac{1}{2}}$
- the variance parameters b_k are updated by $b_k^{(q)} = \text{trace}\left(W^{\frac{1}{2}} C_k^{(q)} W^{\frac{1}{2}}\right) - \sum_{j=1}^{d_k} \hat{a}_{kj}^{(q)}$

Summary of EM Algorithm

- fun HDDC models and clusters time series objects through their projections in group-specific functional principal subspaces.
- These group=specific functional principal subspaces are obtained by performing FPCA conditionally on the posterior probabilities t_{ik}
- No discriminative information is lost since the b_k term models the variance outside the subspaces

Hyper-parameters

- We do not estimate d_k or K and they cannot be found from maximizing the likelihood since they control the model complexity
- Can use BIC criterion to select both hyperparameters.

Classifying Observations

- The last step is to add a classification for each observation
- We do this using the maximum a posteriori (MAP) rule
- Assing an observation $\gamma_i \in \mathbb{R}^p$ to the group for which γ_i has the highest posterior probability $P(Z_{ik} = 1|\gamma_i)$
- Assign the observation γ_i to the group with the highest $t_{ik}^{(q_f)}$ where q_f is the last iteration of the algorrith before its convergence.

Convergence

- Since it is an EM-based algorithm we are guaranteed convergence to a local maximum
- Can execute the algorithm several times from random initializations to try and find a global maximum

Examples

Example 1 National Basketball Association (NBA)

- In their paper "Possession Sketches: Mapping NBA Strategies", [Miller and Bornn, 2017] use this idea to cluster similar movement patterns of NBA players to develop a dictionary of actions that players make throughout a possesion
- They then use this dictionary as a vocabulary to do topic analysis on NBA possession.
- They are able to then group together possession with similar offensive structure

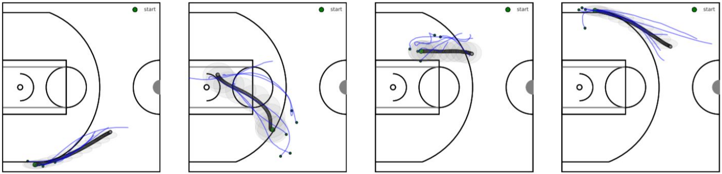

Figure: Example of Movement Clusters in NBA Possessions

Example 2 Canadian Temperatures

- The original paper uses an example with the Canadian temperature data provided in the *fda* package
- Use a basis of 20 natural cubic splines
- Use BIC to select 4 clusters

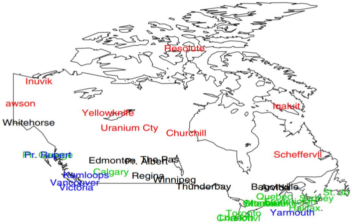

Figure: Colors indicating the cluster membership plotted with respect to geographical positions

R Demo

References

Bouveyron, C. and Jacques, J. (2011).
Model-based clustering of time series in group-specific functional subspaces.
Advances in Data Analysis and Classification, 5(4):281–300.

Miller, A. C. and Bornn, L. (2017).
Possession sketches : Mapping nba strategies.
In *Proceedings of the 2017 MIT Sloan Sports Analytics Conference*.