Hierarchical dynamical mixtures for functional data clustering and segmentation

Faicel Chamroukhi

Joint Statistics Seminar

KU LEUVEN

December 1st, 2016
Temporal data

Temporal data with regime changes

- Data with regime changes over time
- Abrupt and/or smooth regime changes

Objectives

Temporal data modeling and segmentation
Functional data

Many curves to analyze

Railway switch curves

Yeast cell cycle curves

Phonemes curves

Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes ➔ Curve segmentation
Scientific context

■ The area of statistical learning and analysis of complex data.

■ **Data**: Complex data $\rightarrow$ heterogeneous, temporal/dynamical, high-dimensional/functional, incomplete, ...

■ **Objective**: Transform the data into knowledge:
  $\leftarrow$ Reconstruct hidden structure/information, groups/hierarchy of groups, summarizing prototypes, underlying dynamical processes, etc

Modeling framework

■ **Latent variable models**: $f(x|\theta) = \int_z f(x, z|\theta)dz$
  
  **Generative formulation**: $z \sim q(z|\theta)$  
  $x|z \sim f(x|z, \theta)$

$\leftarrow$ **Mixture models**: $f(x|\theta) = \sum_{k=1}^{K} p(z = k)f(x|z = k, \theta_k)$ and extensions
Mixture modeling framework

- Mixture density: \( f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k) \)

- Generative model

\[
\begin{align*}
z & \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K) \\
x|z & \sim f(x|\theta_z)
\end{align*}
\]

→ Algorithms for inferring \( \boldsymbol{\theta} \) from the data
Outline

1. Mixture models for temporal data segmentation

2. Mixture models for functional data analysis
Outline

1. Mixture models for temporal data segmentation
   - Regression with hidden logistic process

2. Mixture models for functional data analysis

Temporal data with regime changes

- Railway data
- Energy data
Mixture models for temporal data segmentation

\[ y = (y_1, \ldots, y_n) \] a time series of \( n \) univariate observations \( y_i \in \mathbb{R} \) observed at the time points \( t = (t_1, \ldots, t_n) \)

### Times series segmentation context

- Time series segmentation is a popular problem with a broad literature
- Common problem for different communities, including statistics, detection, signal processing, machine learning, finance

- The observed time series is generated by an underlying process
  \( \iff \) segmentation \( \equiv \) recovering the parameters the process’ states.

- Conventional solutions are subject to limitations in the control of the transitions between these states

- Propose generative latent data modeling for segmentation and approximation

  \( \iff \) segmentation \( \equiv \) inferring the model parameters and the underlying
Regression with hidden logistic process

Let \( y = (y_1, \ldots, y_n) \) be a time series of \( n \) univariate observations \( y_i \in \mathbb{R} \) observed at the time points \( t = (t_1, \ldots, t_n) \) governed by \( K \) regimes.

**The Regression model with Hidden Logistic Process (RHLP) \([1]\)**

\[
y_i = \beta_{z_i}^T x_i + \sigma_{z_i} \epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad (i = 1, \ldots, n)
\]

\[
Z_i \sim \mathcal{M}(1, \pi_1(t_i; w), \ldots, \pi_K(t_i; w))
\]

Polynomial segments \( \beta_{z_i}^T x_i \) with \( x_i = (1, t_i, \ldots, t_i^p)^T \) with logistic probabilities

\[
\pi_k(t_i; w) = \mathbb{P}(Z_i = k | t_i; w) = \frac{\exp(w_{k1} t_i + w_{k0})}{\sum_{\ell=1}^K \exp(w_{\ell1} t_i + w_{\ell0})}
\]

\[
f(y_i | t_i; \theta) = \sum_{k=1}^K \pi_k(t_i; w) \mathcal{N}(y_i; \beta_k^T x_i, \sigma_k^2)
\]

- Both the mixing proportions and the component parameters are time-varying
- Parameter vector of the model : \( \theta = (w^T, \beta_1^T, \ldots, \beta_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T \)
Modeling with the logistic distribution allows activating simultaneously and preferentially several regimes during time

\[ \pi_k(t_i; w) = \frac{\exp(\lambda_k(t_i + \gamma_k))}{\sum_{\ell=1}^{K} \exp(\lambda_\ell(t_i + \gamma_\ell))} \]

⇒ The parameter \( w_{k1} \) controls the quality of transitions between regimes

⇒ The parameter \( w_{k0} \) is related to the transition time point

Ensure time series segmentation into contiguous segments
$K = 5$ polynomial components of degree $p = 2$
Parameter estimation: MLE via EM: EM-RHLP

- Parameter vector: \( \theta = (w^T, \beta_1^T, \ldots, \beta_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T \)
- Maximize the observed-data log-likelihood:
  \[
  \log L(\theta; \mathbf{y}, \mathbf{t}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}(y_i; \beta_k^T \mathbf{x}_i, \sigma_k^2)
  \]
- Complete-data log-likelihood
  \[
  \log L_c(\theta; \mathbf{y}, \mathbf{t}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log [\pi_k(t_i; \mathbf{w}) \mathcal{N}(y_i; \beta_k^T \mathbf{x}_i, \sigma_k^2)]
  \]
  \(Z_{ik} = 1\) if \(Z_i = k\) (i.e., when \(y_i\) belongs to the \(k\)th component)
- The \(Q\)-function
  \[
  Q(\theta, \theta^{(q)}) = \mathbb{E} \left[ \log L_c(\theta; \mathbf{y}, \mathbf{t}, \mathbf{z}) | \mathbf{y}, \mathbf{t}; \theta^{(q)} \right]
  \]
  \[
  = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \left[ \log \pi_k(t_i; \mathbf{w}) \mathcal{N}(y_i; \beta_k^T \mathbf{x}_i, \sigma_k^2) \right]
  \]
**E-Step:** compute the posterior component memberships:

\[
\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | y_i, t_i; \theta^{(q)}) = \frac{\pi_k(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i, \beta_k^T \mathbf{x}_i, \sigma_k^2)}{\sum_{k=1}^K \pi_k(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i, \beta_k^T \mathbf{x}_i, \sigma_k^2)}.
\]

**M-Step:** compute the parameter update \( \theta^{(q+1)} = \arg \max_{\theta} Q(\theta, \theta^{(q)}) \)

\[
\beta_k^{(q+1)} = \left[ \sum_{i=1}^n \tau_{ik}^{(q)} \mathbf{x}_i \mathbf{x}_i^T \right]^{-1} \sum_{i=1}^n \tau_{ik}^{(q)} y_i \mathbf{x}_i \quad \text{weighted polynomial regression}
\]

\[
\sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^n \tau_{ik}^{(q)}} \sum_{i=1}^n \tau_{ik}^{(q)} (y_i - \beta_k^{T(q+1)} \mathbf{x}_i)^2
\]

\[
\mathbf{w}^{(q+1)} = \arg \max_{\mathbf{w}} \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(q)} \log \pi_k(t_i; \mathbf{w}) \quad \text{weighted logistic regression}
\]
EM-RHLP algorithm

M-Step: Weighted multi-class logistic regression

\[ w^{(q+1)} = \arg \max_w \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(t_i; w) \]

- A convex optimization problem
- Solved with a multi-class Iteratively Reweighted Least Squares (IRLS) algorithm (Newton-Raphson)

\[ w^{(l+1)} = w^{(l)} - \left[ \frac{\partial^2 Q_w(w, \theta^{(q)})}{\partial w \partial w^T} \right]^{-1} \frac{\partial Q_w(w, \theta^{(q)})}{\partial w} \bigg|_{w=w^{(l)}} \]

- Analytic calculation of the Hessian and the gradient
- EM-RHLP algorithm complexity: \( O(I_{EM}I_{IRLS}K^3p^3n) \) (more advantageous than dynamic programming).
Approximation: a prototype mean curve

\[ \hat{y}_i = \mathbb{E}[y_i | t_i; \hat{\theta}] = \sum_{k=1}^{K} \pi_k(t_i; \hat{\mathbf{w}}) \hat{\beta}_k^T \mathbf{x}_i \]

A smooth and flexible approximation thanks to the logistic weights.

The RHLP can be used as nonlinear regression model

\[ y_i = f(t_i; \theta) + \epsilon_i \]

by covering functions of the form

\[ f(t_i; \theta) = \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \beta_k^T \mathbf{x}_i \quad [3] \]

Curve segmentation:

\[ \hat{z}_i = \arg \max_{1 \leq k \leq K} \mathbb{E}[z_i | t_i; \hat{\mathbf{w}}] = \arg \max_{1 \leq k \leq K} \pi_k(t_i; \hat{\mathbf{w}}) \]

Model selection Application of BIC, ICL

\[ \text{BIC}(K, p) = \log L(\hat{\theta}) - \frac{\nu_{\theta} \log(n)}{2}; \quad \text{ICL}(K, p) = \log L_c(\hat{\theta}) - \frac{\nu_{\theta} \log(n)}{2} \]

where \( \nu_{\theta} = K(p + 4) - 2. \)
Approximation error as a function of the speed of transitions

Computing time
Evaluation in approximation and segmentation

varying $m$

varying $\sigma$

Faicel Chamroukhi
Application to real data

![Graphs showing application to real data with frequency on the x-axis and power or impedance on the y-axis.](image_url)
Outline

1 Mixture models for temporal data segmentation

2 Mixture models for functional data analysis
   - Mixture of piecewise regressions
   - Mixture of hidden logistic process regressions
   - Functional discriminant analysis
Functional data analysis context

Many curves to analyze

Railway switch curves

Yeast cell cycle curves

Phonemes curves

Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes → Curve segmentation
Functional data analysis context

Data
- The individuals are entire functions (e.g., curves, surfaces)
- A set of $n$ univariate curves $((x_1, y_1), \ldots, (x_n, y_n))$
- $(x_i, y_i)$ consists of $m_i$ observations $y_i = (y_{i1}, \ldots, y_{im_i})$ observed at the independent covariates, (e.g., time $t$ in time series), $(x_{i1}, \ldots, x_{imi})$

Objectives: exploratory or decisional
1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [4] [9], [C11] [16]
2. Discriminant analysis of functional data: [2], [5]

Functional data clustering/classification tools
- A broad literature (Kmeans-type, Model-based, etc)
  ⇒ Mixture-model based cluster and discriminant analyzes
Mixture modeling framework for functional data

- The functional mixture model:

\[
 f(y|x; \Psi) = \sum_{k=1}^{K} \alpha_k f_k(y|x; \Psi_k)
\]

- \(f_k(y|x)\) are tailored to functional data: can be polynomial (B-)spline regression, regression using wavelet bases etc, or Gaussian process regression, functional PCA

\[\rightarrow\] more tailored to approximate smooth functions

\[\rightarrow\] do not account for segmentation

Here \(f_k(y|x)\) itself exhibits a clustering property via hidden variables (regimes):

1. Riecewise regression model (PWR)
2. Regression model with a hidden process (RHLP)
Piecewise regression mixture model (PWRM) [9]

- A probabilistic version of the $K$-means-like approach of (Hébrail et al., 2010)

$$f(y_i|x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \beta^T_{kr} x_{ij}, \sigma^2_{kr})$$

$I_{kr} = (\xi_{kr}, \xi_{k,r+1}]$ are the element indexes of segment $r$ for component $k$

- Simultaneously accounts for curve clustering and segmentation
- Parameter vector $\Psi = (\alpha_1, \ldots, \alpha_{K-1}, \theta^T_1, \ldots, \theta^T_K, \xi^T_1, \ldots, \xi^T_K)^T$ with $\theta_k = (\beta^T_{k1}, \ldots, \beta^T_{kR_k}, \sigma^2_{k1}, \ldots, \sigma^2_{kR_k})^T$ and $\xi_k = (\xi_{k1}, \ldots, \xi_{k,R_k+1})^T$

Parameter estimation

1. Maximum likelihood estimation: EM-PWRM
2. Maximum classification likelihood estimation: CEM-PWRM
Maximum likelihood estimation via EM: EM-PWRM

- Maximize the observed-data log-likelihood:

\[
\log L(\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \beta_{kr}^T x_{ij}, \sigma_{kr}^2)
\]

- The complete-data log-likelihood

\[
\log L_c(\Psi, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} Z_{ik} \log \mathcal{N}(y_{ij}; \beta_{kr}^T x_{ij}, \sigma_{kr}^2)
\]

- The conditional expected complete-data log-likelihood

\[
Q(\Psi, (\Psi)^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} \tau_{ik}^{(q)} \log \mathcal{N}(y_{ij}; \beta_{kr}^T x_{ij}, \sigma_{kr}^2)
\]
EM-PWRM algorithm

**E-step:** Compute the $Q$-function

$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|y_i, x_i; \Psi^{(q)}) = \frac{\alpha_k^{(q)} f_k(y_i|x_i; \Psi_k^{(q)})}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} f_{k'}(y_i|x_i; \Psi_{k'}^{(q)})}$

**M-step:** Compute the update $\Psi^{(q+1)} = \arg\max_{\Psi} Q(\Psi, \Psi^{(q)})$

- $\alpha_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)}$, $(k = 1, \ldots, K)$
- Maximization w.r.t the piecewise regression parameters $\{\xi_{kr}, \beta_{kr}, \sigma_{kr}^2\} \rightarrow$ a weighted piecewise regression problem \rightarrow dynamic programming:

$$\beta_{kr}^{(q+1)} = \left[ \sum_{i=1}^{n} \tau_{ik}^{(q)} X_{ir}^T X_{ir} \right]^{-1} \sum_{i=1}^{n} X_{ir} y_{ir}$$

$$\sigma_{kr}^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j \in I_{kr}^{(q)}} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \| y_{ir} - X_{ir} \beta_{kr}^{(q+1)} \|^2$$

$y_{ir}$ are the observations of segment $r$ of the $i$th curve and $X_{ir}$ its design matrix
Maximum classification likelihood estimation: CEM-PWRM

- Maximize the complete-data log-likelihood w.r.t \((\Psi, z)\) simultaneously
- C-step: Bayes’ optimal allocation rule: 
  \[
  \hat{z}_i = \arg\max_{1 \leq k \leq K} \tau_{ik}(\hat{\Psi})
  \]

CEM-PWRM is equivalent to the \(K\)-means-like algorithm of Hébrail et al. (2010):

\[
\log L_c(z, \Psi) \propto \mathcal{J}(z, \{\mu_{kr}, I_{kr}\}) = \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{i | Z_i = k} \sum_{j \in I_{kr}} (y_{ij} - \mu_{kr})^2
\]

if the following conditions hold:

- \(\alpha_k = \frac{1}{K} \) \(\forall K\) (identical mixing proportions);
- \(\sigma_{kr}^2 = \sigma^2 \) \(\forall r\) and \(\forall k\); (isotropic and homoskedastic model);
- \(\mu_{kr}\): piecewise constant regime approximation

- Curve clustering: 
  \[
  \hat{z}_i = \arg\max_k \tau_{ik}(\hat{\Psi}) \text{ with } \tau_{ik}(\hat{\Psi}) = \mathbb{P}(Z_i | x_i, y_i; \hat{\Psi})
  \]
- Model selection: Application of BIC, ICL
- Complexity in \(O(I_{EM}KRnm^2p^3)\): Significant computational load for large \(m\)
Simulation results

Figure: Misclassification error rate versus the noise level variation.
Application to switch operation curves

Data set: \( n = 146 \) real curves of \( m = 511 \) observations.
Each curve is composed of \( R = 6 \) electromechanical phases (regimes)
The Tecator data set\(^1\) contains \(n = 240\) spectra with \(m = 100\) observations for each spectrum. Data considered in the same setting as in Hébrail et al. (2010) (six clusters, each cluster is approximated by five linear segments (\(R = 5, p = 1\))).

\(^1\)Tecator data are available at [http://lib.stat.cmu.edu/datasets/tecator](http://lib.stat.cmu.edu/datasets/tecator).
Figure: Clusters and the corresponding piecewise prototypes for each cluster obtained with the CEM-PWRM algorithm for the Tecator data set.
Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data\(^2\) contains \(n = 472\) waveforms of the measured echoes, sampled at \(m = 70\) (number of echoes)

We considered the same number of clusters (twenty) and a piecewise linear approximation of four segments per cluster as in Hébrail et al. (2010).

CEM-PWRM clustering
Summary

- Probabilistic approach to the simultaneous curve clustering and optimal segmentation
- Two algorithms: EM-PWRM and CEM-PWRM
- CEM-PWRM is a probabilistic-based version of the $K$-means-like algorithm Hébrail et al. (2010)

- If the aim is density estimation, the EM version is suggested (CEM provides biased estimators but is well-tailored to the segmentation/clustering end)
- For continuous functions the PWRM in its current formulation, may lead to discontinuities between segments for the piecewise approximation.
- This may be avoided by posterior interpolation as in Hébrail et al. (2010).
- May lead to significant computational load especially for large time series. However, for quite reasonable dimensions, the algorithms remain usable
The mixture of regressions with hidden logistic processes (MixRHLP):

\[
f(y_i | x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2)
\]

\[
\pi_{kr}(x_j; w_k) = \mathbb{P}(H_{ij} = r | Z_i = k, x_j; w_k) = \frac{\exp (w_{kr0} + w_{kr1} x_j)}{\sum_{r' = 1}^{R_k} \exp (w_{kr'0} + w_{kr'1} x_j)}
\]

Two types of component memberships:

\(\leftrightarrow\) cluster memberships (global) \(Z_{ik} = 1\) iff \(Z_i = k\)

\(\leftrightarrow\) regime memberships for a given cluster (local): \(H_{ijr} = 1\) iff \(H_{ij} = r\)

MixRHLP deals better with the quality of regime changes

Parameter estimation via the EM algorithm: EM-MixRHLP
MLE estimation via the EM algorithm

- The observed-data log-likelihood

\[ \log L(\Psi) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2) \right) \]

- The complete-data log-likelihood:

\[ \log L_c(\Psi) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} Z_{ik} H_{ijr} \log \left[ \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2) \right]\]

- The conditional expected complete-data log-likelihood

\[ Q(\Psi, \Psi^{(q)}) = \mathbb{E} \left[ \log L_c(\Psi) | D; \Psi^{(q)} \right] \]

\[ = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \tau_{ik}^{(q)} \gamma_{ijr}^{(q)} \log \left[ \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2) \right] \]
EM-MixRHLP algorithm

E-step

- The posterior cluster memberships:

\[ \tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | y_i, x_i; \Psi_k^{(q)}) = \frac{\alpha_k^{(q)} f(y_i | Z_i = k, x_i; \Psi_k^{(q)})}{\sum_{k' = 1}^{K} \alpha_{k'}^{(q)} f(y_i | Z_i = k', x_i; \Psi_{k'}^{(q)})} \]

- the posterior regime memberships:

\[ \gamma_{ijr}^{(q)} = \mathbb{P}(H_{ij} = r | Z_i = k, y_{ij}, t_j; \Psi_k^{(q)}) = \frac{\pi_k r(x_j; \mathbf{w}_k^{(q)}) N(y_{ij}; \beta_{kr}^{T(q)} \mathbf{x}_j, \sigma_{kr}^{2(q)})}{\sum_{r' = 1}^{R_k} \pi_k r'(x_j; \mathbf{w}_k^{(q)}) N(y_{ij}; \beta_{kr'}^{T(q)} \mathbf{x}_j, \sigma_{kr'}^{2(q)})} \]

Computed directly (i.e., without a forward-backward recursion as in the Markovian model).
M-step of the EM-MixRHLP

**M-step:** calculate the update $\Psi^{(q+1)} = \arg \max_{\Psi} Q(\Psi, \Psi^{(q)})$.

- **Mixing proportions update:** standard

  $$\alpha_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(q)}, \quad (k = 1, \ldots, K).$$

- **Regression parameters update:** Analytic weighted least-squares problems

  $$\beta_{kr}^{(q+1)} = \left[ \sum_{i=1}^{n} \tau_{ik}^{(q)} X_i^T W_{ikr}^{(q)} X_i \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} X_i^T W_{ikr}^{(q)} y_i,$$

  $$\sigma_{kr}^{2(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \| \sqrt{W_{ikr}^{(q)}} (y_i - X_i \beta_{kr}^{(q+1)})\|^2}{\sum_{i=1}^{n} \tau_{ik}^{(q)} \text{trace}(W_{ikr}^{(q)})},$$

  where $W_{ikr}^{(q)} = \text{diag}(\gamma_{ijr}^{(q)}; j = 1, \ldots, m_i)$.

- **Maximization w.r.t the logistic processes’ parameters $\{w_k\}$:** solving multinomial logistic regression problems $\Rightarrow$ IRLS

  $\Leftarrow$ EM-MixRHLP has complexity in $O(I_{EM} I_{IRLS} K R^3 n m p^3)$ ($K$-means like algo. for PWR is in $O(I_{KM} K R n m^2 p^3)$ $\Leftarrow$ computationally attractive for large $m$ with moderate value of $R$.)

---

**Curve approximation, segmentation and model selection**

- Each cluster $k$ is summarized by approximating it by a single “mean” curve, which we denote by $\hat{y}_k$. Each point $\hat{y}_{kj}$ of this mean curve is defined by the conditional expectation $\hat{y}_{kj} = \mathbb{E}[y_{ij} | Z_i = k, t_j; \Psi_k]$ given by:

  $$\hat{y}_{kj} = \sum_{R_k r = 1}^{R_k} \pi_{kr} (x_j; \hat{w}_k) \hat{\beta}_{kr}^T x_j$$

  which is a sum of polynomials weighted by the logistic probabilities $\pi_{kr}$ that model the regime variability over time and which constitutes a smooth flexible approximation.

- The number of mixture components $K$, the number regimes $R_k$ and the polynomial degree $p$ can be estimated by maximizing some information criteria such the BIC. The number of free parameters of the MixRHLP model $\nu_\Psi = K - 1 + \sum_{k=1}^{K} \nu_{\Psi_k}$ with $\nu_{\Psi_k} = (p + 4) R_k - 2$ represents the number of free parameters of each RHLP component.
EM-MixRHLP clustering of simulated data

Faicel Chamroukhi
Hierarchical dynamical mixtures for functional data clustering and segmentation
Clustering switch operations

Clustering real curves of switch operations The data set contains 115 curves of $R = 6$ operations electromechanical process $K = 2$ clusters: operating state without/with possible defect
Clustering switch operations

Clustering real curves of switch operations The data set contains 115 curves of $R = 6$ operations electromechanical process $K = 2$ clusters: operating state without/with possible defect
Functional discriminant analysis

Supervised classification context

- Data: a training set of labeled functions \(((x_1, y_1, c_1), \ldots, (x_n, y_n, c_n))\)
  where \(c_i \in \{1, \ldots, G\}\) is the class label of the \(i\)th curve
- Problem: predict the class label \(c_i\) for a new unlabeled function \((x_i, y_i)\)

Tool: Discriminant analysis

Use the Bayes’ allocation rule

\[
\hat{c}_i = \arg \max_{1 \leq g \leq G} \frac{\mathbb{P}(C_i = g) f(y_i | x_i; \Psi_g)}{\sum_{g'=1}^G \mathbb{P}(C_i = g') f(y_i | x_i; \Psi_{g'})},
\]

based on a generative model \(f(y_i | x_i; \Psi_g)\) for each group \(g\)

- Homogeneous classes: Functional Linear Discriminant Analysis [8]
Applications to switch curves

<table>
<thead>
<tr>
<th>Approach</th>
<th>Classification error rate (%)</th>
<th>Intra-class inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLDA-PR</td>
<td>11.5</td>
<td>$10.7350 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-SR</td>
<td>9.53</td>
<td>$9.4503 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-RHLP</td>
<td>8.62</td>
<td>$8.7633 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-PRM</td>
<td>9.02</td>
<td>$7.9450 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-SRM</td>
<td>8.50</td>
<td>$5.8312 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-MixRHLP</td>
<td><strong>6.25</strong></td>
<td><strong>3.2012 \times 10^9</strong></td>
</tr>
</tbody>
</table>
A full generative model for curve clustering and segmentation

The segmentation is smoothly controlled by logistic functions

An alternative to the previously described mixture of piecewise regressions

more advantageous compared to approaches involving dynamic programming namely when using piecewise regression especially for large samples.

Could be extended to the multivariate case without a major effort
Some ongoing research and perspectives

- Model-based co-clustering for high-dimensional functional data

**Functional latent block model (FLBM)** available soon on arXiv

Data: \( Y = (y_{ij}) \): \( n \) individuals defined on a set \( I \) with \( d \) continuous functional variables defined on a set \( J \) where \( y_{ij}(t) = \mu(x_{ij}(t); \beta) + \varepsilon(t) \), \( t \) defined on \( T \).

- FLBM model:

\[
f(Y|X; \Psi) = \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} P(Z, W)f(Y|X, Z, W; \theta) = \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_{\ell}^{w_{j\ell}} \prod_{i,j,k,\ell} f(y_{ij}|x_{ij}; \theta_{k\ell})^{z_{ik} w_{j\ell}}.
\]

- An RHLP is used as a conditional block distribution \( f(y_{ij}|x_{ij}; \theta_{k\ell}) \)

- Model inference using Stochastic EM
Some ongoing research and perspectives

MixtComp Software

Mixtures for massive data

- Mixture density estimation for massive data clustering
- Use ensemble methods to distribute the data
  - Bag of Little Boostraps (BLB) (Kleiner et al., 2014)
  - Aggregate local estimators from BLB sub-samples: Hierarchical (mixture) of experts aggregation
References


[15] F. Chamroukhi. Robust mixture of experts modeling using the skew-$t$ distribution. 2015d. under review
References I


F. Chamroukhi. Robust mixture of experts modeling using the skew-$t$ distribution. 2015d. under review.


References II


Thank you for your attention!
Identifiability of the RHLP model

- \( f(\cdot; \Psi) \) is identifiable when \( f(\cdot; \Psi) = f(\cdot; \Psi^*) \) if and only if \( \Psi = \Psi^* \).

- via Lemma 2 of Jiang and Tanner (1999) for Mixture of Experts, we have any ordered and initialized irreducible RHLP is identifiable (up to a permutation).

- Ordered implies that there exist a certain ordering relationship such that
  \[
  (\beta_1^T, \sigma_1^2)^T < \ldots < (\beta_K^T, \sigma_K^2)^T;
  \]

- initialized implies that \((w_{K0}, w_{K1}) = (0, 0)\)

- irreducible implies that if \( k \neq k' \), then one of the following conditions holds:
  \( \beta_k \neq \beta_{k'} \) or \( \sigma_k \neq \sigma_{k'} \)

- The set \( \{N(y; \mu(x; \beta_1), \sigma_1^2), \ldots, N(y; \mu(x; \beta_{2K}), \sigma_{2K}^2)\} \) contains \( 2K \) linearly independent functions of \( y \), for any \( 2K \) distinct pair \((\beta_k, \sigma_k^2)\) for \( k = 1, \ldots, 2K \).