Hierarchical dynamical mixtures for functional data clustering and segmentation

Faicel Chamroukhi

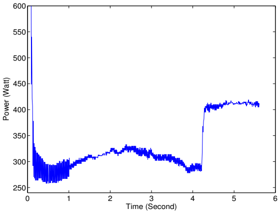
Joint Statistics Seminar

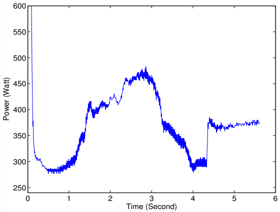
KU LEUVEN

December 1st, 2016

Temporal data

Temporal data with regime changes





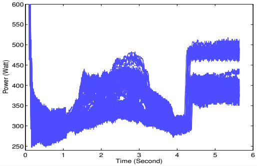
- Data with regime changes over time
- Abrupt and/or smooth regime changes

Objectives

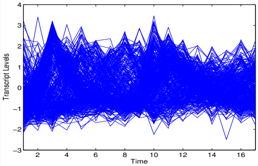
Temporal data modeling and segmentation

Functional data

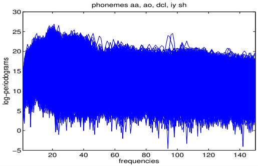
Many curves to analyze



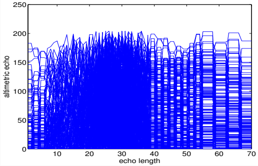
Railway switch curves



Yeast cell cycle curves



Phonemes curves



Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes $\hookrightarrow$ Curve segmentation

Scientific context

- The area of statistical learning and analysis of complex data.
- **Data :** Complex data $\hookrightarrow$ *heterogeneous, temporal/dynamical, high-dimensional/functional, incomplete,...*
- **Objective:** Transform the data into knowledge :
  $\hookrightarrow$ Reconstruct hidden structure/information, groups/hierarchy of groups, summarizing prototypes, underlying dynamical processes, etc

Modeling framework

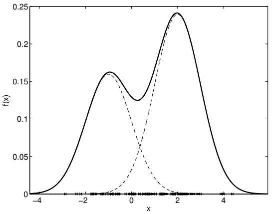
- Latent variable models : $f(x|\boldsymbol{\theta}) = \int_z f(x, z|\boldsymbol{\theta})\mathrm{d}z$
  Generative formulation :
  $$\begin{aligned} z &\sim q(z|\boldsymbol{\theta}) \\ x|z &\sim f(x|z, \boldsymbol{\theta}) \end{aligned}$$

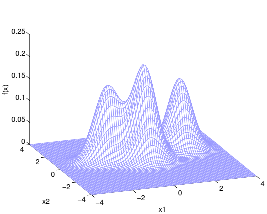
$\hookrightarrow$ Mixture models : $f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{P}(z = k) f(x|z = k, \boldsymbol{\theta}_k)$ and extensions

Mixture modeling framework

Mixture modeling framework

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





- Generative model

$$
\begin{aligned}
z &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_K) \\
x|z &\sim f(x|\boldsymbol{\theta}_z)
\end{aligned}
$$

$\hookrightarrow$ Algorithms for inferring $\boldsymbol{\theta}$ from the data

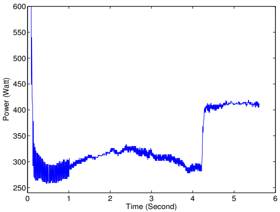
Outline

Outline

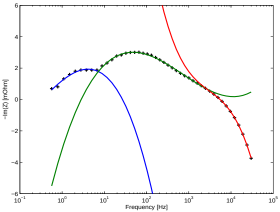
1. Mixture models for temporal data segmentation
   - Regression with hidden logistic process

2. Mixture models for functional data analysis

Temporal data with regime changes



Railway data



Energy data

Mixture models for temporal data segmentation

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

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
  $\hookrightarrow$ segmentation $\equiv$ recovering the parameters the process' states.
- Conventional solutions are subject to limitations in the control of the transitions between these states

- $\hookrightarrow$ Propose generative latent data modeling for segmentation and approximation
- $\hookrightarrow$ segmentation $\equiv$ inferring the model parameters and the underlying

Regression with hidden logistic process

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

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

$$\begin{array}{rcl} y_i & = & \boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i + \sigma_{z_i} \epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0,1), \quad (i = 1, \ldots, n) \\ Z_i & \sim & \mathcal{M}(1, \pi_1(t_i; \mathbf{w}), \ldots, \pi_K(t_i; \mathbf{w})) \end{array}$$

Polynomial segments $\boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i$ with $\boldsymbol{x}_i = (1, t_i, \ldots, t_i^p)^T$ with logistic probabilities

$$\pi_k(t_i; \mathbf{w}) = \mathbb{P}(Z_i = k | t_i; \mathbf{w}) = \frac{\exp\left(w_{k1} t_i + w_{k0}\right)}{\sum_{\ell=1}^{K} \exp\left(w_{\ell 1} t_i + w_{\ell 0}\right)}$$

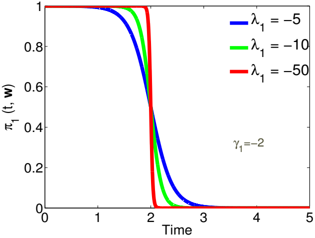
$$f(y_i | t_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)$$

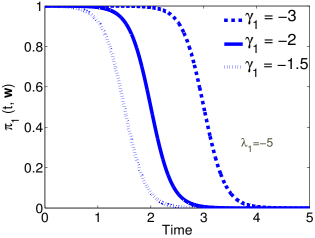
- Both the mixing proportions and the component parameters are time-varying
- Parameter vector of the model :$\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T$

Illustration

- Modeling with the logistic distribution allows activating simultaneously and preferentially several regimes during time

$$\pi_k(t_i; \mathbf{w}) = \frac{\exp\left(\lambda_k(t_i + \gamma_k)\right)}{\sum_{\ell=1}^{K} \exp\left(\lambda_\ell(t_i + \gamma_\ell)\right)}$$



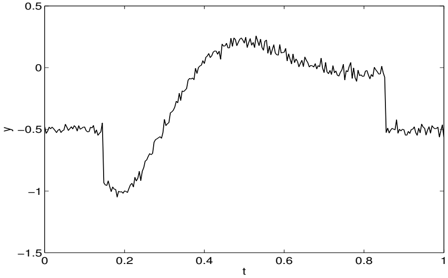


⇒ The parameter w_{k1} controls the quality of transitions between regimes

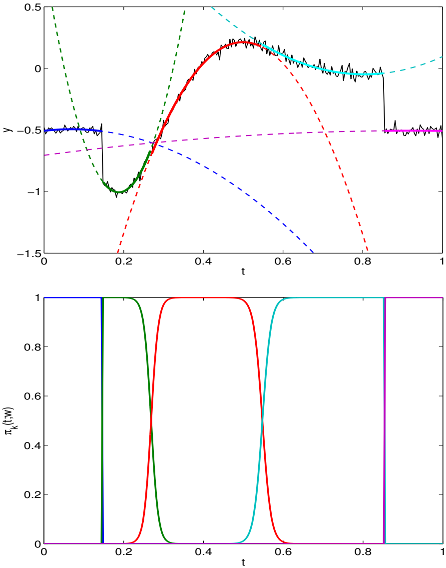
⇒ The parameter w_{k0} is related to the transition time point

- Ensure time series segmentation into contiguous segments

Illustration



Illustration



$K = 5$ polynomial components of degree $p = 2$

Parameter estimation: MLE via EM: EM-RHLP

- Parameter vector: $\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T$
- Maximize the observed-data log-likelihood:

$$\log L(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}\big(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\big)$$

- Complete-data log-likelihood

$$\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log[\pi_k(t_i; \mathbf{w}) \mathcal{N}\big(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\big)]$$

$Z_{ik} = 1$ if $Z_i = k$ (i.e., when y_i belongs to the kth component)

- The Q-function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathbb{E}\left[\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) | \boldsymbol{y}, \mathbf{t}; \boldsymbol{\theta}^{(q)}\right]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \left[\log \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)\right]$$

EM-RHLP

- **E-Step**: compute the posterior component memberships:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|y_i, t_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t_i; \mathbf{w}^{(q)})\mathcal{N}(y_i; \boldsymbol{\beta}_k^{T(q)}\boldsymbol{x}_i, \sigma_k^{2(q)})}{\sum_{\ell=1}^{K} \pi_\ell(t_i; \mathbf{w}^{(q)})\mathcal{N}(y_i; \boldsymbol{\beta}_\ell^{T(q)}\boldsymbol{x}_i, \sigma_\ell^{2(q)})}\cdot$$

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

$$\boldsymbol{\beta}_k^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} y_i \boldsymbol{x}_i$$ weighted polynomial regression

$$\sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{\beta}_k^{T(q+1)} \boldsymbol{x}_i)^2$$

$$\mathbf{w}^{(q+1)} = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(t_i; \mathbf{w})$$ weighted logistic regression

EM-RHLP algorithm

M-Step: Weighted multi-class logistic regression

$$\mathbf{w}^{(q+1)} = \arg\max_{\mathbf{w}} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(t_i;\mathbf{w})$$

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

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \left[\frac{\partial^2 Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial\mathbf{w}\partial\mathbf{w}^T}\right]^{-1}_{\mathbf{w}=\mathbf{w}^{(l)}} \frac{\partial Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial\mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(l)}}$$

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

Time series approximation and segmentation

1. Approximation: a prototype mean curve

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

↪ A smooth and flexible approximation thanks to the the logistic weights

↪ The RHLP can be used as nonlinear regression model $y_i = f(t_i;\boldsymbol{\theta}) + \epsilon_i$ by covering functions of the form $f(t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ [3]

2. Curve segmentation:

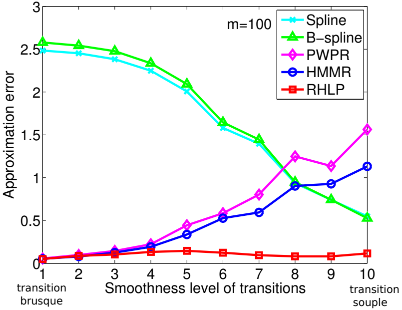
$$\hat{z}_i = \arg\max_{1\leq k\leq K} \mathbb{E}[z_i|t_i;\hat{\mathbf{w}}] = \arg\max_{1\leq k\leq K} \pi_k(t_i;\hat{\mathbf{w}})$$

3. **Model selection** Application of BIC, ICL

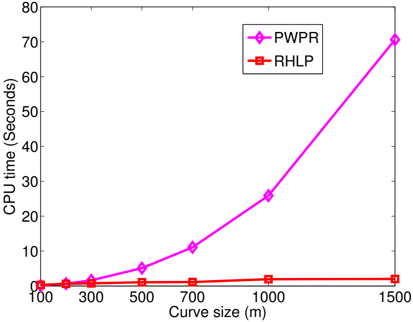
   BIC$(K,p) = \log L(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$; ICL$(K,p) = \log L_c(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$ where $\nu_{\boldsymbol{\theta}} = K(p+4) - 2$.

Evaluation in modeling and segmentation

Approximation error as a function of the speed of transitions

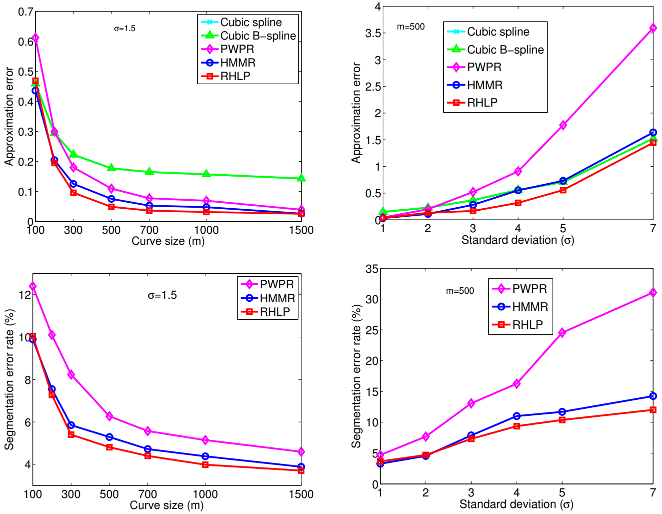


Computing time

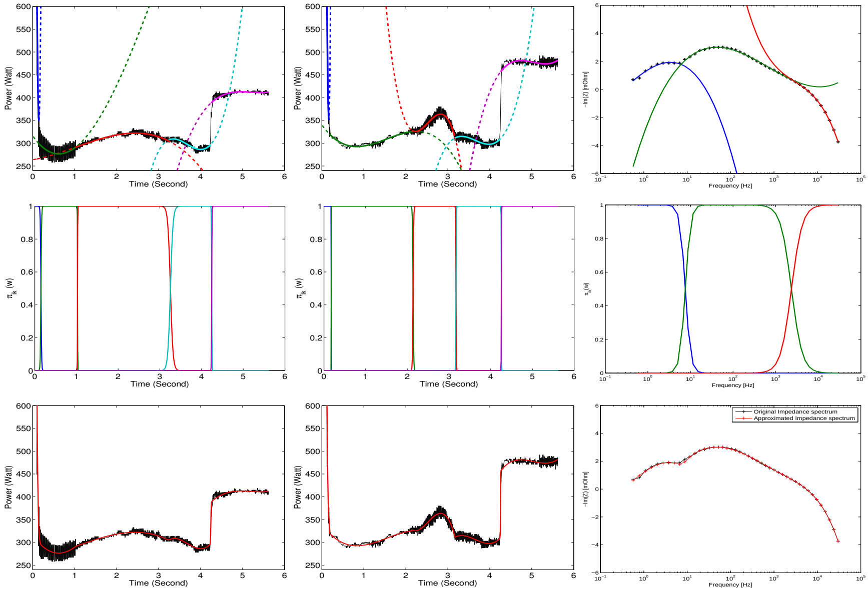


Evaluation in approximation and segmentation

varying m | varying σ



Application to real data



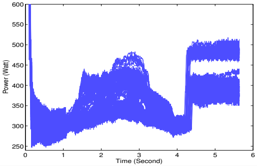
Outline

1. Mixture models for temporal data segmentation

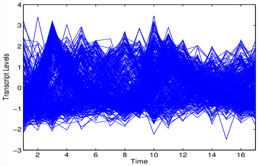
2. Mixture models for functional data analysis
    - Mixture of piecewise regressions
    - Mixture of hidden logistic process regressions
    - Functional discriminant analysis

Functional data analysis context

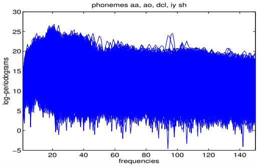
Many curves to analyze



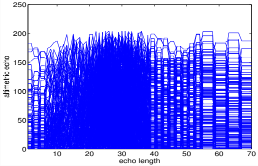
Railway switch curves



Yeast cell cycle curves



Phonemes curves



Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes $\hookrightarrow$ Curve segmentation

Functional data analysis context

Data

- The individuals are entire functions (e.g., curves, surfaces)
- A set of n univariate curves $((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$
- $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ consists of m_i observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ observed at the independent covariates, (e.g., time t in time series), $(x_{i1}, \ldots, x_{im_i})$

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)

  $\Rightarrow$ Mixture-model based cluster and discriminant analyzes

Mixture modeling framework for functional data

- The functional mixture model:

$$f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Psi}_k)$$

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

  $\hookrightarrow$ more tailored to approximate smooth functions

  $\hookrightarrow$ do not account for segmentation

Here $f_k(y|\boldsymbol{x})$ 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)

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}) = \sum_{k=1}^{K}\alpha_k \underbrace{\prod_{r=1}^{R_k}\prod_{j\in I_{kr}}\mathcal{N}(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_{ij},\sigma_{kr}^2)}_{\text{PWR}}$$

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

- $\hookrightarrow$ Simultaneously accounts for curve clustering and segmentation
- Parameter vector $\boldsymbol{\Psi} = (\alpha_1,\ldots,\alpha_{K-1},\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_K^T,\boldsymbol{\xi}_1^T,\ldots,\boldsymbol{\xi}_K^T)^T$ with $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_{k1}^T,\ldots,\boldsymbol{\beta}_{kR_k}^T,\sigma_{k1}^2,\ldots,\sigma_{kR_k}^2)^T$ and $\boldsymbol{\xi}_k = (\xi_{k1},\ldots,\xi_{k,R_k+1})^T$

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:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2\right)$$

- The complete-data log-likelihood

$$\log L_c(\boldsymbol{\Psi}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} Z_{ik} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2)$$

- The conditional expected complete-data log-likelihood

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} \tau_{ik}^{(q)} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2)$$

EM-PWRM algorithm

E-step: Compute the Q−function

$\hookrightarrow$ Compute the posterior probability that the ith curve belongs to component k:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(q)}) = \frac{\alpha_k^{(q)} f_k\left(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}_k^{(q)}\right)}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} f_{k'}\left(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}_{k'}^{(q)}\right)}$$

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

- $\alpha_k^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n}, \quad (k = 1, \ldots, K)$
- maximization w.r.t the piecewise regression parameters $\{\boldsymbol{\xi}_{kr}, \boldsymbol{\beta}_{kr}, \sigma^2_{kr}\} \hookrightarrow$ a weighted piecewise regression problem $\hookrightarrow$ dynamic programming:

$$\boldsymbol{\beta}_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_{ir}^T \mathbf{X}_{ir}\Big]^{-1} \sum_{i=1}^{n} \mathbf{X}_{ir} \boldsymbol{y}_{ir}$$

$$\sigma_{kr}^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j \in I_{kr}^{(q)}} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \|\boldsymbol{y}_{ir} - \mathbf{X}_{ir} \boldsymbol{\beta}_{kr}^{(q+1)}\|^2$$

$\boldsymbol{y}_{ir}$ are the observations of segment r of the ith curve and $\mathbf{X}_{ir}$ its design matrix

Maximum classification likelihood estimation: CEM-PWRM

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

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

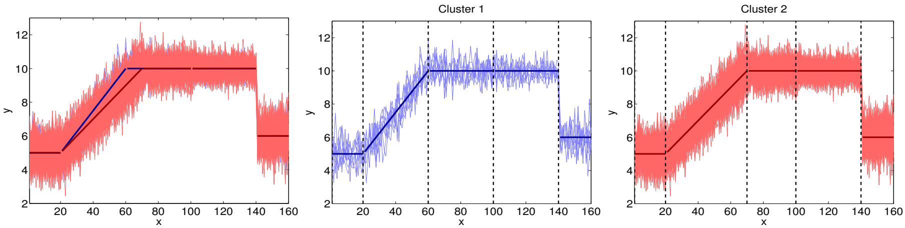
$$\log L_c(\mathbf{z}, \boldsymbol{\Psi}) \propto \mathcal{J}\big(\mathbf{z}, \{\mu_{kr}, I_{kr}\}\big) = \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{i|Z_i=k} \sum_{j \in I_{kr}} \big(y_{ij} - \mu_{kr}\big)^2$$

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);
- μ_{kr}: piecewise *constant* regime approximation

- Curve clustering: $\hat{z}_i = \arg\max_k \tau_{ik}(\hat{\boldsymbol{\Psi}})$ with $\tau_{ik}(\hat{\boldsymbol{\Psi}}) = \mathbb{P}(Z_i | \boldsymbol{x}_i, \boldsymbol{y}_i; \hat{\boldsymbol{\Psi}})$
- Model selection: Application of BIC, ICL
- Complexity in $\mathcal{O}(I_{\mathsf{EM}} K R n m^2 p^3)$: Significant computational load for large m

Simulation results



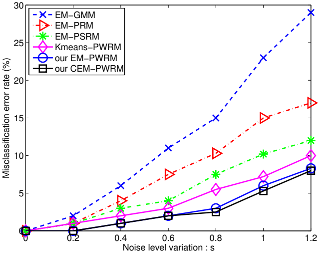
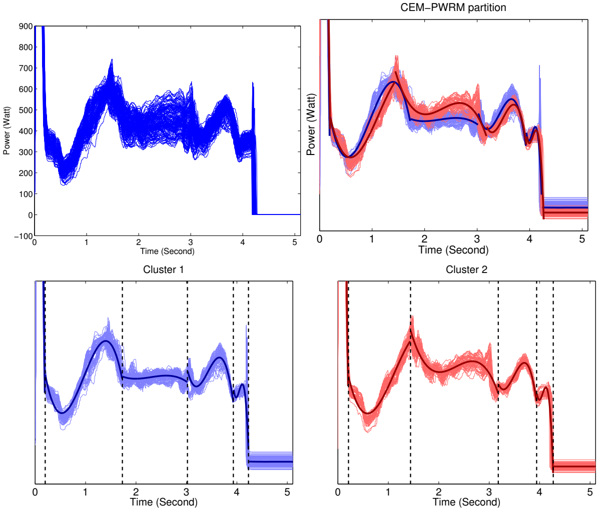


Figure: Misclassification error rate versus the noise level variation.

Application to switch operation curves

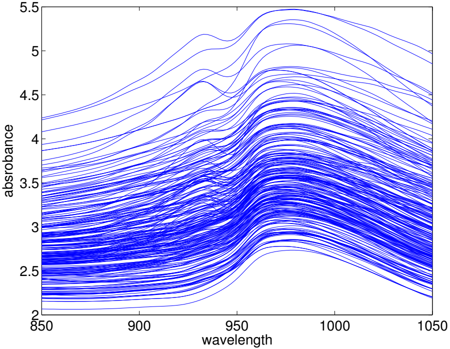
Data set: $n = 146$ real curves of $m = 511$ observations.
Each curve is composed of $R = 6$ electromechanical phases (regimes)



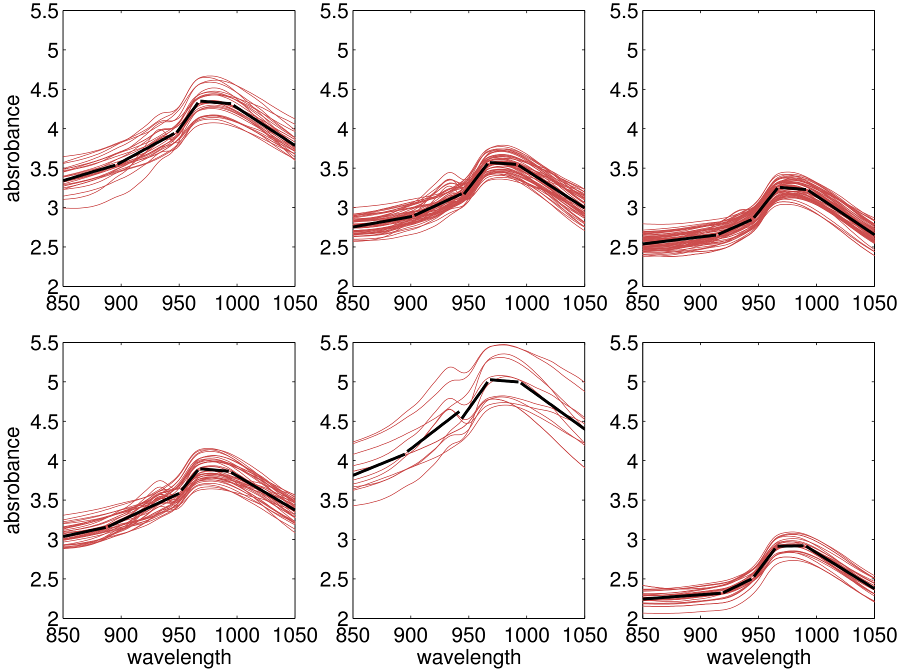
Application to Tecator data

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 ($R = 5, p = 1$))

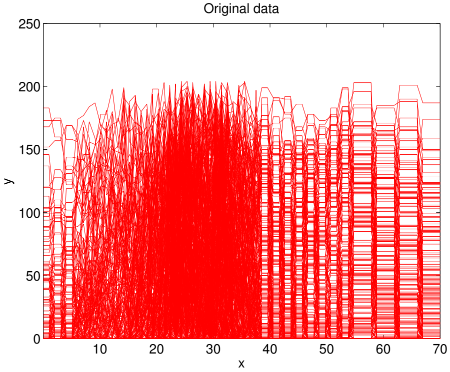


[1] Tecator data are available at `http://lib.stat.cmu.edu/datasets/tecator`.


absrobance
wavelength
5.5
5
4.5
4
3.5
3
2.5
2
850
900
950
1000
1050

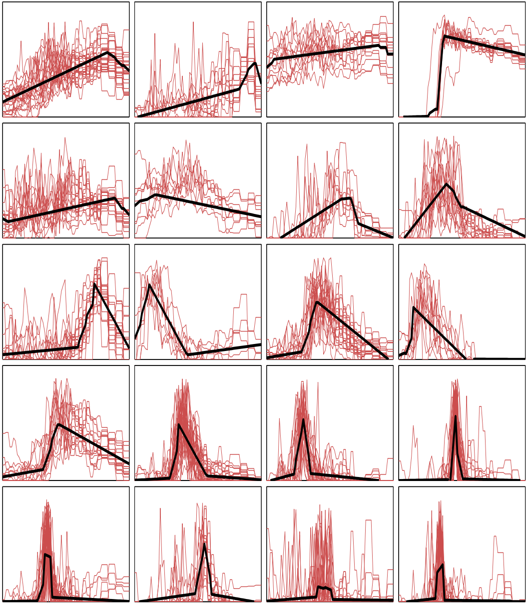
Topex/Poseidon satellite data

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).



[2]Satellite data are available at
http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html.

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

Mixture of hidden logistic process regressions [4]

- The mixture of regressions with hidden logistic processes (MixRHLP):

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \underbrace{\prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\left(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\right)}_{\text{RHLP}}$$

$$\pi_{kr}(x_j;\mathbf{w}_k) = \mathbb{P}(H_{ij}=r|Z_i=k,x_j;\mathbf{w}_k) = \frac{\exp\left(w_{kr0}+w_{kr1}x_j\right)}{\sum_{r'=1}^{R_k}\exp\left(w_{kr'0}+w_{kr'1}x_j\right)},$$

- Two types of component memberships:

  ↪ cluster memberships (global) $Z_{ik}=1$ iff $Z_i=k$

  ↪ 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

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\right)$$

- The complete-data log-likelihood:

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} Z_{ik} H_{ijr} \log \left[\pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\right)\right]$$

- The conditional expected complete-data log-likelihood

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) &= \mathbb{E}\left[\log L_c(\boldsymbol{\Psi}) \middle| \mathcal{D}; \boldsymbol{\Psi}^{(q)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \tau_{ik}^{(q)} \gamma_{ijr}^{(q)} \log \left[\pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\right)\right]. \end{aligned}$$

EM-MixRHLP algorithm

E-step

- The posterior cluster memberships:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)}) = \frac{\alpha_k^{(q)} f(\boldsymbol{y}_i|Z_i = k, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)})}{\sum_{k\prime=1}^{K} \alpha_{k'}^{(q)} f(\boldsymbol{y}_i|Z_i = k\prime, \boldsymbol{x}_i; \boldsymbol{\Psi}_{k\prime}^{(q)})}$$

- the posterior regime memberships:

$$\gamma_{ijr}^{(q)} = \mathbb{P}(H_{ij} = r|Z_i = k, y_{ij}, t_j; \boldsymbol{\Psi}_k^{(q)}) = \frac{\pi_{kr}(x_j; \mathbf{w}_k^{(q)})\mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^{T(q)}\boldsymbol{x}_j, \sigma_{kr}^{2(q)})}{\sum_{r\prime=1}^{R_k} \pi_{kr\prime}(x_j; \mathbf{w}_k^{(q)})\mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr\prime}^{T(q)}\boldsymbol{x}_j, \sigma_{kr\prime}^{2(q)})}$$

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 $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$.

- 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

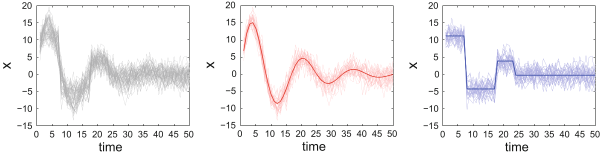
$$\boldsymbol{\beta}_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \mathbf{X}_i\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \boldsymbol{y}_i,$$

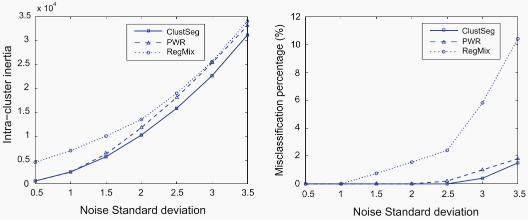
$$\sigma_{kr}^{2\;(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \|\sqrt{\mathbf{W}_{ikr}^{(q)}}(\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{\beta}_{kr}^{(q+1)})\|^2}{\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathrm{trace}(\mathbf{W}_{ikr}^{(q)})},$$

where $\mathbf{W}_{ikr}^{(q)} = \mathrm{diag}(\gamma_{ijr}^{(q)}; j = 1, \ldots, m_i)$.

- Maximization w.r.t the logistic processes' parameters $\{\mathbf{w}_k\}$: solving multinomial logistic regression problems $\Rightarrow$ IRLS
- $\hookrightarrow$ EM-MixRHLP has complexity in $\mathcal{O}(I_{\text{EM}} I_{\text{IRLS}} K R^3 n m p^3)$ (K-means like algo. for PWR is in $\mathcal{O}(I_{\text{KM}} K R n m^2 p^3)$ $\hookrightarrow$ computationally attractive for large m with moderate value of R.

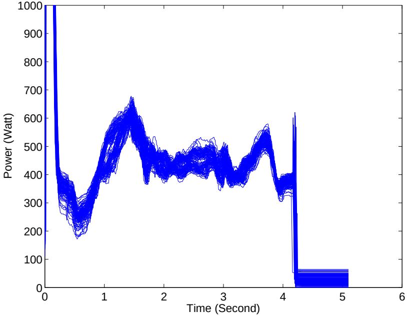
EM-MixRHLP clustering of simulated data





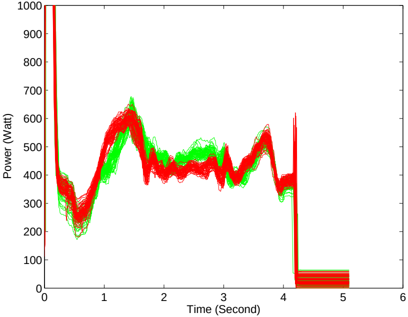
Clustering switch operations

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 $((\boldsymbol{x}_1, \boldsymbol{y}_1, c_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n, c_n))$ where $c_i \in \{1, \ldots, G\}$ is the class label of the ith curve
- Problem: predict the class label c_i for a new unlabeled function $(\boldsymbol{x}_i, \boldsymbol{y}_i)$

Tool: Discriminant analysis

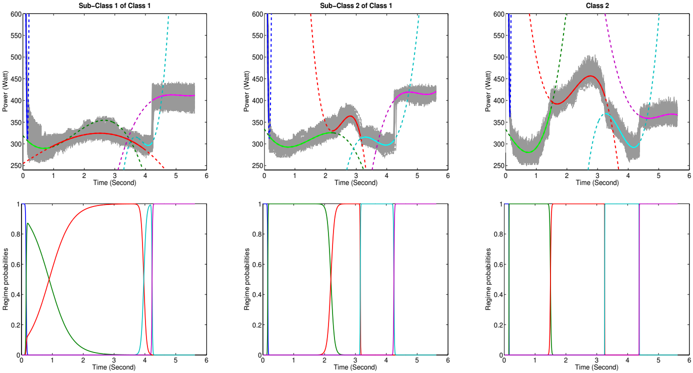
Use the Bayes' allocation rule

$$\hat{c}_i = \arg \max_{1 \leq g \leq G} \frac{\mathbb{P}(C_i = g) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}_g)}{\sum_{g'=1}^{G} \mathbb{P}(C_i = g') f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}_{g'})},$$

based on a generative model $f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}_g)$ for each group g

- Homogeneous classes: Functional Linear Discriminant Analysis [8]
- Dispersed classes: Functional Mixture Discriminant Analysis [5]

Applications to switch curves



Approach	Classification error rate (%)	Intra-class inertia
FLDA-PR	11.5	10.7350×10^9
FLDA-SR	9.53	9.4503×10^9
FLDA-RHLP	8.62	8.7633×10^9
FMDA-PRM	9.02	7.9450×10^9
FMDA-SRM	8.50	5.8312×10^9
FMDA-MixRHLP	**6.25**	$\mathbf{3.2012 \times 10^9}$

Summary

- 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: $\boldsymbol{Y} = (\boldsymbol{y}_{ij})$: n individuals defined on a set $\mathcal{I}$ with d continuous functional variables defined on a set $\mathcal{J}$ where $y_{ij}(t) = \mu(x_{ij}(t); \boldsymbol{\beta}) + \epsilon(t)$, t defined on$\mathcal{T}$.

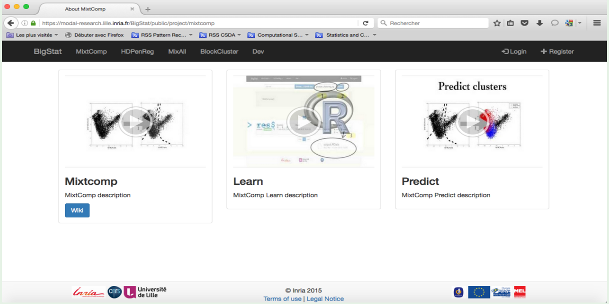
- FLBM model:

$$
\begin{aligned}
f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Psi}) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\mathbf{Z},\mathbf{W}) f(\boldsymbol{Y}|\boldsymbol{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) \\
&= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})^{z_{ik}w_{j\ell}}.
\end{aligned}
$$

- An RHLP is used as a conditional block distribution $f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})$
- 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

[1] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009

[2] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010

[3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011

[4] A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011

[5] F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a

[6] F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b

[7] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE TASE*, 3(10):829–335, 2013

[8] F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015c. doi: 10.1080/00949655.2015.1109096. 05 Nov 2015

[9] F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015

[10] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 33, 2016a. doi: 10.1007/s00357-. In Press

[11] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 2016b. In press

[12] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, 2015. In revision

[13] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a. (v1) submitted

[14] F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015b. Report (61 pages)

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

References I

F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015.

F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a. (v1) submitted.

F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015b. Report (61 pages).

F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015c. doi: 10.1080/00949655.2015.1109096. 05 Nov 2015.

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

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 33, 2016a. doi: 10.1007/s00357-. In Press.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 2016b. In press.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011.

F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a.

F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b.

F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, 2015. In revision.

References II

G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7-9):1125–1141, March 2010.

Wenxin Jiang and Martin A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12:197–220, 1999.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, September 2014.

A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE TASE*, 3(10):829–335, 2013.

Thank you for your attention!

Identifiability of the RHLP model

- $f(.;\boldsymbol{\Psi})$ is identifiable when $f(.;\boldsymbol{\Psi}) = f(.;\boldsymbol{\Psi}^\star)$ if and only if $\boldsymbol{\Psi} = \boldsymbol{\Psi}^\star$.
- 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 $(\boldsymbol{\beta}_1^T, \sigma_1^2)^T \prec \ldots \prec (\boldsymbol{\beta}_K^T, \sigma_K^2)^T$;
- initialized implies that $(w_{K0}, w_{k1}) = (0, 0)$
- irreducible implies that if $k \neq k\prime$, then one of the following conditions holds: $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k\prime}$ or $\sigma_k \neq \sigma_{k\prime}$
- The set $\{\mathcal{N}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_1), \sigma_1^2), \ldots, \mathcal{N}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_{2K}), \sigma_{2K}^2)\}$ contains $2K$ linearly independent functions of y, for any $2K$ distinct pair $(\boldsymbol{\beta}_k, \sigma_k^2)$ for $k = 1, \ldots, 2K$.