# SMILES

FAICEL CHAMROUKHI

Journée NormaSTIC et NorMAth
Rouen, 12 octobre 2018

# SMILES (nov 2018, for 42 months)

Statistical Modeling and Inference for unsupervised Learning at LargE-Scale

**Partenaires :**

| | |
|---|---|
| | F. Chamroukhi (coord.), A. Sesboüé, J. Fadili |
| | G. Chagny, Antoine Channarond, N. Vergne, C. Bérard, A. Roche |
| | H. Glotin, S. Paris, J. Razik, M. Richard |
| | C. Biernacki, V. Vandewalle |

# Research Axes

1. Task 1 : Models and inference for unsupervised large-scale data classification.
   - Sub-task 1.1 : Large-scale model-based clustering. (LMNO, INRIA) :
   - Sub-task 1.2 : Large-scale LDM and inference for functional data. (LMNO, INRIA) :
   - Sub-task 1.3 : Large-Scale LDM and inference for discrete data. (LMRS, LMNO, INRIA) :

# Research Axes

1. Task 1 : Models and inference for unsupervised large-scale data classification.
   - Sub-task 1.1 : Large-scale model-based clustering. (LMNO, INRIA) :
   - Sub-task 1.2 : Large-scale LDM and inference for functional data. (LMNO, INRIA) :
   - Sub-task 1.3 : Large-Scale LDM and inference for discrete data. (LMRS, LMNO, INRIA) :
2. Task 2 : Models and inference for large-scale data representation.
   - Sub-task 2.1 : High-dimensional (non-)parametric sparse regression for large-scale representation (LMRS, LMNO) :
   - Sub-task 2.2 : Unsupervised large-scale multimodal data representation (LIS, LMNO) :

# Research Axes

1. Task 1 : Models and inference for unsupervised large-scale data classification.
   - Sub-task 1.1 : Large-scale model-based clustering. (LMNO, INRIA) :
   - Sub-task 1.2 : Large-scale LDM and inference for functional data. (LMNO, INRIA) :
   - Sub-task 1.3 : Large-Scale LDM and inference for discrete data. (LMRS, LMNO, INRIA) :
2. Task 2 : Models and inference for large-scale data representation.
   - Sub-task 2.1 : High-dimensional (non-)parametric sparse regression for large-scale representation (LMRS, LMNO) :
   - Sub-task 2.2 : Unsupervised large-scale multimodal data representation (LIS, LMNO) :
3. Task 3 : Validation and applications.
   - i) Large-scale functional data analysis (LMNO-INRIA-LMRS) of heterogeneous multivariate times series and fMRI images
   - ii) Large-scale Bioacoustical data analysis (LIS-LMNO) for environmental survey
   - iii) Large-scale biological data analysis (LMRS) by inferring large-scale biological sequences from high-throughput sequencing

# Objectifs

- Données complexes $\hookrightarrow$ *hétérogènes, temporelles dynamiques, fonctionnelles, incomplètes, de grande dimension, et disponibles en masse*
- **Objectif :** Transformation de telles données en connaissances :
  $\hookrightarrow$ Reconstruction/révélation de structures cachées, i.e, (hiérarchie de) groupes ; sélection de variables et prédiction, etc

$\hookrightarrow$ SMILES vise à élaborer un cadre scientifique et technique pour traiter et analyser des données massives hétérogènes et peu ou non-annotées

$\hookrightarrow$ Avec une visibilité à l'international

## Axes du projet

1. Modélisation non supervisée par des modèles à variables latentes (MVL)
2. Inférence efficace non supervisée à grande échelle des MVL

# Modélisation statistique par des MVL à l'échelle

## Cadre scientifique général

$\hookrightarrow$ **Modèles** statistiques à variables latente : $f(x|\boldsymbol{\theta}) = \int_{\mathcal{Z}} f(x, z|\boldsymbol{\theta})\mathrm{d}z$

$\hookrightarrow$ **Inférence** à grande échelle par régularisation et échantillonnage :
$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) - \mathrm{Pen}_{\lambda}(\boldsymbol{\theta})$

## Modélisation statistique non supervisée à grande échelle par des MVL

- Apprentissage génératif, via des modèles à variables latentes (**régression** et **clustering**).
- représenter explicitement la structure des données brutes et la révéler

  $\hookrightarrow$ $\exists$ fondement théorique solide

  $\hookrightarrow$ Outils afférents d'estimation et de choix de modèle
- $\Rightarrow$ n'ont pas été considérés avec succès pour une analyse à grande échelle

# Inférence non supervisée à grande échelle des MVL

## Inférence en grande dimension

- L'inférence se ramène en général à l'optimisation de problèmes non linéaires complexes. à grande échelle :

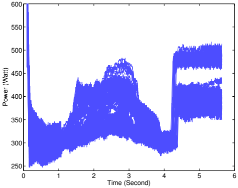
  $\hookrightarrow$ suggère de nouvelles stratégies de **régularisation** pour pallier la grande dimension

  $\hookrightarrow$ Méthodes **parcimonieuses** pour une meilleure représentation

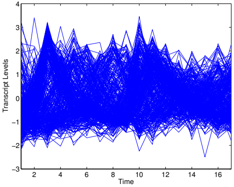
## Données de gros volume

- la distribution des calculs est une façon naturelle de s'y prendre
- méthodo : échantillonnage et inférence des modèles agrégés à partir d'un gros volume de données
- $\hookrightarrow$ Nouvelles stratégies d'**agrégation d'estimateurs** et de **sélection de modèle**

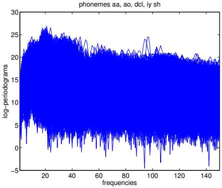
# Données longitudinales de plus en plus fréquentes



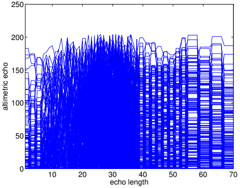

Railway switch curves




Yeast cell cycle curves

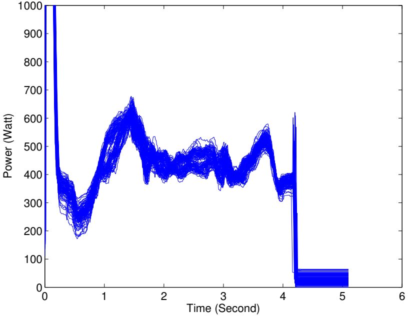



Phonemes curves

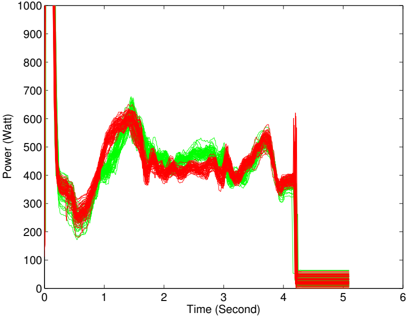



Satellite waveforms

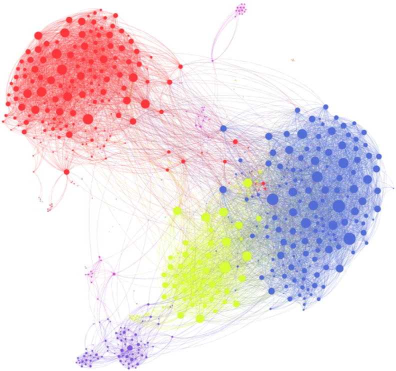
# Clustering/segmentation de données temporelles



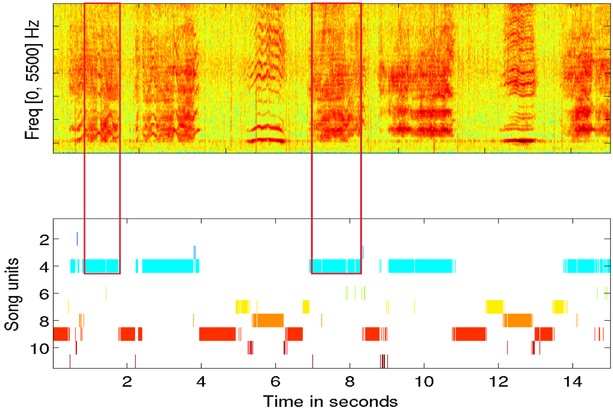
# Clustering/segmentation de données temporelles



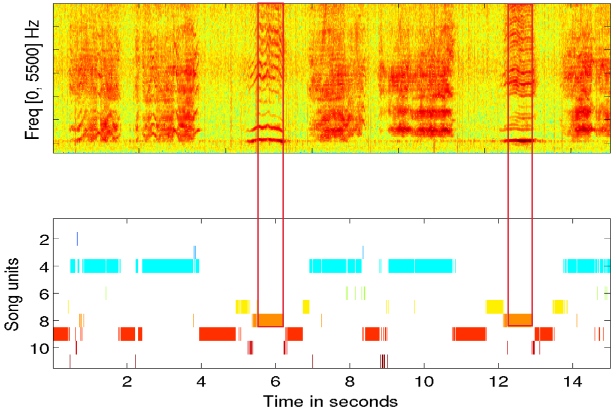
# Clustering de données représentées par des graphes



# Décomposition parcimonieuse non-supervisée



# Décomposition parcimonieuse non-supervisée



# Outline

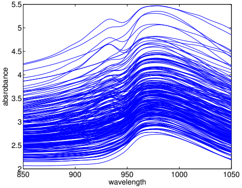
Model-Based Co-Clustering of Multivariate Functional Data

Joint work with Christophe Biernacki, INRIA-Lille

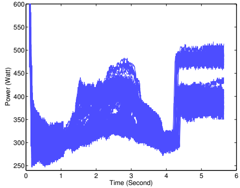
# Outline

# Functional data are increasingly frequent

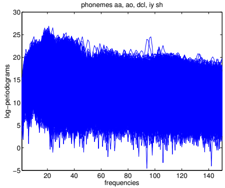
[James and Hastie, 2001; James and Sugar, 2003]
[Ramsay and Silverman, 2005]
[Chamroukhi et al., 2010]
[Bouveyron and Jacques, 2011]
[Samé et al., 2011]
[Jacques and Preda, 2014]
[Bouveyron et al., 2018]
[Chamroukhi and Nguyen, 2018]



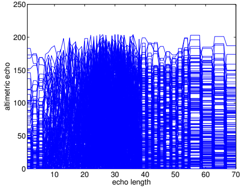

Tecator data




Railway switch curves




Phonemes curves




Satellite waveforms

# Clustering of functional data

$\hookrightarrow$ a growing investigation of Model-Based Clustering (MBC) for functional data

Some Reviews on MBC for functional data : [Jacques and Preda, 2014; Chamroukhi and Nguyen, 2018]

# Clustering of functional data

$\hookrightarrow$ a growing investigation of Model-Based Clustering (MBC) for functional data

Some Reviews on MBC for functional data : [Jacques and Preda, 2014; Chamroukhi and Nguyen, 2018]

Tecator data set [1] : $n = 240$ spectra with $m = 10($

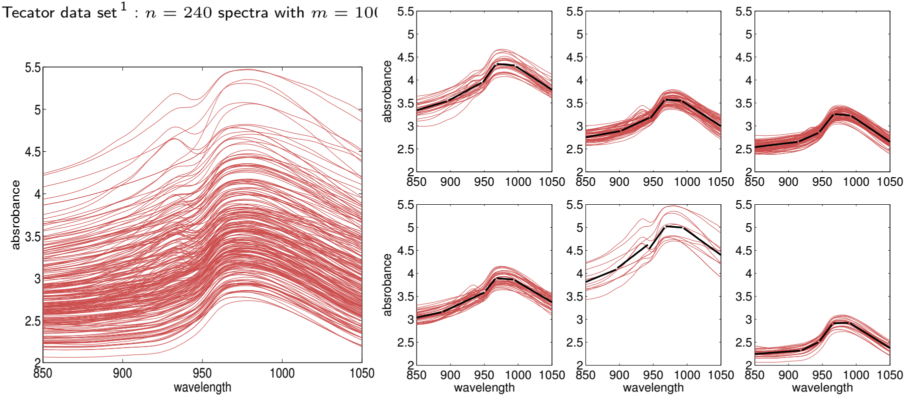



FIGURE – Original data and clustering results from Chamroukhi [2016b] for the data considered in the same setting as in Hébrail et al. [2010] (six clusters, each cluster is approximated by five linear segments $(R = 5, p = 1)$)

# Clustering of functional data

Topex/Poseidon satellite data [2] : $n = 472$ waveforms of $m = 70$ measured echoes

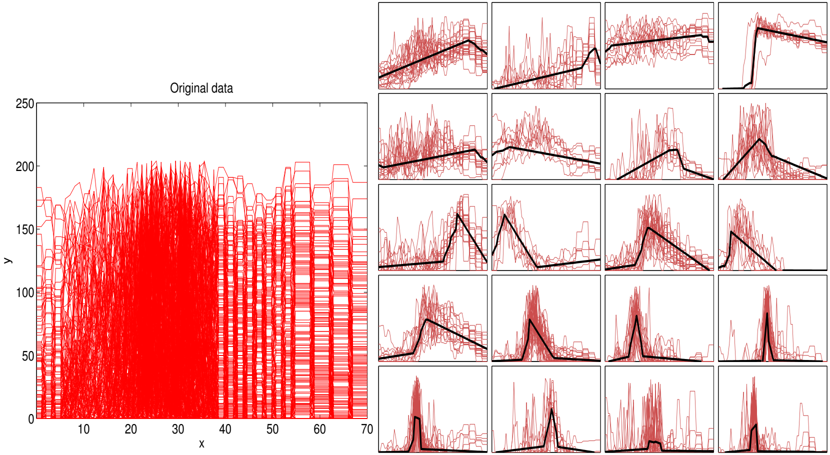



FIGURE – Original data and clustering results from Chamroukhi [2016b] with the same setting as in Hébrail et al. [2010] : twenty clusters and a piecewise linear approximation of four segments.

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

# Clustering of functional data

Phonemes data set [3] : $n = 1000$ log-periodograms for $m = 150$ frequencies

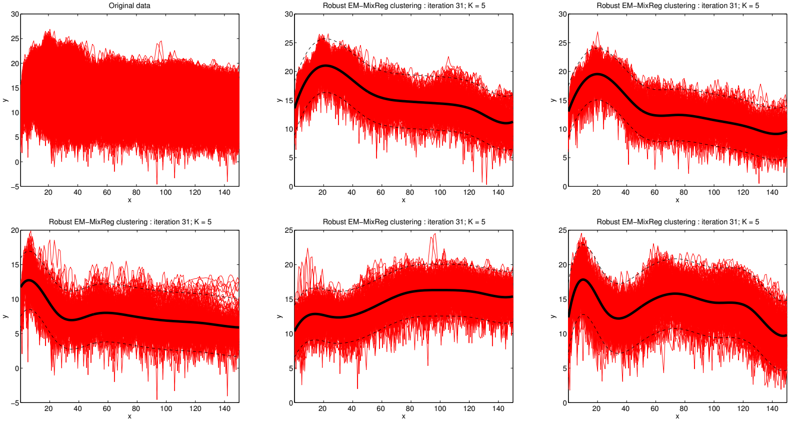



FIGURE – Original data and clustering results from Chamroukhi [2016b]

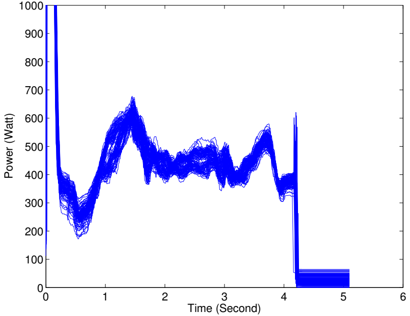
3. Data from `http://www.math.univ-toulouse.fr/staph/npfda/`, used in Ferraty and Vieu [2003]

# Clustering of functional data

**Clustering real curves of high-speed railway-switch operations**

Data : $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters : operating state without/with possible defect

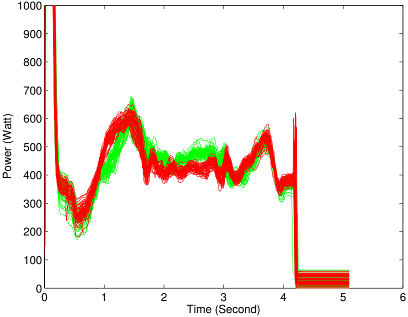


# Clustering switch operations

**Clustering real curves of high-speed railway-switch operations**

Data : $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters : operating state without/with possible defect

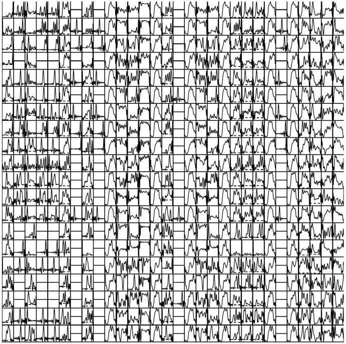


# Outline

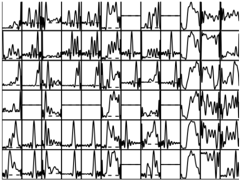
# This talk : Multivariate functional data clustering

- Multivariate functional data are increasingly present
- e.g : Data continuously recorded for different subjects from multiple subject' sensors

$\hookrightarrow$ Measurements collected from different network elements (transceivers, cells, sites. . . ) :



Data



Zoom

FIGURE – An example with $d = 30$ and $n = 20$ daily observations [Ben Slimen et al., 2016].

# This talk

## Questioning

Clustering of highly multivariate functional data with two guidelines :

- (1) Mathematical guideline : warranty for estimation and selection
- (2) User guideline : keep a user-friendly meaning of the process

Both are important because clustering is a highly risky task...

## Proposed answering

(1) Model-based co-clustering with (2) temporal curve segmentation

Novelty corresponds to combining both (1) and (2)

# Difference between clustering and co-clustering

- Simultaneous clustering of lines/indiv. ($\boldsymbol{Z}$) and columns/var. ($\boldsymbol{W}$)
- Can be used as a way to reduce dimensionality (var. $\rightarrow \boldsymbol{W}$)

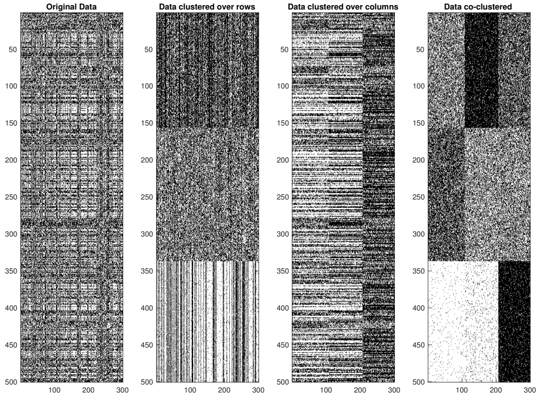



FIGURE – Binary data set with $n = 500$, $d = 300$, $K = M = 3$

# Latent block model for co-clustering

## The Latent Block Model [Govaert and Nadif, 2013]

$$f(\boldsymbol{X};\boldsymbol{\Psi}) \quad = \quad \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\pi},\boldsymbol{\rho})\,\underbrace{f(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{data kind dependent}}$$

## Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent : $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid :
  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$
  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence : $x_{ij}|(z_i,w_j) \perp x_{i\prime j\prime}|(z_{i\prime},w_{j\prime})$

# Latent block model for co-clustering

## The Latent Block Model [Govaert and Nadif, 2013]

$$f(\boldsymbol{X};\boldsymbol{\Psi}) \quad = \quad \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\pi},\boldsymbol{\rho})\,\underbrace{f(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{data kind dependent}}$$

## Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent : $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid :
  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$
  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence : $x_{ij}|(z_i,w_j) \perp x_{i'j'}|(z_{i'},w_{j'})$

$\hookrightarrow$ binary data : binary [Govaert and Nadif, 2003, 2008; Keribin et al., 2012],
$\hookrightarrow$ categorical data : multinomial [Keribin et al., 2014]
$\hookrightarrow$ contingency table : Poisson [Govaert and Nadif, 2003, 2006, 2008]
$\hookrightarrow$ continuous data : Gaussian [Lomet, 2012; Govaert and Nadif, 2013]
$\hookrightarrow$ functional data : functional PCA + Gaussian, see further [Ben Slimen et al., 2016]

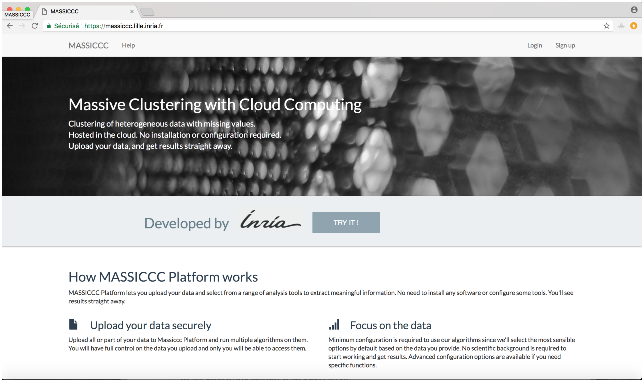
# Inference for the latent block model

## Inference of the latent block model

- variational block EM (VBEM) for maximum likelihood estimation and fuzzy co-clustering [Govaert and Nadif, 2006, 2008].
- block classification EM (CEM) algorithm for maximum classification likelihood and hard co-clustering [Govaert and Nadif, 2003, 2006, 2008]
- Bayesian inference [Keribin et al., 2012, 2014] : Bayesian latent block mixtures for binary data and categorical data & a variational Bayesian inference and Gibbs sampling.
- Number of blocks estimation : ICL criterion [Lomet, 2012; Keribin et al., 2014]

# Package blockcluster on the cloud

massiccc.lille.inria.fr



# Functional data notation

- Data : (discretized) values of underlying smooth functions, not just vectors
- Data : A sample of $n$ heterogeneous 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), $(x_{i1}, \ldots, x_{im_i})$

# Functional data modeling : "classical" approach

[Ramsay and Silverman, 2005] and many others

- Step 1 : $(\boldsymbol{x}, \boldsymbol{y})$ decomposed into a finite basis of function (B-spline. . . ) : $Y_i(t) \approx \sum_{r=1}^{d} c_{ir}\phi_r(x_i(t))$ with $\mathbf{c}$ estimated by OLS
- Step 2 : functional principal components analysis (PCA) which is performed as a usual PCA of the basis expansion coefficients $\mathbf{c}$ using a metric defined by the inner products between the basis functions
- Step 3 : set a probability distribution on $\mathbf{c}$, typically Gaussian

It defines a distribution on $\mathbf{c}$ instead of $\boldsymbol{y}$. . .

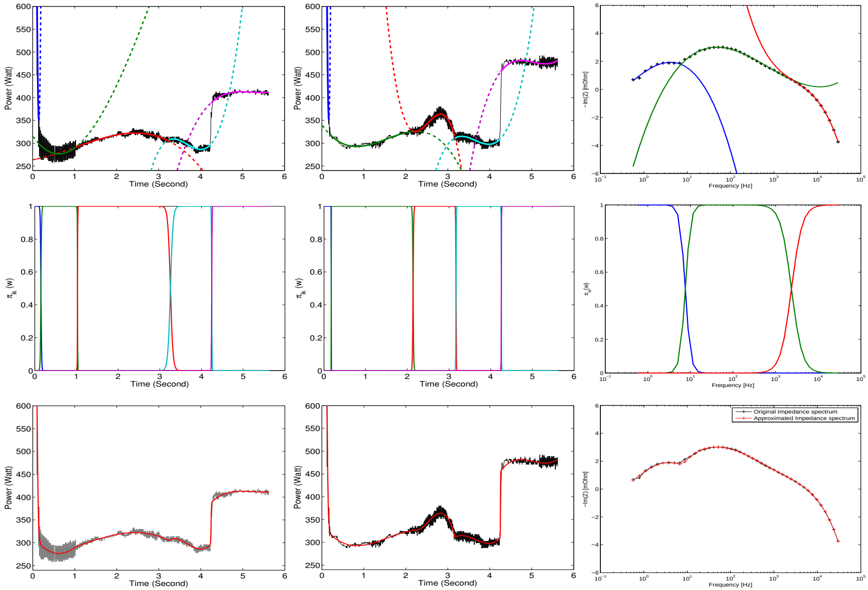
# Functional data modeling : regression RHLP

Alternatively, use a segmentation via generative piecewise polynomial regression modeling of $f(\boldsymbol{y}|\boldsymbol{x})$ [Chamroukhi et al.])

$\hookrightarrow$ Regression with Hidden Logistic Process (RHLP)
$\hookrightarrow$ See formula later

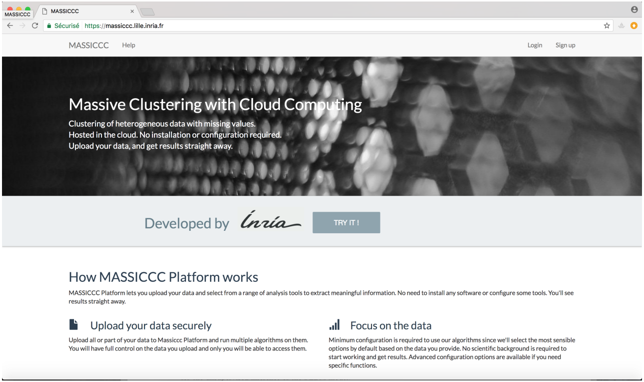
It gives a distribution on $\boldsymbol{y}$ and also a meaningful segmentation of the curve

# RHLP for modeling different types of functions



# Package mixtcomp on the cloud

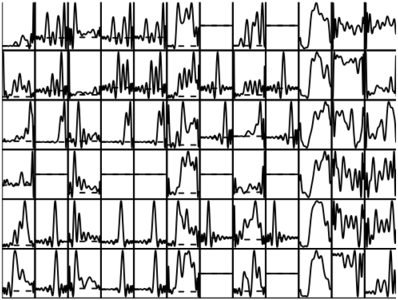
massiccc.lille.inria.fr



# Multivariate functional data co-clustering

[Chamroukhi and Biernacki, 2017]

- Data : $\boldsymbol{Y} = (\boldsymbol{y}_{ij})$ a data sample matrix of $n$ individuals defined on a set $\mathcal{I}$ and $d$ continuous functional variables defined on a set $\mathcal{J}$.
- Each variable $\boldsymbol{y}_{ij}$ is an univariate curve $\boldsymbol{y}_{ij} = \big(y_{ij}(t_1), \ldots, y_{ij}(t_{T_{ij}})\big)$ of $T_{ij}$ observations $y(t) \in \mathbb{R}$ linked to covariates $\boldsymbol{x}_{ij} = \big(x_{ij}(t_1), \ldots, x_{ij}(t_{T_{ij}})\big)$ at the points $(t_1, \ldots, t_{T_{ij}})$, typically a sampling time



# Embedding RHLP in co-clustering

[Chamroukhi and Biernacki, 2017]

- Functional Latent Block Model for Co-clustering :

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

with parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\pi}^T, \boldsymbol{\rho}^T, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)^T$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{11}^T, \ldots, \boldsymbol{\theta}_{k\ell}^T, \ldots, \boldsymbol{\theta}_{KM}^T)^T$.

# Embedding RHLP in co-clustering

- RHLP [Chamroukhi et al., 2009] : model the conditional data distribution for each block $kl$, assuming that each functional variable $\boldsymbol{y}_{ij}$ is governed by an $S_{k\ell}$-state hidden process of $y_{ij}$ :

$$f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell}) = \prod_{t=1}^{T_{ij}} \sum_{r=1}^{S_{k\ell}} \alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^T \boldsymbol{x}_{ij}(t),\sigma_{k\ell r}^2\right)$$

where the dynamical weights $\alpha'$s are given by the multinomial logistic :

$$\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell}) = \frac{\exp\left(\xi_{k\ell r0} + \xi_{kr\ell 1} t\right)}{1 + \sum_{r'=1}^{S_{k\ell}-1} \exp\left(\xi_{k\ell r'0} + \xi_{k\ell r'1} t\right)}.$$

$\hookrightarrow$ Can be seen as a generative piecewise polynomial regression model where the transition points are <u>smoothly</u> controlled by logistic weights

$\hookrightarrow$ a particular mixture-of-experts model [Jacobs et al., 1991; Jordan and Jacobs, 1994]/(parametric) mixture of regressions with predictor-dependent mixing proportions [Young and Hunter, 2010]

# Block mean curve approximation and segmentation

- Approximation : a prototype mean curve

$$y_t|(z_i, w_j) \approx \widehat{y}_t = \mathbb{E}[Y(t)|z_i, w_j, x(t); \widehat{\boldsymbol{\Psi}}] = \sum_{s=1}^{S_{kl}} \alpha_{k\ell r}(t; \widehat{\boldsymbol{\xi}}_{k\ell})\widehat{\boldsymbol{\beta}}_{k\ell r}^T \boldsymbol{x}_i(t)$$

$\hookrightarrow$ A smooth and flexible approximation thanks to the the logistic weights

- Curve segmentation :

$$\widehat{h}_t|(z_i, w_j) = \arg \max_{1 \leq s \leq S_{kl}} \mathbb{E}[H_t|z_i, w_j, x_{ij}(t); \widehat{\boldsymbol{\xi}}] = \arg \max_{1 \leq k \leq K} \alpha_{k\ell r}(t; \widehat{\boldsymbol{\xi}}_{k\ell})$$

# Parameter estimation : EM not feasible

EM algorithm :

$$\boldsymbol{\Psi}^{(q+1)} \in \arg\max_{\boldsymbol{\Psi}} \mathbb{E}\left[\log L_c(\boldsymbol{\Psi})|\mathcal{D}, \boldsymbol{\Psi}^{(q)}\right]$$

- The complete-data log-likelihood :

$$\begin{aligned}
\log L_c(\boldsymbol{\Psi}) &= \log f(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{H}|\boldsymbol{X}; \boldsymbol{\Psi}) \\
&= \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell \\
&\quad + \sum_{i,j,k,\ell,t,r} z_{ik} w_{j\ell} h_{tr} \log\left[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^T \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^2\right)\right]
\end{aligned}$$

where $(h_{tr}; t = 1, \ldots, T_{ij}, r = 1, \ldots, S_{k\ell})$ is a binary variable indicating from which state the observation $y_{ij}(t)$ within the block cluster $k\ell$ is originated

# Parameter estimation : EM not feasible

- The E-Step computes the expected complete-data log-likelihood, given the observed curves $(\boldsymbol{X}, \boldsymbol{Y})$, and the current parameter estimation $\boldsymbol{\Psi}^{(q)}$

$$
\begin{aligned}
Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) =& \mathbb{E}\left[\log L_c(\boldsymbol{\Psi}) \middle| \boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Psi}^{(q)}\right] \\
=& \sum_{i,k} \mathbb{P}(z_{ik} = 1 | \boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \log \pi_k + \sum_{j,\ell} \mathbb{P}(w_{j\ell} = 1 | \boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \log \rho_\ell \\
&+ \sum_{i,j,k,\ell,t,r} \mathbb{P}(z_{ik} w_{j\ell} = 1 | \boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \mathbb{P}(h_{tr} = 1 | z_{ik}, w_{j\ell}, y_{ij}(t), x_{ij}(t)) \times \\
& \log\left[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^T \boldsymbol{x}_{ij}(t), \sigma^2_{k\ell r}\right)\right]
\end{aligned}
$$

↪ Requires the calculation of the posterior joint distribution $\mathbb{P}(z_{ik} w_{j\ell} = 1 | \boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$

↪ does not factorize due to the conditional dependence on the observed curves of the row and the column labels

⇒ [Govaert and Nadif, 2008, 2013] proposed a variational approximation by relying on the Neal and Hinton's interpretation of the EM algorithm [Neal and Hinton, 1998].

↪ We adopt this variational approximation in our context

# Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

# Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$$

**Initialization** : start from an initial solution at iteration $q=0$, and then alternate at the $(q+1)$th iteration between the following variational E- and M- steps until convergence :

**VE Step** Estimate the variational approximated posterior memberships :

1. $\tilde{z}_{ik}^{(q+1)} \propto$ $\pi_k^{(q)} \exp\Big(\sum_{j,\ell,t,r} \tilde{w}_{j\ell}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^{T^{(q)}}\boldsymbol{x}_{ij}(t),{\sigma_{k\ell r}^{(q)}}^2\right)\Big]\Big)$
2. $\tilde{w}_{j\ell}^{(q+1)} \propto$ $\rho_\ell^{(q)} \exp\Big(\sum_{i,k,t,r} \tilde{z}_{ik}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^{T^{(q)}}\boldsymbol{x}_{ij}(t),{\sigma_{k\ell r}^{(q)}}^2\right)\Big]\Big)$
3. $\tilde{h}_{tr}^{(q+1)} \propto \alpha_{k\ell r}^{(q)}(t;\boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t);{\boldsymbol{\beta}_{k\ell r}^{(q)}}^T\boldsymbol{x}_{ij}(t),{\sigma_{k\ell r}^{(q)}}^2\right)$

where :

- $\tilde{z}_{ik} = \mathbb{P}(z_{ik}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$,
- $\tilde{w}_{j\ell} = \mathbb{P}(w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$,
- $\tilde{h}_{tr} = \mathbb{P}(h_{tr}=1|z_i,w_j,y_{ij}(t),x_{ij}(t))$

# Variational block EM algorithm

**M Step** update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$ :

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

# Variational block EM algorithm

**M Step** update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$ :

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules :

3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

# Variational block EM algorithm

**M Step** update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$ :

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules :

3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

   The regression parameters updates consist in analytic WLS problems :

4. $\boldsymbol{\beta}_{k\ell r}^{(q+1)} = \left[\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \mathbf{X}_{ij}\right]^{-1} \sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \boldsymbol{y}_{ij}$
5. $\sigma_{k\ell r}^{2(q+1)} = \frac{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \|\sqrt{\boldsymbol{\Lambda}_{ijkr}^{(q)}}(\boldsymbol{y}_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_{kr}^{(q+1)})\|^2}{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \ \mathrm{trace}(\boldsymbol{\Lambda}_{ijkr}^{(q)})}$ where $\mathbf{X}_{ij}$ is the design matrix for the $i$th curve, $\boldsymbol{\Lambda}_{ijkr}^{(q)}$ is the diagonal matrix whose diagonal elements are the posterior segment memberships $\{\tilde{h}_{ijtr}^{(q)}; t = 1, \ldots, T_{ij}\}$.

$\hookrightarrow$ It is also possible to use the Classification EM (CEM) approximation of EM [Celeux and Govaert, 1992].

## Parameter estimation by an SEM algorithm : SEM-FLBM

- $\hookrightarrow$ The SEM algorithm [Celeux and Diebolt, 1985] allows to overcome some drawbacks of the variational-EM algorithm, including its sensitivity to starting values ; SEM does not use an approximation.
- Eg. SEM for latent block models for categorical data [Keribin et al., 2012, 2014]
- The formulas of VEM-FLBM and SEM-FLBM are essentially the same, except that we incorporate a stochastic step consisting of sampling binary indicator variables $z_{ik}$, $w_{j\ell}$ and $h_{tr}$ according to $\tilde{z}_{ik}$, $\tilde{w}_{j\ell}$ and $\tilde{h}_{tr}$.

# Conclusion and perspectives

## Conclusion

- A full generative framework for the cluster analysis and segmentation of high-dimensional non-stationary functional data
- The model inference can be performed by a variational EM algorithm or SEM

## Perspectives

- Numerical experiments
- Package

# References I


Y. Ben Slimen, S. Allio, and J. Jacques. Model-Based Co-clustering for Functional Data. HAL preprint hal-01422756, December 2016. URL `https://hal.inria.fr/hal-01422756`.

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

C. Bouveyron, L. Bozzi, J. Jacques, and F.-X. Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society, Series C*, 2018.

G. Celeux and J. Diebolt. The SEM algorithm a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1) :73–82, 1985.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14 :315–332, 1992.

F. Chamroukhi. Robust mixture of experts modeling using the $t$-distribution. *Neural Networks - Elsevier*, 79 :20–36, 2016a. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/TMoE.pdf`.

F. Chamroukhi. Skew $t$ mixture of experts. *Neurocomputing - Elsevier*, 266 :390–408, 2017. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/STMoE.pdf`.

F. Chamroukhi and C. Biernacki. Model-Based Co-Clustering of Multivariate Functional Data. In *ISI 2017 - 61st World Statistics Congress*, Marrakech, Morocco, Jul 2017. URL `https://hal.archives-ouvertes.fr/hal-01653782`.

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. PDF.

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. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_neucomp_2010.pdf`.

Faicel Chamroukhi. Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification*, 33(3) :374–411, 2016b. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi-PWRM-JournalClassif-2016.pdf`.

# References II


Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. 2018. URL https://chamroukhi.users.lmno.cnrs.fr/papers/MBCC-FDA.pdf. arXiv :1803.00276v2.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B*, 39(1) :1–38, 1977.

F. Ferraty and P. Vieu. Curves discrimination : a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2) :161–173, 2003.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1, 2010.

G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2) :463 – 473, 2003. Biometrics.

G. Govaert and M. Nadif. Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5) : 415–422, 2006.

G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6) :3233 –3245, 2008.

G. Govaert and M. Nadif. *Co-Clustering*. Computer engineering series. Wiley-ISTE, November 2013. 256 pages.

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.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1) : 79–87, 1991.

Julien Jacques and Cristian Preda. Functional data clustering : A survey. *Adv. Data Anal. Classif.*, 8(3) :231–255, September 2014. ISSN 1862-5347. doi: 10.1007/s11634-013-0158-y. URL http://dx.doi.org/10.1007/s11634-013-0158-y.

G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B*, 63 :533–550, 2001.

G. M. James and C. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98 (462), 2003.

# References III


M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6 :181–214, 1994.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, 2012.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014. ISSN 0960-3174. doi: 10.1007/s11222-014-9472-2. URL `http://dx.doi.org/10.1007/s11222-014-9472-2`.

A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4) : 519–539, 2010.

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical association*, 102(479) :1025–1038, 2007.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3) :1420–1443, 2014.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient $l_1$ regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

A. Lomet. *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne, 2012.

R. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. Dordrecht : Kluwer Academic Publishers, 1998.

Hien D. Nguyen and Faicel Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling : An overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, pages e1246–n/a, Feb 2018. ISSN 1942-4795. doi: 10.1002/widm.1246. URL `http://dx.doi.org/10.1002/widm.1246`.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, June 2005.

A. Samé, F. Chamroukhi, G. Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, pages 1–21, 2011. ISSN 1862-5347.

# References IV


P. Tseng. Coordinate ascent for maximizing nondifferentiable concave functions. 1988.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3) :475–494, 2001.

DS Young and DR Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, 55(10) :2253–2266, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005.

Thank you for your attention !