

Statistical learning of latent variable models for complex data analysis

FAICEL CHAMROUKHI

LSIS, UMR CNRS 7296

LPP, UMR CNRS 8524 / Inria-Modal

Presentation, LORIA-ABC UMR CNRS 7503

March 17, 2016

Research interests

- The area of **statistical learning** and **analysis of complex data**.
- Acquiring knowledge from such data:
 - ↔ exploratory analysis
 - ↔ decisional analysis: make decision and prediction for future data

Scientific context

- density estimation
- regression
- classification/segmentation

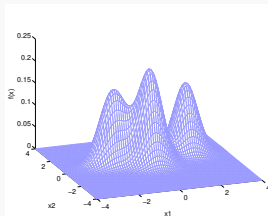
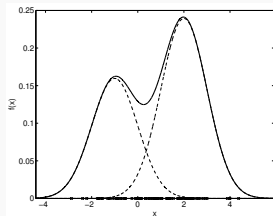
Goals and tools

- define generative probabilistic models
- propose estimation procedures

Mixture modeling framework

Mixture modeling framework

- Mixture density: $f(x) = \sum_{k=1}^K \mathbb{P}(z = k) f(x|z = k) = \sum_{k=1}^K \pi_k f_k(x)$



- Generative model

$$z \sim \mathcal{M}(1; \pi_1, \dots, \pi_K)$$
$$x|z \sim f(x|z)$$

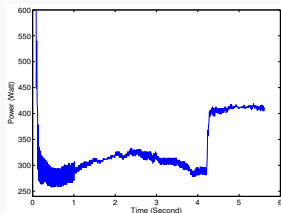
- Fitting such models for the analysis

Outline

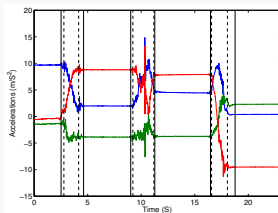
- 1 Mixture models for temporal data segmentation
- 2 Mixture models for functional data analysis
- 3 Bayesian (non-)parametric mixtures for spatial and multivariate data

Temporal data

Temporal data with regime changes



Railway data



Human activity data

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

Objectives

Temporal data modeling and segmentation

Outline

- 1 Mixture models for temporal data segmentation
 - Regression with hidden logistic process
 - Multiple hidden process regression
 - Non-normal mixtures of experts
- 2 Mixture models for functional data analysis
- 3 Bayesian (non-)parametric mixtures for spatial and multivariate data

Mixture models for temporal data segmentation

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

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
↔ 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
- ↔ segmentation \equiv inferring the model parameters and the underling process

Regression with hidden logistic process

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

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

$$y_i = \beta_{z_i}^T \mathbf{x}_i + \sigma_{z_i} \epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad (i = 1, \dots, n)$$
$$Z_i \sim \mathcal{M}(1, \pi_1(t_i; \mathbf{w}), \dots, \pi_K(t_i; \mathbf{w}))$$

Polynomial segments $\beta_{z_i}^T \mathbf{x}_i$ with $\mathbf{x}_i = (1, t_i, \dots, t_i^p)^T$ with logistic probabilities

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

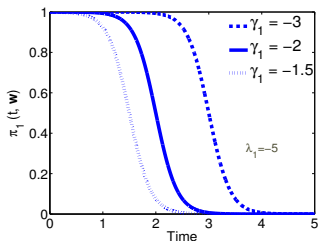
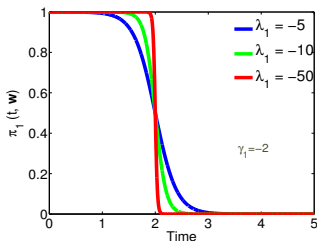
$$f(y_i | t_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i; \mathbf{w}) \mathcal{N}(y_i; \beta_k^T \mathbf{x}_i, \sigma_k^2)$$

- Both the mixing proportions and the component parameters are time-varying

Model properties

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

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



⇒ 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

EM-RHLP

Parameter estimation via a the EM algorithm: EM-RHLP

- Parameter estimation via a the EM algorithm (EM-RHLP)

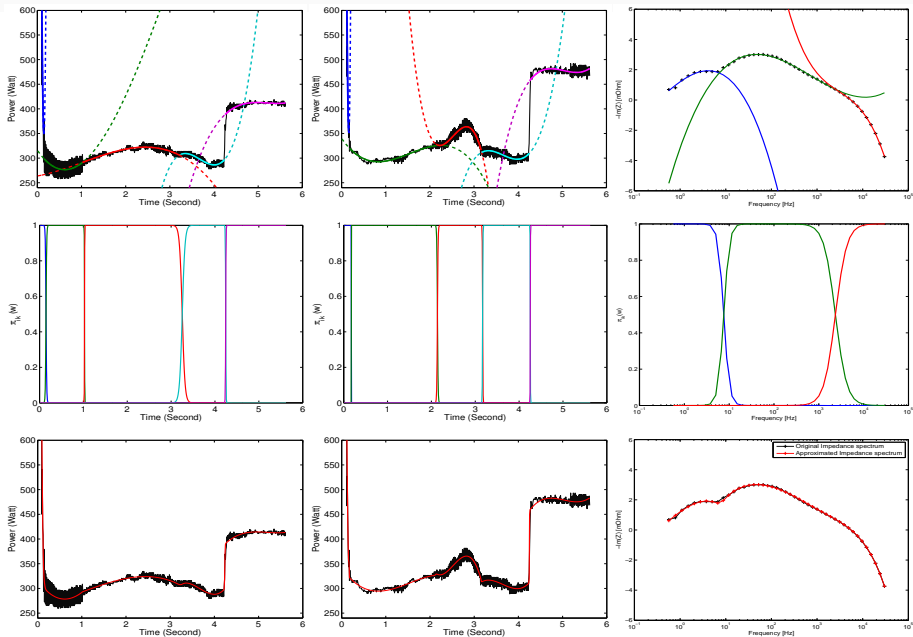
M-Step: **includes a weighted logistic regression problem** \leftrightarrow IRLS
(and weighted polynomial regressions)

- EM-RHLP algorithm complexity: $\mathcal{O}(I_{EM}I_{IRLS}K^3p^3n)$ (more advantageous than dynamic programming).

Time series approximation and segmentation

- 1** Approximation: a curve prototype $\hat{y}_i = \mathbb{E}[y_i|t_i; \hat{\boldsymbol{\theta}}] = \sum_{k=1}^K \pi_k(t_i; \hat{\mathbf{w}}) \hat{\boldsymbol{\beta}}_k^T \mathbf{x}_i$
 \leftrightarrow The RHLP can be used as nonlinear regression model $y_i = f(t_i; \boldsymbol{\theta}) + \epsilon_i$
by covering functions of the form $f(t_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i; \mathbf{w}) \boldsymbol{\beta}_k^T \mathbf{x}_i$ [J-3]
- 2** Curve segmentation:
 $\hat{z}_i = \arg \max_{1 \leq k \leq K} \mathbb{E}[z_i|t_i; \hat{\mathbf{w}}] = \arg \max_{1 \leq k \leq K} \pi_k(t_i; \hat{\mathbf{w}})$
Model selection: Application of BIC, ICL ($\nu_{\boldsymbol{\theta}} = K(p+4) - 2.$)

Application to real data



Joint segmentation of multivariate time series

Multiple hidden process regression

- Data: $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ a time series of n multidimensional observations $\mathbf{y}_i = (y_i^{(1)}, \dots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at instants $\mathbf{t} = (t_1, \dots, t_n)$.
- Model

$$\begin{aligned} y_i^{(1)} &= \beta_{z_i}^{(1)T} \mathbf{x}_i + \sigma_{z_i}^{(1)} \epsilon_i \\ &\vdots \\ y_i^{(d)} &= \beta_{z_i}^{(d)T} \mathbf{x}_i + \sigma_{z_i}^{(d)} \epsilon_i \end{aligned}$$

Vectorial form: $\mathbf{y}_i = \mathbf{B}_{z_i}^T \mathbf{x}_i + \mathbf{e}_i$; $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{z_i})$, ($i = 1, \dots, n$)

- The latent process $\mathbf{z} = (z_1, \dots, z)$ simultaneously governs the univariate time series components

PhD of Dorra Trabelsi 2010-2013^a

^aD. Trabelsi. *Contribution à la reconnaissance non-intrusive d'activités humaines*. Ph.D. thesis, Université Paris-Est Créteil, Laboratoire Images, Signaux et Systèmes Intelligents (LiSSI), June 2013

↔ Multiple regression with hidden logistic process: Multiple RHLP [J-6]

↔ Multiple Hidden Markov model regression (MHMMR) [J-7]

Multiple hidden Markov model regression

- MHMMR: Estimation by the EM algorithm (as for HMMs)
 - ↪ Solve multiple regression problems

Application to human activity time series

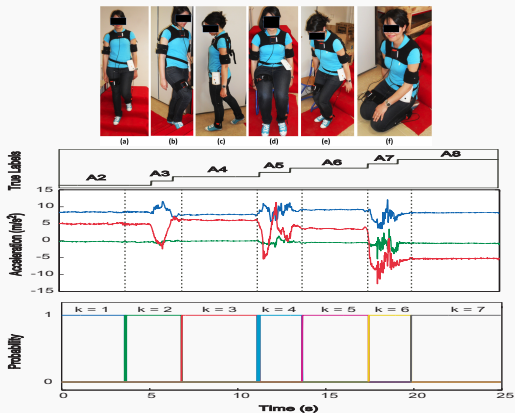


Figure: MHMMR Segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Multiple regression with hidden logistic process

- MRHLP: Estimation by the EM algorithm (as for the RHLP)
 - ↔ Solve multiple regression problems

Application to human activity time series

Problem: Activity recognition from multivariate acceleration time series

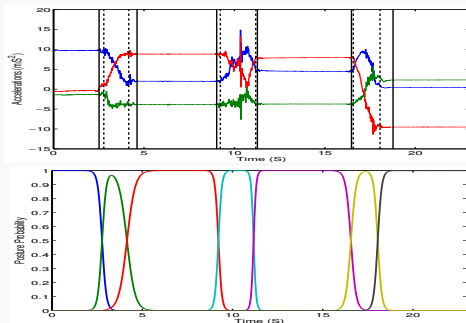


Figure: MRHLP segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Data with atypical features

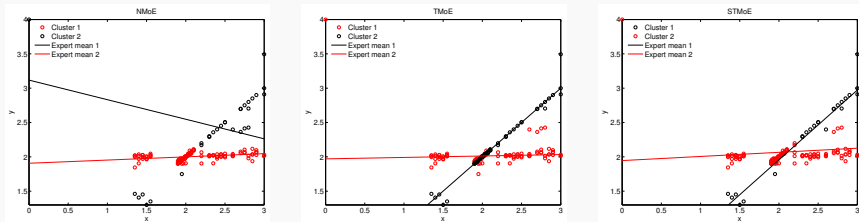


Figure: Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

- Data with possible atypical observations
- Data with possibly asymmetric and heavy-tailed distributions

Objectives

- Derive robust models to fit at best the data
- Deal with other possible features like skewness, heavy tails

Mixture of Experts (MoE) modeling framework

- Observed pairs of data (\mathbf{x}, y) where $y \in \mathbb{R}$ is the response for some covariate $\mathbf{x} \in \mathbb{R}^p$ governed by a hidden categorical random variable Z
- Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994) :

$$f(y|\mathbf{x}; \Psi) = \sum_{k=1}^K \underbrace{\pi_k(\mathbf{r}; \boldsymbol{\alpha})}_{\text{Gating network}} \underbrace{f_k(y|\mathbf{x}; \Psi_k)}_{\text{Experts}}$$

- Gating function of some predictors $\mathbf{r} \in \mathbb{R}^q$: $\pi_k(\mathbf{r}; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha}_k^T \mathbf{r})}{\sum_{\ell=1}^K \exp(\boldsymbol{\alpha}_\ell^T \mathbf{r})}$
- MoE for regression usually use normal experts $f_k(y|\mathbf{x}; \Psi_k)$

Objectives

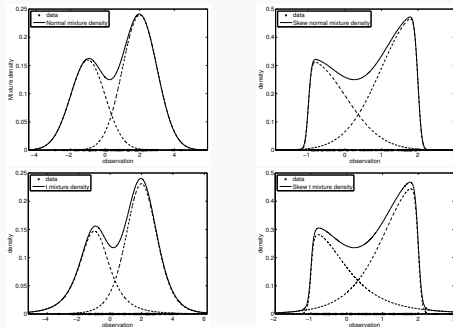
- Overcome (well-known) limitations of modeling with the normal distribution.
↪ Not adapted For a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations

Non-normal mixtures of experts

Non-normal mixtures of experts (NNMoE)

- 1 the t MoE (TMoE) (Robustness, heavy tails) [J-11]
- 2 the skew-normal MoE (SNMoE) (skewness) [J-14]
- 3 the skew- t MoE (STMoE) (skewness, robustness, heavy tails) [J-15]

Non-normal mixtures



$$\pi_k = [0.4, 0.6], \mu_k = [-1, 2]; \sigma_k = [1, 1]; \nu_k = [3, 7]; \lambda_k = [14, -12];$$

The skew t mixture of experts (STMoE) model

- A K -component mixture of skew t experts (STMoE) is defined by:

$$f(y|\mathbf{r}, \mathbf{x}; \Psi) = \sum_{k=1}^K \pi_k(\mathbf{r}; \alpha) \text{ST}(y; \mu(\mathbf{x}; \beta_k), \sigma_k^2, \lambda_k, \nu_k)$$

- k th expert: has a skew t distribution (Azzalini and Capitanio, 2003):

$$f(y|\mathbf{x}; \mu(\mathbf{x}; \beta_k), \sigma^2, \lambda, \nu) = \frac{2}{\sigma} t_\nu(d_y(\mathbf{x})) T_{\nu+1} \left(\lambda d_y(\mathbf{x}) \sqrt{\frac{\nu+1}{\nu+d_y^2(\mathbf{x})}} \right)$$

Model characteristics

↔ For $\{\nu_k\} \rightarrow \infty$, the STMoE reduces to the SNMoE

↔ For $\{\lambda_k\} \rightarrow 0$, the STMoE reduces to the TMoE.

↔ For $\{\nu_k\} \rightarrow \infty$ and $\{\lambda_k\} \rightarrow 0$, it approaches the NMoE.

↔ The STMoE is flexible as it generalizes the previously described models to accommodate situations with asymmetry, heavy tails, and outliers.

Parameter estimation via the ECM algorithm

1 E-Step: requires the following conditional expectations:

$$\begin{aligned}\tau_{ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [Z_{ik} | y_i, \mathbf{x}_i, \mathbf{r}_i], \\ w_{ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i | y_i, Z_{ik} = 1, \mathbf{x}_i, \mathbf{r}_i], \\ e_{1,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i U_i | y_i, Z_{ik} = 1, \mathbf{x}_i, \mathbf{r}_i], \\ e_{2,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i U_i^2 | y_i, Z_{ik} = 1, \mathbf{x}_i, \mathbf{r}_i], \\ e_{3,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [\log(W_i) | y_i, Z_{ik} = 1, \mathbf{x}_i, \mathbf{r}_i].\end{aligned}$$

↔ Calculated analytically except $e_{3,ik}^{(m)}$ ↔ I adopted a one-step-late (OSL) approach as in Lee and McLachlan (2014)

↔ Note that Lee and McLachlan (2015) presented an exact series-based truncation approach for the multivariate skew t mixture models

2 CM-Steps: **Include weighted logistic regressions and linear regressions**

↔ Predicted response: $\hat{y} = \mathbb{E}_{\hat{\Psi}}(Y | \mathbf{r}, \mathbf{x})$ with

$$\mathbb{E}_{\hat{\Psi}}(Y | \mathbf{r}, \mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{r}; \hat{\alpha}_n) \mathbb{E}_{\hat{\Psi}}(Y | Z = k, \mathbf{x})$$

↔ Predicted class: $\hat{z} = \arg \max_{k=1}^K \mathbb{E}[Z | \mathbf{r}, \mathbf{x}; \hat{\Psi}]$

↔ Model selection: Choose (K, p) using BIC or ICL

Tone perception data set

- Recently studied by Bai et al. (2012) and Song et al. (2014) by using, respectively, robust t regression mixture and Laplace regression mixture
- Data consist of $n = 150$ pairs of “tuned” variables, considered here as predictors (x), and their corresponding “stretch ratio” variables considered as responses (y).

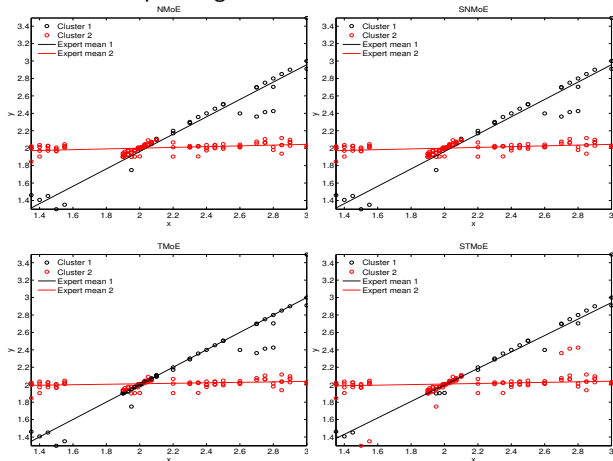


Figure: Fitting the MoE models to the tone data set

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

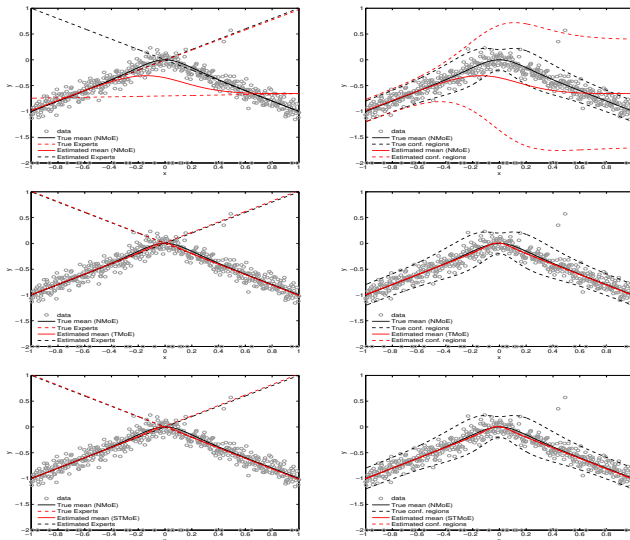


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers ($x; y = -2$): NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Tone perception data set (noisy case)

- Consider the same scenario used in Bai et al. (2012) and Song et al. (2014) (the last and more difficult scenario) by adding 10 identical pairs $(0, 4)$

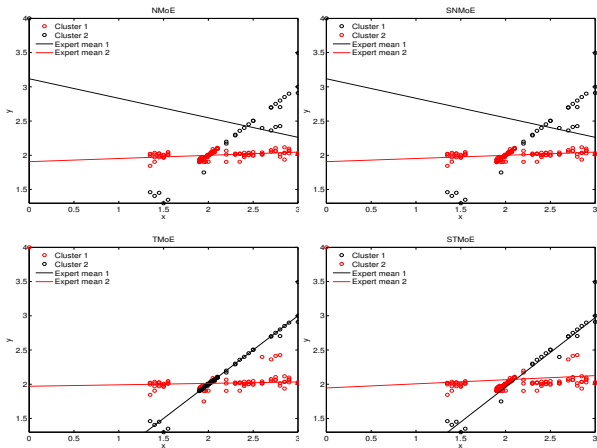
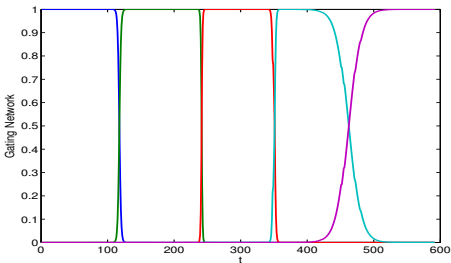
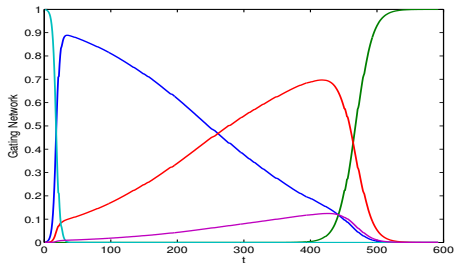
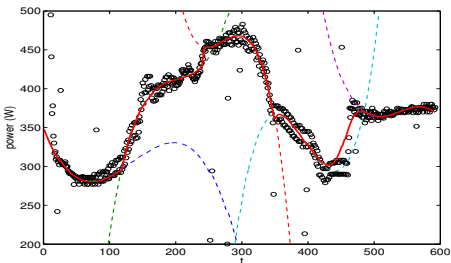
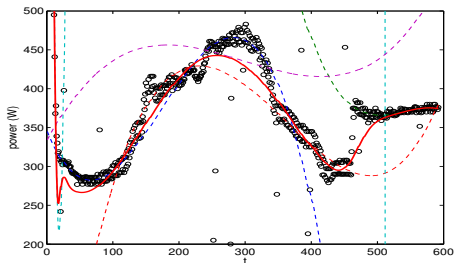


Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

↪ In this noisy case the t mixture of regressions fails (is affected severely by the outliers) as showed in Song et al. (2014)

Temporal railway data segmentation

- $n = 562$ temporal data
- 30 added artificial outliers

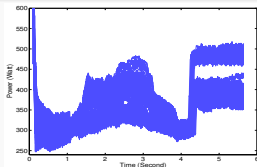


Outline

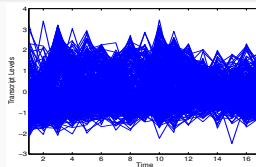
- 1 Mixture models for temporal data segmentation
- 2 Mixture models for functional data analysis
 - Mixture of piecewise regressions
 - Mixture of hidden Markov model regressions
 - Mixture of hidden logistic process regressions
 - Functional discriminant analysis
 - Regularized regression mixtures for functional data
- 3 Bayesian (non-)parametric mixtures for spatial and multivariate data

Functional data analysis context

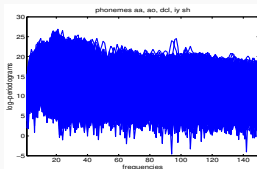
Many curves to analyze



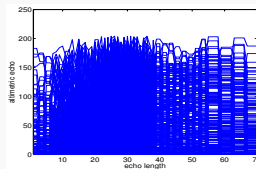
Railway switch curves



Yeast cell cycle curves



Phonemes curves



Satellite waveforms

Objectives

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

Functional data analysis context

Data

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

Objectives: exploratory or decisional

- 1 Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [J-4] [J-9], [C-11] [J-16]
- 2 Discriminant analysis of functional data: [J-2], [J-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:

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

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

↔ more tailored to approximate smooth functions

↔ do not account for the segmentation

Here $f_k(\mathbf{y}|\mathbf{x})$ itself exhibits a clustering property due to regimes:

- 1 Riecewise regression model (PWR)
- 2 Regression model with a hidden Markov process (HMMR)
- 3 Regression model with hidden logistic process (RHLP)

Piecewise regression mixture model (PWRM) [J-9]

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

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

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

- \leftrightarrow Simultaneously accounts for curve clustering and segmentation

Parameter estimation

1 Maximum likelihood estimation: EM-PWRM

2 Maximum classification likelihood estimation: CEM-PWRM

\leftrightarrow a generalization of the K -means-like algorithm of Hébrail et al. (2010):

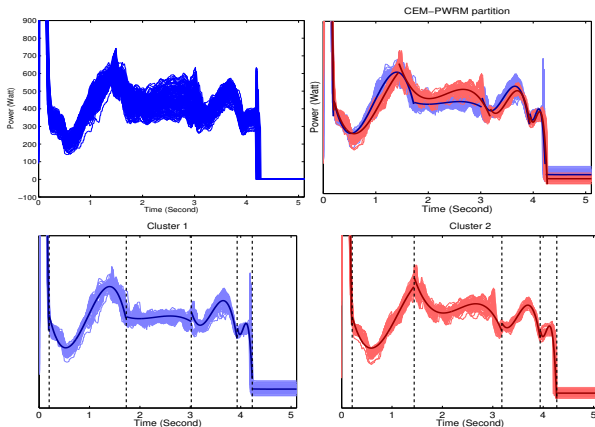
M-step: includes weighted piecewise regression problems \leftrightarrow **dynamic programming**

Complexity in $\mathcal{O}(I_{EM} K R n m^2 p^3)$: Significant computational load for very large m

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)



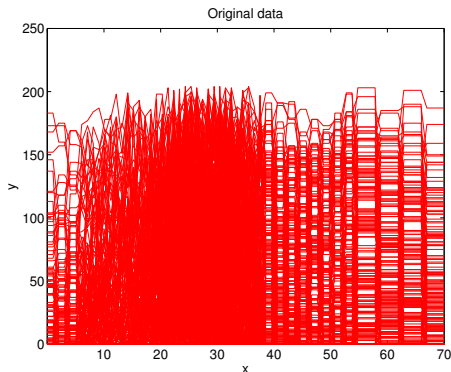
EM-GMM	EM-PRM	EM-PSRM	K -means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

Table: Estimated intra-cluster inertia for the switch curves.

Application to Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data¹ contains $n = 472$ waveforms of the measured echoes, sampled at $m = 70$ (number of echoes)

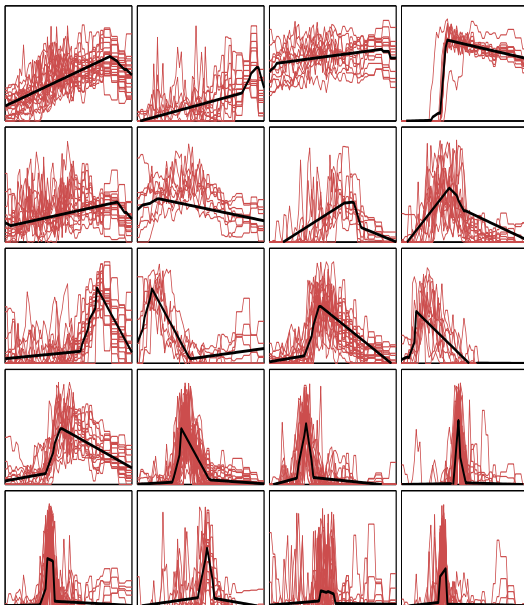
We considered the same number of clusters (twenty) and a piecewise linear approximation of four segments per cluster as in Hébrail et al. (2010).



¹Satellite data are available at

<http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html>.

CEM-PWRM clustering of the satellite data



Mixture of hidden logistic process regressions [J-4]

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

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

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

- Two types of component memberships:

↔ 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
- EM-MixRHLP has complexity in $\mathcal{O}(I_{EM} I_{IRLS} K R^3 n m p^3)$ (K -means type for piecewise regression is in $\mathcal{O}(I_{KM} K R n m^2 p^3)$) ↔ EM-MixRHLP is computationally attractive for large values of m and moderate values of R .

Functional discriminant analysis

Supervised classification context

- Data: a training set of labeled functions $((\mathbf{x}_1, y_1, c_1), \dots, (\mathbf{x}_n, y_n, c_n))$ where $c_i \in \{1, \dots, G\}$ is the class label of the i th curve
- Problem: predict the class label c_i for a new unlabeled function $(\mathbf{x}_i, \mathbf{y}_i)$

Tool: Discriminant analysis

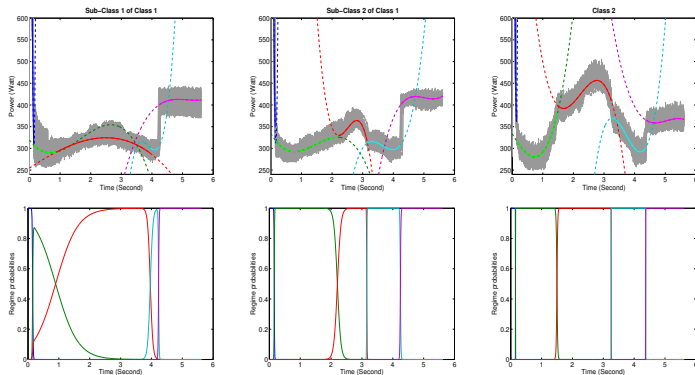
Use the Bayes' allocation rule

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

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

- Homogeneous classes: Functional Linear Discriminant Analysis [J-2]
- Dispersed classes: Functional Mixture Discriminant Analysis [J-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	3.2012×10^9

Regularized regression mixtures

The finite Gaussian regression mixture model

$$f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_{m_i})$$

- The parameter $\boldsymbol{\theta}$ is usually estimated by ML: $\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta})$
- the EM algorithm is the usual tool

↔ requires careful initialization (Biernacki et al., 2003)

↔ requires the number of components K to be supplied by the user (or BIC, ICL etc)

Idea of the proposed approach [J-8]

↔ A fully unsupervised fitting of regression mixtures

↔ EM-like algorithm which is robust with regard initialization and infers the number of components from the data

Regularized regression mixtures [J-8]

- Penalized log-likelihood criterion:

$$\begin{aligned}\mathcal{J}(\lambda, \Psi) &= \log L(\Psi) - \lambda H(\mathbf{z}), \quad \lambda \geq 0 \\ &= \sum_{i=1}^n \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{1}_m) + \lambda n \sum_{k=1}^K \pi_k \log \pi_k\end{aligned}$$

- $H(\mathbf{Z}) = -\mathbb{E}[\log \mathbb{P}(\mathbf{Z})]$: - entropy accounting for model complexity
- $\lambda \geq 0$ is a smoothing parameter

EM-like algorithm for unsupervised learning [J-8]

initialization : $K^{(0)} = n$; $\pi_k^{(0)} = \frac{1}{K^{(0)}}$, $(\boldsymbol{\beta}_k^{(0)}, \sigma_k^{2(0)})$: polynomial regression

1 E-step: Posterior component memberships $\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | \mathbf{x}_i, \mathbf{y}_i; \hat{\Psi})$

2 M-step: $\pi_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^n \tau_{ik}^{(q)} + \lambda \pi_k^{(q)} \left(\log \pi_k^{(q)} - \sum_{h=1}^K \pi_h^{(q)} \log \pi_h^{(q)} \right)$

$$\boldsymbol{\beta}_k^{(q+1)} = \left[\sum_{i=1}^n \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{X}_i \right]^{-1} \sum_{i=1}^n \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{y}_i \quad \sigma_k^{2(q+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(q)} \|\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_k\|^2}{m \sum_{i=1}^n \tau_{ik}^{(q)}}$$

The penalization coefficient λ is set in an adaptive way

↪ However, does not guarantee the ascent property of the objective function

Phonemes data

Phonemes data set used in Ferraty and Vieu (2003)²
1000 log-periodograms (200 per cluster)

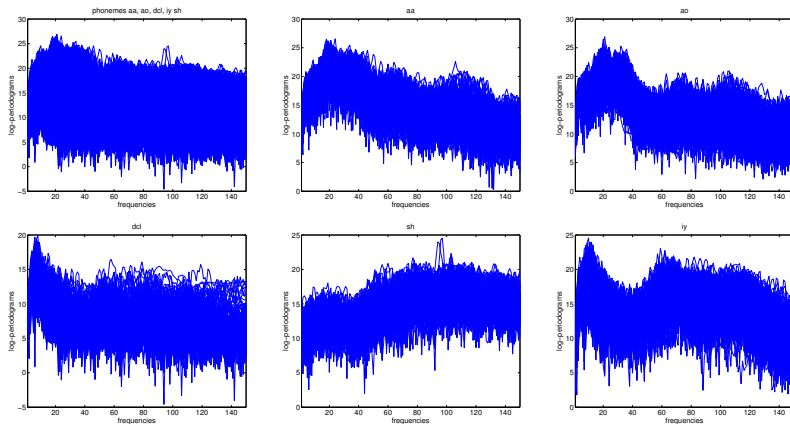


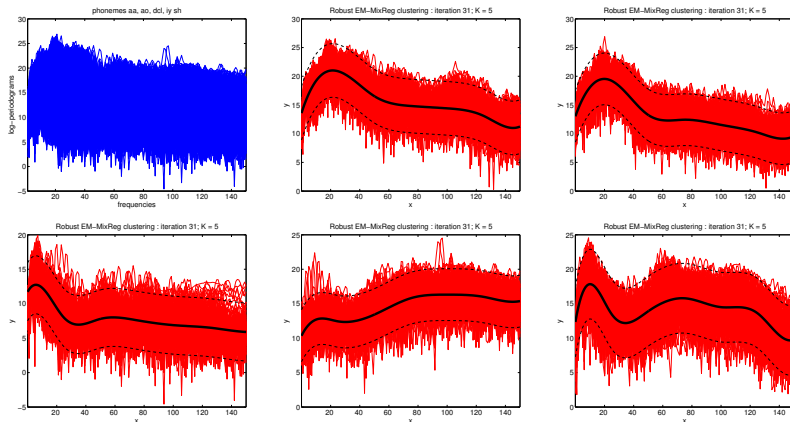
Figure: Original phoneme data and curves of the five classes: "ao", "aa", "iy", "dcl", "sh".

²Data from <http://www.math.univ-toulouse.fr/staph/npfda/>

EM-like clustering results for Phonemes

Phonemes data set used in Ferraty and Vieu (2003)³

1000 log-periodograms (200 per cluster)



	EM-PRM	EM-SRM	EM-bSRM
Estimated K	5	5	5
Misc. error rate	14.29 %	14.09 %	14.2 %

³Data from <http://www.math.univ-toulouse.fr/staph/pnfda/>

Yeast cell cycle data

- Time course Gene expression data as in Yeung et al. (2001) ⁴
- 384 genes expression levels over 17 time points.

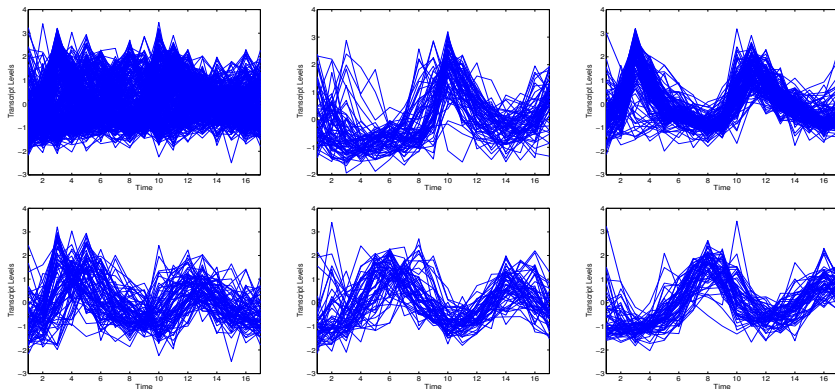


Figure: The five “actual” clusters of the used yeast cell cycle data according to Yeung et al. (2001).

⁴

<http://faculty.washington.edu/kayee/model/>

EM-like clustering results for yeast cell cycle data

- Time course Gene expression data as in Yeung et al. (2001)
- 384 genes expression levels over 17 time points.

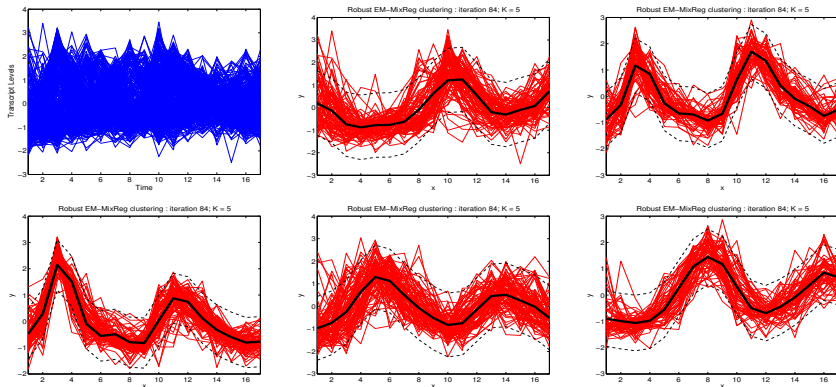


Figure: EM-like clustering results with the bSRM model.

Rand index: 0.7914 which indicates that the partition is quite well defined.

Outline

- 1 Mixture models for temporal data segmentation
- 2 Mixture models for functional data analysis
- 3 Bayesian (non-)parametric mixtures for spatial and multivariate data
 - Bayesian spatial spline regression with mixed-effects
 - Bayesian mixture of spatial spline regressions with mixed-effects
 - Dirichlet Process Parsimonious Mixtures for multivariate data clustering
 - Application to whale song decomposition

Bayesian spatial spline regression with mixed-effects

- Data: $((\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n))$ a sample of n surfaces $\mathbf{y}_i = (y_{i1}, \dots, y_{im_i})^T$ and their spatial coordinates $\mathbf{x}_i = ((x_{i11}, x_{i12}), \dots, (x_{im_i1}, x_{im_i2}))^T$.
- Propose regression and regression mixtures, with three additional features:
 - 1 Include random effects
 - 2 Models for spatial functional data
 - 3 A full Bayesian inference

Bayesian spatial spline regression with mixed-effects [Esann 2016, J-13]

$$\mathbf{y}_i = \mathbf{S}_i(\boldsymbol{\beta} + \mathbf{b}_i) + \mathbf{e}_i, \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{m_i}), \quad (i = 1, \dots, n)$$

- $\boldsymbol{\beta}$: fixed-effects regression coefficients
- \mathbf{b}_i : random subject-specific regression coefficients $\mathbf{b}_i \perp \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \xi^2 \mathbf{I}_{m_i})$
- \mathbf{S}_i is a spatial design matrix.

- \mathbf{S}_i constructed from the Nodal basis functions (NBF) (Malfait and Ramsay, 2003) used in (Ramsay et al., 2011; Sangalli et al., 2013; Nguyen et al., 2014)
- NBFs extend the univariate B-spline bases to bivariate surfaces.

$$\mathbf{S}_i = \begin{pmatrix} s(\mathbf{x}_1; \mathbf{c}_1) & s(\mathbf{x}_1; \mathbf{c}_2) & \cdots & s(\mathbf{x}_1; \mathbf{c}_d) \\ s(\mathbf{x}_2; \mathbf{c}_1) & s(\mathbf{x}_2; \mathbf{c}_2) & \cdots & s(\mathbf{x}_2; \mathbf{c}_d) \\ \vdots & \vdots & \ddots & \vdots \\ s(\mathbf{x}_{m_i}; \mathbf{c}_1) & s(\mathbf{x}_{m_i}; \mathbf{c}_2) & \cdots & s(\mathbf{x}_{m_i}; \mathbf{c}_d) \end{pmatrix}$$

d : number of basis functions d

$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})$ the two spatial coordinates of y_{ij}

$\mathbf{c} = (c_1, c_2)$ is a node center parameter, with v/h shape parameters δ_1 and δ_2

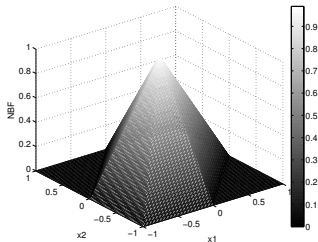


Figure: Nodal basis function $s(\mathbf{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$.

Bayesian spatial spline regression with mixed-effects

Under the BSRR model, the density of the observation \mathbf{y}_i is given by

$$f(\mathbf{y}_i | \mathbf{S}_i; \Psi) = \mathcal{N}(\mathbf{y}_i; \mathbf{S}_i \boldsymbol{\beta}, \xi^2 \mathbf{S}_i \mathbf{S}_i^T + \sigma^2 \mathbf{I}_{m_i}).$$

Conjugate prior distributions

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \mathbf{b}_i | \xi^2 &\sim \mathcal{N}(\mathbf{0}_d, \xi^2 \mathbf{I}_d) \\ \xi^2 &\sim \mathcal{IG}(a_0, b_0) \\ \sigma^2 &\sim \mathcal{IG}(g_0, h_0)\end{aligned}$$

Bayesian inference using Gibbs sampling

- Sample from the full conditional posterior distributions (analytic)

$$\begin{aligned}\boldsymbol{\beta} | \dots &\sim \mathcal{N}(\boldsymbol{\nu}_0, \mathbf{V}_0) \\ \mathbf{b}_i | \dots &\sim \mathcal{N}(\boldsymbol{\nu}_1, \mathbf{V}_1) \\ \sigma^2 | \dots &\sim \mathcal{IG}(g_1, h_1) \\ \xi^2 | \dots &\sim \mathcal{IG}(a_1, b_1)\end{aligned}$$

Illustration on simulated surfaces' approximation

A sample of 100 simulated noisy surfaces from $\mu(\mathbf{x}) = \frac{\sin(\sqrt{1+x_1^2+x_2^2})}{\sqrt{1+x_1^2+x_2^2}}$

The simulated data include mixed effects.

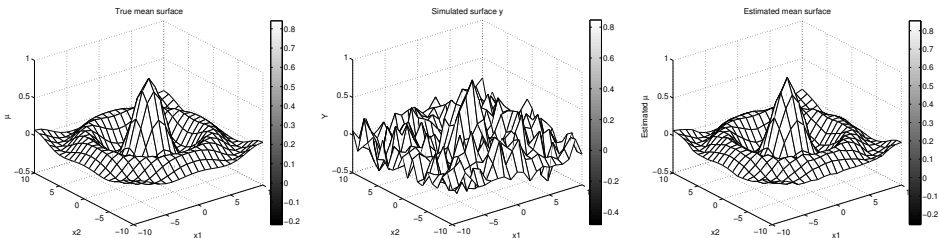


Figure: True mean surface (left), an example of noisy surface (middle), A BSSR fit $\hat{\mu}(\mathbf{x}) = \mathbf{S}_i \hat{\beta}$ from 100 surfaces using 15×15 NBFs (right).

Empirical sum of squared error: $SSE = \sum_{j=1}^m (\mu_j(\mathbf{x}) - \hat{\mu}_j(\mathbf{x}))^2$ ($m = 441$ here):
0.0865 (a very reasonable fit)

Bayesian mixture of spatial spline regressions

Data: A sample of n surfaces $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ and their spatial covariates $(\mathbf{S}_1, \dots, \mathbf{S}_n)$ issued from K sub-populations

- Bayesian mixture of spatial spline regression models with mixed-effects (BMSSR):

$$f(\mathbf{y}_i | \mathbf{S}_i; \Psi) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{S}_i(\boldsymbol{\beta}_k + \mathbf{b}_{ik}), \sigma_k^2 \mathbf{I}_{m_i})$$

↔ Useful for density estimation and model-based clustering of heterogeneous surfaces

Hierarchical prior for the BMSSR

$$\begin{aligned} \boldsymbol{\pi} &\sim \mathcal{D}(\alpha_1, \dots, \alpha_K) \\ \boldsymbol{\beta}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \\ \mathbf{b}_{ik} | \xi_k^2 &\sim \mathcal{N}(\mathbf{0}_d, \xi_k^2 \mathbf{I}_d) \\ \xi_k^2 &\sim \mathcal{IG}(a_0, b_0) \\ \sigma_k^2 &\sim \mathcal{IG}(g_0, h_0). \end{aligned}$$

Bayesian inference of the BMSSR

- For the BMSSR, the parameter Ψ is augmented by the unknown components labels $\mathbf{z} = (z_1, \dots, z_n)$

Bayesian inference of the BMSSR using Gibbs sampling

- Sample from the analytic full conditional distributions:

$$Z_i | \dots \sim \mathcal{M}(1; \tau_{i1}, \dots, \tau_{iK}) \text{ with } \tau_{ik} (1 \leq k \leq K) = \mathbb{P}(Z_i = k | \mathbf{y}_i, \mathbf{S}_i; \Psi)$$

$$\boldsymbol{\pi} | \dots \sim \mathcal{D}(\alpha_1 + n_1, \dots, \alpha_K + n_K)$$

$$\boldsymbol{\beta}_k | \dots \sim \mathcal{N}(\boldsymbol{\nu}_0, \mathbf{V}_0)$$

$$\mathbf{b}_{ik} | \dots \sim \mathcal{N}(\boldsymbol{\nu}_1, \mathbf{V}_1)$$

$$\sigma_k^2 | \dots \sim \mathcal{IG}(g_1, h_1)$$

$$\xi_k^2 | \dots \sim \mathcal{IG}(a_1, b_1)$$

- relabel the obtained posterior parameter samples if label switching by the K-means-like algorithm of (Celeux, 1999; Celeux et al., 2000).

Handwritten digit clustering using the BMSSR

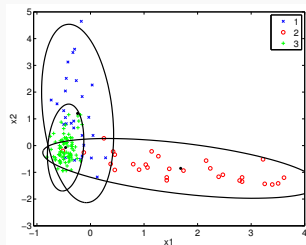
- BMSSR applied on a subset of the ZIPcode data set (issued from MNIST)
- Each individual \mathbf{y}_i contains $m_i = 256$ observations
A subset of 1000 digits randomly chosen from the test set



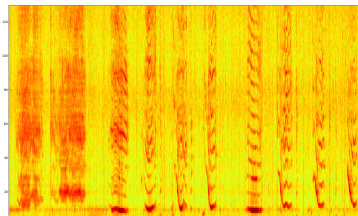
Figure: Cluster mean images obtained by the BMSSR model with 12 mixture components.

The best solution is selected in terms of the Adjusted Rand Index (ARI) values, which promotes a partition with $K = 12$ clusters (ARI: 0.5238).

Multivariate data



Diabetes Benchmark



Spectrum of bioacoustic data

Objectives

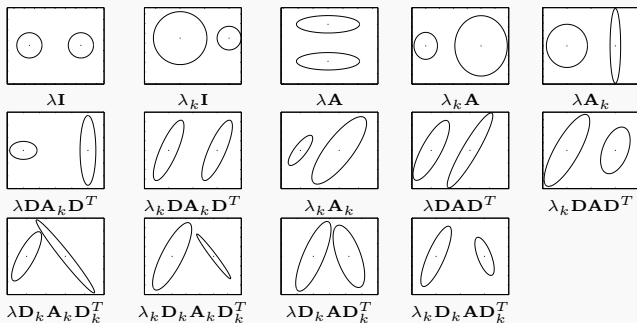
- Clustering
- Dimensionality reduction

Model-Based clustering of multidimensional data

- Data: $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ A sample of n i.i.d observations in \mathbb{R}^d from K sub-populations, with K possibly unknown
- Objective: clustering and dimensionality reduction

Parsimonious mixtures

- Finite Gaussian mixtures: $f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Eigenvalue decomposition of the covariance matrix^a $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$



^aCeleux and Govaert (1995); Banfield and Raftery (1993)

Dirichlet Process Parsimonious Mixtures

- Bayesian parametric inference: (Bensmail, 1995; Bensmail and Celeux, 1996; Bensmail et al., 1997; Bensmail and Meulman, 2003)

PhD thesis of Marius Bartcus, 2012- Oct.2015^a

^aM. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015

- Mixture models for multivariate data in a fully Bayesian framework
- Dirichlet Process and Parsimonious Mixtures [C-5,6,8], [J-11]

Dirichlet Processes (DP)

$DP(\alpha, G_0)$ (Ferguson, 1973) is a distribution over distributions:

$$\tilde{\theta}_i | G \sim G ; \quad G | \alpha, G_0 \sim DP(\alpha, G_0), i = 1, 2, \dots$$

Pólya urn representation (Blackwell and MacQueen, 1973)

$$\tilde{\theta}_i | \tilde{\theta}_1, \dots, \tilde{\theta}_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta_{\theta_k}$$

DP places its probability mass on an infinite mixture of Dirac deltas

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad \theta_k | G_0 \sim G_0, k = 1, 2, \dots, \text{ with } \sum_{k=1}^{\infty} \pi_k = 1$$

$$G|\alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

$$\tilde{\theta}_i|G \sim G$$

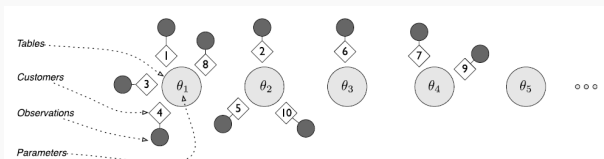
$$\mathbf{x}_i|\tilde{\theta}_i \sim f(\cdot|\tilde{\theta}_i)$$

Chinese Restaurant Process mixtures (Pitman, 2002; Samuel and Blei, 2012)

- Latent variables (z_1, \dots, z_n)

- Predictive distribution:

$$p(z_i = k|z_1, \dots, z_{i-1}; \alpha) = \frac{\alpha}{\alpha + i - 1} \delta(z_i, K_{i-1} + 1) + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta(z_i, k) \cdot$$



- Generative model:

$$z_i|\alpha \sim \text{CRP}(\mathbf{z}_{\setminus i}; \alpha)$$

$$\theta_{z_i}|G_0 \sim G_0$$

$$\mathbf{x}_i|\theta_{z_i} \sim f(\cdot|\theta_{z_i})$$

Implemented parsimonious models

Decomposition	Model-Type	Prior	Applied to
$\lambda \mathbf{I}$	Spherical	\mathcal{IG}	λ
$\lambda_k \mathbf{I}$	Spherical	\mathcal{IG}	λ_k
$\lambda \mathbf{A}$	Diagonal	\mathcal{IG}	each diagonal element of $\lambda \mathbf{A}$
$\lambda_k \mathbf{A}$	Diagonal	\mathcal{IG}	each diagonal element of $\lambda_k \mathbf{A}$
$\lambda \mathbf{DAD}^T$	General	\mathcal{IW}	$\Sigma = \lambda \mathbf{DAD}^T$
$\lambda_k \mathbf{DAD}^T$	General	\mathcal{IG} and \mathcal{IW}	λ_k and $\Sigma = \mathbf{DAD}^T$
$\lambda \mathbf{DA}_k \mathbf{D}^{T*}$	General	\mathcal{IG}	each diagonal element of $\lambda \mathbf{A}_k$
$\lambda_k \mathbf{DA}_k \mathbf{D}^{T*}$	General	\mathcal{IG}	each diagonal element of $\lambda_k \mathbf{A}_k$
$\lambda \mathbf{D}_k \mathbf{AD}_k^T$	General	\mathcal{IG}	each diagonal element of $\lambda \mathbf{A}$
$\lambda_k \mathbf{D}_k \mathbf{AD}_k^T$	General	\mathcal{IG}	each diagonal element of $\lambda_k \mathbf{A}$
$\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{T*}$	General	\mathcal{IG} and \mathcal{IW}	λ and $\Sigma_k = \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{T*}$
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{T*}$	General	\mathcal{IW}	$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{T*}$

Bayesian inference using Gibbs sampling

- Posterior distribution for the component labels:

$$p(z_i = k | \mathbf{z}_{-i}, \mathbf{X}, \Theta, \alpha) \propto p(\mathbf{x}_i | z_i; \Theta) p(z_i | \mathbf{z}_{-i}; \alpha) \text{ with } p(z_i | \mathbf{z}_{-i}; \alpha) \text{ the CRP prior}$$

- Posterior distribution for the component parameters:

$$p(\theta_k | \mathbf{z}, \mathbf{X}, \Theta_{-k}, \alpha; H) \propto \prod_{i|z_i=k} p(\mathbf{x}_i | z_i = k; \theta_k) p(\theta_k; H) \text{ with } p(\theta_k; H) : \text{Prior distribution over } \theta_k$$

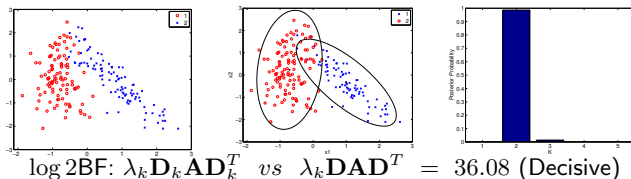
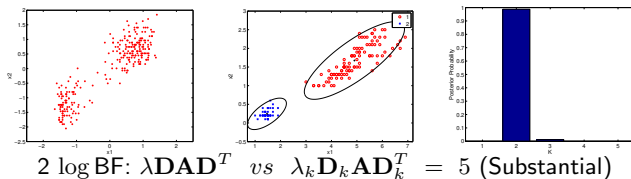
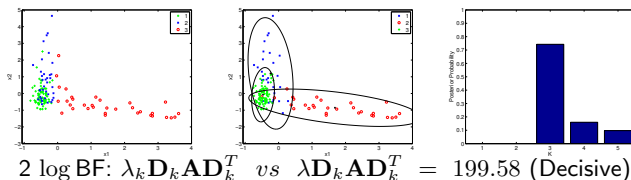
Bayesian model comparison by using Bayes Factors

$$BF_{12} = \frac{p(\mathbf{X}|M_1)p(M_1)}{p(\mathbf{X}|M_2)p(M_2)} \approx \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)} \text{ with the Laplace-Metropolis approximation}$$

$$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\theta_m, M_m) p(\theta_m | M_m) d\theta_m \approx (2\pi)^{\frac{\nu_m}{2}} |\hat{\mathbf{H}}|^{\frac{1}{2}} p(\mathbf{X}|\hat{\theta}_m, M_m) p(\hat{\theta}_m | M_m)$$

Clustering of benchmarks

Diabetes data set, Geyser data set, Crabs data set



Humpback whale song decomposition

- Real fully unsupervised problem
- Data: 8.6 minutes of a Humpback whale song recording (with MFCC)

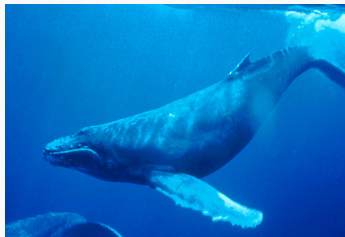


Figure: Humpback Whale.

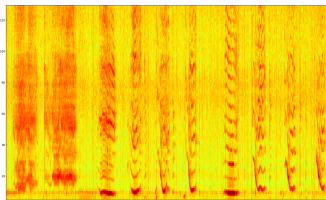
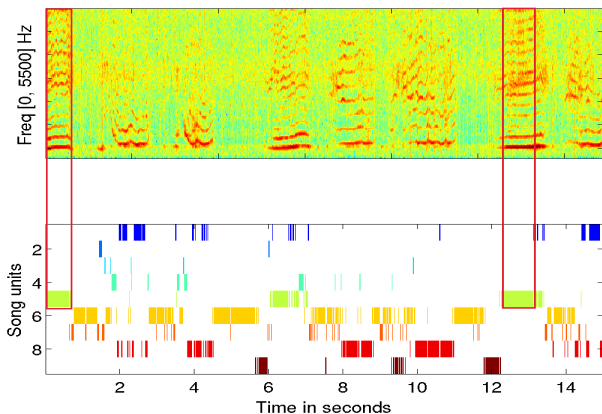


Figure: Spectrum of a signal (20 s).

Objectives

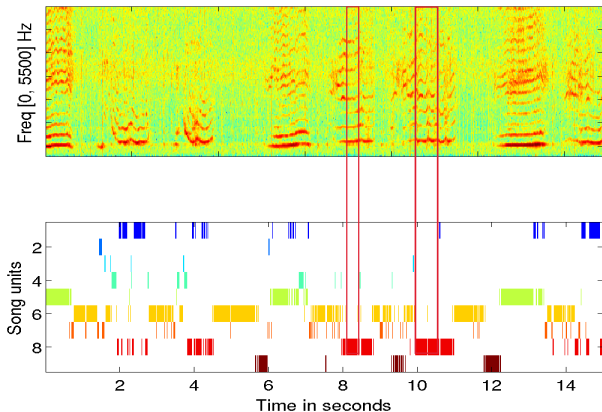
- Discovering “call units”, which can be considered as a whale “alphabet”
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.

Unsupervised decomposition of whale song signals



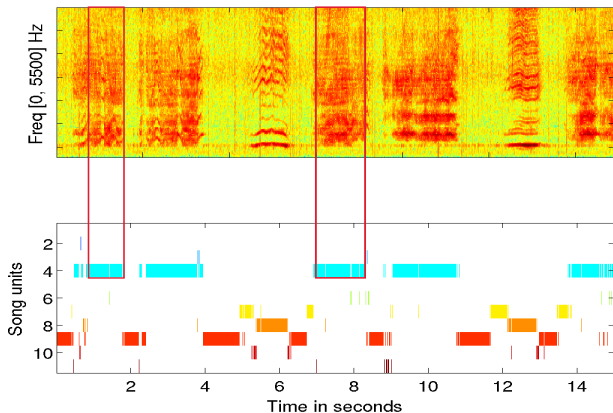
- Sound demo of Unit 5 DPPM λ I: (sec. 0) (sec. 12)

Unsupervised decomposition of whale song signals



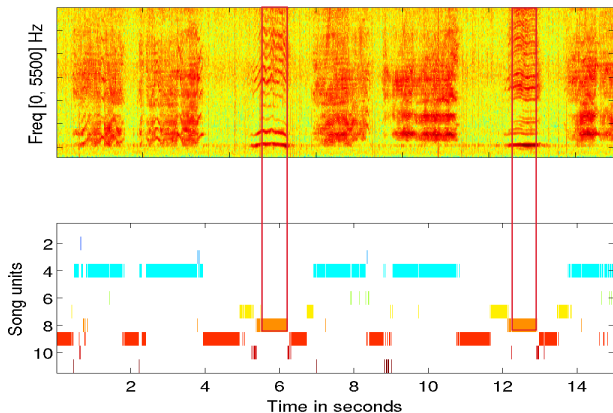
- Sound demo of Unit 8 DPPM λ I: (sec. 8) (sec. 10)

Unsupervised decomposition of whale song signals



- Sound demo of Unit 4 DPPM $\lambda_k \mathbf{A}$: (sec. 1) (sec. 7)

Unsupervised decomposition of whale song signals



- Sound demo of Unit 8 DPPM $\lambda_k \mathbf{A}$: (sec. 6) (sec. 12)

Ongoing research and perspectives

- Advanced mixtures for complex data (My ongoing CNRS leave project)
- Model-based co-clustering for high-dimensional functional data

Functional latent block model (FLBM) available soon on arXiv

Data: $\mathbf{Y} = (\mathbf{y}_{ij})$: 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} .

FLDM model:

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

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

Model inference using Stochastic EM

(Other things: Two ongoing PhD (co-direction with M. Quafafou) on Multilabel learning (funding: Indonesia) and on spatio-temporal analysis of tweets (funding: Algeria))

Variational Learning of Dirichlet Process Parsimonious Mixtures

[Ongoing M2 Internship + expected PACA-PME PhD grant (with H. Glotin)]

- Dirichlet Process parsimonious mixtures (DPPM) and Variational Bayesian learning for DPM (Blei and Jordan, 2006)
- DPPM Clustering for signal segmentation, and hierarchical DPPM for signals clustering (source separation (Moulines et al., 1997; Attias, 1999; Hyvärinen et al., 2001))

<http://chamroukhi.univ-tln.fr//phd-training-positions/FChamroukhi-M2Internship-Variational-DPPM.pdf>

Hierarchical Mixture of Non-Normal Experts [PhD grant (Vietnam) 2016-2019]

- Mixture of experts are universal approximators (Nguyen et al., 2016).
 - ↔ Non-normal (skewed, heavy-tailed) MoE regression, clustering and segmentation
 - ↔ MoE to construct Fisher vectors (Sanchez et al., 2013) for classification
- Hierarchical MoE for unsupervised learning of feature hierarchies:
 - ↔ Hierarchical (deep) MoE as in Eigen et al. (2014)

Possible applications to Biological problems

Probabilistic biological sequence segmentation

- Hidden process regression for sequence segmentation
- e.g., Segmentation of CGH profiles (or other technologies (SNP6 ?))

Functional data analysis models for clustering and segmentation of biological sequences

- Mixture of functional regression models with possibly mixed effects
- Clustering and segmentation of e.g., CGH profiles with individual variability

Latent block models for clustering of biological data

- Latent block model for co-clustering of differentially expressed genes (e.g. time series)
- Latent block model for clustering of biological networks (graph adjacency Matrix)

Bayesian non-parametrics (with variational learning) can probably help to determine the data structure (number of segments, clusters, blocks)

Perspectives

Bayesian learning of sparse representations [Requested PhD grant (Mexico)]

- Predictive Sparse Decomposition (PSD) (Kavukcuoglu et al., 2008; Kavukcuoglu, 2011) which jointly learns a dictionary and approximates the sparse representations by a predictive function (rather than computing exact sparse representations).
- Bayesian Predictive Sparse Decomposition (BPSD):

$$\mathcal{L}(\mathbf{x}, \mathbf{z}; B, \boldsymbol{\theta}) \propto \underbrace{\frac{1}{2} \|\mathbf{x} - B\mathbf{z}\|_2^2}_{\text{Likelihood term}} + \underbrace{\lambda \|\mathbf{z}\|_1}_{\text{Laplacian Prior}} + \underbrace{\alpha \|\mathbf{z} - f(\mathbf{x}; \boldsymbol{\theta})\|_2^2}_{\text{Gaussian Prior centred at } f}$$

Prior term

- \Rightarrow The model parameters : (B, \mathbf{z}) and the **hyperparameters** : $(\alpha, \lambda, \boldsymbol{\theta})$

<http://chamroukhi.univ-tln.fr/FChamroukhi-PhD-Proposal-BPSD.pdf>

Aggregation of mixtures for massive data

\Rightarrow Density estimation and collaborative clustering of massive data

- Consider that the global data distribution is a mixture distribution
- Use ensemble methods to distribute the data
- Bag of Little Bootstraps (BLB) (Kleiner et al., 2014)
- Aggregate local estimators from BLB sub-samples: Hierarchical (mixture) of

Reference papers

Published papers

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

Submitted papers

- [J-12] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, 2015. In revision
- [J-13] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a. (v1) submitted
- [J-14] F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015b. Report (61 pages)
- [J-15] F. Chamroukhi. Robust mixture of experts modeling using the skew- t distribution. 2015d. under review

Thank you for your attention!

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.
- H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.
- A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B*, 65:367–389, 2003.
- Xiuqin Bai, Weixin Yao, and John E. Boyer. Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, 56(7):2347 – 2359, 2012.
- Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.
- M. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015.
- H. Bensmail and Jacqueline J. Meulman. Model-based Clustering with Noise: Bayesian Inference and Estimation. *Journal of Classification*, 20(1):049–076, 2003.
- H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.
- Halima Bensmail. *Modèles de régularisation en discrimination et classification bayésienne*. PhD thesis, Université Paris 6, 1995.
- Halima Bensmail and Gilles Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American statistical Association*, 91(436):1743–1748, 1996.
- C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.
- D. Blackwell and J. MacQueen. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1:353–355, 1973.
- David M. Blei and Michael I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.
- G. Celeux. Bayesian inference for mixture: the label switching problem. Technical report, INRIA Rhone-Alpes, 1999.
- G. Celeux and G. Govaert. Gaussian Parsimonious Clustering Models. *Pattern Recognition*, 28(5):781–793, 1995.

References II

- G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.
- 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. Akin. 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. Akin. 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. Akin. 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 III

- David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskeve. Learning factored representations in a deep mixture of experts. *arXiv:1312.4314v3*, March 2014.
- Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, 2003.
- G. Hébrail, B. Huguency, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7-9):1125–1141, March 2010.
- Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, editors. *Independent Component Analysis*. Wiley, 2001.
- 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.
- M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.
- Koray Kavukcuoglu. *Learning Feature Hierarchies for Object Recognition*. PhD thesis, Department of Computer Science, New York University, 2011.
- Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. Fast inference in sparse coding algorithms with applications to object recognition. Technical Report CBLL-TR-2008-12-01, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 2008.
- 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.
- Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of multivariate skew t -distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202, 2014.
- Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of canonical fundamental skew t -distributions. *Statistics and Computing (To appear)*, 2015. doi: 10.1007/s11222-015-9545-x.
- N. Malfait and J. O. Ramsay. The historical functional linear model. *The Canadian Journal of Statistics*, 31(2), 2003.

References IV

- Eric Moulines, Jean-François Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997*, pages 3617–3620. IEEE Computer Society, 1997. doi: 10.1109/ICASSP.1997.604649. URL <http://dx.doi.org/10.1109/ICASSP.1997.604649>.
- Hien D. Nguyen and Geoffrey J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93: 177–191, 2016. doi: <http://dx.doi.org/10.1016/j.csda.2014.10.016>.
- Hien D. Nguyen, Geoffrey J. McLachlan, and Ian A. Wood. Mixtures of spatial spline regressions for clustering and classification. *Computational Statistics and Data Analysis*, 2014. doi: <http://dx.doi.org/10.1016/j.csda.2014.01.011>.
- Hien D Nguyen, Luke R Lloyd-Jones, and Geoffrey J McLachlan. A universal approximation theorem for mixture of experts models. *arXiv*, Feb 2016. arXiv:1602.03683.
- Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A probabilistic theory of deep learning. Technical Report Technical Report No 2015-1, Rice University Electrical and Computer Engineering Dept., April 2015. URL <http://arxiv.org/abs/1504.00641v1>.
- J.O. Ramsay, T.O. Ramsay, and L.M. Sangalli. *Spatial functional data analysis*, pages 269–275. Springer, 2011.
- 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.
- Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal on Computer Vision (IJCV)*, 105(3):222–245, 2013.
- L.M. Sangalli, J.O. Ramsay, and T.O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society (Series B)*, 75:681–703, 2013.
- Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71(0):128 – 137, 2014.
- 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.
- Ka Yee Yeung, Chris Fraley, A. Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.