On some mixtures for modeling complex data

Faicel Chamroukhi

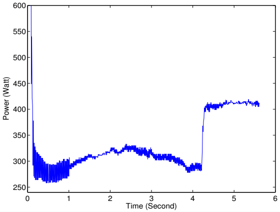
LSIS, UMR CNRS 7296
LPP, UMR CNRS 8524

Modal Seminar, INRIA Lille-Europe

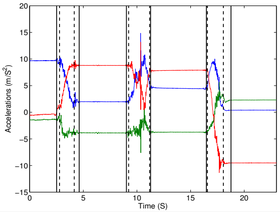
January 12, 2016

Temporal data

Temporal data with regime changes



Railway data



Human activity data

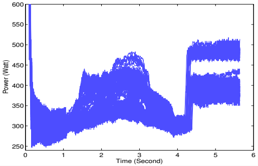
- Data with regime changes over time
- Abrupt and/or smooth regime changes
- Multidimensional temporal data

Objectives

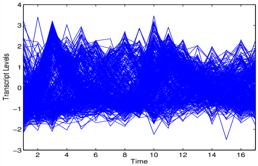
Temporal data modeling and segmentation

Functional data analysis context

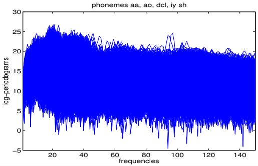
Many curves to analyze



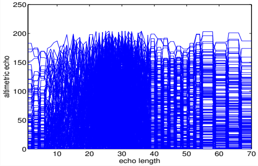
Railway switch curves



Yeast cell cycle curves



Phonemes curves

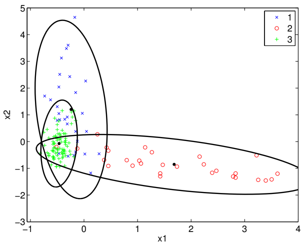


Satellite waveforms

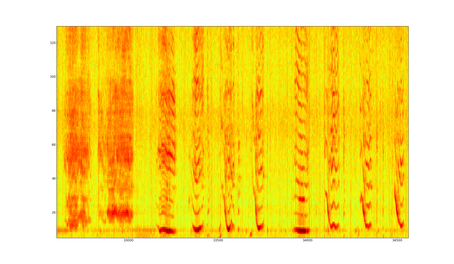
Objectives

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

Multivariate data



Diabetes Benchmark



Spectrum of bioacoustic data

Objectives

- Clustering
- Dimensionality reduction

Data with atypical features

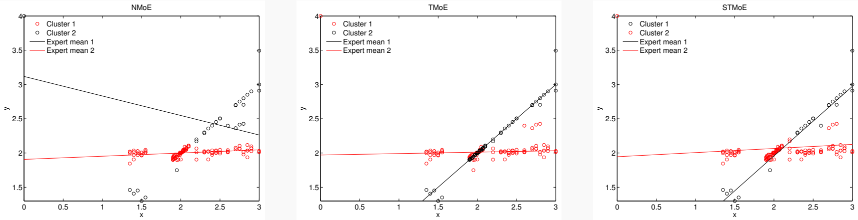


Figure: Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

- 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 modeling framework

Topics

$\hookrightarrow$ exploratory analysis (segmentation/clustering)

$\hookrightarrow$ decisional analysis: make decision and prediction for future data (regression/classification)

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

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

- Fitting such models is in the core of the analysis task

Outline

Outline

1. Mixture models for temporal data segmentation
   - Regression with hidden logistic process
   - Multiple hidden process regression
2. Mixture models for functional data analysis
3. Bayesian regularization of mixtures for functional data
4. Bayesian non-parametric parsimonious mixtures for multivariate data
5. Non-normal mixtures of experts

Mixture models for temporal data segmentation

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

Times series segmentation context

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

- The observed time series is generated by an underlying process
  $\hookrightarrow$ segmentation $\equiv$ recovering the parameters the process' states.
- Conventional solutions are subject to limitations in the control of the transitions between these states

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

Regression with hidden logistic process

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

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

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

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

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

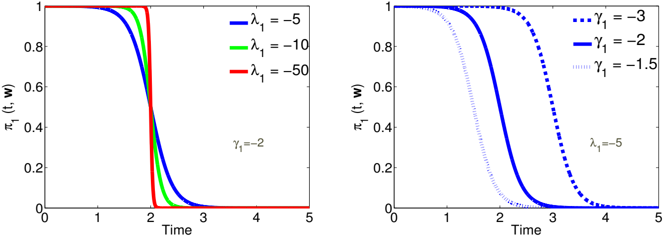
$$f(y_i | t_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)$$

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

Model properties

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

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



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

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

- Ensure time series segmentation into contiguous segments

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 $\hookrightarrow$ IRLS (and weighted polynomial regressions)

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

Time series approximation and segmentation

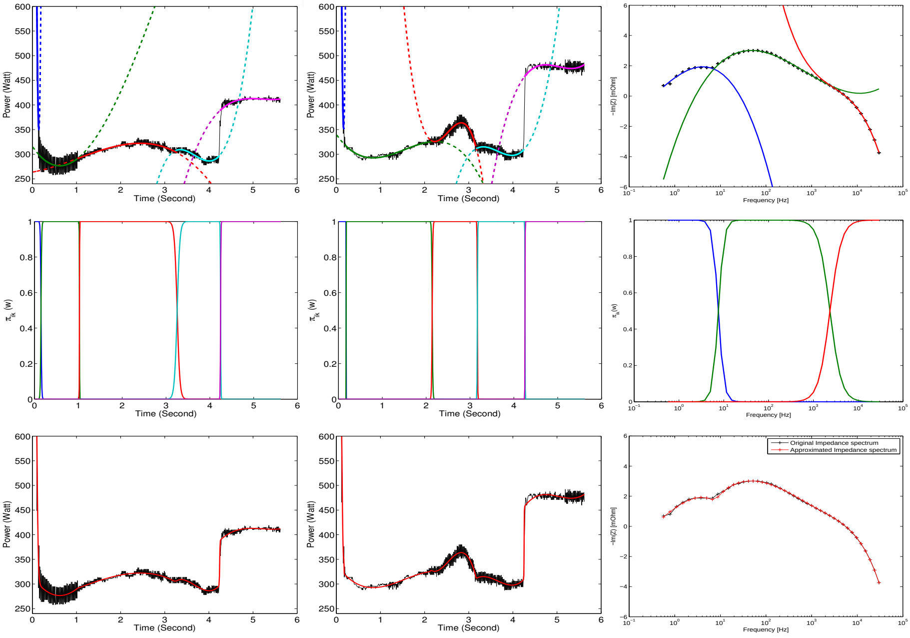
1. Approximation: a curve prototype $\hat{y}_i = \mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i$

   $\hookrightarrow$ The RHLP can be used as nonlinear regression model $y_i = f(t_i;\boldsymbol{\theta}) + \epsilon_i$ by covering functions of the form $f(t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ [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: $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ a time series of n multidimensional observations $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at instants $\mathbf{t} = (t_1, \ldots, t_n)$.
- Model

$$
\begin{aligned}
y_i^{(1)} &= \boldsymbol{\beta}_{z_i}^{(1)T} \boldsymbol{x}_i + \sigma_{z_i}^{(1)} \epsilon_i \\
\vdots \quad & \quad \vdots \\
y_i^{(d)} &= \boldsymbol{\beta}_{z_i}^{(d)T} \boldsymbol{x}_i + \sigma_{z_i}^{(d)} \epsilon_i
\end{aligned}
$$

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

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

PhD of Dorra Trabelsi 2010-2013[a]

[a] D. 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)

  $\hookrightarrow$ Solve multiple regression problems

Application to human activity time series

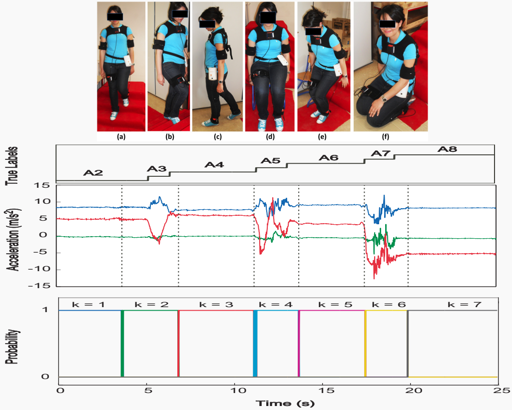


Figure: MHMMR Segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Multiple regression with hidden logistic process

- MRHLP: Estimation by the EM algorithm (as for the RHLP)

  $\hookrightarrow$ Solve multiple regression problems

Application to human activity time series

Problem: Activity recognition from multivariate acceleration time series

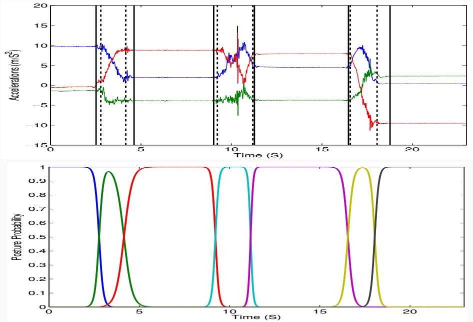


Figure: MRHLP segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Outline

Functional data analysis context

Data

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

Objectives: exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [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)

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

Mixture modeling framework for functional data

- The functional mixture model:

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

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

  $\hookrightarrow$ more tailored to approximate smooth functions

  $\hookrightarrow$ do not account for the segmentation

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

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

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

- ↪ Simultaneously accounts for curve clustering and segmentation

Parameter estimation

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

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

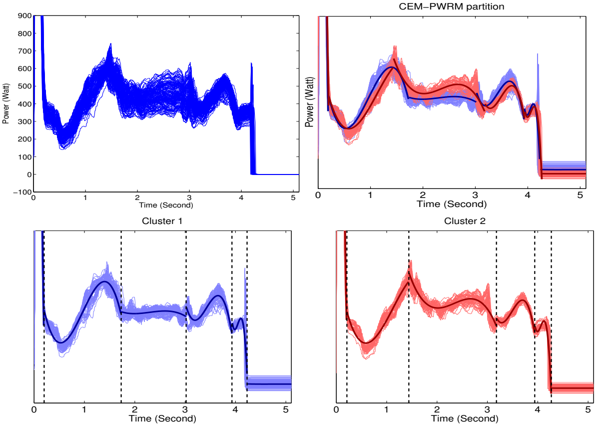
M-step: includes wighted piecewise regression problems ↪ dynamic programming

Complexity in $\mathcal{O}(I_{\text{EM}} K R n m^2 p^3)$: 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)

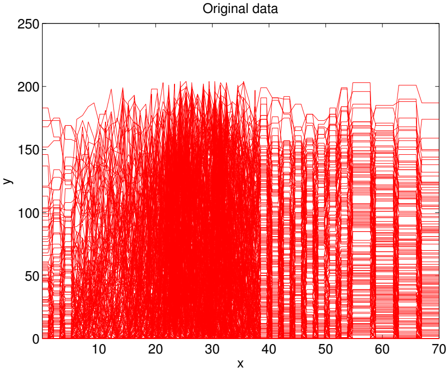


EM-GMM	EM-PRM	EM-PSRM	K-means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

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

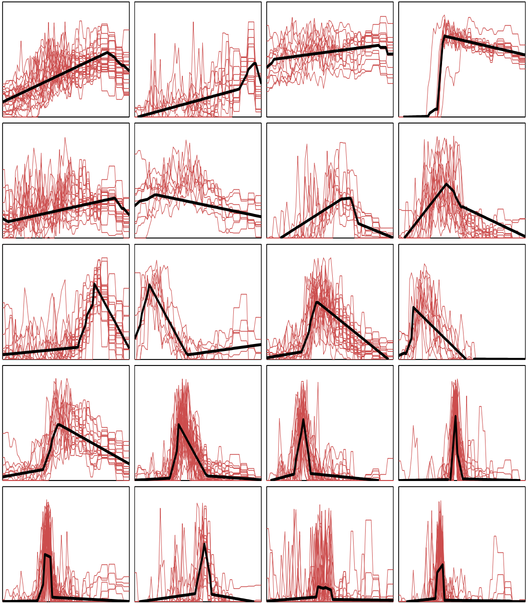
Application to Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data[1] 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).



[1]Satellite 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):

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

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

- Two types of component memberships:

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

  $\hookrightarrow$ 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 $\mathcal{O}(I_{\text{EM}} I_{\text{IRLS}} K R^3 n m p^3)$ (K-means type for piecewise regression is in $\mathcal{O}(I_{\text{KM}} K R n m^2 p^3)$ $\hookrightarrow$ 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 $((\boldsymbol{x}_1, y_1, c_1), \ldots, (\boldsymbol{x}_n, y_n, c_n))$ where $c_i \in \{1, \ldots, G\}$ is the class label of the ith curve
- Problem: predict the class label c_i for a new unlabeled function $(\boldsymbol{x}_i, \boldsymbol{y}_i)$

Tool: Discriminant analysis

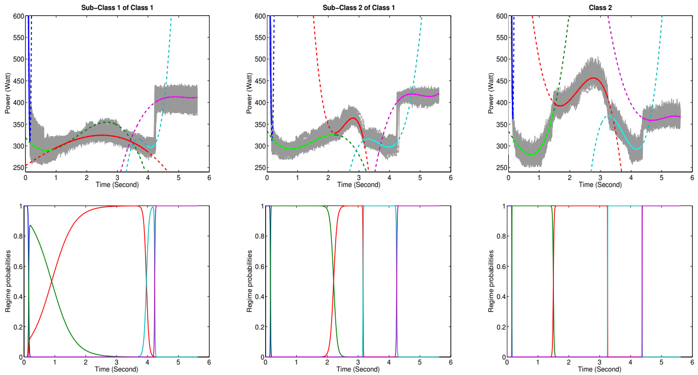
Use the Bayes' allocation rule

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

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

- Homogeneous classes: Functional Linear Discriminant Analysis [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**	$\mathbf{3.2012 \times 10^9}$

Outline

1. Mixture models for temporal data segmentation
2. Mixture models for functional data analysis
3. Bayesian regularization of mixtures for functional data
   - Regularized regression mixtures for functional data
   - Bayesian spatial spline regression with mixed-effects
   - Bayesian mixture of spatial spline regressions with mixed-effects
4. Bayesian non-parametric parsimonious mixtures for multivariate data
5. Non-normal mixtures of experts

The finite Gaussian regression mixture model

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

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

↪ requires careful initialization

↪ requires the number of mixture components K to be supplied by the user

- Initialization strategies Biernacki et al. (2003)
- An afterward model selection procedures: BIC, AIC, 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:

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

- $H(\mathbf{Z}) = -\mathbb{E}[\log \mathbb{P}(\mathbf{Z})]$: - 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 $\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | \boldsymbol{x}_i, \boldsymbol{y}_i; \widehat{\boldsymbol{\Psi}})$
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

$\hookrightarrow$ 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)

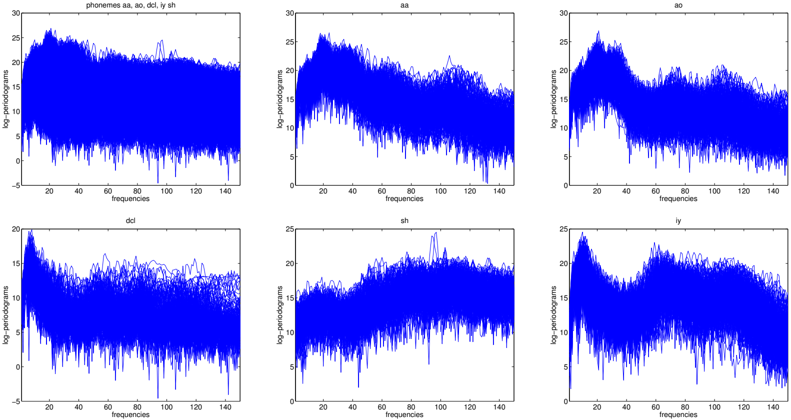
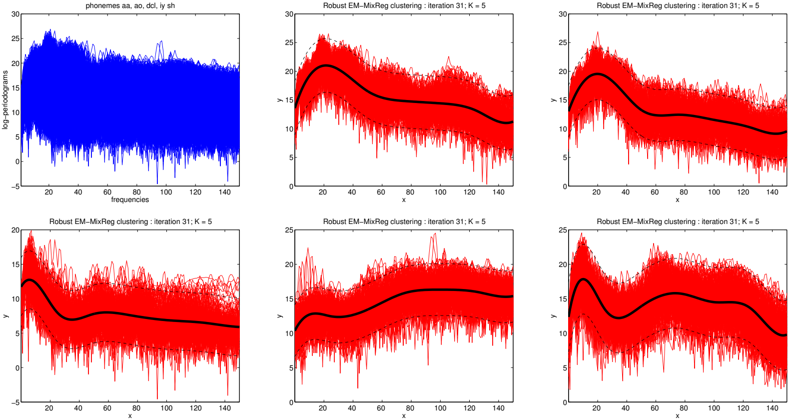


Figure: Original phoneme data and curves of the five classes: "ao", "aa", "yi", "dcl", "sh".

[2] Data 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 %

[3] Data from http://www.math.univ-toulouse.fr/staph/npfda/

Yeast cell cycle data

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

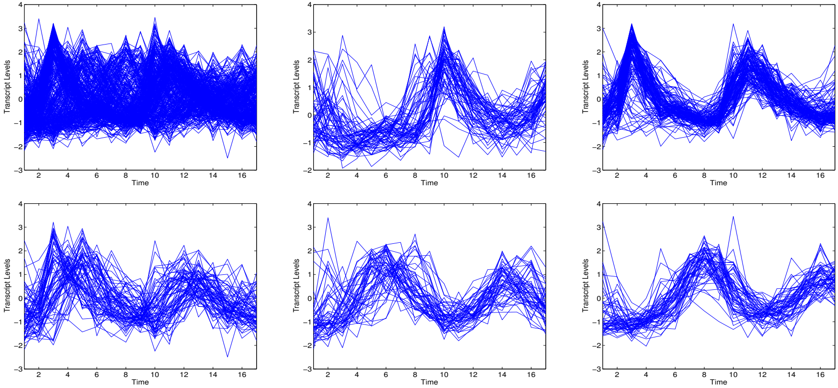


Figure: The five "actual" clusters of the used yeast cell cycle data according to Yeung et al. (2001).

[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.

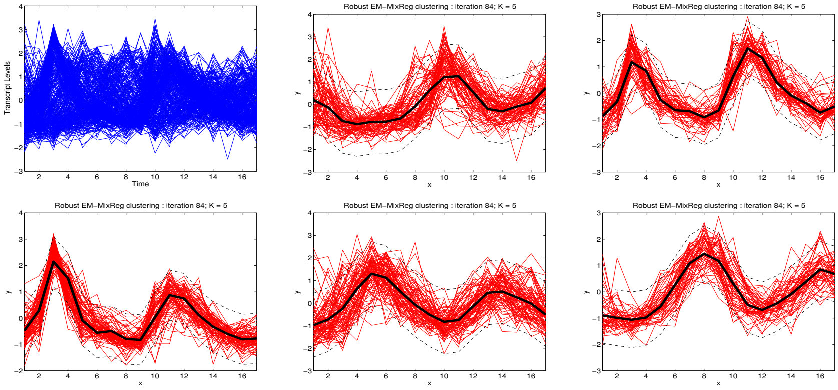


Figure: EM)like clustering results with the bSRM model.

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

Bayesian spatial spline regression with mixed-effects

- Data: $((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ a sample of n surfaces $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^T$ and their spatial coordinates $\boldsymbol{x}_i = ((x_{i11}, x_{i12}), \ldots, (x_{im_i1}, x_{im_i2}))^T$.
- 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 formulation of the models of Nguyen et al. (2014)

Bayesian spatial spline regression with mixed-effects

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

- $\boldsymbol{\beta}$: 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.

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

d: number of basis functions d

$\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2})$ 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_1

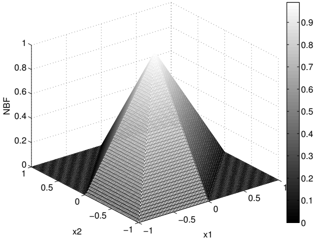


Figure: Nodal basis function $s(\boldsymbol{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0,0)$ and $\delta_1 = \delta_2 = 1$.

Bayesian spatial spline regression with mixed-effects

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

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

Conjugate prior distributions

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

Bayesian inference using Gibbs sampling

- Sample from the full conditional posterior distributions (analytic)

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

Illustration on simulated surfaces' approximation

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

The simulated data include mixed effects.

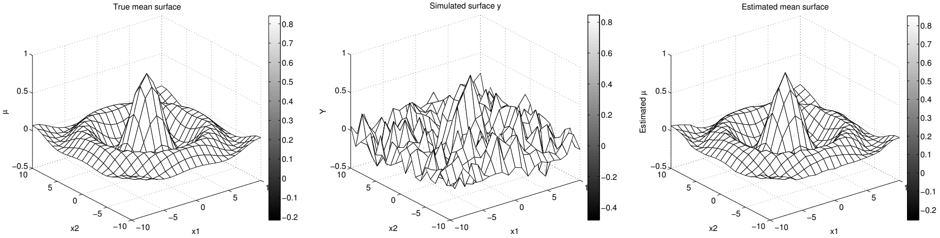


Figure: True mean surface (left), an example of noisy surface (middle), A BSSR fit $\hat{\mu}(\boldsymbol{x}) = \mathbf{S}_i\hat{\boldsymbol{\beta}}$ from 100 surfaces using 15×15 NBFs (right).

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

Bayesian mixture of spatial spline regressions

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

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

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

$\hookrightarrow$ Useful for density estimation and model-based clustering of heterogeneous surfaces

Hierarchical prior from for the BMSSR

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

Bayesian inference of the BMSSR

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

Bayesian inference of the BMSSR using Gibbs sampling

- Sample from the analytic full conditional distributions:

$$Z_i|... \sim \mathcal{M}(1; \tau_{i1}, \ldots, \tau_{iK}) \text{ with } \tau_{ik}(1 \leq k \leq K) = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \mathbf{S}_i; \boldsymbol{\Psi})$$
$$\boldsymbol{\pi}|... \sim \mathcal{D}\left(\alpha_1 + n_1, \ldots, \alpha_K + n_K\right)$$
$$\boldsymbol{\beta}_k|... \sim \mathcal{N}(\boldsymbol{\nu}_0, \mathbf{V}_0)$$
$$\mathbf{b}_{ik}|... \sim \mathcal{N}(\boldsymbol{\nu}_1, \mathbf{V}_1)$$
$$\sigma_k^2|... \sim \mathcal{IG}(g_1, h_1)$$
$$\xi_k^2|... \sim \mathcal{IG}\left(a_1, b_1\right)$$

- 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 $\boldsymbol{y}_i$ 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).

Outline

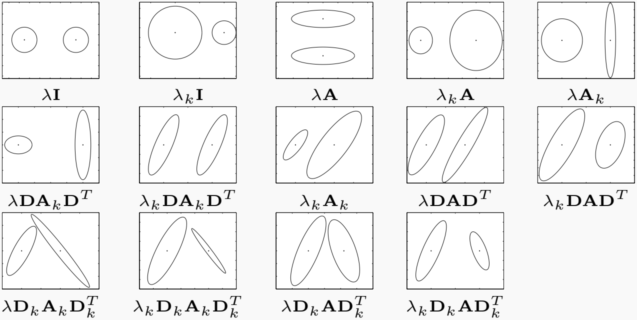
1. Mixture models for temporal data segmentation
2. Mixture models for functional data analysis
3. Bayesian regularization of mixtures for functional data
4. Bayesian non-parametric parsimonious mixtures for multivariate data
   - Dirichlet Process Parsimonious Mixtures
   - Application to whale song decomposition
5. Non-normal mixtures of experts

Model-Based clustering of multidimensional data

- Data: $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ 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: $f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Eigenvalue decomposition of the covariance matrix[a] $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$



[a] Celeux 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]

[a] 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

- 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(α, G_0) (Ferguson, 1973) is a distribution over distributions:

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

Pólya urn representation (Blackwell and MacQueen, 1973)

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

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

$$G = \sum^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \quad \boldsymbol{\theta}_k | G_0 \sim G_0, \; k = 1, 2, \ldots, \text{ with} \sum^{\infty} \pi_k = 1$$

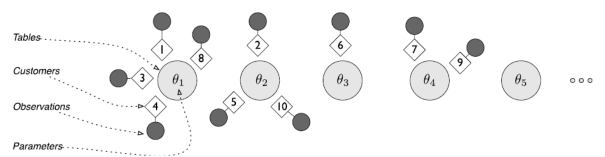
DPM: Generative model

$$\begin{aligned} G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \\ \tilde{\boldsymbol{\theta}}_i|G &\sim G \\ \boldsymbol{x}_i|\tilde{\boldsymbol{\theta}}_i &\sim f(.|\tilde{\boldsymbol{\theta}}_i) \end{aligned}$$

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

- Latent variables $(z_1, \ldots, z_n)$
- Predictive distribution:

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



- Generative model:

$$\begin{aligned} z_i|\alpha &\sim \mathrm{CRP}(\mathbf{z}_{\setminus i}; \alpha) \\ \boldsymbol{\theta}_{z_i}|G_0 &\sim G_0 \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim f(.|\boldsymbol{\theta}_{z_i}) \end{aligned}$$

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	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$
$\lambda\mathbf{DA}_k\mathbf{D}^T*$	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 $\mathbf{\Sigma}_k = \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Bayesian inference using Gibbs sampling

- Posterior distribution for the component labels:
  $p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \mathbf{\Theta}, \alpha) \propto p(\mathbf{x}_i|z_i; \mathbf{\Theta})p(z_i|\mathbf{z}_{-i}; \alpha)$ with $p(z_i|\mathbf{z}_{-i}; \alpha)$ the CRP prior
- Posterior distribution for the component parameters:
  $p(\boldsymbol{\theta}_k|\mathbf{z}, \mathbf{X}, \mathbf{\Theta}_{-k}, \alpha; H) \propto \prod_{i|z_i=k} p(\mathbf{x}_i|z_i = k; \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k; H)$ with $p(\boldsymbol{\theta}_k; H)$: Prior distribution over $\boldsymbol{\theta}_k$

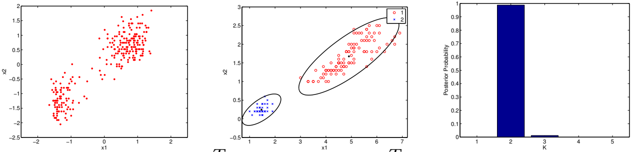
Bayesian model comparison by using Bayes Factors

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

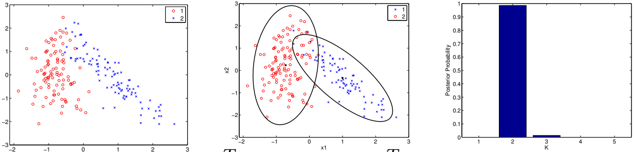
$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)\mathrm{d}\boldsymbol{\theta}_m \approx (2\pi)^{\frac{\nu_m}{2}}|\hat{\mathbf{H}}|^{\frac{1}{2}}p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m)$

Clustering of benchmarks

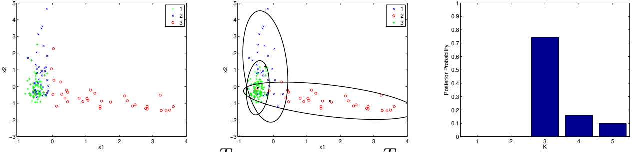
Geyser data set, Crabs data set, Diabetes data set



$2\log$ BF: $\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T \quad vs \quad \lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}_k^T \;=\;$ 5 (Substantial)



$\log 2$BF: $\lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}_k^T \quad vs \quad \lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T \;=\; 36.08$ (Decisive)



$2\log$ BF: $\lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}_k^T \quad vs \quad \lambda \mathbf{D}_k \mathbf{A}\mathbf{D}_k^T \;=\; 199.58$ (Decisive)

Humpback whale song decomposition

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

Figure: Humpback Whale.

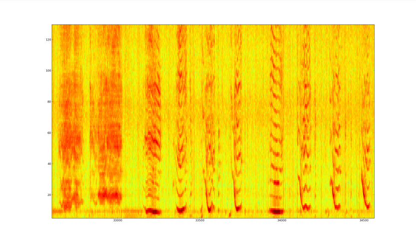
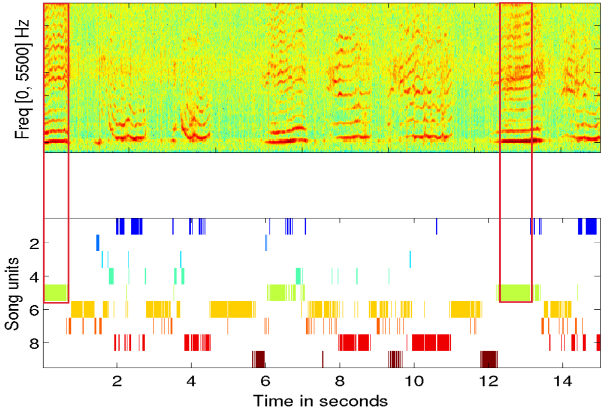


Figure: Spectrum of a signal (20 s).

Objectives

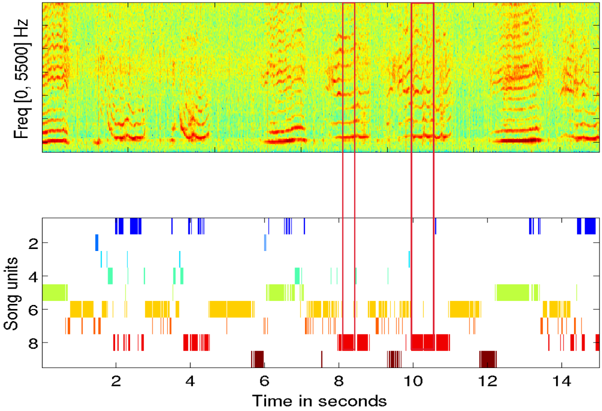
- Discovering "call units", which can be considered as a whale "alphabet"
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.

Unsupervised decomposition of whale song signals



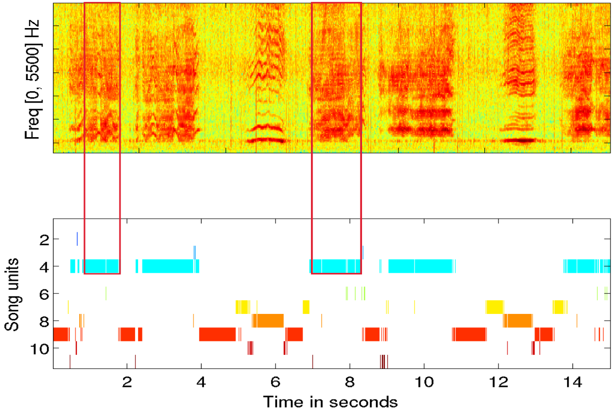
- Sound demo of Unit 5 DPPM $\lambda\mathbf{I}$: (sec. 0) (sec. 12)

Unsupervised decomposition of whale song signals



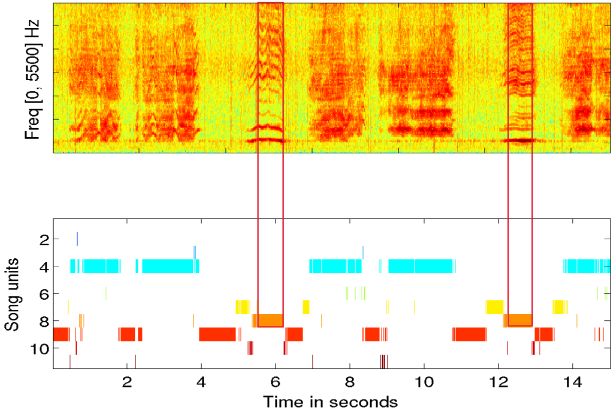
- Sound demo of Unit 8 DPPM $\lambda\mathbf{I}$: (sec. 8) (sec. 10)

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)

Outline

Mixture of Experts (MoE) modeling framework

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

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

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

Objectives

- Overcome (well-known) limitations of modeling with the normal distribution.

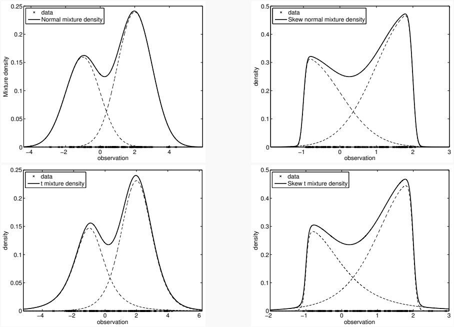
  $\hookrightarrow$ Not adapted For a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations

Non-normal mixtures of experts

Non-normal mixtures of experts (NNMoE)

1. the skew-normal MoE (SNMoE) (skewness) [J-13]
2. the t MoE (TMoE) (Robustness, heavy tails) [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:

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

- kth expert: has skew t distribution (Azzalini and Capitanio, 2003):

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

Model characteristics

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

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

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

$\hookrightarrow$ 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:

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

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

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

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

Temperature anomalies data set

- Data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2016) by using Laplace mixture of linear experts
- $n = 135$ yearly measurements of the global annual temperature anomalies for the period of $1882 - 2012$.

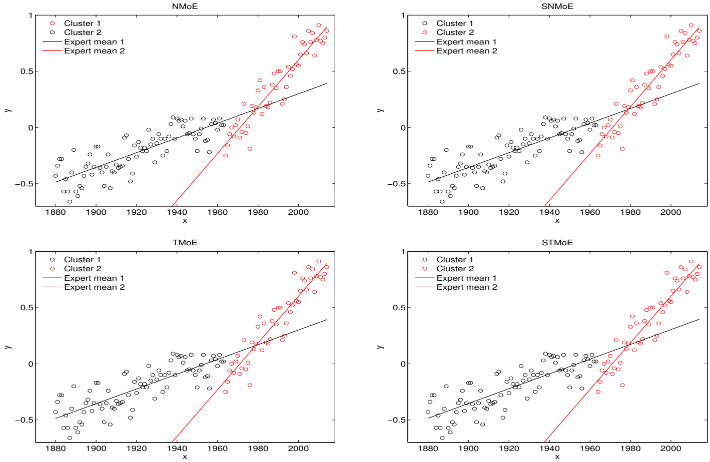


Figure: Fitting the MoLE models to the temperature anomalies data set.

Temperature anomalies data set

- Data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2016) by using Laplace mixture of linear experts
- $n = 135$ yearly measurements of the global annual temperature anomalies for the period of $1882 - 2012$.

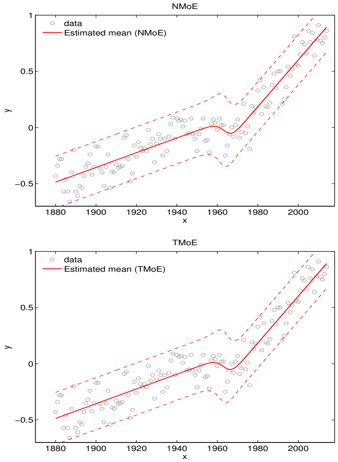


Figure: Fitting the MoLE models to the temperature anomalies data set.

- Both the TMoE and STMoE fits provide a degrees of freedom more than 17, which tends to approach a normal distribution.
- On the other hand, the regression coefficients are also similar to those found by Nguyen and McLachlan (2016) who used a Laplace mixture of linear experts.
- Model selection : Except the result provided by AIC for the NMoE model which overestimates the number of components, all the others results provide evidence for two components in the data.

	NMoE			SNMoE			TMoE			STMoE		
K	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	46.0623	50.4202	46.0623	43.6096	49.4202	43.6096	43.5521	49.3627	43.5521	40.9715	48.2347	40.9715
2	79.9163	91.5374	79.6241	75.0116	89.5380	74.7395	74.7960	89.3224	74.5279	69.6382	87.0698	69.3416
3	71.3963	90.2806	58.4874	63.9254	87.1676	50.8704	63.9709	87.2131	47.3643	54.1267	81.7268	30.6556
4	66.7276	92.8751	54.7524	55.4731	87.4312	41.1699	56.8410	88.7990	45.1251	42.3087	80.0773	20.4948
5	59.5100	92.9206	51.2429	45.3469	86.0207	41.0906	43.7767	84.4505	29.3881	28.0371	75.9742	-8.8817

Table: Choosing the number of expert components K for the temperature anomalies data by using the information criteria BIC, AIC, and 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 "strech ratio" variables considered as responses (y).

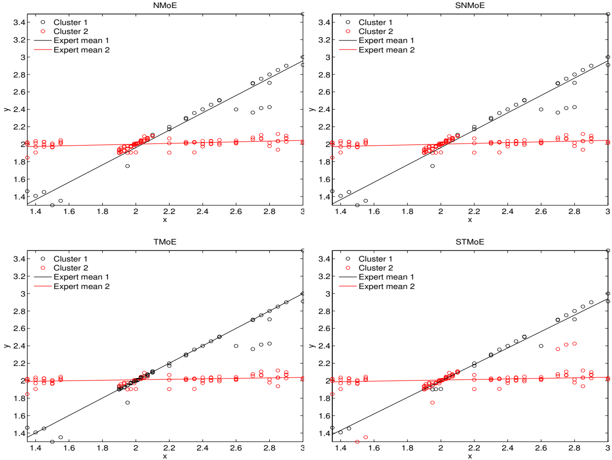


Figure: Fitting the MoE models to the tone data set

Model selection

K	NMoE			SNMoE			TMoE			STMoE		
	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	1.8662	6.3821	1.8662	-0.6391	5.3821	-0.6391	71.3931	77.4143	71.3931	69.5326	77.0592	69.5326
2	122.8050	134.8476	107.3840	122.8725	132.8471	102.4049	204.8241	219.8773	186.8415	92.4352	110.4990	82.4552
3	118.1939	137.7630	76.5249	117.7939	146.9576	98.0442	199.4030	223.4880	183.0389	77.9753	106.5764	52.5642
4	121.7031	148.7989	94.4606	109.5917	142.7087	97.6108	201.8046	234.9216	187.7673	77.7092	116.8474	56.3654
5	141.6961	176.3184	123.6550	107.2795	149.4284	96.6832	187.8652	230.0141	164.9629	79.0439	128.7194	67.7485

Table: Choosing the number of experts K for the original tone perception data.

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

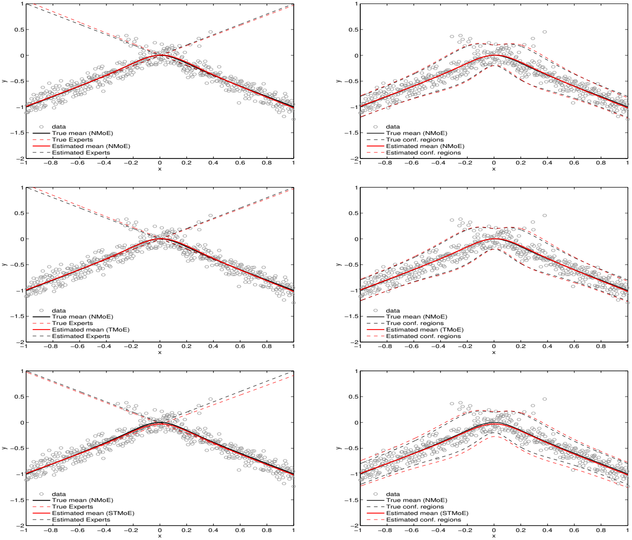


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

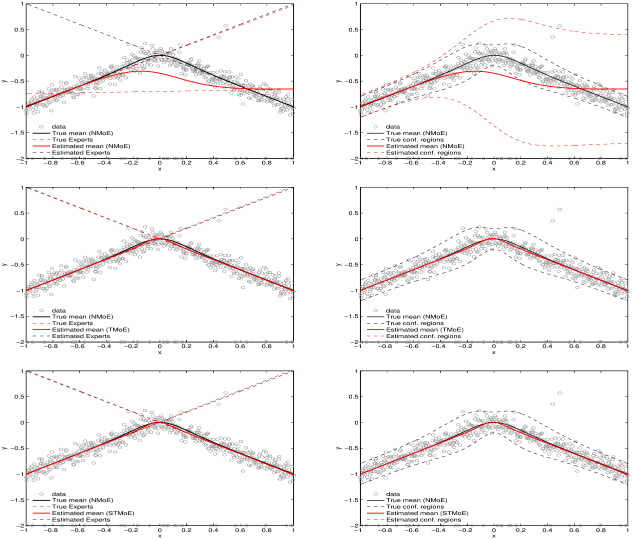


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Robustness of the NNMoE

MSE $\frac{1}{n}\sum_{i=1}^{n}\|\mathbb{E}_{\boldsymbol{\Psi}}(Y_i|\boldsymbol{r}_i,\boldsymbol{x}_i) - \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y_i|\boldsymbol{r}_i,\boldsymbol{x}_i)\|^2$ for different noise levels

	Model | Outliers	0%	1%	2%	3%	4%	5%
NMoE	NMoE	0.0001783	0.001057	0.001241	0.003631	0.013257	0.028966
	SNMoE	0.0001798	0.003479	0.004258	0.015288	0.022056	0.028967
	TMoE	0.0001685	0.000566	0.000464	0.000221	0.000263	0.000045
	STMoE	0.0002586	0.000741	0.000794	0.000696	0.000697	0.000626
SNMoE	NMoE	0.0000229	0.000403	0.004012	0.002793	0.018247	0.031673
	SNMoE	0.0000228	0.000371	0.004010	0.002599	0.018247	0.031674
	TMoE	0.0000325	0.000089	0.000130	0.000513	0.000108	0.000355
	STMoE	0.0000562	0.000144	0.000022	0.000268	0.000152	0.001041
TMoE	NMoE	0.0002579	0.0004660	0.002779	0.015692	0.005823	0.005419
	SNMoE	0.0002587	0.0004659	0.006743	0.015686	0.005835	0.004813
	TMoE	0.0002529	0.0002520	0.000144	0.000157	0.000488	0.000045
	STMoE	0.0002473	0.0002451	0.000173	0.000176	0.000214	0.000291
STMoE	NMoE	0.000710	0.0007238	0.001048	0.006066	0.012457	0.031644
	SNMoE	0.000713	0.0009550	0.001045	0.006064	0.012456	0.031644
	TMoE	0.000279	0.0003808	0.000371	0.000609	0.000651	0.000609
	STMoE	0.000280	0.0001865	0.000447	0.000600	0.000509	0.000602

Table: MSE between the estimated mean function and the true one

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 $(0, 4)$

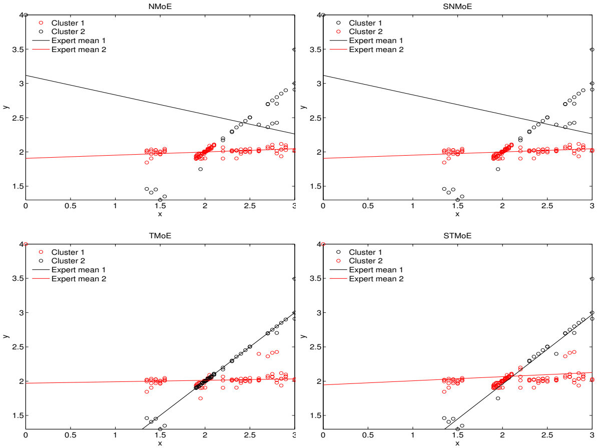
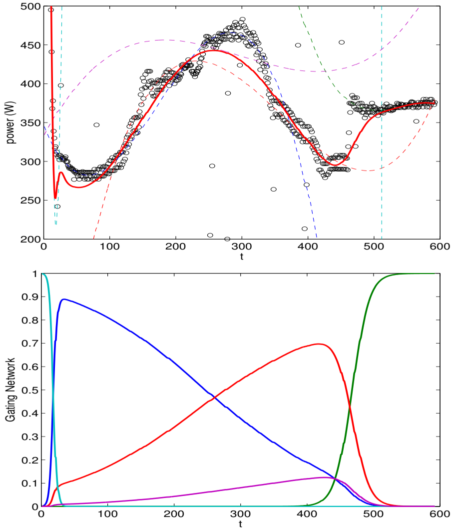


Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

$\hookrightarrow$ 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

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



Ongoing research and perspectives

- Advanced mixtures for complex data, including functional data (My ongoing CNRS leave project)
- LEarning from biG cOmplex FunctIonal daTa - LegoFit (2015 - an ANR proposal, PI with LIPN, IFSTTAR, LIPADE and AIRBUS)

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

- Non-normal mixture modeling
- Feature selection in model-based clustering
- Bayesian latent variable models for sparse representations
- Unsupervised learning of feature hierarchies: Deep learning

  Patel et al. (2015) introduced a probabilistic theory to answer some key questions on deep learning

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 Transactions on Automation Science and Engineering*, 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*, 2015d. doi: 10.1080/00949655.2015.1109096. In Press

[J-9] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015c. Accepted

[J-10] 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

Submitted papers

[J-11] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, January 2015. Submitted

[J-12] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a

[J-13] F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015b. 61 pages

[J-14] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015f. submitted

[J-15] F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015e. submitted

Thank you for your attention!

References I

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

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.

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178, 1985.

A. Azzalini. Further results on a class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 199–208, 1986.

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.

References II

D. Blackwell and J. MacQueen. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1:353–355, 1973.

Richard P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall series in automatic computation. Englewood Cliffs, N.J. Prentice-Hall, 1973. ISBN 0-13-022335-2.

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.

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.

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

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015c. Accepted.

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

F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015e. submitted.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015f. submitted.

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

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

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

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

References III

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*, January 2015. Submitted.

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.

J. Hansen, R. Ruedy, J. Glascoe, and M. Sato. Giss analysis of surface temperature change. *Journal of Geophysical Research*, 104:30997–31022, 1999.

J. Hansen, R. Ruedy, Sato M., M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl. A closer look at united states and global surface temperature change. *Journal of Geophysical Research*, 106:23947–23963, 2001.

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.

Salvatore Ingrassia, Simona Minotti, and Giorgio Vittadini. Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, 29(3):363–401, 2012.

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.

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.

Tsung I. Lin, Jack C. Lee, and Wan J. Hsieh. Robust mixture modeling using the skew *t* distribution. *Statistics and Computing*, 17(2):81–92, 2007a.

References IV

Tsung I. Lin, Jack C. Lee, and Shu Y Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17: 909–927, 2007b.

N. Malfait and J. O. Ramsay. The historical functional linear model. *The Canadian Journal of Statistics*, 31(2), 2003.

Geoffrey J. Mclachlan and David Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, pages 658–666. Springer-Verlag, 1998.

X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2): 267–278, 1993.

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.

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.

D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2):257–286, 1989.

L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986.

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.

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.

References V

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 Transactions on Automation Science and Engineering*, 3(10): 829–335, 2013.

A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

Y. Wei. Robust mixture regression models using t-distribution. Technical report, Master Report, Department of Statistics, Kansas State University, 2012.

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.