Statistical learning of generative models for signal processing

FAICEL CHAMROUKHI

LSIS, UMR CNRS 7296
LPP, UMR CNRS 8524 / Inria-Modal

Seminar, SIGMA - CRISTAL UMR CNRS 9189

March 07, 2016

Research interests

- The area of statistical learning and analysis of complex signals.

  ↪ exploratory analysis of non-stationary signals

  ↪ decisional analysis: make prediction for future observation

Scientific context

- density estimation
- regression
- classification/segmentation

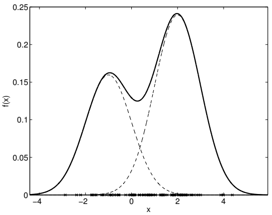
Goals and tools

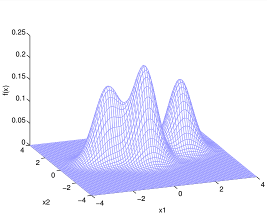
- propose generative probabilistic models
- derive (unsupervised) inference procedures

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

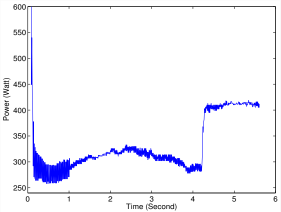
$$
\begin{aligned}
z &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_k) \\
x|z &\sim f(x|z)
\end{aligned}
$$

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

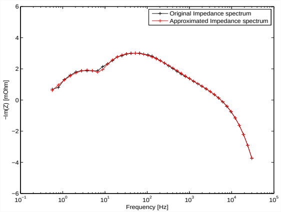
Outline

Non-stationary signals

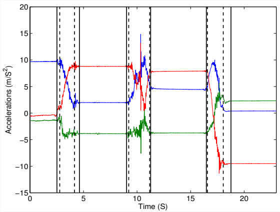
Signals with regime changes



Railway signal



Energy signal



Human activity signal

- Signals with regime changes
- Abrupt and/or smooth regime changes
- Mono-dimensional and Multidimensional signals

Objectives

Signal modeling and segmentation

Outline

1. Mixture modeling for signal approximation and segmentation
   - Regression with hidden logistic process
   - Multiple hidden process regression
   - Non-normal mixtures of experts
2. Mixture modeling for signals classification and segmentation
3. Bayesian (non-)parametric mixtures for surfaces and multivariate signals

Mixture models for signal segmentation

$\boldsymbol{y} = (y_1, \ldots, y_n)$ a signal (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 signal 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 signal 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

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

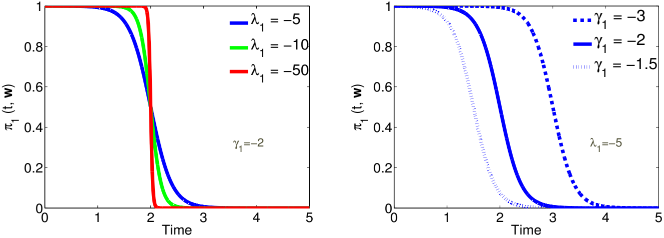
$$f(y_i | t_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)$$

- 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 signal segmentation into contiguous segments

Parameter estimation via a the EM algorithm

Parameter estimation via a the EM algorithm: EM-RHLP

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

  M-Step includes a weighted multinomial logistic regression problem $\hookrightarrow$ IRLS and K 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).

Signal approximation and segmentation

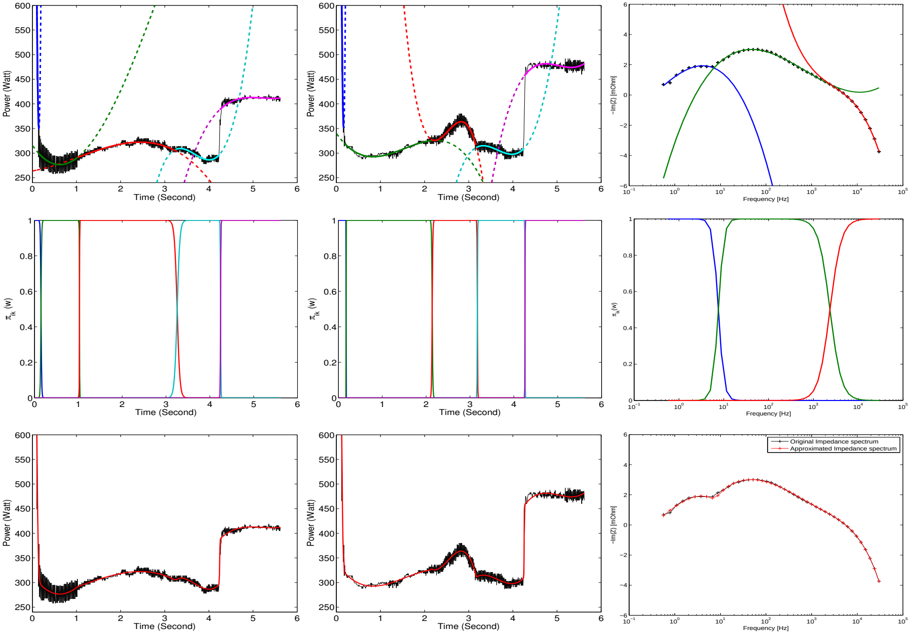
1. Approximation: a signal prototype $\hat{y}_i = \mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i$

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

3. Model selection: Application of BIC, ICL ($\nu_{\boldsymbol{\theta}} = K(p+4) - 2.$)

Application to real-world signals



Joint segmentation of multivariate signals

Multiple hidden process regression

- Data: $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ a signal 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 signal 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

$\hookrightarrow$ Multiple regression with hidden logistic process: Multiple RHLP [J-6]

$\hookrightarrow$ 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

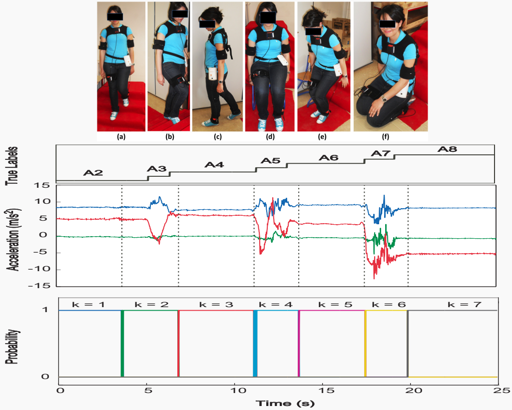


Figure: MHMMR Segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Multiple regression with hidden logistic process

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

  $\hookrightarrow$ Solve multiple regression problems

Application to human activity time series

Problem: Activity recognition from multivariate acceleration time series

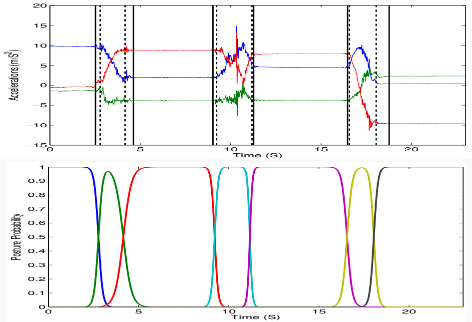
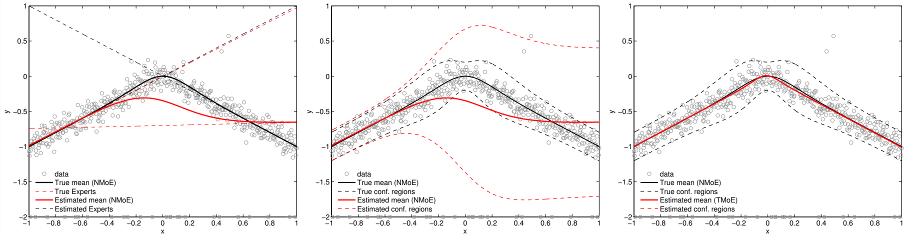


Figure: MRHLP segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)

Signals with atypical characteristics



- Signals 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 $(\boldsymbol{x}, y)$ 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.

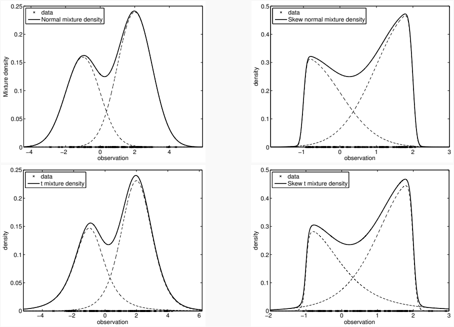
$\hookrightarrow$ Not adapted For a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations

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:

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

- 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:

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

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

$\hookrightarrow$ 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 polynomial regressions

3. Model selection and prediction

$\hookrightarrow$ Model selection: Choose (K, p) using BIC or ICL

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

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

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

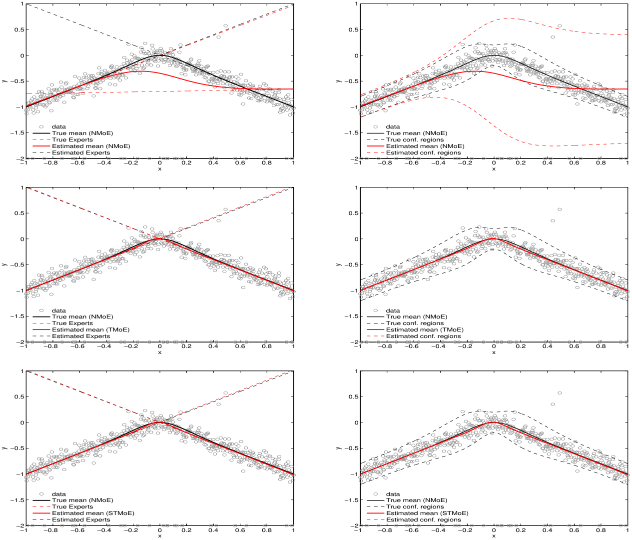
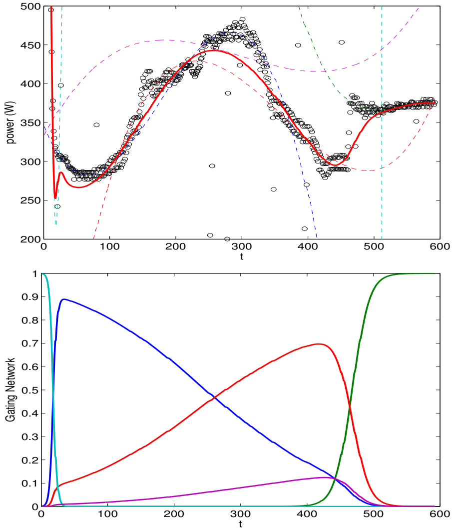


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Segmentation of a noisy railway signal

- $n = 562$ values of the time consumed power
- 30 added artificial outliers



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 $(0, 4)$

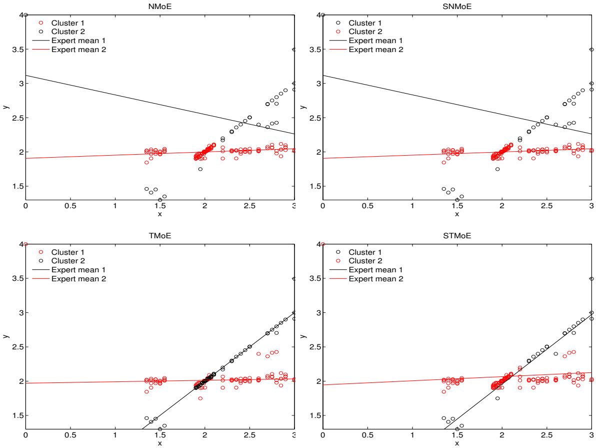


Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

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

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 $1882 - 2012$.

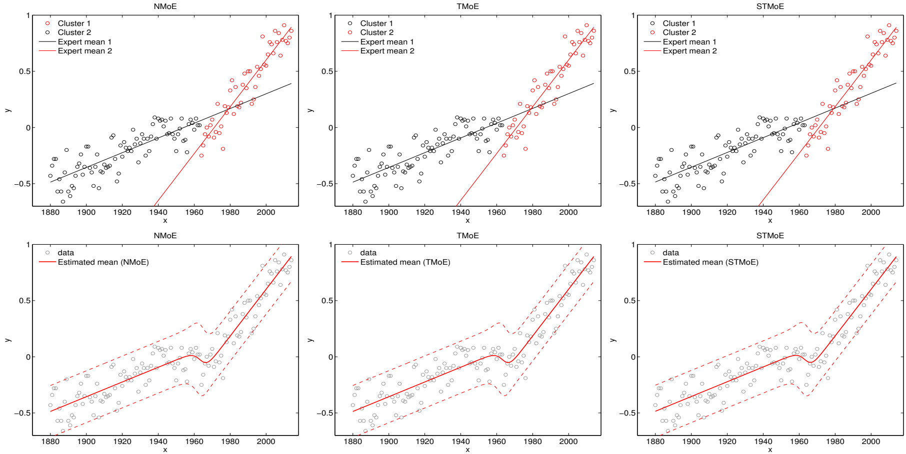
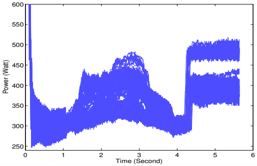


Figure: Fitting the MoLE models to the temperature anomalies data set.

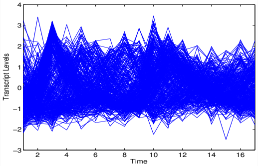
Outline

Functional data analysis context

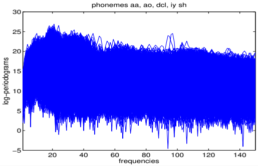
Many signals to analyze simultaneously



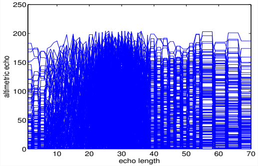
Railway switch signals



Yeast cell cycle curves



Phonemes signals



Satellite waveforms

Objectives

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

Functional data analysis context

Data

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

Objectives: exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly signals 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:

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

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

- $\hookrightarrow$ Simultaneously accounts for curve clustering and segmentation

Parameter estimation

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

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

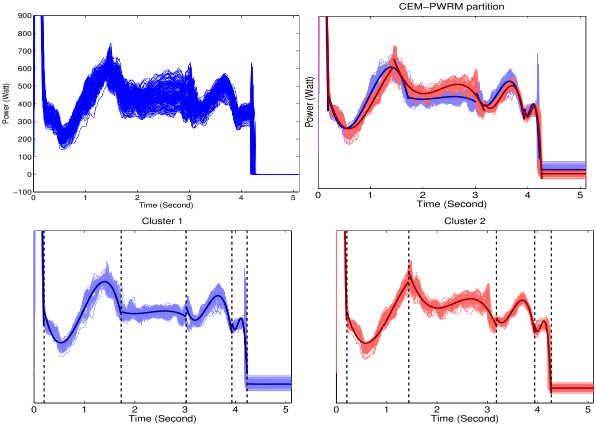
M-step: includes wighted piecewise regression problems $\hookrightarrow$ dynamic programming

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 signals

Data set: $n = 146$ real-word signals of $m = 511$ observations.

Each curve is composed of $R = 6$ electromechanical phases (regimes)

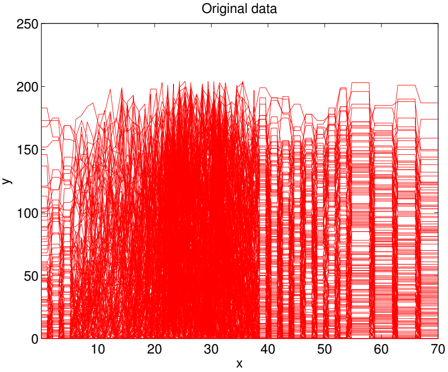


EM-GMM	EM-PRM	EM-PSRM	K-means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

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

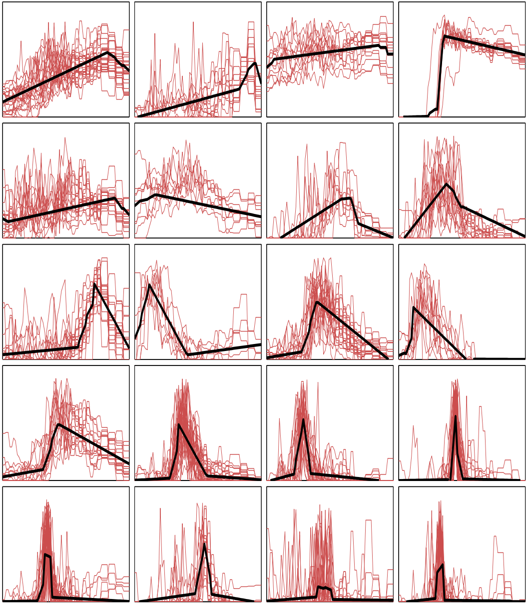
Application to Topex/Poseidon satellite data

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 $\mathcal{O}(I_{\mathsf{EM}} I_{\mathsf{IRLS}} K R^3 n m p^3)$ (K-means type for piecewise regression is in $\mathcal{O}(I_{\mathsf{KM}} K R n m^2 p^3)$ $\hookrightarrow$ 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

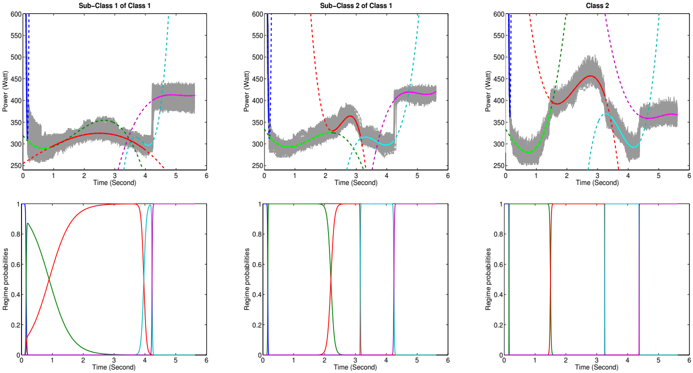
Use the Bayes' allocation rule

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

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**	$\mathbf{3.2012 \times 10^9}$

Regularized regression mixtures

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 (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:

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

- $H(\mathbf{Z}) = -\mathbb{E}[\log \mathbb{P}(\mathbf{Z})]$: - 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)

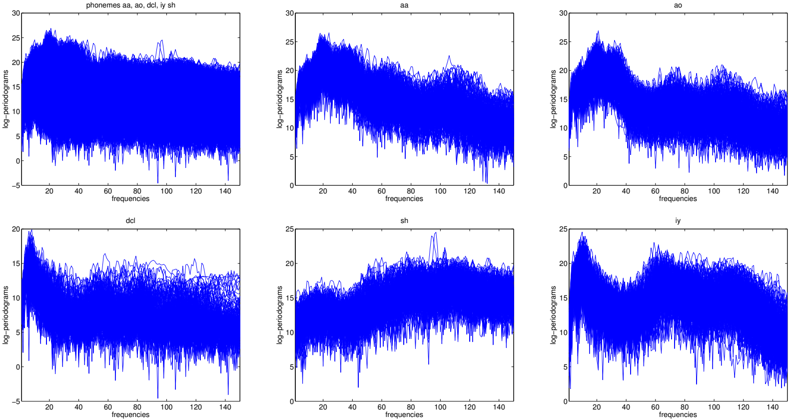


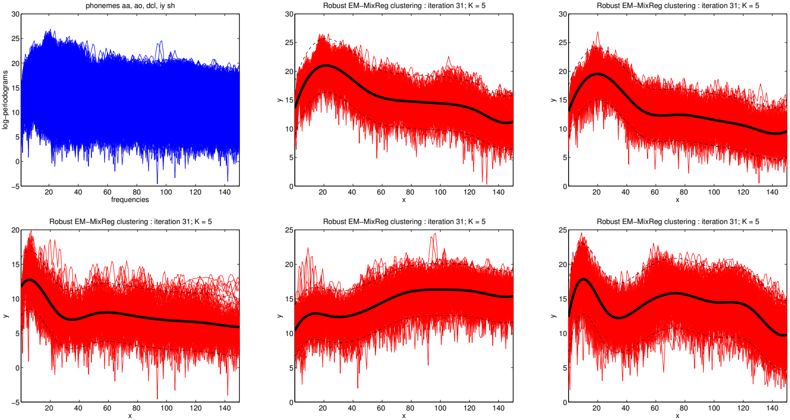
Figure: Original phoneme data and curves of the five classes: "ao", "aa", "yi", "dcl", "sh".

[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.

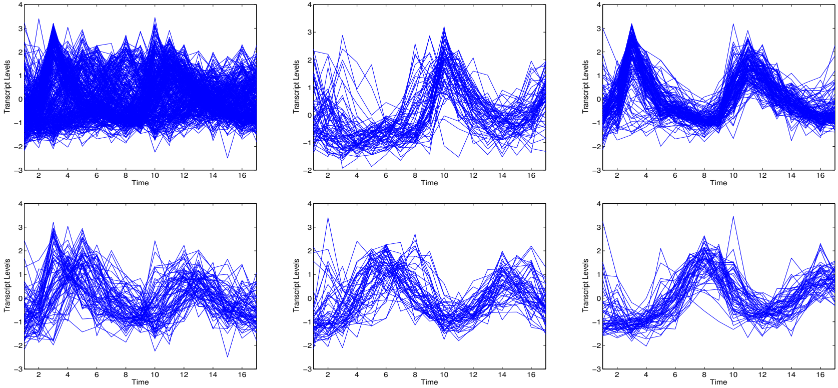


Figure: The five "actual" clusters of the used yeast cell cycle data according to Yeung et al. (2001).

[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.

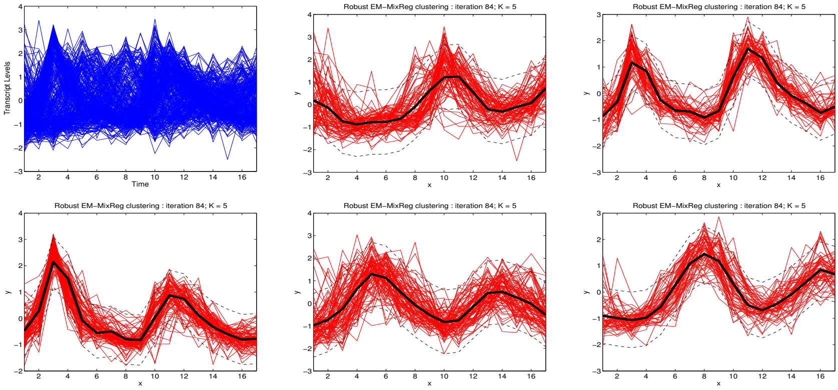


Figure: EM)like clustering results with the bSRM model.

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

Outline

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 spatial spline regression with mixed-effects [Esann 2016, J-12]

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

- $\boldsymbol{\beta}$: 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

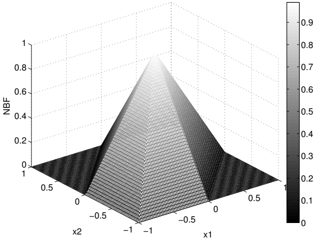


Figure: Nodal basis function $s(\boldsymbol{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$.

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.

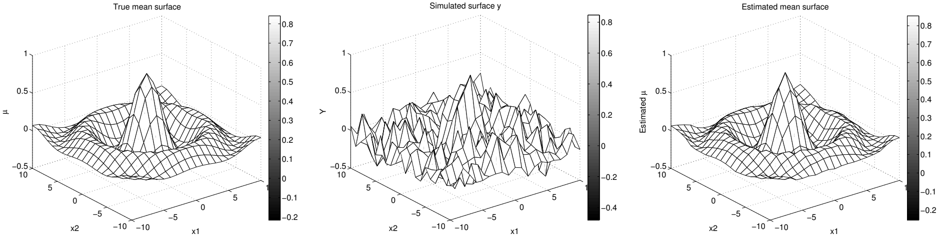


Figure: True mean surface (left), an example of noisy surface (middle), A BSSR fit $\hat{\mu}(\boldsymbol{x}) = \mathbf{S}_i \hat{\boldsymbol{\beta}}$ from 100 surfaces using 15×15 NBFs (right).

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 [Esann 2016, J-12]

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

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:

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

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

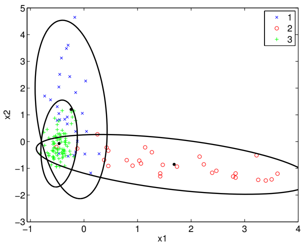
Handwritten digit clustering using the BMSSR

- 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

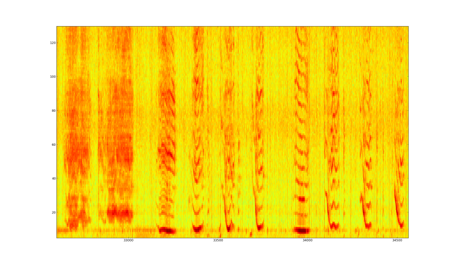
Figure: Cluster mean images obtained by the BMSSR model with 12 mixture components.

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).

Multivariate data



Diabetes Benchmark



Spectrum of bioacoustic data

Objectives

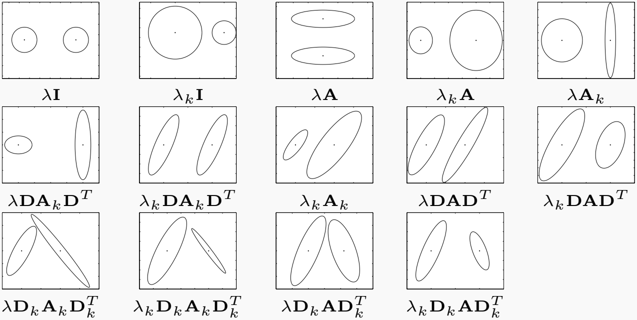
- Clustering
- Dimensionality reduction

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

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

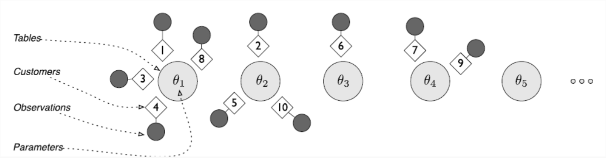
DPM: Generative model

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

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:

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

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 $\boldsymbol{\theta}_k$

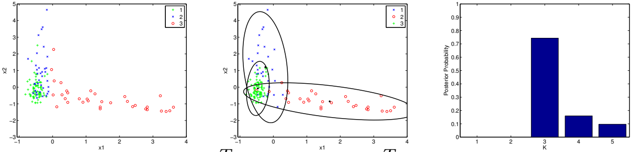
Bayesian model comparison by using Bayes Factors

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

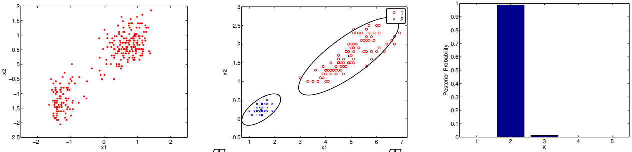
$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)\mathrm{d}\boldsymbol{\theta}_m \approx (2\pi)^{\frac{\nu_m}{2}} |\hat{\mathbf{H}}|^{\frac{1}{2}} p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m)$

Clustering of benchmarks

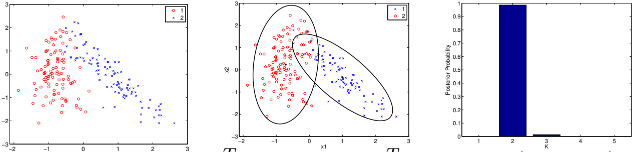
Diabetes data set, Geyser data set, Crabs data set



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



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



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

Humpback whale song decomposition

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

Figure: Humpback Whale.

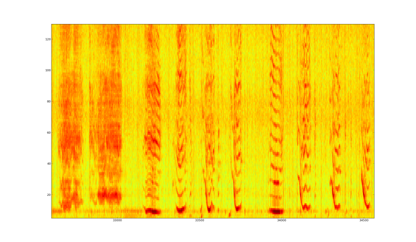
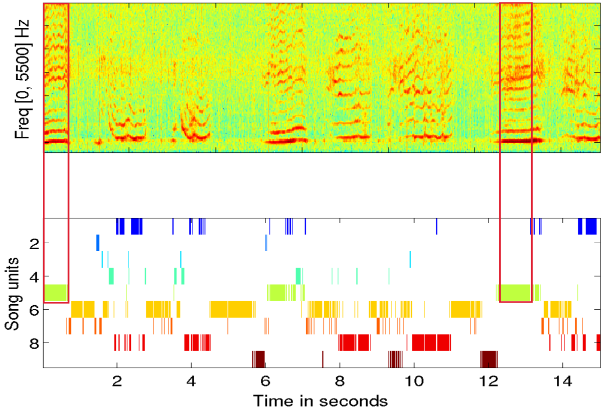


Figure: Spectrum of a signal (20 s).

Objectives

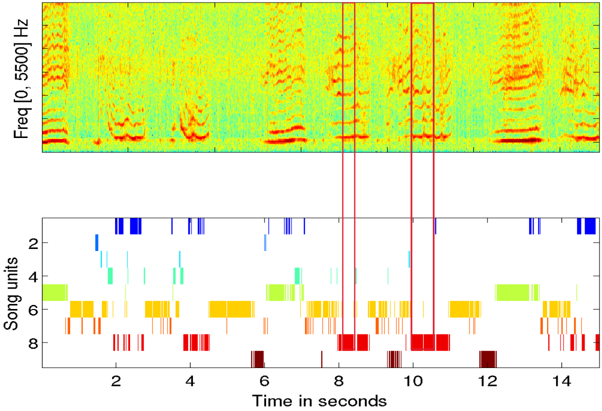
- Discovering "call units", which can be considered as a whale "alphabet"
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.

Unsupervised decomposition of whale song signals



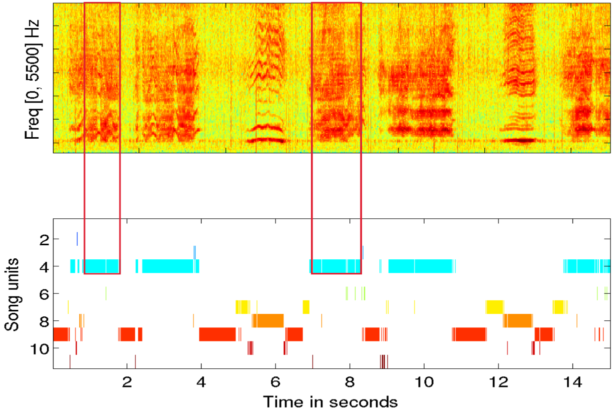
- Sound demo of Unit 5 DPPM $\lambda \mathbf{I}$: (sec. 0) (sec. 12)

Unsupervised decomposition of whale song signals



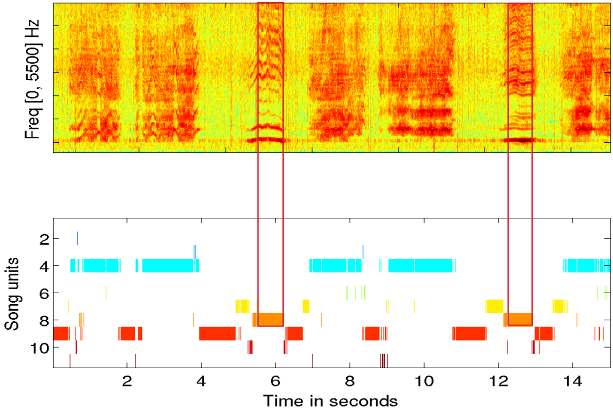
- Sound demo of Unit 8 DPPM $\lambda\mathbf{I}$: (sec. 8) (sec. 10)

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)

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

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

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

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

Perspectives

Hierarchical mixture of experts for signal representation and classification [PhD grant (Vietnam) 2016-2019]

- Mixture of experts are universal approximators (Nguyen et al., 2016).

  $\rightarrow$Consider using MoE in the Fisher space for image/audio classification: Fisher vectors (Sanchez et al., 2013).
- Latent variable models for unsupervised learning of feature hierarchies:

  $\rightarrow$ consider hierarchical (deep) mixtures of experts (MoE) as in Eigen et al. (2014)

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

Perspectives

Variational Learning of Dirichlet Process Parsimonious Mixtures

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

On Variational Bayesian learning for DPM (Blei and Jordan, 2006)

- Consider Dirichlet Process parsimonious mixtures (DPPM)
- Signals decomposition using DPPMs
- Source separation (Moulines et al., 1997; Attias, 1999; Hyvärinen et al., 2001) and signal decomposition using hierarchical DPPMs

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

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

- Consider the problem of learning sparse representations
- 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)
- Application to sounds and/or images representation for recognition.

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

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*, 2015c. 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*, 2016. In press

[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*, 2016. In revision

[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. 2015e. submitted

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

Thank you for your attention!

Perspectives

Mixtures for massive data

$\Rightarrow$ Mixtures for collaborative clustering of massive data

For distributed massive data

- Consider that the global distribution is a mixture distribution
- Probabilisitc aggregation of locally estimated mixtures on distributed data
- e.g. use as a similarity measure the KL divergence

For non-distributed massive data

- Use ensemble methods to distribute the data:
- Bag of Little Boostraps (BLB) (Kleiner et al., 2014)
- Construct local mixture estimators using classical EM of other techniques on each BLB sub-sample

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.

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

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

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

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2016. In press.

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

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

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

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

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

F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, 2016. 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.

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.

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.

References IV

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.

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.

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.

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.