Statistical learning of latent variable models for complex data analysis

Faicel Chamroukhi

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

Seminar, LMNO UMR CNRS 6139

March 15, 2016
Research interests

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

Scientific context

- density estimation
- regression
- classification/segmentation

Goals and tools

- define generative probabilistic models
- propose estimation procedures
Mixture modeling framework

- Mixture density: \( f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \)

- Generative model

\[
\begin{align*}
  z & \sim \mathcal{M}(1; \pi_1, \ldots, \pi_k) \\
  x | z & \sim f(x | z)
\end{align*}
\]

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

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

Temporal data with regime changes

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

Objectives

Temporal data modeling and segmentation
1 Mixture models for temporal data segmentation
   ■ Regression with hidden logistic process
   ■ Multiple hidden process regression
   ■ Non-normal mixtures of experts

2 Mixture models for functional data analysis

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

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

**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
  \( \iff \) segmentation \( \equiv \) recovering the parameters the process’ states.

- Conventional solutions are subject to limitations in the control of the transitions between these states

- Propose generative latent data modeling for segmentation and approximation
  \( \iff \) segmentation \( \equiv \) inferring the model parameters and the underling process
Regression with hidden logistic process

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

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

\[
y_i = \beta_{z_i}^T x_i + \sigma_{z_i} \epsilon_i ; \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad (i = 1, \ldots, n)
\]

\[
Z_i \sim \mathcal{M}(1, \pi_1(t_i; \mathbf{w}), \ldots, \pi_K(t_i; \mathbf{w}))
\]

Polynomial segments \( \beta_{z_i}^T x_i \) with \( x_i = (1, t_i, \ldots, t_i^p)^T \) with logistic probabilities

\[
\pi_k(t_i; \mathbf{w}) = \mathbb{P}(Z_i = k|t_i; \mathbf{w}) = \frac{\exp (w_{k1} t_i + w_{k0})}{\sum_{\ell=1}^{K} \exp (w_{\ell1} t_i + w_{\ell0})}
\]

\[
f(y_i|t_i; \theta) = \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}(y_i; \beta_k^T x_i, \sigma_k^2)
\]

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

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

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

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

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

- Ensure time series segmentation into contiguous segments
**E-Step:** compute the posterior component memberships:

\[
\tau^{(q)}_{ik} = \mathbb{P}(Z_i = k | y_i, t_i; \theta^{(q)}) = \frac{\pi_k(t_i; w^{(q)}) N(y_i; \beta_k^{T(q)} x_i, \sigma_k^{2(q)})}{\sum_{\ell=1}^{K} \pi_{\ell}(t_i; w^{(q)}) N(y_i; \beta_{\ell}^{T(q)} x_i, \sigma_{\ell}^{2(q)})}.
\]

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

- \( w^{(q+1)} = \arg\max_{w} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau^{(q)}_{ik} \log \pi_k(t_i; w) \)  
  weighted logistic regression
- \( \beta_k^{(q+1)} = \left[ \sum_{i=1}^{n} \tau^{(q)}_{ik} x_i x_i^T \right]^{-1} \sum_{i=1}^{n} \tau^{(q)}_{ik} y_i x_i \)  
  weighted polynomial regression
- \( \sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \tau^{(q)}_{ik}} \sum_{i=1}^{n} \tau^{(q)}_{ik} (y_i - \beta_k^{T(q+1)} x_i)^2 \)
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 $\rightarrow$ IRLS (and weighted polynomial regressions)
- EM-RHLP algorithm complexity: $O(I_{\text{EM}}I_{\text{IRLS}}K^3p^3n)$ (more advantageous than dynamic programming).

Time series approximation and segmentation

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

$\leftrightarrow$ The RHLP can be used as nonlinear regression model $y_i = f(t_i; \theta) + \epsilon_i$ by covering functions of the form $f(t_i; \theta) = \sum_{k=1}^{K} \pi_k(t_i; w)\beta_k^T x_i$ [J-3]

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

Model selection: Application of BIC, ICL ($\nu_{\theta} = K(p + 4) - 2.$)
Application to real data
Joint segmentation of multivariate time series

Multiple hidden process regression

- Data: \((y_1, \ldots, y_n)\) a time series of \(n\) multidimensional observations
  \(y_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d\) observed at instants \(t = (t_1, \ldots, t_n)\).

- Model

\[
\begin{align*}
y_i^{(1)} &= \beta_{z_i}^{(1)}^T x_i + \sigma_{z_i}^{(1)} \epsilon_i \\
\vdots & \quad \vdots \\
y_i^{(d)} &= \beta_{z_i}^{(d)}^T x_i + \sigma_{z_i}^{(d)} \epsilon_i
\end{align*}
\]

Vectorial form: \(y_i = B_{z_i}^T x_i + e_i\); \(e_i \sim \mathcal{N}(0, \Sigma_i)\), \((i = 1, \ldots, n)\)

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

PhD of Dorra Trabelsi 2010-2013


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

\leftrightarrow Multiple Hidden Markov model regression (MHMMMR) [J-7]
Multiple hidden Markov model regression

- MHMMR: Estimation by the EM algorithm (as for HMMs)

→ Solve multiple regression problems

Application to human activity time series

![Figure: MHMMR Segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)]
Multiple regression with hidden logistic process

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

⇒ Solve multiple regression problems

Application to human activity time series

Problem: Activity recognition from multivariate acceleration time series

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

- 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

Figure: Fitting MoLE to the tone data set with ten outliers \((0, 4)\).
Mixture of Experts (MoE) modeling framework

- Observed pairs of data \((x, y)\) where \(y \in \mathbb{R}\) is the response for some covariate \(x \in \mathbb{R}^p\) governed by a hidden categorical random variable \(Z\).

- Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994):

\[
f(y|x; \Psi) = \sum_{k=1}^{K} \pi_k(r; \alpha) \cdot f_k(y|x; \Psi_k)
\]

- Gating function of some predictors \(r \in \mathbb{R}^q\): \(\pi_k(r; \alpha) = \frac{\exp(\alpha_k^T r)}{\sum_{\ell=1}^{K} \exp(\alpha_{\ell}^T r)}\)

- MoE for regression usually use normal experts \(f_k(y|x; \Psi_k)\)

Objectives

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

\(\leftrightarrow\) 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-14]
2. the $t$ MoE (TMoE) (Robustness, heavy tails) [J-11]
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:

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

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

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

**Model characteristics**

- For $\{\nu_k\} \rightarrow \infty$, the STMoE reduces to the SNMoE
- For $\{\lambda_k\} \rightarrow 0$, the STMoE reduces to the TMoE.
- For $\{\nu_k\} \rightarrow \infty$ and $\{\lambda_k\} \rightarrow 0$, it approaches the NMoE.
- The STMoE is flexible as it generalizes the previously described models to accommodate situations with asymmetry, heavy tails, and outliers.
Parameter estimation via the ECM algorithm

1. **E-Step:** requires the following conditional expectations:

\[ \tau^{(m)}_{ik} = \mathbb{E}_{\Psi^{(m)}} \left[ Z_{ik} | y_i, x_i, r_i \right], \]
\[ w^{(m)}_{ik} = \mathbb{E}_{\Psi^{(m)}} \left[ W_i | y_i, Z_{ik} = 1, x_i, r_i \right], \]
\[ e^{(m)}_{1,ik} = \mathbb{E}_{\Psi^{(m)}} \left[ W_i U_i | y_i, Z_{ik} = 1, x_i, r_i \right], \]
\[ e^{(m)}_{2,ik} = \mathbb{E}_{\Psi^{(m)}} \left[ W_i U_i^2 | y_i, Z_{ik} = 1, x_i, r_i \right], \]
\[ e^{(m)}_{3,ik} = \mathbb{E}_{\Psi^{(m)}} \left[ \log(W_i) | y_i, Z_{ik} = 1, x_i, r_i \right]. \]

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

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

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

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

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

→ Model selection: Choose \((K, p)\) using BIC or ICL
Tone perception data set

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

Figure: Fitting the MoE models to the tone data set
Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

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

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

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

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

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

1 Mixture models for temporal data segmentation

2 Mixture models for functional data analysis
   - Mixture of piecewise regressions
   - Mixture of hidden Markov model regressions
   - Mixture of hidden logistic process regressions
   - Functional discriminant analysis
   - Regularized regression mixtures for functional data

3 Bayesian (non-)parametric mixtures for spatial and multivariate data
Functional data analysis context

Many curves to analyze

Railway switch curves

Yeast cell cycle curves

Phonemes curves

Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes → Curve segmentation
Functional data analysis context

Data

- The individuals are entire functions (e.g., curves, surfaces)
- A set of $n$ univariate curves $((x_1, y_1), \ldots, (x_n, y_n))$
- $(x_i, y_i)$ consists of $m_i$ observations $y_i = (y_{i1}, \ldots, y_{imi})$ observed at the independent covariates, (e.g., time $t$ in time series), $(x_{i1}, \ldots, x_{imi})$

Objectives: exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [J-4] [J-9], [C-11] [J-16]
2. Discriminant analysis of functional data: [J-2], [J-5]

Functional data clustering/classification tools

- A broad literature (Kmeans-type, Model-based, etc)

  ⇒ Mixture-model based cluster and discriminant analyzes
Mixture modeling framework for functional data

- The functional mixture model:

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

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

→ more tailored to approximate smooth functions

→ do not account for the segmentation

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

1. Piecewise 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)

$$f(y_i|x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} N(y_{ij}; \beta_{kr}^T x_{ij}, \sigma_{kr}^2)$$

$I_{kr} = (\xi_{kr}, \xi_{k,r+1}]$ 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 weighted piecewise regression problems → dynamic programming

Complexity in $O(I_{EM}KRnm^2p^3)$: 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)

\[
\begin{array}{cccc}
\text{EM-GMM} & \text{EM-PRM} & \text{EM-PSRM} & \text{K-means-like} \\
721.46 & 738.31 & 734.33 & 704.64 & \text{CEM-PWRM} & 703.18 \\
\end{array}
\]

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):

$$f(y_i|x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2)$$

- Two types of component memberships:
  - Cluster memberships (global): $Z_{ik} = 1$ iff $Z_i = k$
  - Regime memberships for a given cluster (local): $H_{ijr} = 1$ iff $H_{ij} = r$

MixRHLP deals better with the quality of regime changes

- Parameter estimation via the EM algorithm: EM-MixRHLP

- EM-MixRHLP has complexity in $O(I_{EM}I_{IRLS}KR^3nmp^3)$ ($K$-means type for piecewise regression is in $O(I_{KM}KRnm^2p^3) \leftrightarrow EM$-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 \( ((x_1, y_1, c_1), \ldots, (x_n, y_n, c_n)) \)
  where \( c_i \in \{1, \ldots, G\} \) is the class label of the \( i \)th curve
- Problem: predict the class label \( c_i \) for a new unlabeled function \( (x_i, y_i) \)

Tool: Discriminant analysis

Use the Bayes’ allocation rule

\[
\hat{c}_i = \arg \max_{1 \leq g \leq G} \frac{\mathbb{P}(C_i = g) f(y_i | x_i; \Psi_g)}{\sum_{g' = 1}^{G} \mathbb{P}(C_i = g') f(y_i | x_i; \Psi_{g'})},
\]

based on a generative model \( f(y_i | x_i; \Psi_g) \) for each group \( g \)

- Homogeneous classes: Functional Linear Discriminant Analysis [J-2]
- Dispersed classes: Functional Mixture Discriminant Analysis [J-5]
Applications to switch curves

<table>
<thead>
<tr>
<th>Approach</th>
<th>Classification error rate (%)</th>
<th>Intra-class inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLDA-PR</td>
<td>11.5</td>
<td>$10.7350 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-SR</td>
<td>9.53</td>
<td>$9.4503 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-RHLP</td>
<td>8.62</td>
<td>$8.7633 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-PRM</td>
<td>9.02</td>
<td>$7.9450 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-SRM</td>
<td>8.50</td>
<td>$5.8312 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-MixRHLP</td>
<td><strong>6.25</strong></td>
<td><strong>3.2012 \times 10^9</strong></td>
</tr>
</tbody>
</table>
Regularized regression mixtures

The finite Gaussian regression mixture model

\[ f(y_i | x_i; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y_i; X_i \beta_k, \sigma_k^2 I_{m_i}) \]

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

\[ \text{requires careful initialization (Biernacki et al., 2003)} \]
\[ \text{requires the number of components } K \text{ to be supplied by the user (or BIC, ICL etc)} \]

Idea of the proposed approach [J-8]

\[ \text{A fully unsupervised fitting of regression mixtures} \]
\[ \text{EM-like algorithm which is robust with regard initialization and infers the number of components from the data} \]
Regularized regression mixtures [J-8]

- Penalized log-likelihood criterion:

\[ J(\lambda, \Psi) = \log L(\Psi) - \lambda H(z), \quad \lambda \geq 0 \]

\[ J = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(y_i; X_i \beta_k, \sigma_k^2 I_m) + \lambda n \sum_{k=1}^{K} \pi_k \log \pi_k \]

- \( H(Z) = -\mathbb{E}[\log \mathbb{P}(Z)]: - \) entropy accounting for model complexity

- \( \lambda \geq 0 \) is a smoothing parameter

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

**Initialization:** \( K^{(0)} = n; \ \pi_k^{(0)} = \frac{1}{K^{(0)}}, \ (\beta_k^{(0)}, \sigma_k^{2(0)}): \) polynomial regression

1. **E-step:** Posterior component memberships \( \tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | x_i, y_i; \hat{\Psi}) \)

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

\( \beta_k^{(q+1)} = \left[ \sum_{i=1}^{n} \tau_{ik}^{(q)} X_i^T X_i \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} X_i^T y_i \)

\( \sigma_k^{2(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \| y_i - X_i \beta_k \|^2}{m \sum_{i=1}^{n} \tau_{ik}^{(q)}} \)

The penalization coefficient \( \lambda \) is set in an adaptive way

\( \leftrightarrow \) 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".

²Data from http://www.math.univ-toulouse.fr/staph/npfda/
EM-like clustering results for Phonemes

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

1000 log-periodograms (200 per cluster)

<table>
<thead>
<tr>
<th></th>
<th>EM-PRM</th>
<th>EM-SRM</th>
<th>EM-bSRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $K$</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Misc. error rate</td>
<td>14.29 %</td>
<td>14.09 %</td>
<td>14.2 %</td>
</tr>
</tbody>
</table>

³Data 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

1. Mixture models for temporal data segmentation

2. Mixture models for functional data analysis

3. Bayesian (non-)parametric mixtures for spatial and multivariate data
   - Bayesian spatial spline regression with mixed-effects
   - Bayesian mixture of spatial spline regressions with mixed-effects
   - Dirichlet Process Parsimonious Mixtures for multivariate data clustering
   - Application to whale song decomposition
Bayesian spatial spline regression with mixed-effects

- Data: \(((x_1, y_1), \ldots, (x_n, y_n))\) a sample of \(n\) surfaces \(y_i = (y_{i1}, \ldots, y_{im_i})^T\) and their spatial coordinates \(x_i = ((x_{i11}, x_{i12}), \ldots, (x_{im_i1}, x_{im_i2}))^T\).

- 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-13]

\[ y_i = S_i (\beta + b_i) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2 I_{m_i}), \quad (i = 1, \ldots, n) \]

- \(\beta\): fixed-effects regression coefficients
- \(b_i\): random subject-specific regression coefficients \(b_i \perp e_i \sim \mathcal{N}(0, \xi^2 I_{m_i})\)
- \(S_i\) is a spatial design matrix.
- $S_i$ 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.

$$S_i = \begin{pmatrix}
    s(x_1; c_1) & s(x_1; c_2) & \cdots & s(x_1; c_d) \\
    s(x_2; c_1) & s(x_2; c_2) & \cdots & s(x_2; c_d) \\
    \vdots & \vdots & \ddots & \vdots \\
    s(x_{m_i}; c_1) & s(x_{m_i}; c_2) & \cdots & s(x_{m_i}; c_d)
\end{pmatrix}
$$

$d$: number of basis functions $d$

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

c = $(c_1, c_2)$ is a node center parameter, with v/h shape parameters $\delta_1$ and $\delta_1$

Figure: Nodal basis function $s(\mathbf{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$. 
Bayesian spatial spline regression with mixed-effects

Under the BSRR model, the density of the observation $y_i$ is given by

$$f(y_i|S_i; \Psi) = \mathcal{N}(y_i; S_i\beta, \xi^2 S_i S_i^T + \sigma^2 I_{m_i}).$$

**Conjugate prior distributions**

- $\beta \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $b_i | \xi^2 \sim \mathcal{N}(0_d, \xi^2 I_d)$
- $\xi^2 \sim \mathcal{IG}(a_0, b_0)$
- $\sigma^2 \sim \mathcal{IG}(g_0, h_0)$

**Bayesian inference using Gibbs sampling**

- Sample from the full conditional posterior distributions (analytic)

  - $\beta | ... \sim \mathcal{N}(\nu_0, V_0)$
  - $b_i | ... \sim \mathcal{N}(\nu_1, V_1)$
  - $\sigma^2 | ... \sim \mathcal{IG}(g_1, h_1)$
  - $\xi^2 | ... \sim \mathcal{IG}(a_1, b_1)$
Illustration on simulated surfaces’ approximation

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

The simulated data include mixed effects.

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

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

Data: A sample of \( n \) surfaces \((y_1, \ldots, y_n)\) and their spatial covariates \((S_1, \ldots, S_n)\) issued from \( K \) sub-populations

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

\[
f(y_i | S_i; \Psi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y_i; S_i(\beta_k + b_{ik}), \sigma_k^2 I_{m_i})
\]

\( \rightarrow \) Useful for density estimation and model-based clustering of heterogeneous surfaces

Hierarchical prior from for the BMSSR

\[
\begin{align*}
\pi & \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \\
\beta_k & \sim \mathcal{N}(\mu_0, \Sigma_0) \\
b_{ik} | \xi_k^2 & \sim \mathcal{N}(0_d, \xi_k^2 I_d) \\
\xi_k^2 & \sim \mathcal{IG}(a_0, b_0) \\
\sigma_k^2 & \sim \mathcal{IG}(g_0, h_0).
\end{align*}
\]
Bayesian inference of the BMSSR

- For the BMSSR, the parameter $\Psi$ 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:
  
  $Z_i|\ldots \sim \mathcal{M}(1; \tau_{i1}, \ldots, \tau_{iK})$ with $\tau_{ik}(1 \leq k \leq K) = \mathbb{P}(Z_i = k|y_i, S_i; \Psi)$
  
  $\pi|\ldots \sim \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$
  
  $\beta_k|\ldots \sim \mathcal{N}(\nu_0, V_0)$
  
  $b_{ik}|\ldots \sim \mathcal{N}(\nu_1, V_1)$
  
  $\sigma_k^2|\ldots \sim \mathcal{IG}(g_1, h_1)$
  
  $\xi_k^2|\ldots \sim \mathcal{IG}(a_1, b_1)$

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

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

![Cluster mean images obtained by the BMSSR model with 12 mixture components.](image)

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: \((x_1, \ldots, x_n)\) 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: 
  \[ f(x_i; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \]

- Eigenvalue decomposition of the covariance matrix: 
  \[ \Sigma_k = \lambda_k D_k A_k D_k^T \]

\[ a \] 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)


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

Dirichlet Processes (DP)

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

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

Pólya urn representation (Blackwell and MacQueen, 1973)

\[
\tilde{\theta}_i | \tilde{\theta}_1, \ldots, \tilde{\theta}_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{k=1}^{K_i-1} \frac{n_k}{\alpha + i - 1} \delta \theta_k
\]

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

\[
G = \sum_{k=1}^{\infty} \pi_k \delta \theta_k \quad \theta_k | G_0 \sim G_0, \quad k = 1, 2, \ldots, \quad \text{with} \quad \sum_{k=1}^{\infty} \pi_k = 1
\]
DPM: Generative model

\[ G | \alpha, G_0 \sim \text{DP}(\alpha, G_0) \]
\[ \tilde{\theta}_i | G \sim G \]
\[ x_i | \tilde{\theta}_i \sim f(. | \tilde{\theta}_i) \]

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

- Latent variables \((z_1, \ldots, z_n)\)
- Predictive distribution:
  \[
  p(z_i = k | z_1, \ldots, z_{i-1}; \alpha) = \frac{\alpha}{\alpha + i - 1} \delta(z_i, K_{i-1} + 1) + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta(z_i, k)
  \]

- Generative model:
  \[
  z_i | \alpha \sim \text{CRP}(z_i; \alpha) \\
  \theta_{z_i} | G_0 \sim G_0 \\
  x_i | \theta_{z_i} \sim f(. | \theta_{z_i})
  \]
Implemented parsimonious models

<table>
<thead>
<tr>
<th>Decomposition</th>
<th>Model-Type</th>
<th>Prior</th>
<th>Applied to</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda I$</td>
<td>Spherical</td>
<td>$\mathcal{IG}$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>$\lambda_k I$</td>
<td>Spherical</td>
<td>$\mathcal{IG}$</td>
<td>$\lambda_k$</td>
</tr>
<tr>
<td>$\lambda A$</td>
<td>Diagonal</td>
<td>$\mathcal{IG}$</td>
<td>each diagonal element of $\lambda A$</td>
</tr>
<tr>
<td>$\lambda_k A$</td>
<td>Diagonal</td>
<td>$\mathcal{IG}$</td>
<td>each diagonal element of $\lambda_k A$</td>
</tr>
<tr>
<td>$\lambda D A D^T$</td>
<td>General</td>
<td>$\mathcal{IW}$</td>
<td>$\Sigma = \lambda D A D^T$</td>
</tr>
<tr>
<td>$\lambda_k D A D^T$</td>
<td>General</td>
<td>$\mathcal{IG}$</td>
<td>$\lambda_k$ and $\Sigma = D A D^T$</td>
</tr>
<tr>
<td>$\lambda D A_k A D^T$</td>
<td>General</td>
<td>$\mathcal{IG}$</td>
<td>each diagonal element of $\lambda A_k$</td>
</tr>
<tr>
<td>$\lambda_k D A_k A D^T$</td>
<td>General</td>
<td>$\mathcal{IG}$</td>
<td>each diagonal element of $\lambda_k A_k$</td>
</tr>
<tr>
<td>$\lambda D_k A D_k^T$</td>
<td>General</td>
<td>$\mathcal{IW}$</td>
<td>$\lambda$ and $\Sigma_k = D_k A_k D_k^T$</td>
</tr>
<tr>
<td>$\lambda_k D_k A_k D_k^T$</td>
<td>General</td>
<td>$\mathcal{IW}$</td>
<td>$\Sigma_k = \lambda_k D_k A_k D_k^T$</td>
</tr>
</tbody>
</table>

Bayesian inference using Gibbs sampling

- Posterior distribution for the component labels:
  \[
p(z_i = k | z_{-i}, X, \Theta, \alpha) \propto p(x_i | z_i; \Theta)p(z_i | z_{-i}; \alpha) \quad \text{with} \; p(z_i | z_{-i}; \alpha) \; \text{the CRP prior}
\]

- Posterior distribution for the component parameters:
  \[
p(\theta_k | z, X, \Theta_{-k}, \alpha; H) \propto \prod_i z_i = k \; p(x_i | z_i = k; \theta_k)p(\theta_k; H) \quad \text{with} \; p(\theta_k; H) : \text{Prior distribution over} \; \theta_k
\]

Bayesian model comparison by using Bayes Factors

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

\[
p(X | M_m) = \int p(X | \theta_m, M_m)p(\theta_m | M_m)d\theta_m \approx (2\pi)^{-\nu_m/2} |\hat{H}|^{-1/2} p(X | \hat{\theta}_m, M_m)p(\hat{\theta}_m | M_m)
\]
Clustering of benchmarks

Diabetes data set, Geyser data set, Crabs data set

2 log BF: $\lambda_k D_k A D_k^T$ vs $\lambda D_k A D_k^T = 199.58$ (Decisive)

2 log BF: $\lambda D A D^T$ vs $\lambda_k D_k A D_k^T = 5$ (Substantial)

$log 2BF: \lambda_k D_k A D_k^T$ vs $\lambda_k D A D^T = 36.08$ (Decisive)
Humpback whale song decomposition

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

![Image of a Humpback Whale](image1.png)
![Image of Spectrum of a signal (20 s)](image2.png)

**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.
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 $\lambda I$: (sec. 8) (sec. 10)
Unsupervised decomposition of whale song signals

Sound demo of Unit 4 DPPM $\lambda_k A$: (sec. 1) (sec. 7)
Unsupervised decomposition of whale song signals

Sound demo of Unit 8 DPPM $\lambda_k A$: (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: \( Y = (y_{ij}) \): \( n \) individuals defined on a set \( \mathcal{I} \) with \( d \) continuous functional variables defined on a set \( \mathcal{J} \) where \( y_{ij}(t) = \mu(x_{ij}(t); \beta) + \epsilon(t) \), \( t \) defined on \( \mathcal{T} \).

FLDM model:

\[
f(Y|X; \Psi) = \sum_{(z,w) \in Z \times W} \mathbb{P}(Z, W) f(Y|X, Z, W; \theta)
= \sum_{(z,w) \in Z \times W} \prod_{i,k} \pi_{ik}^{z_{ik}} \prod_{j,\ell} \rho_{j\ell}^{w_{j\ell}} \prod_{i,j,k,\ell} f(y_{ij}|x_{ij}; \theta_{k\ell})^{z_{ik}w_{j\ell}}.
\]

An RHLP is used as a conditional block distribution \( f(y_{ij}|x_{ij}; \theta_{k\ell}) \)

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

Variational Learning of Dirichlet Process Parsimonious Mixtures

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

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


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

- Mixture of experts are universal approximators (Nguyen et al., 2016).
  - Consider MoE to construct Fisher vectors (Sanchez et al., 2013)
  - Consider non-normal (skewed, heavy-tailed) MoE.
- Latent variable models for unsupervised learning of feature hierarchies:
  - consider hierarchical (deep) MoE as in Eigen et al. (2014)
- Patel et al. (2015) introduced a probabilistic theory to answer some questions on deep learning
Perspectives

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

Aggregation of mixtures for massive data

⇒ Density estimation and collaborative clustering of massive data

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

Published papers


Submitted papers


[J-15] F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015d. under review
Thank you for your attention!
References


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


References III


