On some statistical data analysis and learning problems in Data Science

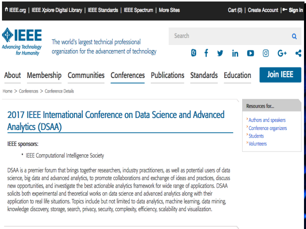
Faicel Chamroukhi

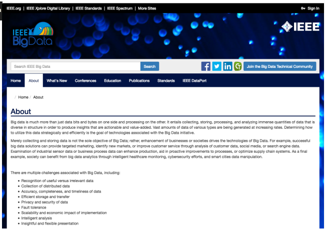
Normastic

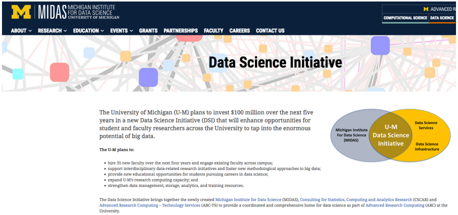
FÉDÉRATION NORMANDE
DE RECHERCHE EN SCIENCES
ET TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

Axe Données, Apprentissage, Connaissances
Journée du 30 mars 2017

- The term "Data Science" has surged in popularity
- Data science is increasingly commonly used with "big data."
- Data science, including Big Data has recently attracted an enormous interest from the scientific community

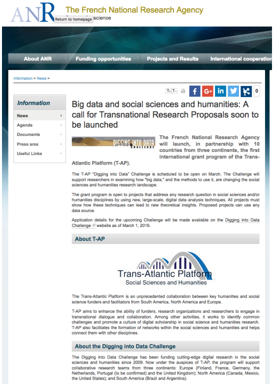


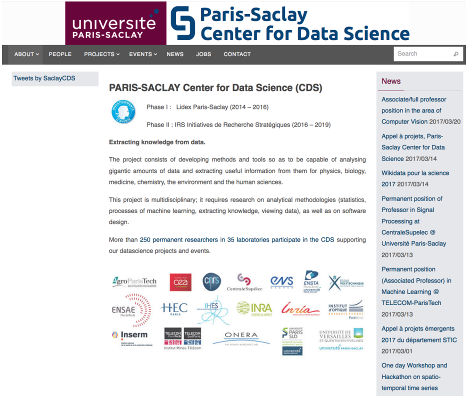





Le CNRS | Annuaires | Mots-Clefs CNRS | Autres sites
Mission interdisciplinarité
Centre national de la recherche scientifique
La Mission pour l'interdisciplinarité
Actualités interdisciplinaires
Les défis
Actions de site
PEPS en réseau
Plateforme réseaux
Colloques et manifestations
Espace communication
Archives
Rechercher
Accueil du site > Les défis > MASTODONS
Grandes masses de données scientifiques - MASTODONS
Année de démarrage : 2012
Contexte :
Nos capacités de production, d'acquisition et d'accès aux données dépassent largement nos capacités de traitement, d'analyse et d'archivage. De nombreux domaines scientifiques (ex : cartographie du génome, observatoire virtuel en astronomie, simulation en physique et énergie), économiques (ex : e-commerce, systèmes décisionnels), ou sociétaux (ex : réseaux sociaux, bibliothèques numériques, distribution de contenus multimédia) produisent et consomment des volumes de données considérables. Le CNRS, à travers ses unités de recherche et ses grands instruments, capitalise plusieurs centaines, voire milliers, de bases de données et de corpus d'informations dont les volumes croissent de façon exponentielle.
Objectifs :
Le but est d'identifier et de soutenir des actions de recherche dont les résultats ne pourraient être obtenus sans une fertilisation croisée des disciplines et sans une synergie effective entre chercheurs. Ces actions de recherche doivent couvrir aussi bien les verrous scientifiques posés par l'utilisation à grande échelle des données (sémantique, stockage, recherche, visualisation, …) que l'impact sociétal qu'induisent les technologies proposées (protection de la vie privée, préservation de la connaissance, …).
Les projets sélectionnés couvrent les axes de recherche suivants :
1. Stockage, indexation et accès aux données (en particulier dans le Cloud),
2. Extraction de connaissances, apprentissage et la visualisation de grandes masses de données,
3. Qualité des données, confidentialité et sécurité des données.
4. Problèmes de propriété, de droit d'usage, droit à l'oubli.
5. Préservation/archivage des données pour les générations futures.
Colloque de restitution du défi Mastodons
Défi Mastodons - Les Big Data en recherche - AAP 2017
Grandes masses de données scientifiques - MASTODONS
AAP 2016 - La qualité des données dans les Big Data : résultats


MaDICS
Masses de Données, Informations et Connaissances en Sciences
Big Data - Data Science
Accueil
Se Connecter
MaDICS On Twitter
Prochainement


ANR The French National Research Agency
Return to homepage
About ANR | Funding opportunities | Projects and Results | International cooperation
Information
News
Agenda
Documents
Press area
Useful Links
Big data and social sciences and humanities: A call for Transnational Research Proposals soon to be launched
The French National Research Agency will launch, in partnership with 10 countries from three continents, the first international grant program of the Trans-Atlantic Platform (T-AP).
The T-AP "Digging into Data" Challenge is scheduled to be open on March. The Challenge will support researchers in examining how "big data," and the methods to use it, are changing the social sciences and humanities research landscape.
The grant program is open to projects that address any research question in social sciences and/or humanities disciplines by using new, large-scale, digital data analysis techniques. All projects must show how these techniques can lead to new theoretical insights. Proposed projects can use any data source.
Application details for the upcoming Challenge will be made available on the Digging into Data Challenge website as of March 1, 2016.
About T-AP
Trans-Atlantic Platform
Social Sciences and Humanities
The Trans-Atlantic Platform is an unprecedented collaboration between key humanities and social science funders and facilitators from South America, North America and Europe.
About the Digging into Data Challenge


université PARIS-SACLAY
Paris-Saclay Center for Data Science
ABOUT | PEOPLE | PROJECTS | EVENTS | NEWS | JOBS | CONTACT
Tweets by SaclayCDS
PARIS-SACLAY Center for Data Science (CDS)
Phase I : Lidex Paris-Saclay (2014 – 2016)
Phase II : IRS Initiatives de Recherche Stratégiques (2016 – 2019)
Extracting knowledge from data.
The project consists of developing methods and tools so as to be capable of analysing gigantic amounts of data and extracting useful information from them for physics, biology, medicine, chemistry, the environment and the human sciences.
This project is multidisciplinary; it requires research on analytical methodologies (statistics, processes of machine learning, extracting knowledge, viewing data), as well as on software design.
More than 250 permanent researchers in 35 laboratories participate in the CDS supporting our datascience projects and events.
News
Associate/full professor position in the area of Computer Vision 2017/03/20
Appel à projets, Paris-Saclay Center for Data Science 2017/03/14
Wikidata pour la science 2017 2017/03/14
Permanent position of Professor in Signal Processing at CentraleSupelec @ Université Paris-Saclay 2017/03/13
Permanent position (Associated Professor) in Machine Learning @ TELECOM-ParisTech 2017/03/13
Appel à projets émergents 2017 du département STIC 2017/03/01
One day Workshop and Hackathon on spatio-temporal time series

- What does Data Science mean?
- What about Statistics in the Data Science “area” ?
- There is not yet a consensus on what precisely constitutes Data Science

CONTRIBUTED ARTICLES

Data Science and Prediction

By Vasant Dhar
Communications of the ACM, Vol. 56 No. 12, Pages 64-73
10.1145/2500499
Comments (2)

VIEW AS: SHARE:

Use of the term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

Back to Top

Key Insights

- Data science is the study of the generalizable extraction of knowledge from data.
- A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to explain the past.
- A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.

ASA | Amstat News | ASA Community | The World of Sta

AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME | ABOUT | EDITORIAL CALENDAR | PDF ARCHIVES | ADVERTISE | STATISTICIANS IN HISTORY

Home » Featured

ASA Statement on the Role of Statistics in Data Science

1 OCTOBER 2015 6,956 VIEWS 13 COMMENTS

Statement Contributors
David van Dyk, *Imperial College* (chair)
Montse Fuentes, *NCSU*
Michael I. Jordan, *UC Berkeley*
Michael Newton, *University of Wisconsin*
Bonnie K. Ray, *Pegged Software*
Duncan Temple Lang, *UC Davis*
Hadley Wickham, *RStudio*

The rise of data science, including Big Data and data analytics, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

- For a review, see the report of D. Donoho (2015): “50 years of Data Science”

INSTITUT DE FRANCE
Académie des sciences

recherche ...

NOUS CONNAÎTRE | VIE SCIENTIFIQUE | ENSEIGNEMENT DES SCIENCES | DIFFUSION DES CONNAISSANCES | COLLABORATIONS INTERNATIONALES | EXPERTISE ET CONSEIL

18 FÉV 2014

La datamasse : directions et enjeux pour les données massives

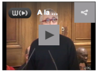
Publié dans Colloques, conférences et débats

Conférence-débat de l'Académie des sciences

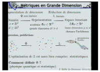
Nous vivons dans une 'société de l'information' dont les avancées scientifiques et techniques rapides, associées au développement d'usages nouveaux, conduisent à produire des quantités toujours plus gigantesques de données numériques. Cette situation d'abondance ouvre des perspectives nouvelles tant dans les sciences exactes que dans les sciences humaines. L'utilisation de cette 'datamasse' (*Big Data* en anglais) pose des défis considérables : Comment stocker de telles quantités de données, les manipuler, les analyser, les trier,... les valoriser ? Comment concilier leur omniprésence et le respect de la vie privée ? Comment faire qu'elles bénéficient à tous ? Ce sont quelques-uns de ces aspects qui seront mis en avant dans cette rencontre, afin d'en mieux comprendre les possibilités et les limitations, pour en mieux maîtriser les développements.

Introduction
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences et Patrick Flandrin, directeur de recherche CNRS, École normale supérieure de Lyon, membre de l'Académie des sciences

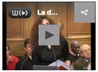
À la découverte des connaissances massives de la Toile
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences



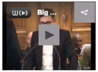
Des mathématiques pour l'analyse de données massives
Stéphane Mallat, professeur à l'École normale supérieure, Paris



La découverte du cerveau grâce à l'exploration de données massives
Anastasia Ailamaki, professeure à l'École polytechnique fédérale de Lausanne



Big Data et Relation Client : quel impact sur les industries et activités de services traditionnelles ?
François Bourdoncle, co-fondateur et CTO d'Exalead, filiale de Dassault Systèmes



Discussion générale et conclusion

Vidéos réalisées par la cellule Webcast CC-IN2P3 du CNRS

- There is not yet a consensus on what precisely constitutes Data Science, but
- Data Science can be seen (defined ?) as[a]:
  - the study of the generalizable extraction of knowledge from data.
  - requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization

[a] Vasant Dhar (2013): Communications of the ACM, Vol. 56 No. 12: 64-73

- Data Science clearly has an interdisciplinary nature and requires substantial collaborative effort
- Databases, statistics and machine learning, and distributed systems are emerging as foundational to data science

(i) Databases: organization of data resources,

(ii) Statistics and Machine Learning: convert data into knowledge,

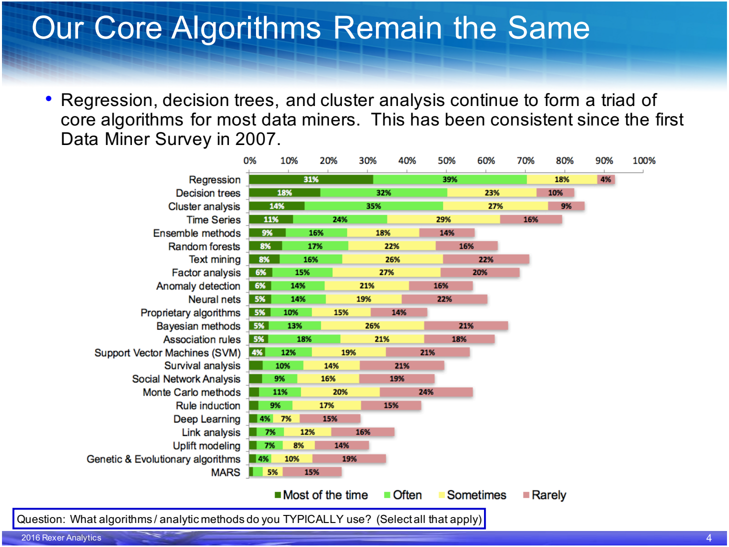
(iii) Distributed and Parallel Systems: computational infrastructure

Statistics play a central role in data science

- Allow to quantify the randmoness component in the data
- A well-established background to deal with uncertainty (probabilistic framework) and to establish generizable methods for prediction and estimation
- allow soft decision: e.g. confidence interval in regression and posterior probabilities in classification
- help for understanding the underlying generative process

Data science models/algorithms

New problems (big data, etc) but ... classical methods ?



Statistical modeling for data science

- The observed data $\{x_1, \ldots, x_n\}$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ are assumed to represent samples from random variables X with unknown probability distribution f
- The main questions are i) how to define flexible and generic models for f ii) construct estimators with desirable properties to learn f from the data iii) to deal with the computational and practical issues for "complex" data
- The area of statistical learning for the analysis of complex data.
- **Data :** Complex data $\hookrightarrow$ *heterogeneous, temporal/dynamical, functional, incomplete, high-dimensional,...*
- **Objective:** Transform the data into knowledge :
  $\hookrightarrow$ Reconstruct hidden structure/information, groups/hierarchy of groupes, summarizing prototypes, underlying dynamical processes, etc

Topics and goals

$\hookrightarrow$ exploratory analysis: segmentation/clustering/dimensionality reduction/vizualisation

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

Modeling framework

- Latent variable models : $f(x|\boldsymbol{\theta}) = \int_z f(x, z|\boldsymbol{\theta})\mathrm{d}z$
  Generative formulation :
$$z \sim q(z|\boldsymbol{\theta})$$
$$x|z \sim f(x|z, \boldsymbol{\theta})$$

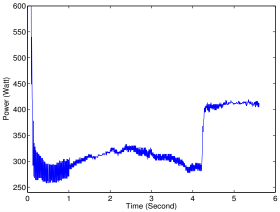
$\hookrightarrow$ Mixture models : $f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{P}(z = k) f(x|z = k, \boldsymbol{\theta}_k)$ and extensions

$\hookrightarrow$ **Algorithms** for inferring $\boldsymbol{\theta}$ from the data

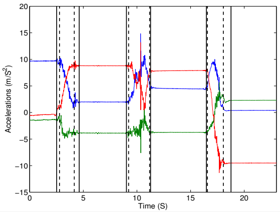
1. Temporal data segmentation
2. Clustering of functional data
3. Bayesian (non-)parametric mixtures for spatial and multivariate data

Temporal data

Temporal data with regime changes



Railway data



Human activity data

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

Objectives

Temporal data modeling and segmentation

Regression with hidden logistic process

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

The Regression model with Hidden Logistic Process (RHLP) [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
- Parameter estimation via a the EM algorithm: EM-RHLP

EM-RHLP

Parameter estimation via a the EM algorithm: EM-RHLP

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

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

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

Time series approximation and segmentation

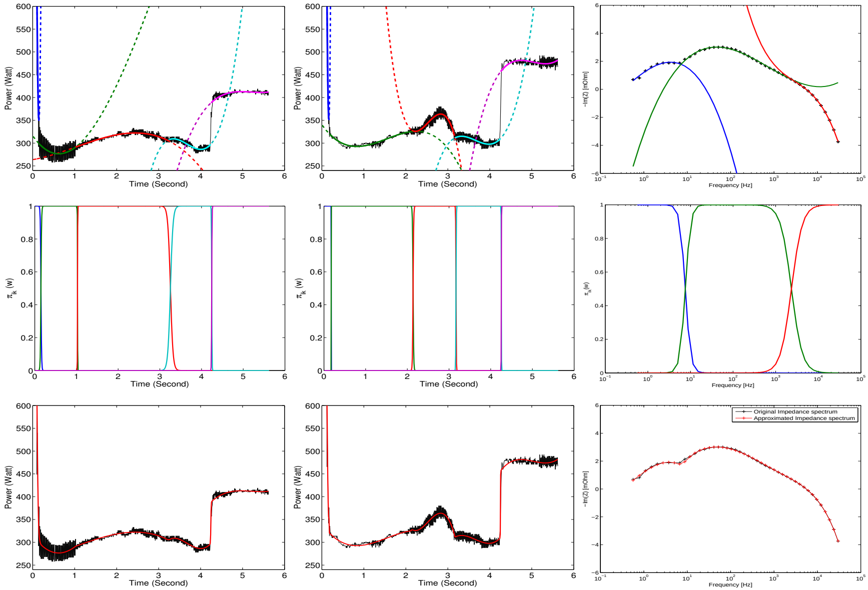
1. Approximation: a curve prototype $\hat{y}_i = \mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i$

   $\hookrightarrow$ The RHLP can be used as nonlinear regression model $y_i = f(t_i;\boldsymbol{\theta}) + \epsilon_i$ by covering functions of the form $f(t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ [3]

2. Curve segmentation: $\hat{z}_i = \arg\max_k \mathbb{E}[z_i|t_i;\hat{\mathbf{w}}] = \arg\max_k \pi_k(t_i;\hat{\mathbf{w}})$

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

Application to temporal data modeling and segmentation



Joint segmentation of multivariate time series

Multiple hidden process regression

- Data: $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ a time series of n multidimensional observations $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at instants $\mathbf{t} = (t_1, \ldots, t_n)$.
- Model $\boldsymbol{y}_i = \mathbf{B}_{z_i}^T \boldsymbol{x}_i + \mathbf{e}_i \quad ; \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{z_i}), \quad (i = 1, \ldots, n)$

  $\mathbf{z} = (z_1, \ldots, z)$ A latent process generating the data

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

$\hookrightarrow$ Multiple Hidden Markov model regression (MHMMR) [7]

Application to human activity time series

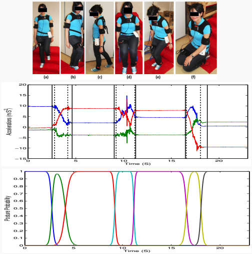
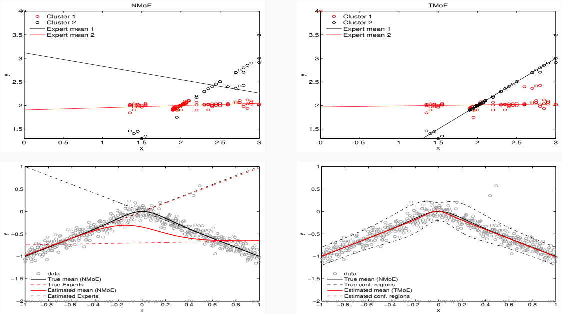


Figure: MRHLP segmentation of acceleration data issued from three body-worn sensors (Data

Data with atypical features



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

Objectives

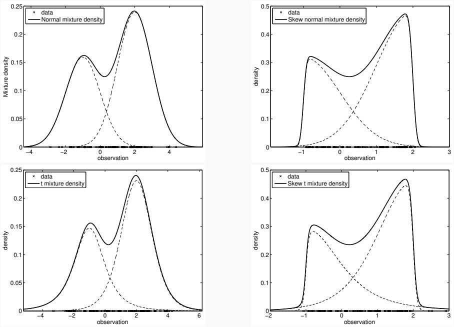
- Derive robust models to fit at best the data
- Deal with other possible features like skewness, heavy tails

Non-normal mixtures of experts

Non-normal mixtures of experts (NNMoE)

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

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

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

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

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

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

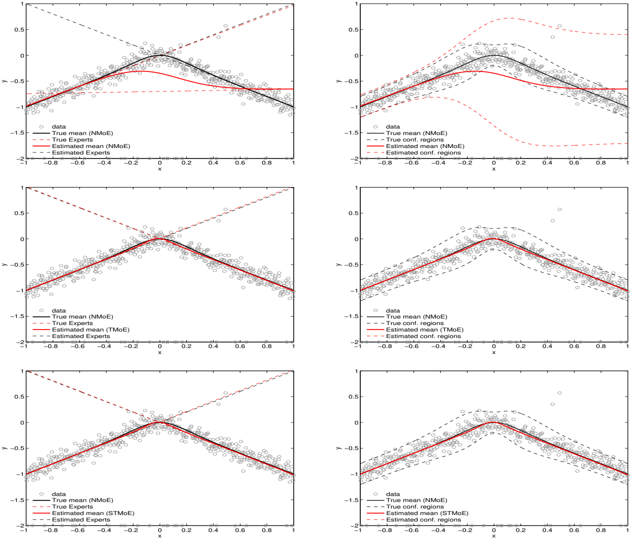


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

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 $(0, 4)$

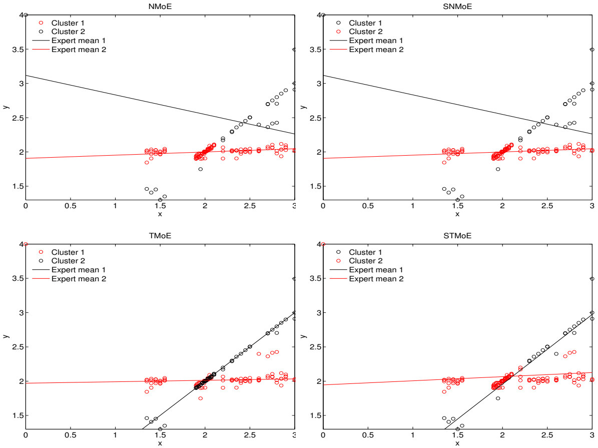


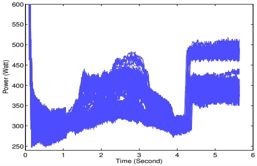
Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

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

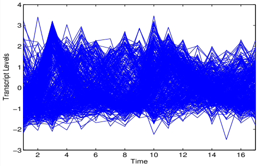
Outline

Functional data analysis context

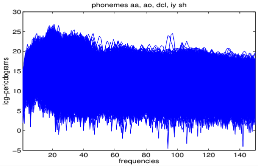
Many curves to analyze



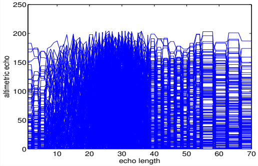
Railway switch curves



Yeast cell cycle curves



Phonemes curves



Satellite waveforms

Objectives

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

Functional data analysis context

Data

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

Objectives: exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [4] [9], [C11] [16]
2. Discriminant analysis of functional data: [2], [5]

Functional data clustering/classification tools

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

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

Mixture modeling framework for functional data

- The functional mixture model:

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

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

  ↪ more tailored to approximate smooth functions

  ↪ do not account for segmentation

Here $f_k(y|\boldsymbol{x})$ itself exhibits a clustering property via hidden variables (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) [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 regressions $\hookrightarrow$ dynamic programming

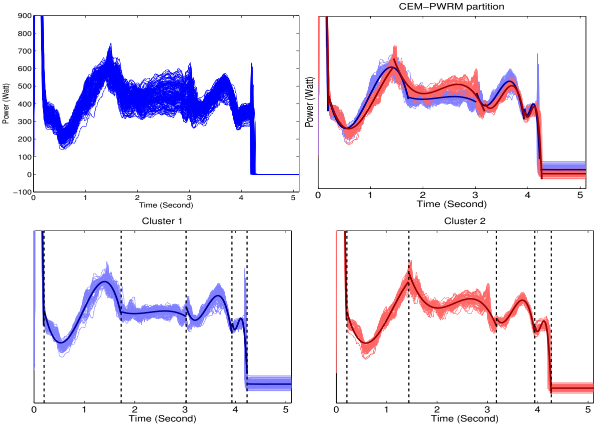
   Complexity in $\mathcal{O}(I_{\text{EM}} K R n m^2 p^3)$: Significant computational load for large m

Curve clustering: $\hat{z}_i = \arg\max_k \tau_{ik}(\hat{\boldsymbol{\Psi}})$ with $\tau_{ik}(\hat{\boldsymbol{\Psi}}) = \mathbb{P}(Z_i|\boldsymbol{x}_i, \boldsymbol{y}_i; \hat{\boldsymbol{\Psi}})$

Application to switch operation curves

Data set: $n = 146$ real curves of $m = 511$ observations.

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

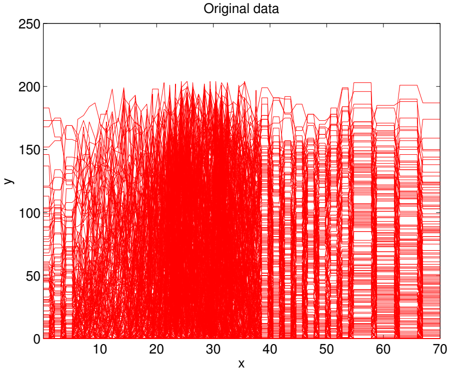


EM-GMM	EM-PRM	EM-PSRM	K-means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

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

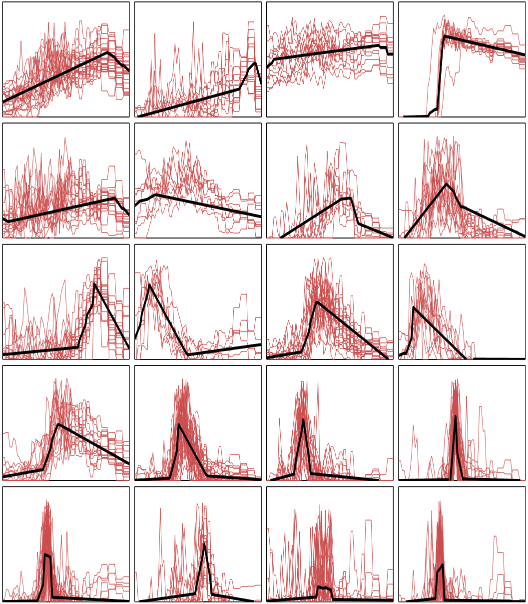
Application to Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data[1] contains $n = 472$ waveforms of the measured echoes, sampled at $m = 70$ (number of echoes)
We considered the same number of clusters (twenty) and a piecewise linear approximation of four segments per cluster as in Hébrail et al. (2010).



[1]Satellite data are available at
http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html.

CEM-PWRM clustering of the satellite data



Mixture of hidden logistic process regressions [4]

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

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

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

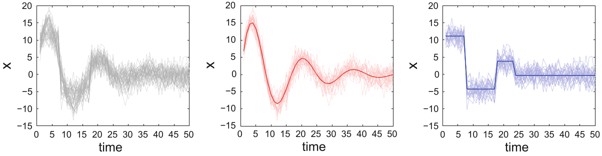
- Two types of component memberships:

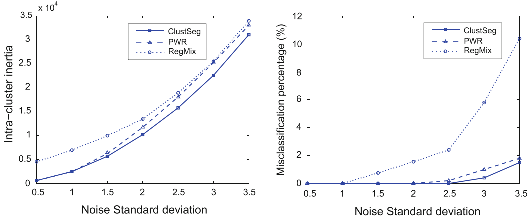
  $\hookrightarrow$ cluster memberships (global) $Z_{ik}=1$ iff $Z_i=k$

  $\hookrightarrow$ regime memberships for a given cluster (local): $H_{ijr}=1$ iff $H_{ij}=r$

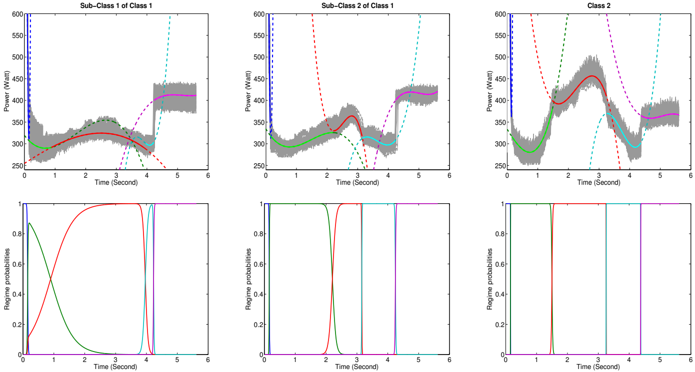
  MixRHLP deals better with the quality of regime changes
- Parameter estimation via the EM algorithm: EM-MixRHLP
- EM-MixRHLP has complexity in $\mathcal{O}(I_{\mathsf{EM}}I_{\mathsf{IRLS}}KR^3nmp^3)$ (K-means type for piecewise regression is in $\mathcal{O}(I_{\mathsf{KM}}KRnm^2p^3)$ $\hookrightarrow$ EM-MixRHLP is computationally attractive for large values of m and moderate values of R.

EM-MixRHLP clustering of simulated data





Functional Linear Discriminant Analysis [8]
Functional Mixture Discriminant Analysis [5]



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 [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 [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)

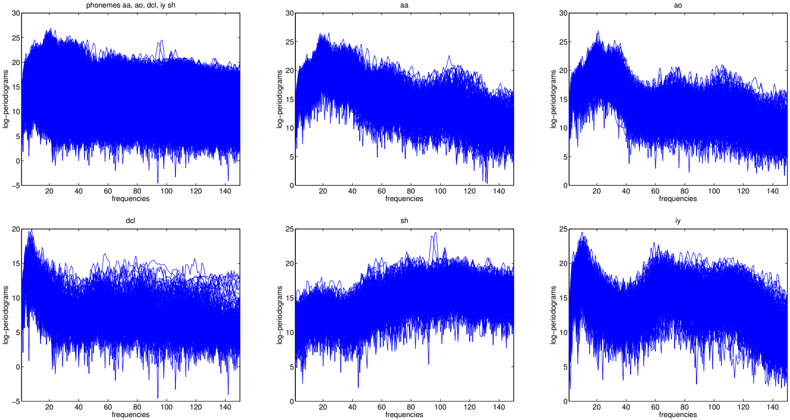
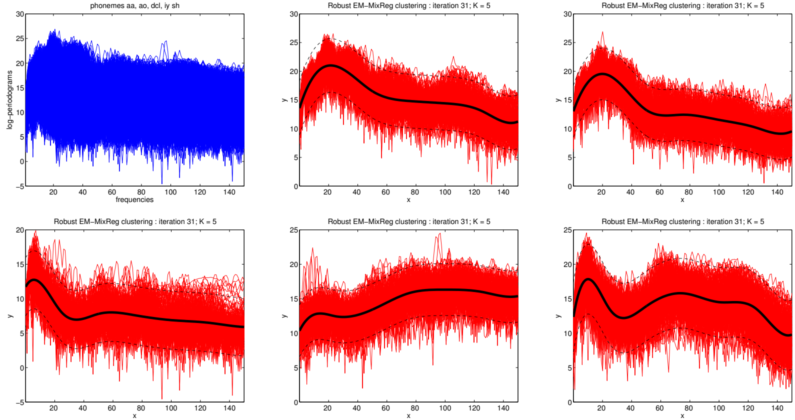


Figure: Original phoneme data and curves of the five classes: "ao", "aa", "yi", "dcl", "sh".

[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/

EM-like clustering results for yeast cell cycle data

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

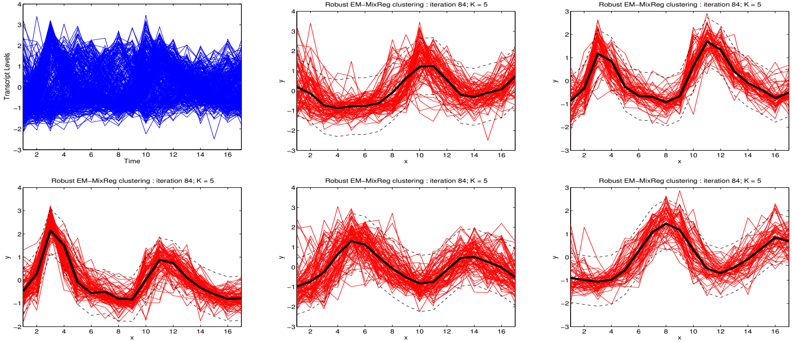


Figure: EM-like clustering results with the bSRM model.

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

[4] http://faculty.washington.edu/kayee/model/

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, 13]

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

- $\boldsymbol{\beta}$: 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

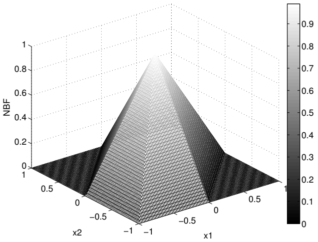


Figure: Nodal basis function $s(\boldsymbol{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$.

Bayesian mixture of spatial spline regressions

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

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

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

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

Hierarchical prior from for the BMSSR

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

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:

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

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

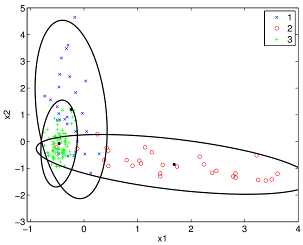
Handwritten digit clustering using the BMSSR

- 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

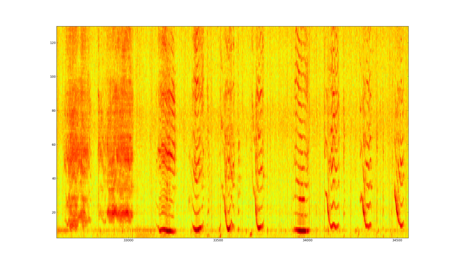
Figure: Cluster mean images obtained by the BMSSR model with 12 mixture components.

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

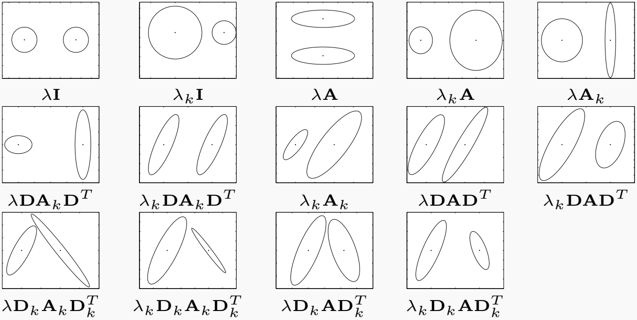
- Clustering
- Dimensionality reduction

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)

- $\hookrightarrow$ Mixture models for multivariate data in a fully Bayesian framework
- $\hookrightarrow$ Dirichlet Process and Parsimonious Mixtures [C5,6,8], [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

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

$\hookrightarrow$ The generated parameters $\tilde{\boldsymbol{\theta}}_i$ for a DP process exhibit a clustering property

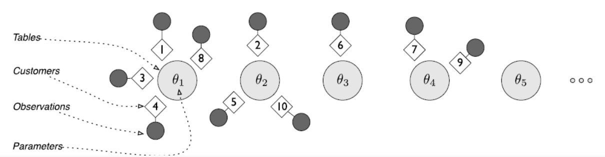
DPM: Generative model

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

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

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

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



- Generative model:

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

Implemented parsimonious models

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

Bayesian inference using Gibbs sampling

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

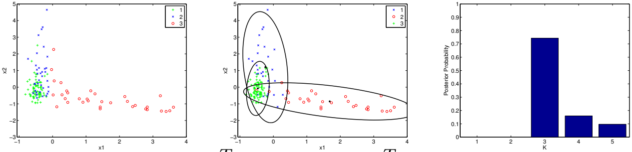
Bayesian model comparison by using Bayes Factors

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

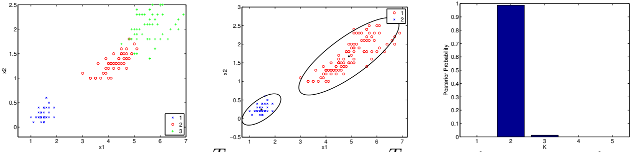
$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)\mathrm{d}\boldsymbol{\theta}_m \approx (2\pi)^{\frac{\nu_m}{2}} |\hat{\mathbf{H}}|^{\frac{1}{2}} p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m)$

Clustering of benchmarks

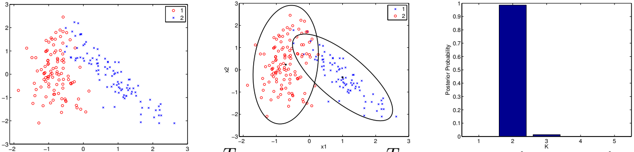
Diabetes data set, Geyser data set, Crabs data set



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



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

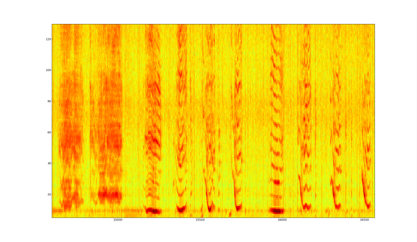


$\log 2\mathrm{BF}$: $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ vs $\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T$ $=$ 36.08 (Decisive)

Humpback whale song decomposition

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

Humpback Whale

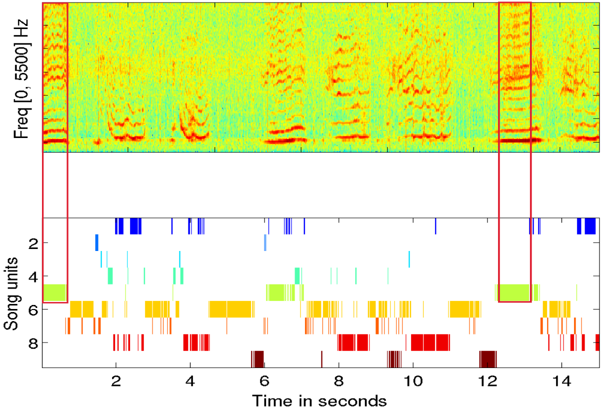


Spectrum of a signal (20 s).

Objectives

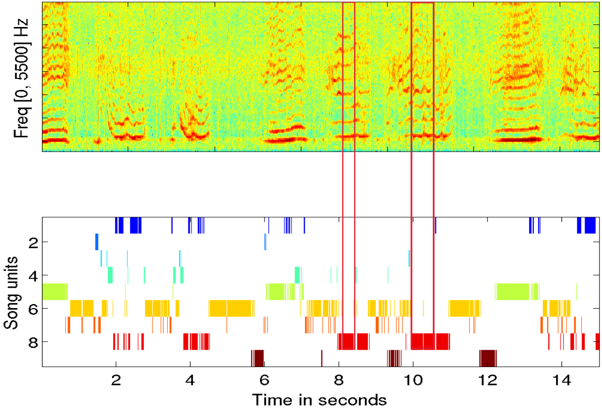
- Discovering "call units", which can be considered as a whale "alphabet"
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.

Unsupervised decomposition of whale song signals



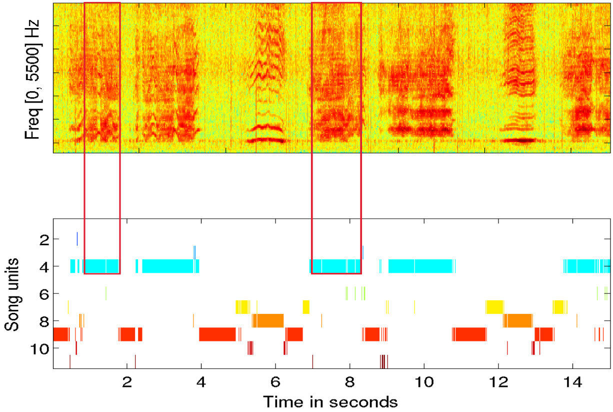
- Sound demo of Unit 5 DPPM $\lambda \mathbf{I}$: (sec. 0) (sec. 12)

Unsupervised decomposition of whale song signals



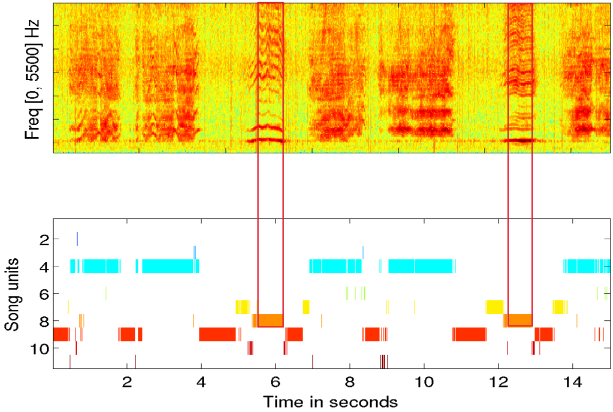
- Sound demo of Unit 8 DPPM $\lambda\mathbf{I}$: (sec. 8) (sec. 10)

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)

Some ongoing research and perspectives

- 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 $\mathcal{T}$.

- FLBM 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

Some ongoing research and perspectives

Mixtures for massive data

- Mixture density estimation for massive data clustering
- Regularized mixture of experts (lasso-like penalties)
- Ensemble methods to distribute data of big volume

  $\hookrightarrow$ Bag of Little Boostraps (BLB) (Kleiner et al., 2014)

  $\hookrightarrow$ Aggregate local estimators from BLB sub-samples: Hierarchical (mixture) of experts aggregation

Latent variable models for unsupervised learning of feature hierarchies

- Hierarchical Mixture of experts for data representation:
- Mixture of experts are universal approximators (Nguyen et al., 2016).

  $\rightarrow$ Hierarchical (deep) mixtures of experts (MoE) Eigen et al. (2014)

  $\rightarrow$ Hierarchical (deep) mixtures of factor analysers (MoE) Tang et al. (ICML,2012)
- Patel et al. (2015) probabilistic answers to some questions on deep learning

References

[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

[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

[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

[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

[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

[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

[7] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE TASE*, 3(10):829–335, 2013

[8] F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015c. doi: 10.1080/00949655.2015.1109096. 05 Nov 2015

[9] F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015

[10] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 33, 2016a. doi: 10.1007/s00357-. In Press

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

[12] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, 2015. In revision

[13] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a. (v1) submitted

[14] F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015b. Report (61 pages)

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

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.

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.

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.

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

G. Celeux. Bayesian inference for mixture: the label switching problem. Technical report, INRIA Rhone-Alpes, 1999.

G. Celeux and G. Govaert. Gaussian Parsimonious Clustering Models. *Pattern Recognition*, 28(5):781–793, 1995.

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

F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a. (v1) submitted.

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

F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015c. doi: 10.1080/00949655.2015.1109096. 05 Nov 2015.

References II

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

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 33, 2016a. doi: $10.1007/s00357$-. In Press.

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

F. Chamroukhi, A. Samé, G. Govaert, and P. 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*, 2015. In revision.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskeve. Learning factored representations in a deep mixture of experts. *arXiv:1312.4314v3*, March 2014.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

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

G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7-9):1125–1141, March 2010.

References III

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.

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

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.

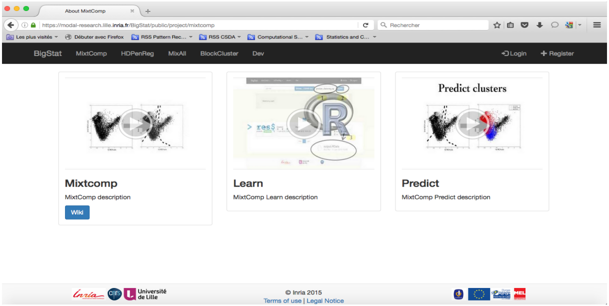
L.M. Sangalli, J.O. Ramsay, and T.O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society (Series B)*, 75:681–703, 2013.

Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71(0):128 – 137, 2014.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE TASE*, 3(10):829–335, 2013.

Ka Yee Yeung, Chris Fraley, A. Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

MASSIC Platfrom



SBDS 2017 : Research Summer School in Statistics & BigData Science (SBDS) 7-9 June @ Caen

- Home
- Lecturers
- Schedule
- Talks & Demos
- Registration
- Practical Information
- Sponsors
- Contact

Christophe Ambroise

- Professor, Evry University, France
- Talk: Statistical learning of stochastic latent block models for networks inference

Peter Tino

- Pofessor, University of Birmingham, UK
- Talk: Probabilistic Modelling in Machine Learning

Ankit B. Patel

- Asst. Professor, Baylor College of Medicine & Rice University, USA
- Talk: Probabilistic Framework for Deep Learning

Jalal Fadili

- Professor, Ensicaen & Insitut Universitaire de France (IUF), France
- Talk: Sparse representation of high dimensional signals and images

Hien Nguyen

- Australian Research Council DECRA Research Fellow, La Trobe University, Australia
- Talk: An introduction to MM algorithms for the machine learning and statistical estimation

Abstract: MM (majorization-minimization) algorithms are an increasingly popular tool for solving optimization problems in machine learning and statistical estimation. This lecture introduces the MM algorithm framework in general and via three commonly considered example applications: Gaussian mixture models, multinomial logistic regressions, and support vector machines. Specific algorithms for these three examples are derived and numerical demonstrations are presented. Theoretical and practical aspects of MM algorithm design are discussed.

Mustapha Lebbah

- Associate Professor, Paris 13 University
- Talk: Scalable machine learning and distributed systems

Faicel Chamroukhi

- Professor, Caen University, France
- Talk: Unsupervised learning of latent variable models from high-dimensional data

Sponsors

LMNO, UNICAEN, CNRS, Normandie - Mathématiques, Inria, Modal: MOdels for Data Analysis and Learning, Caenlamer, SFdS Société Française de Statistique, MaDICS

Thank you for your attention!