Statistical data science and some unsupervised learning problems

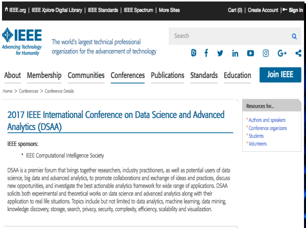
Faicel Chamroukhi

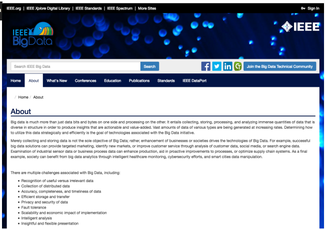
https://chamroukhi.com/

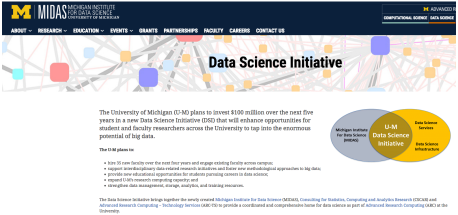
Workshop DSCI, 22-24 dec 2018

Outline

- 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



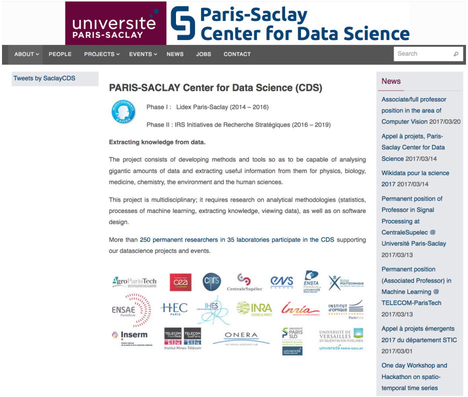





Le CNRS | Annuaires | Mots-Clefs CNRS | Autres sites
Mission interdisciplinarité
Centre national de la recherche scientifique
La Mission pour l'interdisciplinarité
Actualités interdisciplinaires
Les défis
Actions de site
PEPS en réseau
Plateforme réseaux
Colloques et manifestations
Espace communication
Archives
Rechercher
Sur ce site
Sur le Web du CNRS
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 Mastotons
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
Présentation MaDICS
Se Connecter
Username
Password
Login
MaDICS On Twitter
Prochainement


ANR
The French National Research Agency
science
About ANR
Funding opportunities
Projects and Results
International cooperation
Information > News >
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
DIGGING INTO DATA CHALLENGE
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.
T-AP aims to enhance the ability of funders, research organizations and researchers to engage in transnational dialogue and collaboration. Among other activities, it works to identify common challenges and promote a culture of digital scholarship in social science and humanities research. T-AP also facilitates the formation of networks within the social sciences and humanities and helps connect them with other disciplines.
About the Digging into Data Challenge
The Digging into Data Challenge has been funding cutting-edge digital research in the social sciences and humanities since 2009. Now under the auspices of T-AP, the program will support collaborative research teams from three continents: Europe (Finland, France, Germany, the Netherlands, Portugal (to be confirmed) and the United Kingdom); North America (Canada, Mexico, the United States); and South America (Brazil and Argentina).


université PARIS-SACLAY
Paris-Saclay Center for Data Science
ABOUT
PEOPLE
PROJECTS
EVENTS
NEWS
JOBS
CONTACT
Search
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

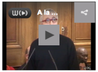
Publié dans Colloques, conférences et débats

Conférence-débat de l'Académie des sciences

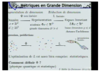
Nous vivons dans une "société de l'information" dont les avancées scientifiques et techniques rapides, associées au développement d'usages nouveaux, conduisent à produire des quantités toujours plus gigantesques de données numériques. Cette situation d'abondance ouvre des perspectives nouvelles tant 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

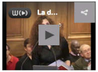
À la découverte des connaissances massives de la Toile
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences



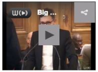
Des mathématiques pour l'analyse de données massives
Stéphane Mallat, professeur à l'École normale supérieure, Paris



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 and Data Science

$\hookrightarrow$ Statistics play a central role in data science

- Allow to quantify the randomness component in the data
- A well-established background to deal with uncertainty (probabilistic frame- work) 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

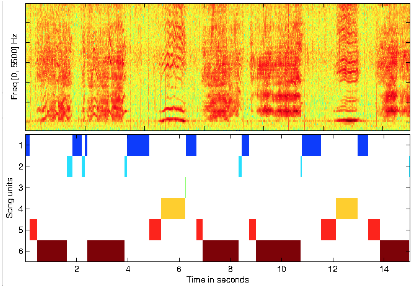
Statistical modeling for data science

- The observed data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ 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.

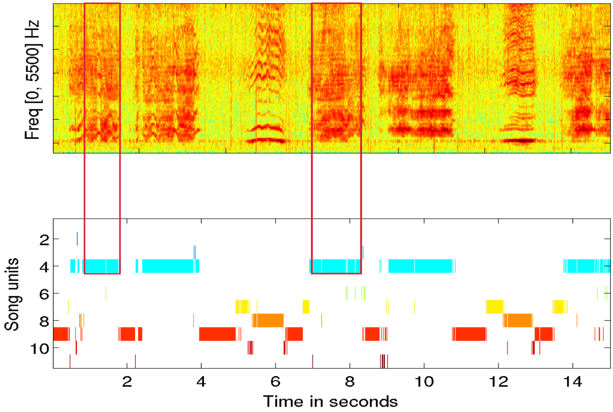
Context and Objectives

- **Context :** Large-scale data are increasingly frequent : Complex data $\hookrightarrow$ *heterogeneous, dynamical (temporal, functional), incomplete, high-dimension, and possibly massive*
- **Objectives :** learn/discern useful information in an unsupervised way from raw data :
  $\hookrightarrow$ Reconstruct/reveal hidden structures, i.e, (hierarchy of) groups ; learn/select relevant features, etc

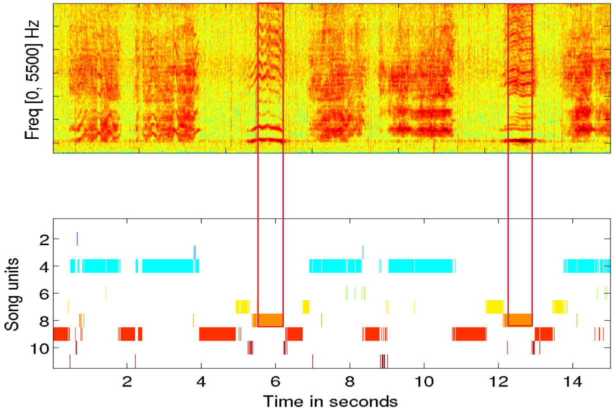
Unsupervised Sparse Signal Decomposition



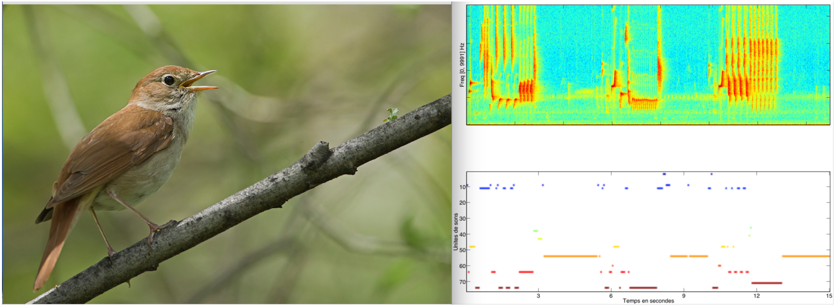
Unsupervised Sparse Signal Decomposition



Unsupervised Sparse Signal Decomposition

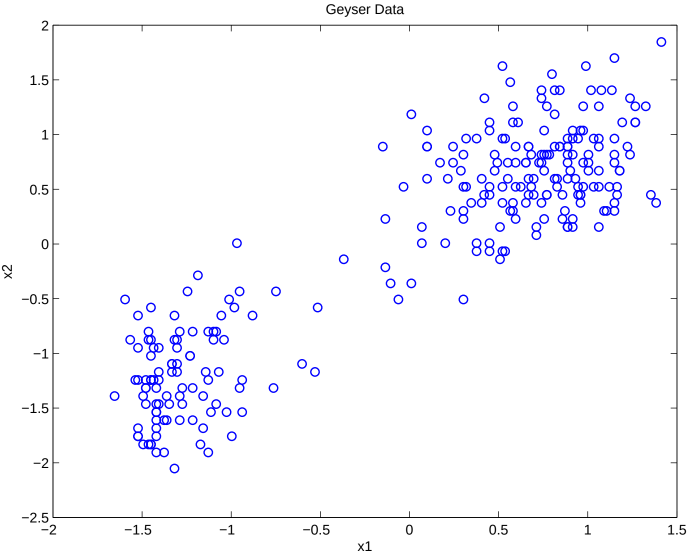


Unsupervised Sparse Signal Decomposition

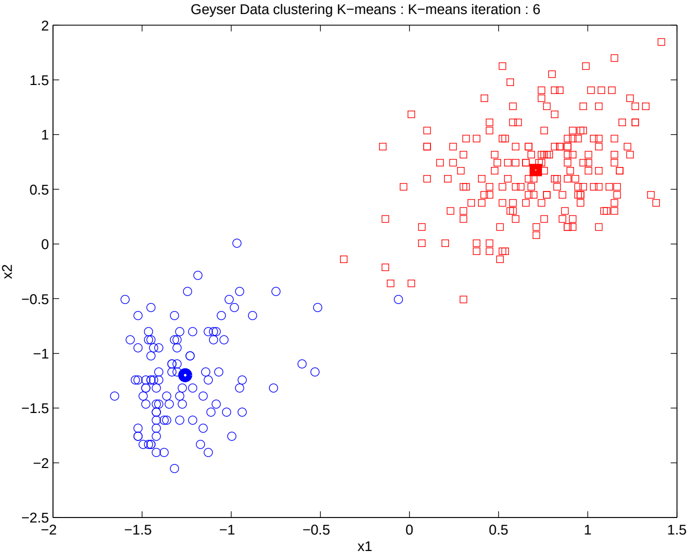


Outline

Clustering of multivariate data



Clustering of multivariate data



K-means

- a straightforward and widely used clustering algorithm, is one of the most important algorithms in unsupervised learning.
- Each cluster is represented by its mean (cluster centroid) $\boldsymbol{\mu}_k$ in $\mathbb{R}^d$.

K-means MacQueen [1967]

$$(\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_K, \widehat{\mathbf{z}}) \in \arg \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}} \mathcal{J}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z})$$

objective function : $\mathcal{J}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}_{z_i}\|^2$

K-means

- a straightforward and widely used clustering algorithm, is one of the most important algorithms in unsupervised learning.
- Each cluster is represented by its mean (cluster centroid) $\boldsymbol{\mu}_k$ in $\mathbb{R}^d$.

K-means MacQueen [1967]

$$(\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_K, \widehat{\mathbf{z}}) \in \arg \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}} \mathcal{J}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z})$$

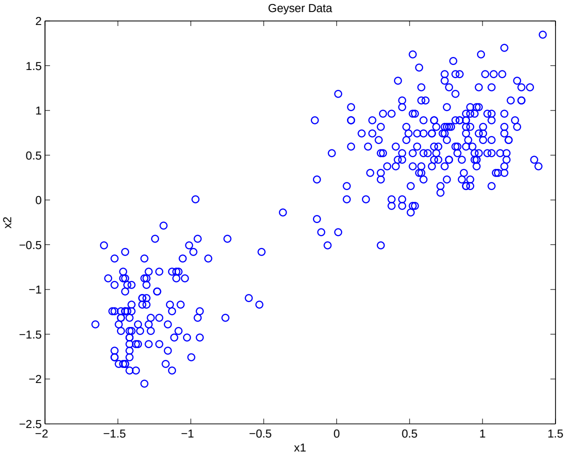
objective function : $\mathcal{J}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}_{z_i}\|^2$

- Initialization : $(\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)})$ (eg, randomly chosen data points)

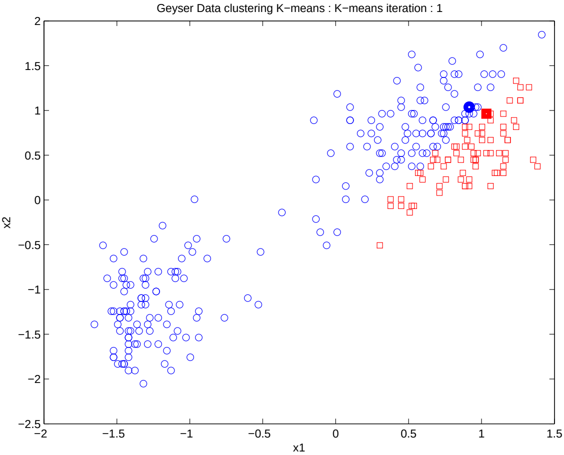
1. **Assignment step** : $z_i^{(q)} = \arg\min_{z \in \mathcal{Z}} \|\mathbf{x}_i - \boldsymbol{\mu}_z\|^2$
2. **Relocation step** : $\boldsymbol{\mu}_k^{(q+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(q)} \mathbf{x}_i}{\sum_{i=1}^{n} z_{ik}^{(q)}}$,

$\Rightarrow$ The K-means algorithm is simple to implement and relatively fast.

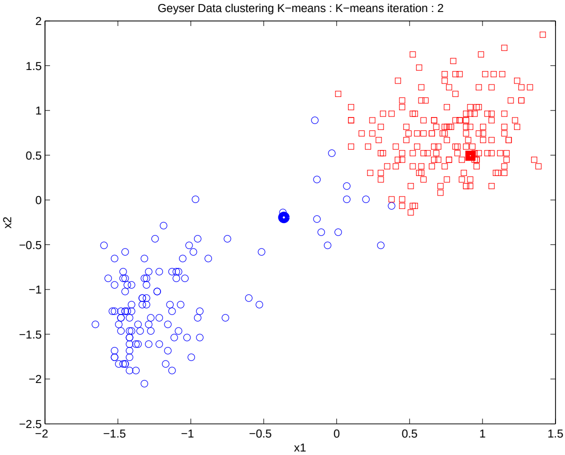
Example



Example

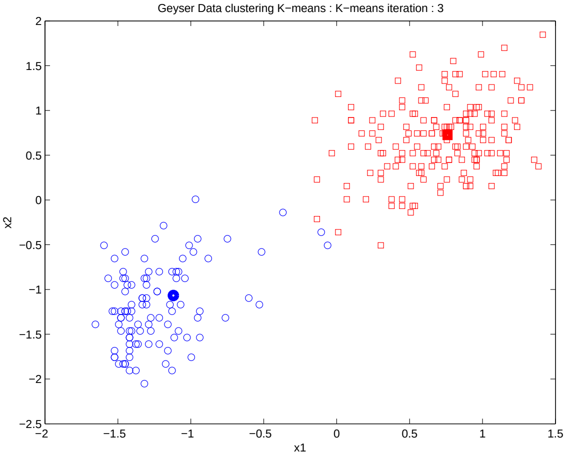

Geyser Data clustering K−means : K−means iteration : 1
x1
x2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
−2.5

Example

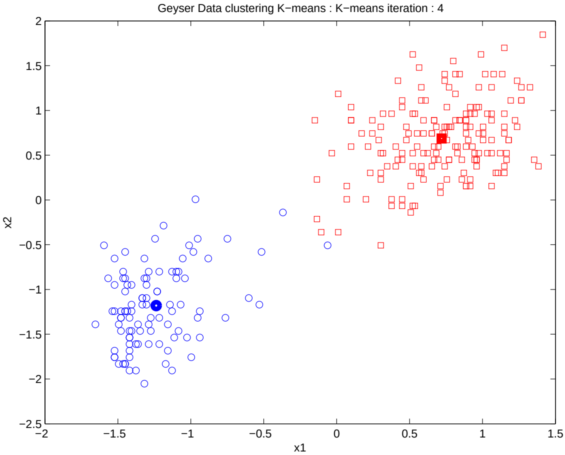


Geyser Data clustering K−means : K−means iteration : 2
x1
x2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
−2.5

Example

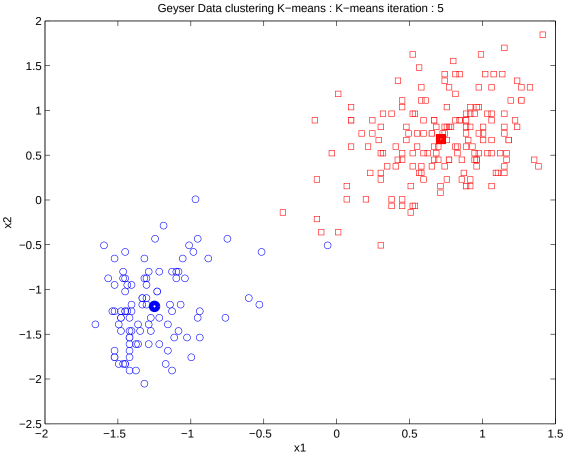

Geyser Data clustering K−means : K−means iteration : 3
x1
x2
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5

Example

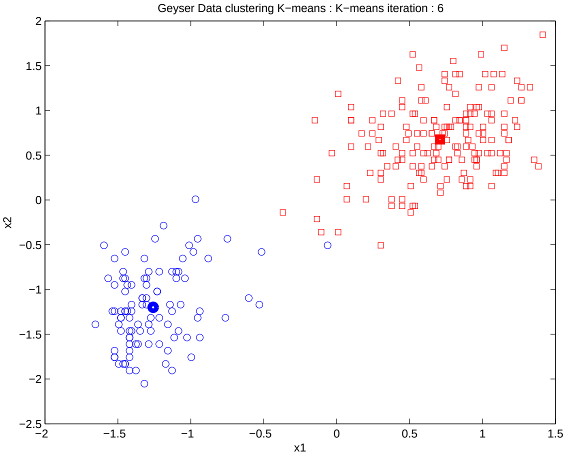


Geyser Data clustering K−means : K−means iteration : 4
x1
x2
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5

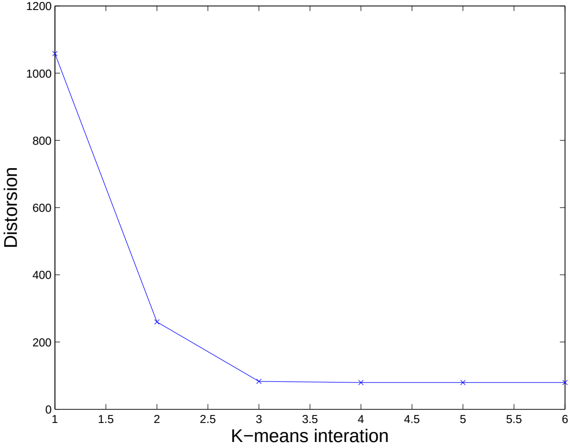
Example


Geyser Data clustering K−means : K−means iteration : 5
x1
x2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
−2.5

Example


Geyser Data clustering K−means : K−means iteration : 6
x1
x2
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5

Example



1200
1000
800
600
400
200
0
Distorsion
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
K−means interation

K-means

How to measure uncertainty ?

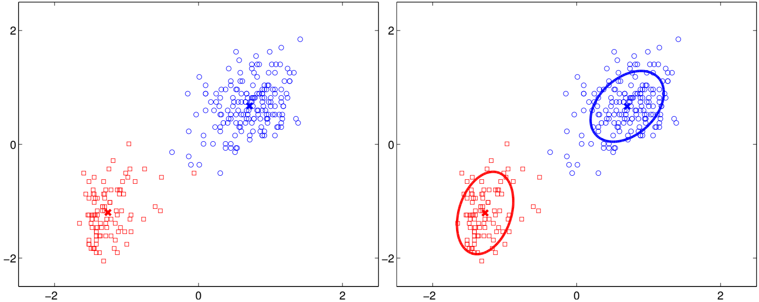


FIGURE – K-means partition (left) vs GMM-EM partition (right).

Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$ with $\pi_k > 0 \; \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$ with $\pi_k > 0 \; \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

Maximum-Likelihood Estimation

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$$

log-likelihood : $\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$.

Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$ with $\pi_k > 0\ \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

Maximum-Likelihood Estimation

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$$

log-likelihood : $\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$.

The EM algorithm [Dempster et al., 1977]

$$\boldsymbol{\theta}^{new} \in \arg\max_{\boldsymbol{\theta}\in\Omega} \mathbb{E}[\ln L_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{old}]$$

complete log-likelihood : $\ln L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log\left[\pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]$ where Z_{ik} is such that $Z_{ik} = 1$ if $Z_i = k$ and $Z_{ik} = 0$ otherwise.

Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$ with $\pi_k > 0\ \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

Maximum-Likelihood Estimation

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$$

log-likelihood : $\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$.

The EM algorithm [Dempster et al., 1977]

$$\boldsymbol{\theta}^{new} \in \arg\max_{\boldsymbol{\theta}\in\Omega} \mathbb{E}[\ln L_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{old}]$$

complete log-likelihood : $\ln L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log\left[\pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]$ where Z_{ik} is such that $Z_{ik} = 1$ if $Z_i = k$ and $Z_{ik} = 0$ otherwise.

Clustering

$$\widehat{z}_i = \arg\max_{1\leq k\leq K} \mathbb{P}(Z_i = k|\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}), \quad (i = 1, \ldots, n)$$

Gaussian mixture models (GMMs)

The finite Gaussian mixture density is defined as :

$$f(\mathbf{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

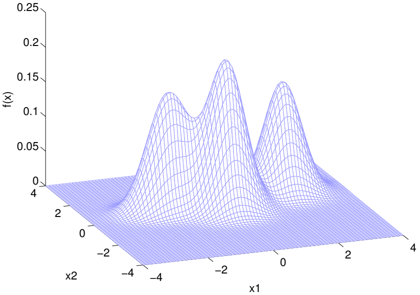


FIGURE – An example of a three-component Gaussian mixture density in $\mathbb{R}^2$.

EM for Gaussian mixture models

1. **E-Step** : calculates the posterior component memberships :

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\mathbf{x}_i, \boldsymbol{\Psi}^{(q)}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(q)})}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_\ell^{(q)}, \boldsymbol{\Sigma}_\ell^{(q)})}$$

   that $\mathbf{x}_i$ originates from the kth component density.

2. **M-Step** : parameter updates :

$$\begin{aligned}
\pi_k^{(q+1)} &= \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n} = \frac{n_k^{(q)}}{n}, \\
\boldsymbol{\mu}_k^{(q+1)} &= \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{x}_i, \\
\boldsymbol{\Sigma}_k^{(q+1)} &= \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}^{(q+1)})^T.
\end{aligned}$$

Examples

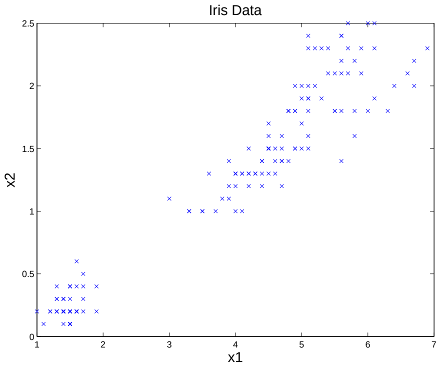


FIGURE – A three-class example of a real data set : Iris data of Fisher.

Examples

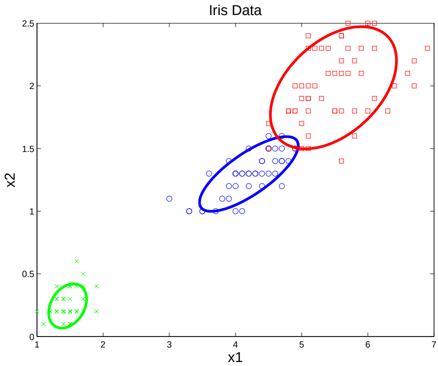


FIGURE – Iris data : Clustering results with EM for a GMM and AIC.

Examples

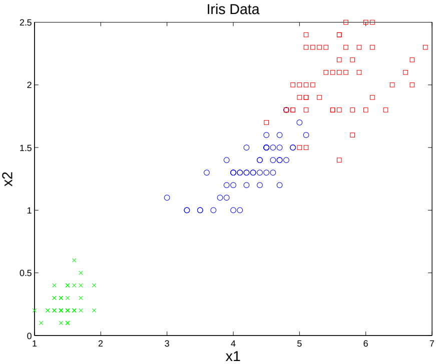
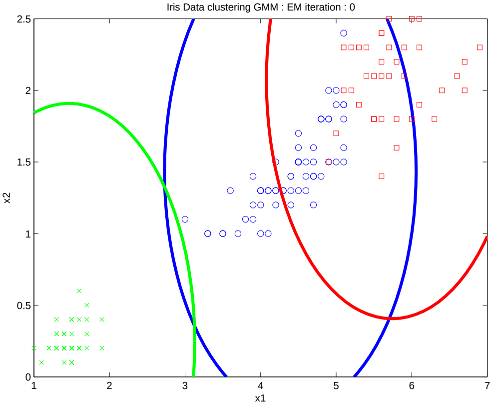
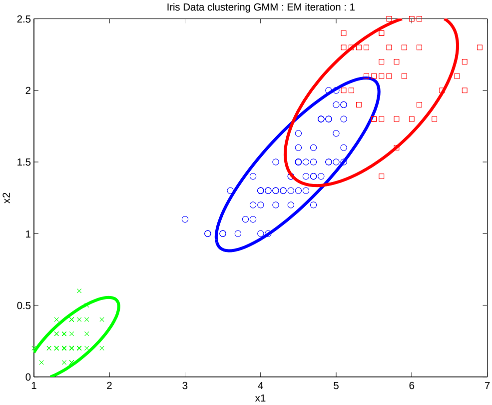


FIGURE – Iris data of Fisher : The data are colored according to the true partition.

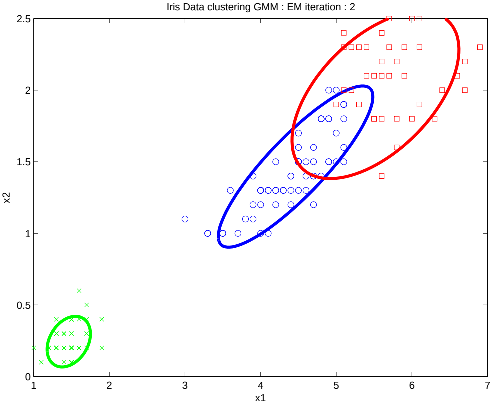
Example


Iris Data clustering GMM : EM iteration : 0
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example

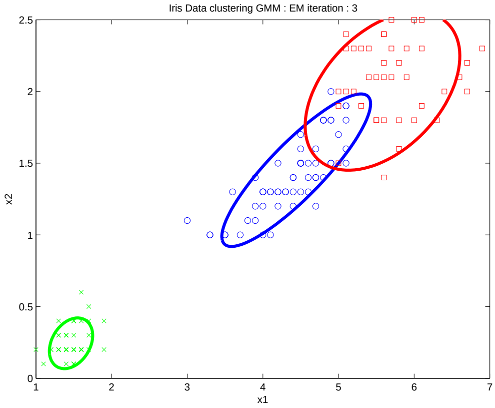

Iris Data clustering GMM : EM iteration : 1
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example



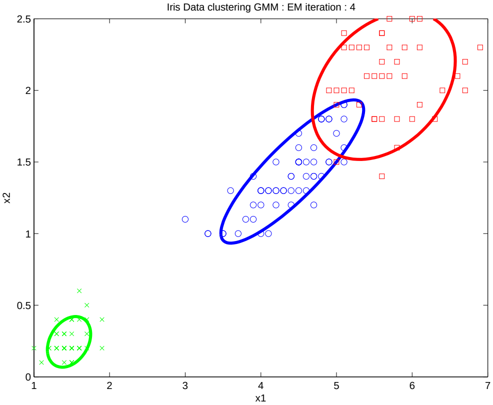
Iris Data clustering GMM : EM iteration : 2
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example



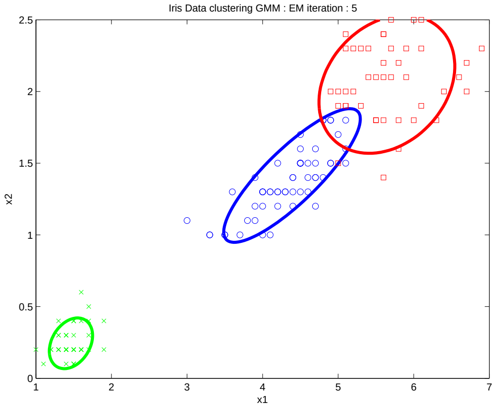
Iris Data clustering GMM : EM iteration : 3
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example



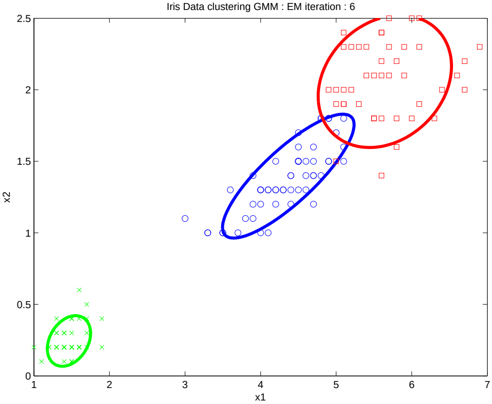
Iris Data clustering GMM : EM iteration : 4
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
x1
x2

Example

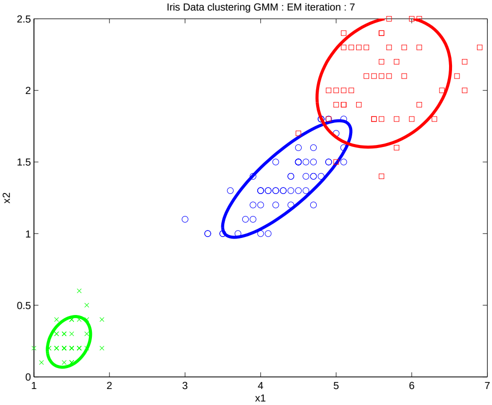


Iris Data clustering GMM : EM iteration : 5
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
x1
x2

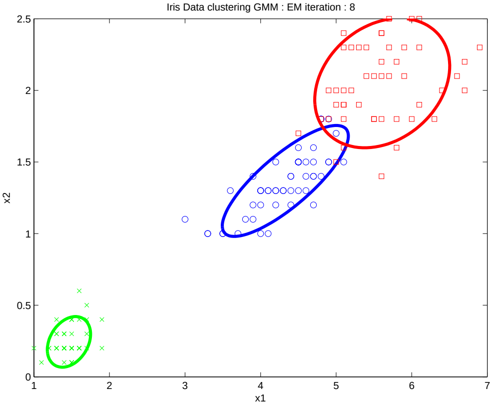
Example


Iris Data clustering GMM : EM iteration : 6
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

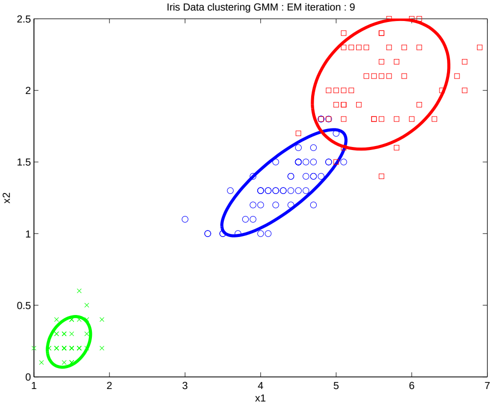
Example


Iris Data clustering GMM : EM iteration : 7
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
x1
x2

Example

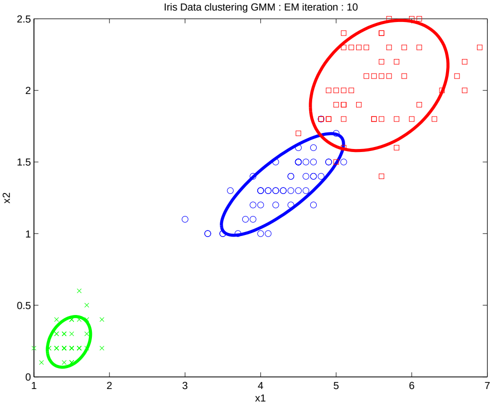


Example

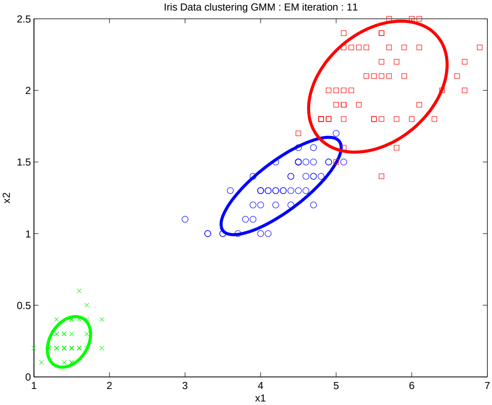


Iris Data clustering GMM : EM iteration : 9
2.5
2
1.5
1
0.5
0
x2
1
2
3
4
5
6
7
x1

Example

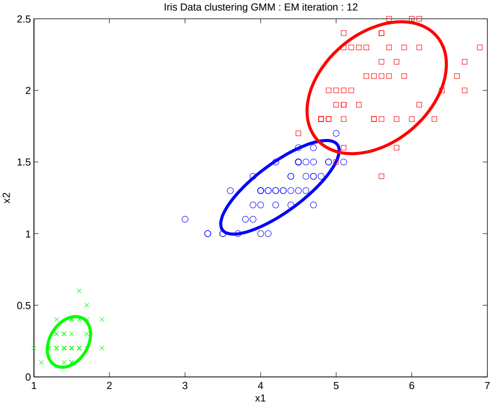

Iris Data clustering GMM : EM iteration : 10
2.5
2
1.5
1
0.5
0
x2
1
2
3
4
5
6
7
x1

Example

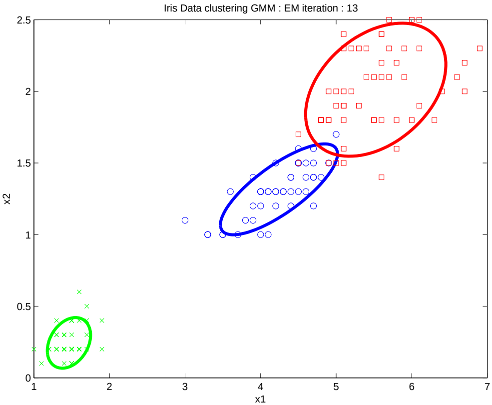


Iris Data clustering GMM : EM iteration : 11
x2
x1
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7

Example

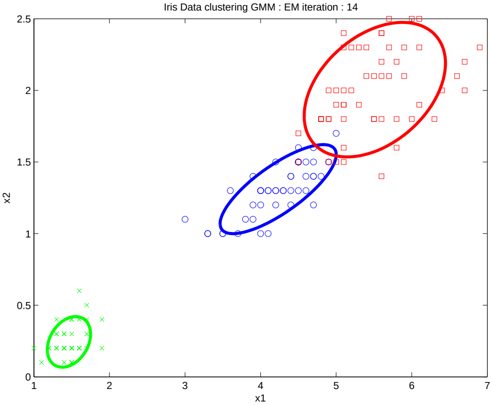

Iris Data clustering GMM : EM iteration : 12
x1
x2
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7

Example

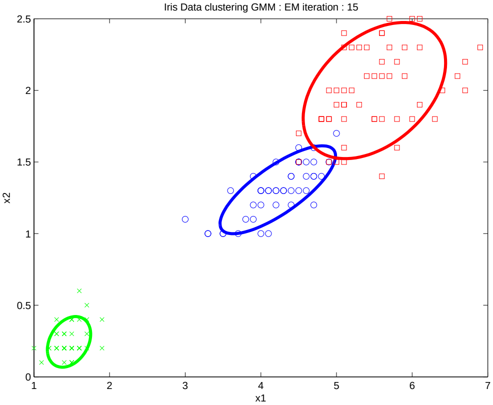


Iris Data clustering GMM : EM iteration : 13
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
x1
x2

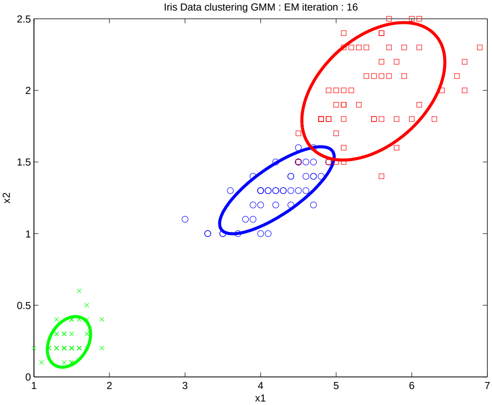
Example


Iris Data clustering GMM : EM iteration : 14
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example

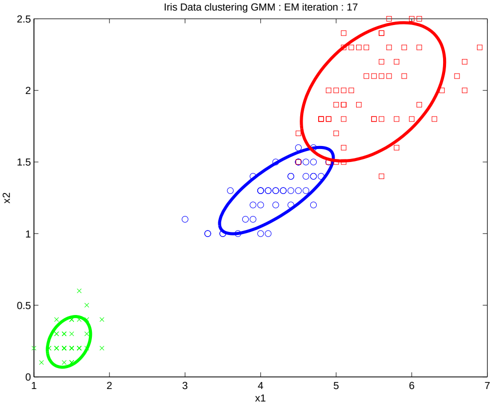

Iris Data clustering GMM : EM iteration : 15
x1
x2
1
2
3
4
5
6
7
0
0.5
1
1.5
2
2.5

Example

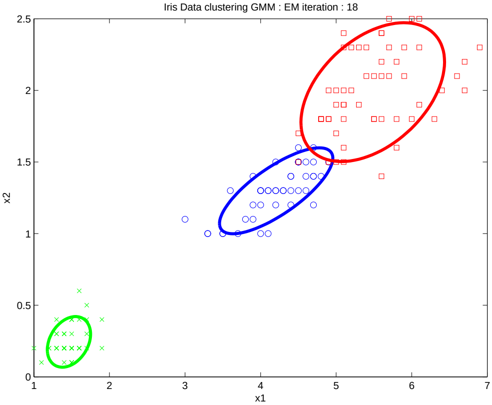


Iris Data clustering GMM : EM iteration : 16
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example

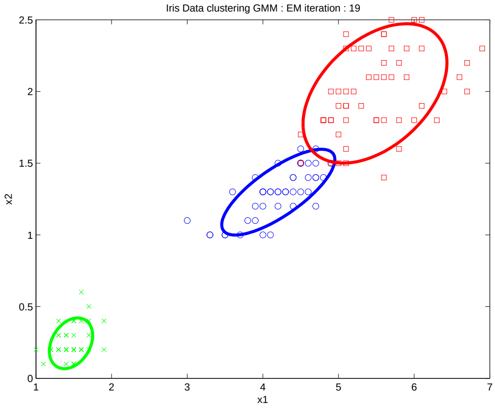

Iris Data clustering GMM : EM iteration : 17
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example



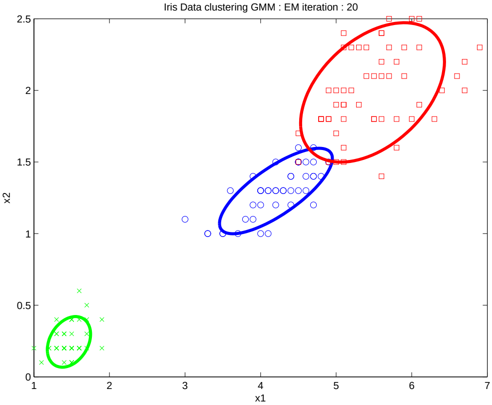
Iris Data clustering GMM : EM iteration : 18
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
x1
x2

Example

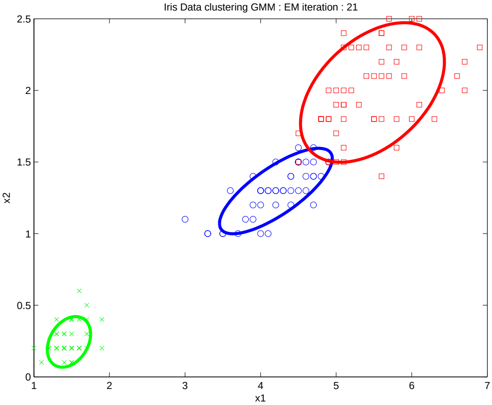


Iris Data clustering GMM : EM iteration : 19
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

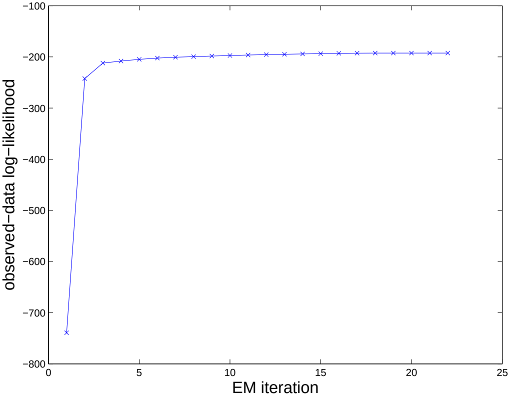
Example


Iris Data clustering GMM : EM iteration : 20
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example


Iris Data clustering GMM : EM iteration : 21
x1
x2
0
0.5
1
1.5
2
2.5
1
2
3
4
5
6
7

Example



observed−data log−likelihood
EM iteration
−100
−200
−300
−400
−500
−600
−700
−800
0
5
10
15
20
25

Parsimonious GMMs for high-dimensional data

- Parsimonious Gaussian mixture models [1] are statistical models that allow for capturing a specific cluster shapes (e.g., clusters having the same shape or different shapes, spherical or elliptical clusters, etc).
- Eigenvalue decomposition of the cluster covariance matrices :

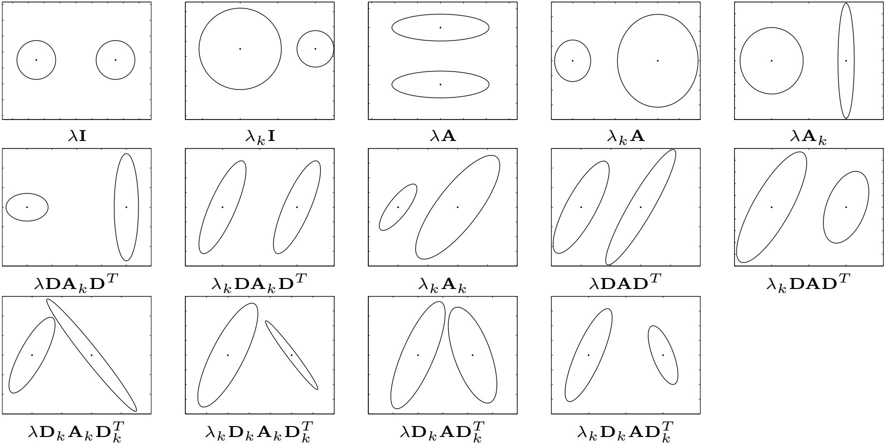
$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$$

where

- λ_k represents the volume of the kth cluster (the amount of space of the cluster).
- $\mathbf{D}_k$ is a matrix with columns corresponding to the eigenvectors of $\mathbf{\Sigma}_k$ that determines the orientation of the cluster.
- $\mathbf{A}_k$ is a diagonal matrix, whose diagonal entries are the normalized eigenvalues of $\mathbf{\Sigma}_k$ arranged in a decreasing order and its determinant is 1. This matrix is associated with the shape of the cluster.

1. Banfield and Raftery [1993], **?**

Parsimonious GMMs for high-dimensional data



Model selection

- The problem of choosing the number of clusters can be seen as a model selection problem.
- The model selection task consists of choosing a suitable compromise between flexibility so that a reasonable fit to the available data is obtained, and over-fitting.
- A common way is to use a criterion (score function) that ensure the compromise.

$$\text{score}(\text{model}) = \text{error}(\text{model}) + \text{penalty}(\text{model complexity})$$

  which will be minimized.
- Here the complexity of a model $\mathcal{M}$ is related to the number of its (free) parameters ν

Model selection

- Akaike Information Criterion (AIC) :

$$\mathsf{AIC}(\mathcal{M}_m) = \ln L(\widehat{\boldsymbol{\theta}}_m) - \nu_m$$

- Bayesian Information Criterion (BIC) :

$$\mathsf{BIC}(\mathcal{M}_m) = \ln L(\widehat{\boldsymbol{\theta}}_m) - \frac{\nu_m \log(n)}{2}$$

- Integrated Classification Likelihood (ICL) :

$$\mathsf{ICL}(\mathcal{M}_m) = \ln L_c(\widehat{\boldsymbol{\theta}}_m) - \frac{\nu_m \log(n)}{2}$$

where $\ln L_c(\widehat{\boldsymbol{\theta}}_m)$ is the complete-data log-likelihood for the model $\mathcal{M}_m$ and ν_m denotes the number of free model parameters. For example, in the case of a d-dimensional Gaussian mixture model we have :

$$\nu = \underbrace{(K-1)}_{\pi_k\text{'s}} + \underbrace{K \times d)}_{\{\boldsymbol{\mu}_k\}} + \underbrace{K \times \frac{d \times (d+1)}{2}}_{\{\boldsymbol{\Sigma}_k\}} = \frac{K \times (d+1) \times (d+2)}{2} - 1.$$

Examples

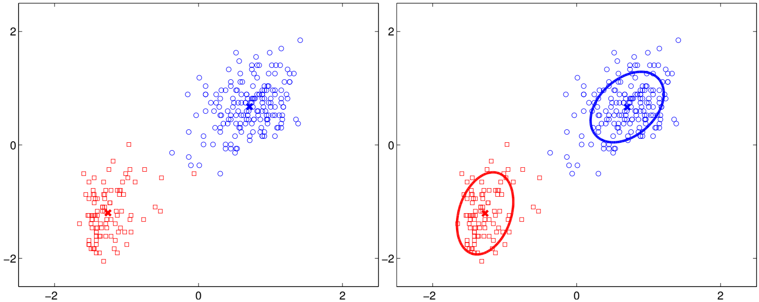


FIGURE – Clustering results obtained with K-means algorithm (left) with $K = 2$ and the EM algorithm (right). The cluster centers are shown by the red and blue crosses and the ellipses are the contours of the Gaussian component densities at level 0.4 estimated by EM. The number of clusters for EM have been chosen by BIC for $K = 1, \dots, 4$.

Examples

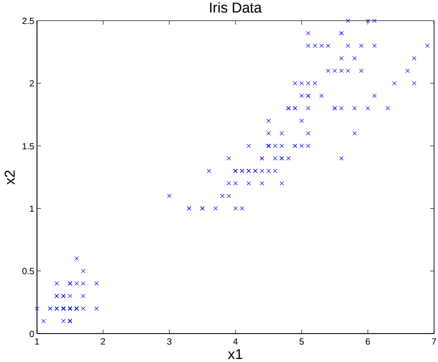


FIGURE – A three-class example of a real data set : Iris data of Fisher.

Examples

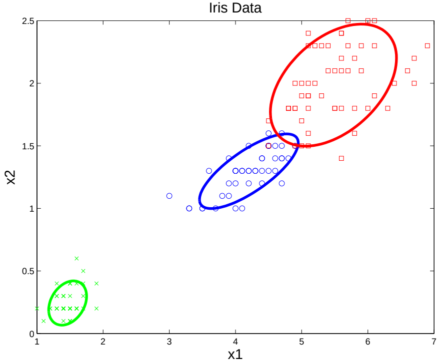


FIGURE – Iris data : Clustering results with EM for a GMM and AIC.

Examples

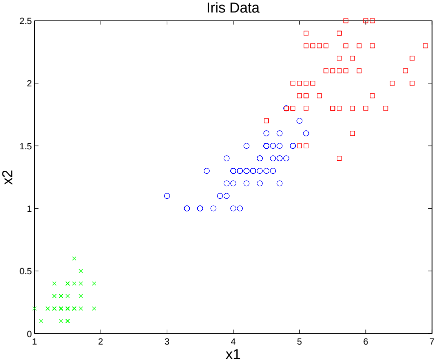


FIGURE – Iris data of Fisher : The data are colored according to the true partition.

Latent data models for dimensionality reduction

- Dimensionality reduction for high dimensional data (for representation/visualization etc)
- Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933],
- Probabilistic PCA [Tipping and Bishop, 1997, 1999, Roweis, 1998]
- Factor Analysis (FA)[Spearman, 1904, Thurstone, 1947],

Principal Component Analysis (PCA)

- PCA is a linear projection which maximizes the variance in the projected space [Hotelling, 1933].

Consider a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$.
$\Rightarrow$ The aim is to project the data onto a space having dimensionality $M < d$ while maximizing the variance of the projected data.

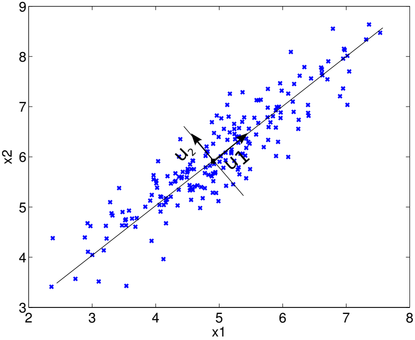
Consider the sample mean vector and the sample covariance matrix :
$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$.
$\Rightarrow$ The variance of the projected data is therefore given by the scalar :

$$v(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i - \mathbf{u}^T\bar{\mathbf{x}})(\mathbf{u}^T\mathbf{x}_i - \mathbf{u}^T\bar{\mathbf{x}})^T = \mathbf{u}^T\mathbf{S}\mathbf{u}\,\cdot \qquad (1)$$

The **principal axes** (the direction vectors) are then given by :

$$\mathbf{u} = \arg\max_{\mathbf{u}\in\mathbb{R}^d} \mathbf{u}^T\mathbf{S}\mathbf{u} \qquad (2)$$

subject to $\mathbf{u}^T\mathbf{u} = 1$ and $\mathbf{u}_j^T\mathbf{u}_k = 0$ for $j \neq k$



9
8
7
6
5
4
3
x2
2
3
4
5
6
7
8
x1
u2
u1

Disadvantage :

The absence of a probability density model and associated likelihood measure.

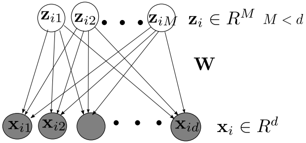
Deriving PCA from the perspective of density estimation would offer a number of important advantages, including the following :

- The likelihood measure allows comparison with other density models
- We can derive EM for PCA and hence deal with possible missing values
- Possibility to perform Bayesian inference (e.g. for model selection)
- Possibility of computing the the posterior class probabilities if PCA is used to model the class-conditional densities in classification,
- The value of the probability density function would give a measure of the novelty of a new data point.
- PCA model could be extended to a mixture framework.

$\Rightarrow$ Use Probabilistic Principal Component Analysis (PPCA)

Probabilistic Principal Component Analysis (PPCA)

Latent variable model for PPCA [Tipping and Bishop, 1997, 1999, Roweis, 1998] :



$$
\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data = linear transf. of } \mathbf{z} + \text{ additive Gaussian noise}\\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \text{ latent variables of the principal component subspace}\\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \text{ conditional density for the observed data}\\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \text{ marginal density for the observed data}
\end{aligned}
$$

$\hookrightarrow (\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$: Parameter estimation using EM [Tipping and Bishop, 1997, 1999, Roweis, 1998]

EM for PPCA

NB : for $\boldsymbol{\mu}$, we get its closed form solution : $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$
Only $\mathbf{W}$ and σ^2 are computed in an iterative way by EM

1. **E-step** : By using the old parameters values, compute

$$\begin{aligned} \mathbb{E}[\mathbf{z}_i] &= (\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (3) \\ \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] &= \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T \quad (4) \end{aligned}$$

2. **M-step**

$$\begin{aligned} \mathbf{W}_{\text{new}} &= \left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\mathbb{E}[\mathbf{z}_i]^T\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\right]^{-1} \quad (5) \\ \sigma^2_{\text{new}} &= \frac{1}{nd}\sum_{i=1}^{n}\left\{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_i]^T\mathbf{W}_{\text{new}}^T(\mathbf{x}_i - \bar{\mathbf{x}}) + trace(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T]\mathbf{W}_{\text{new}}\mathbf{W}_{\text{new}}^T\right\} \\ & \quad (6) \end{aligned}$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}_{\text{old}}]$

Factor Analysis (FA) I

Factor Analysis (FA) [Spearman, 1904, Thurstone, 1947]
FA is closely related to PPCA
The only difference is

$$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ conditional density for the observed data}$$

$\boldsymbol{\Psi}$ is a $d \times d$ digonal matrix ; rather than

$$\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2\mathsf{I}) \text{ conditional density for the observed data}$$

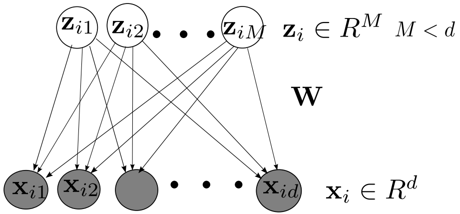
(isotropic covariance matrix).

Factor Analysis (FA) II

Generative model

$$\begin{aligned}
\mathbf{x}_i &= \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \text{ Observed data = linear transf. of } \mathbf{z} + \text{additive Gaussian noise}\\
\mathbf{z}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \text{ latent variables of the principal component subspace}\\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ zero-mean Gaussian noise}
\end{aligned}$$

$$\begin{aligned}
\mathbf{x}_i|\mathbf{z}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ conditional density for the observed data}\\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}) \text{ marginal density for the observed data}
\end{aligned}$$



EM for Facotr Analysis

1. **E-step**

$$
\begin{aligned}
\mathbb{E}[\mathbf{z}_i] &= (\mathbf{I} + \mathbf{W}^T \mathbf{\Psi}^{-1} \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{\Psi}^{-1} \mathbf{x}_i - \bar{\mathbf{x}}) && (7) \\
\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T] &= (\mathbf{I} + \mathbf{W}^T \mathbf{\Psi}^{-1} \mathbf{W})^{-1} + \mathbb{E}[\mathbf{z}_i]\mathbb{E}[\mathbf{z}_i]^T && (8)
\end{aligned}
$$

2. **M-step**

$$
\begin{aligned}
\mathbf{W}_{\text{new}} &= \left[\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}) \mathbb{E}[\mathbf{z}_i]^T\right] \left[\sum_{i=1}^{n} \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T]\right]^{-1} && (9) \\
\mathbf{\Psi}_{\text{new}} &= \text{diag}\left\{\mathbf{S} - \mathbf{W}_{\text{new}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{z}_i] (\mathbf{x}_i - \bar{\mathbf{x}})^T\right\} && (10)
\end{aligned}
$$

NB. Here $\mathbb{E}[.]$ is actually $\mathbb{E}[.|\mathbf{X}, \{\mathbf{W}, \boldsymbol{\mu}, \mathbf{\Psi}\}_{\text{old}}]$

Time series 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

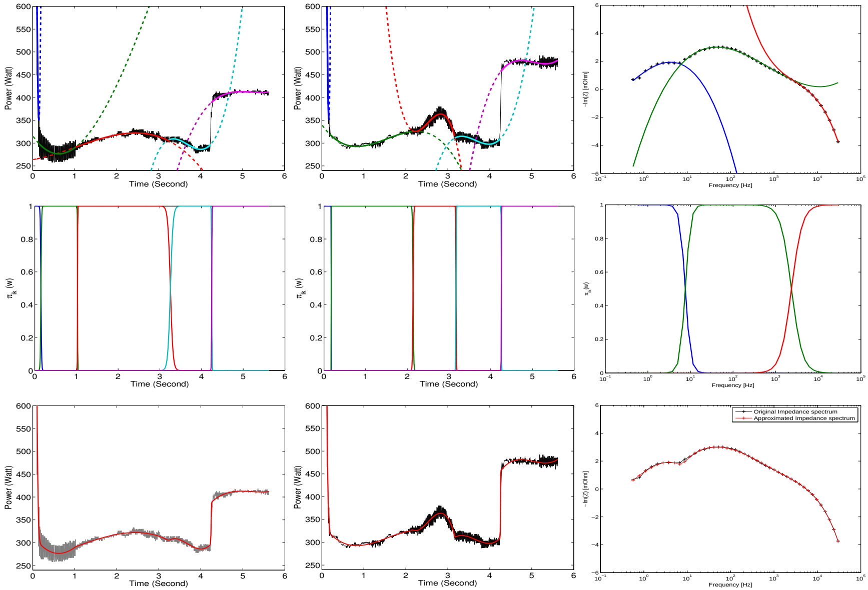
1. Approximation : a curve prototype $\hat{y}_i = \mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i$

   $\hookrightarrow$ 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

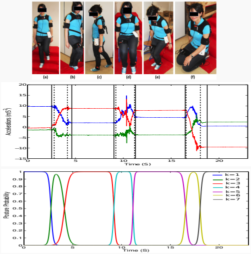
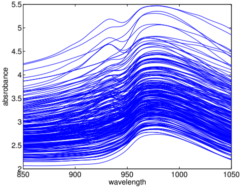


FIGURE – MRHLP segmentation of acceleration data issued from three body-worn sensors

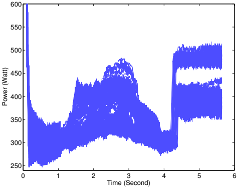
Outline

Functional data are increasingly frequent

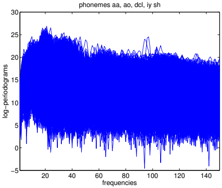
[James and Hastie, 2001, James and Sugar, 2003]
[Ramsay and Silverman, 2005]
[Chamroukhi et al., 2010]
[Bouveyron and Jacques, 2011]
[Samé et al., 2011]
[Jacques and Preda, 2014]
[Bouveyron et al., 2018]
[Chamroukhi and Nguyen, 2018]



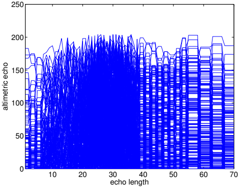
Tecator data



Railway switch curves



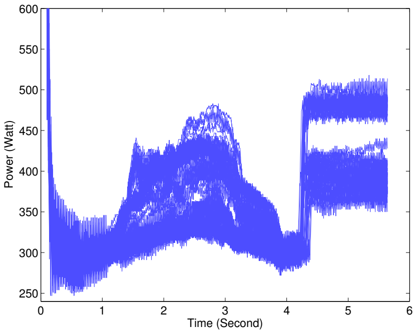
Phonemes curves



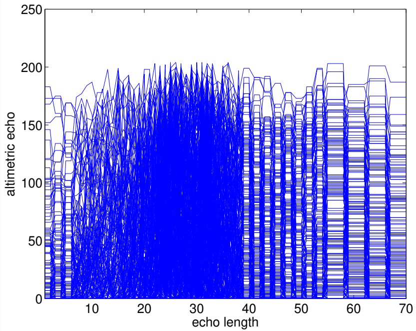
Satellite waveforms

High-dimensional FDA by clustering/segmentation

Non-stationary time series/functions



Railway curves

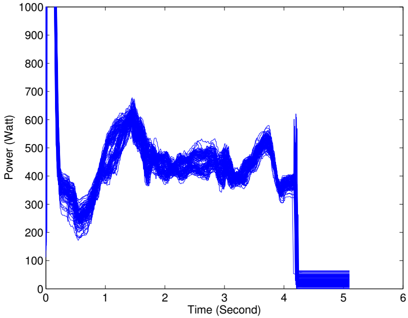


Satellite waveforms

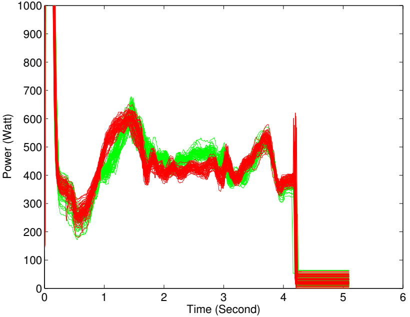
Objectives

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

Functional data clustering



Functional data clustering



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 :

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

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

  $\hookrightarrow$ more tailored to approximate smooth functions

  $\hookrightarrow$ do not account for 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

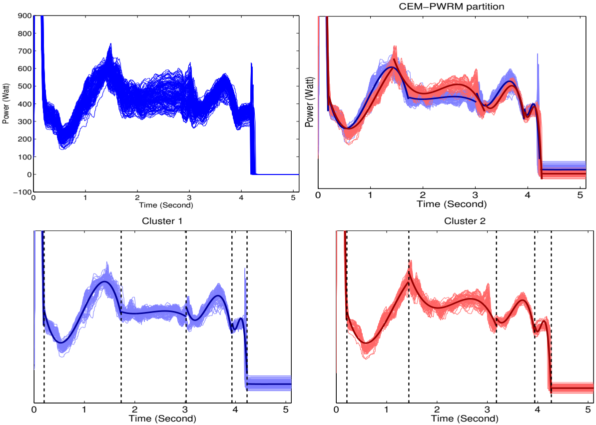
   Complexity in $\mathcal{O}(I_{\text{EM}}KRnm^2p^3)$: An issue for large m

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)

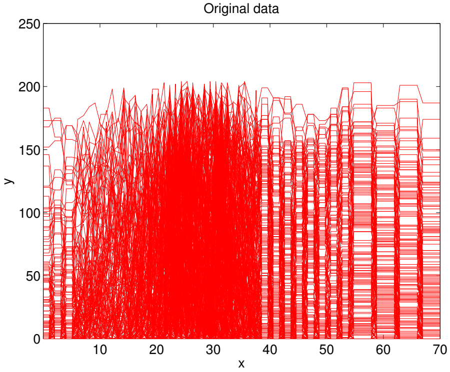


EM-GMM	EM-PRM	EM-PSRM	K-means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

TABLE – Estimated intra-cluster inertia for the switch curves.

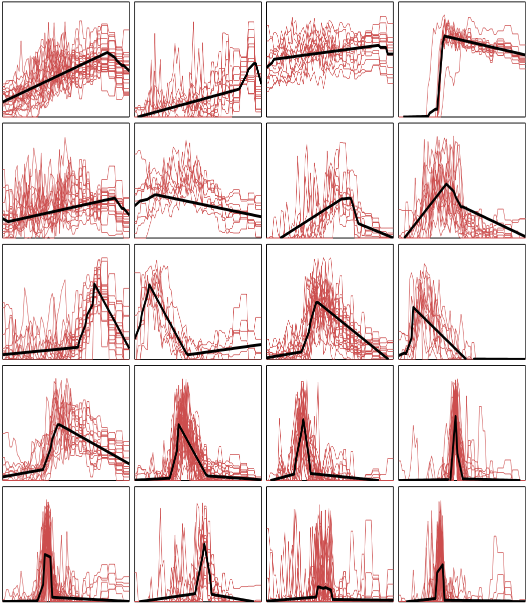
Application to Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data [2] 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].



2. 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 :

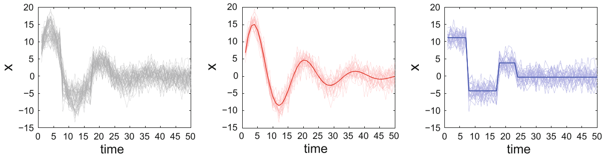
  $\hookrightarrow$ cluster memberships (global) $Z_{ik}=1$ iff $Z_i=k$

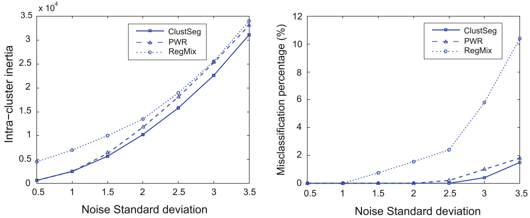
  $\hookrightarrow$ regime memberships for a given cluster (local) : $H_{ijr}=1$ iff $H_{ij}=r$

  MixRHLP deals better with the quality of regime changes

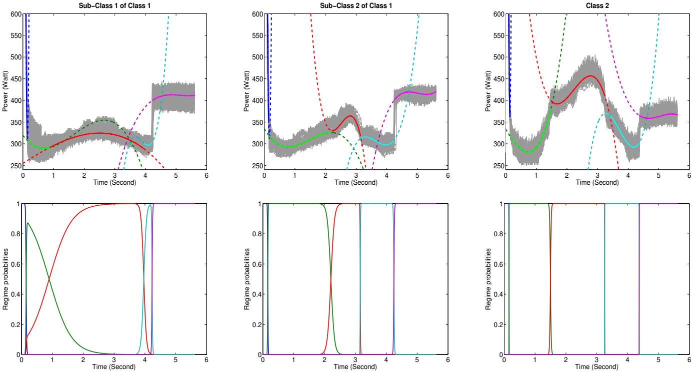
- Parameter estimation via the EM algorithm : EM-MixRHLP
- EM-MixRHLP has complexity in $\mathcal{O}(I_{\text{EM}}I_{\text{IRLS}}KR^3nmp^3)$ (K-means type for piecewise regression is in $\mathcal{O}(I_{\text{KM}}KRnm^2p^3)$ $\hookrightarrow$ EM-MixRHLP is computationally attractive for large values of 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**	$\mathbf{3.2012 \times 10^9}$

Phonemes data

Phonemes data set used in Ferraty and Vieu [2003] [3]
1000 log-periodograms (200 per cluster)

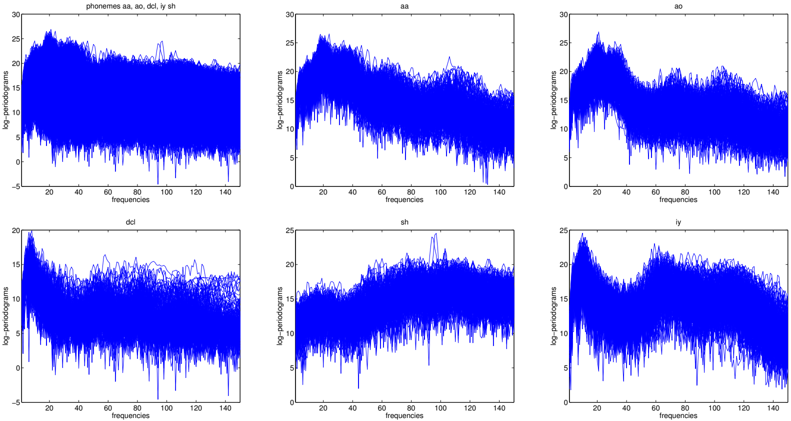


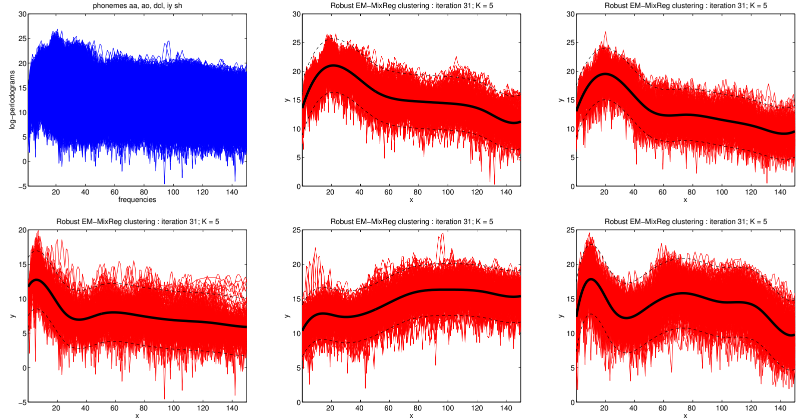
FIGURE – Original phoneme data and curves of the five classes : "ao", "aa", "yi", "dcl", "sh".

3. 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] [4]

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 %

4. 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] [5]
- 384 genes expression levels over 17 time points.

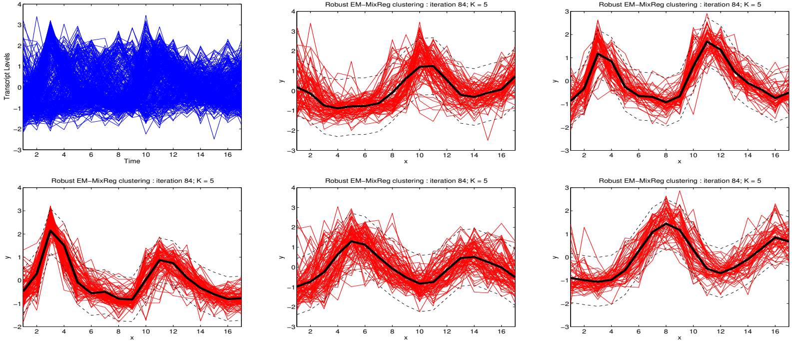


FIGURE – EM-like clustering results with the bSRM model.

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

5. `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.

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 :

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

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

Handwritten digit clustering using the BMSSR

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

FIGURE – Cluster mean images obtained by the BMSSR model with 12 mixture components.

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

Outline

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

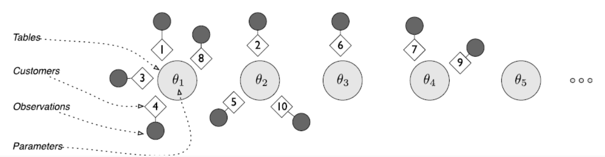
DPM : Generative model

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

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

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

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



- 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 $\boldsymbol{\theta}_k$

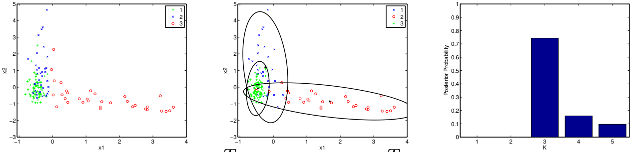
Bayesian model comparison by using Bayes Factors

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

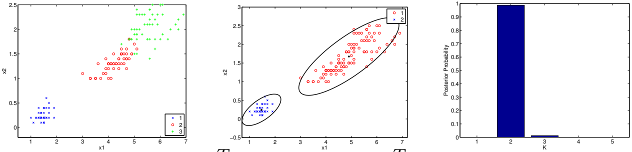
$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)\mathrm{d}\boldsymbol{\theta}_m \approx (2\pi)^{\frac{\nu_m}{2}} |\hat{\mathbf{H}}|^{\frac{1}{2}} p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m)$

Clustering of benchmarks

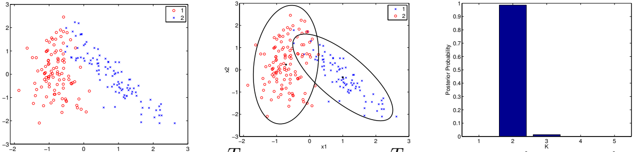
Diabetes data set, Geyser data set, Crabs data set



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



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

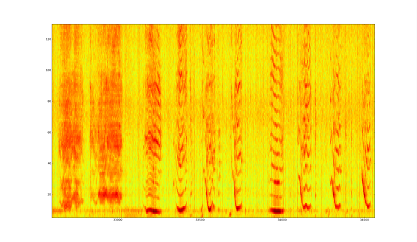


$$\log 2\mathrm{BF} : \lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T \quad vs \quad \lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T \; = \; 36.08 \text{ (Decisive)}$$

Humpback whale song decomposition

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

Humpback Whale

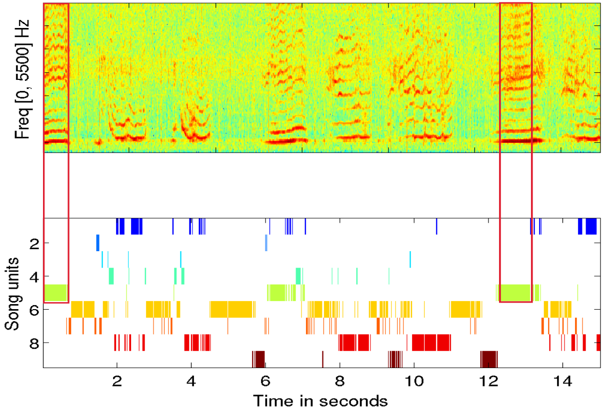


Spectrum of a signal (20 s).

Objectives

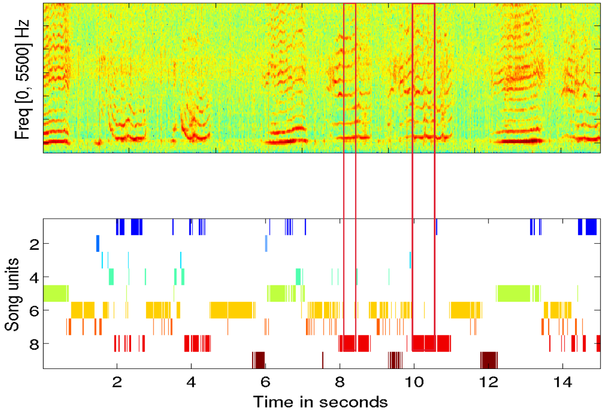
- Discovering “call units”, which can be considered as a whale “alphabet”
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.

Unsupervised decomposition of whale song signals



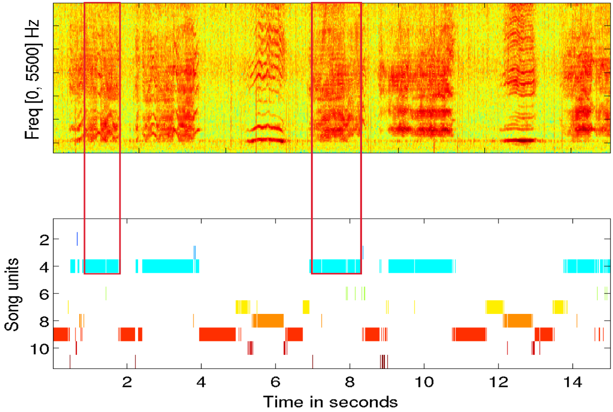
- Sound demo of Unit 5 DPPM $\lambda \mathbf{I}$: (sec. 0) (sec. 12)

Unsupervised decomposition of whale song signals



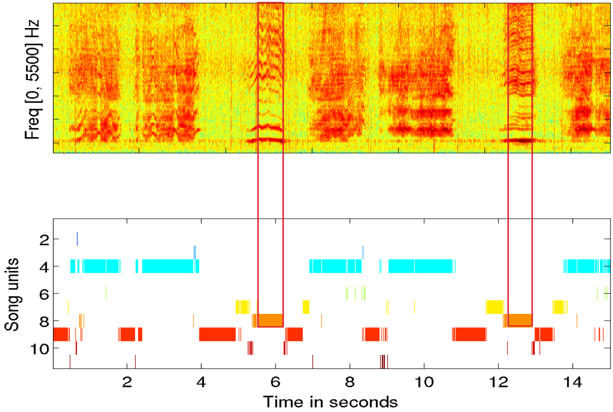
- Sound demo of Unit 8 DPPM $\lambda \mathbf{I}$: (sec. 8) (sec. 10)

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)

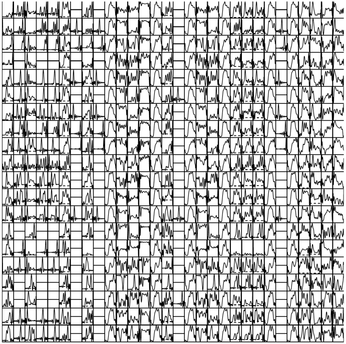
Outline

Outline

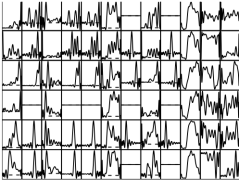
High-dimensional functional data clustering

- Multivariate functional data are increasingly present
- e.g : Data continuously recorded for different subjects from multiple subject' sensors

$\hookrightarrow$ Measurements collected from different network elements (transceivers, cells, sites. . .) :



Data



Zoom

FIGURE – An example with $d = 30$ and $n = 20$ daily observations [Ben Slimen et al., 2016].

High-dimensional functional data clustering

Questioning

Clustering of highly multivariate functional data with two guidelines :

- (1) Mathematical guideline : warranty for estimation and selection
- (2) User guideline : keep a user-friendly meaning of the process

Both are important because clustering is a highly risky task. . .

Proposed answering

(1) Model-based co-clustering with (2) temporal curve segmentation

Novelty corresponds to combining both (1) and (2)

Clustering VS Co-Clustering

- Simultaneous clustering of lines/indiv. ($\boldsymbol{Z}$) and columns/var. ($\boldsymbol{W}$)
- Can be used as a way to reduce dimensionality (var. $\rightarrow \boldsymbol{W}$)

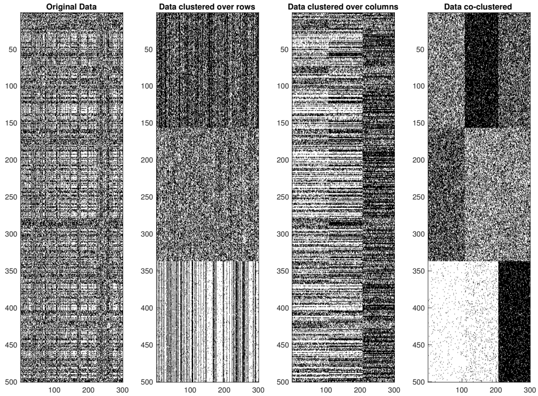


FIGURE – Binary data set with $n = 500$, $d = 300$, $K = M = 3$

Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

$$f(\boldsymbol{X};\boldsymbol{\Psi}) \quad = \quad \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\pi},\boldsymbol{\rho})\, \underbrace{f(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{data kind dependent}}$$

Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent : $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid :
  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$
  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence : $x_{ij}|(z_i,w_j) \perp x_{i'j'}|(z_{i'},w_{j'})$

Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

$$f(\boldsymbol{X};\boldsymbol{\Psi}) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\pi},\boldsymbol{\rho})\underbrace{f(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{data kind dependent}}$$

Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent : $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid :
  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$
  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence : $x_{ij}|(z_i,w_j) \perp x_{i'j'}|(z_{i'},w_{j'})$

↪ binary data : binary [Govaert and Nadif, 2003, 2008, Keribin et al., 2012],
↪ categorical data : multinomial [Keribin et al., 2014]
↪ contingency table : Poisson [Govaert and Nadif, 2003, 2006, 2008]
↪ continuous data : Gaussian [Lomet, 2012, Govaert and Nadif, 2013]
↪ functional data : functional PCA + Gaussian, see further [Ben Slimen et al., 2016]

Inference for the latent block model

Inference of the latent block model

- variational block EM (VBEM) for maximum likelihood estimation and fuzzy co-clustering [Govaert and Nadif, 2006, 2008].
- block classification EM (CEM) algorithm for maximum classification likelihood and hard co-clustering [Govaert and Nadif, 2003, 2006, 2008]
- Bayesian inference [Keribin et al., 2012, 2014] : Bayesian latent block mixtures for binary data and categorical data & a variational Bayesian inference and Gibbs sampling.
- Number of blocks estimation : ICL criterion [Lomet, 2012, Keribin et al., 2014]

Functional data modeling : "classical" approach

[Ramsay and Silverman, 2005] and many others

- Step 1 : $(\boldsymbol{x}, \boldsymbol{y})$ decomposed into a finite basis of function (B-spline. . .) : $Y_i(t) \approx \sum_{r=1}^{d} c_{ir}\phi_r(x_i(t))$ with $\mathbf{c}$ estimated by OLS
- Step 2 : functional principal components analysis (PCA) which is performed as a usual PCA of the basis expansion coefficients $\mathbf{c}$ using a metric defined by the inner products between the basis functions
- Step 3 : set a probability distribution on $\mathbf{c}$, typically Gaussian

It defines a distribution on $\mathbf{c}$ instead of $\boldsymbol{y}$. . .

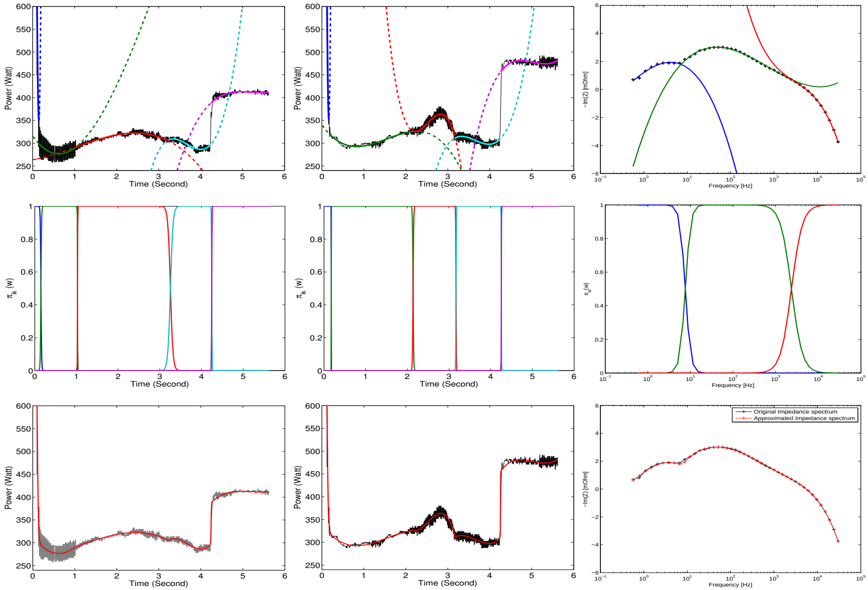
Functional data modeling : regression RHLP

Alternatively, use a segmentation via generative piecewise polynomial regression modeling of $f(\boldsymbol{y}|\boldsymbol{x})$ [Chamroukhi et al.])

$\hookrightarrow$ Regression with Hidden Logistic Process (RHLP)
$\hookrightarrow$ See formula later

It gives a distribution on $\boldsymbol{y}$ and also a meaningful segmentation of the curve

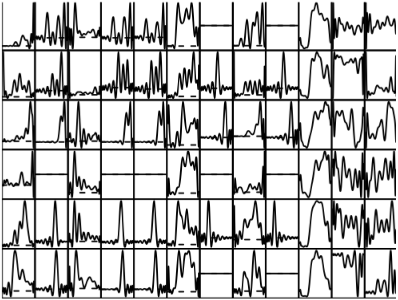
RHLP for modeling different types of functions



Multivariate functional data co-clustering

[Chamroukhi and Biernacki, 2017]

- Data : $\boldsymbol{Y} = (\boldsymbol{y}_{ij})$ a data sample matrix of n individuals defined on a set $\mathcal{I}$ and d continuous functional variables defined on a set $\mathcal{J}$.
- Each variable $\boldsymbol{y}_{ij}$ is an univariate curve $\boldsymbol{y}_{ij} = \big(y_{ij}(t_1), \ldots, y_{ij}(t_{T_{ij}})\big)$ of T_{ij} observations $y(t) \in \mathbb{R}$ linked to covariates $\boldsymbol{x}_{ij} = \big(x_{ij}(t_1), \ldots, x_{ij}(t_{T_{ij}})\big)$ at the points $(t_1, \ldots, t_{T_{ij}})$, typically a sampling time



Embedding RHLP in co-clustering

[Chamroukhi and Biernacki, 2017]

- Functional Latent Block Model for Co-clustering :

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

with parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\pi}^T, \boldsymbol{\rho}^T, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)^T$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{11}^T, \ldots, \boldsymbol{\theta}_{k\ell}^T, \ldots, \boldsymbol{\theta}_{KM}^T)^T$.

Parameter estimation : EM not feasible

- $\hookrightarrow$ Requires the calculation of the posterior joint distribution $\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$
- $\hookrightarrow$ does not factorize due to the conditional dependence on the observed curves of the row and the column labels
- $\Rightarrow$ Variational block EM algorithm : [Govaert and Nadif, 2008, 2013]
- $\hookrightarrow$ We adopt this variational approximation in our context

Variational block EM algorithm

variational approximation

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

Initialization : start from an initial solution at iteration $q = 0$, and then alternate at the $(q+1)$th iteration between the following variational E- and M- steps until convergence :

VE Step Estimate the variational approximated posterior memberships :

1. $\tilde{z}_{ik}^{(q+1)} \propto$
$\pi_k^{(q)} \exp\Big(\sum_{j,\ell,t,r} \tilde{w}_{j\ell}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{T^{(q)}} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\Big]\Big)$
2. $\tilde{w}_{j\ell}^{(q+1)} \propto$
$\rho_\ell^{(q)} \exp\Big(\sum_{i,k,t,r} \tilde{z}_{ik}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{T^{(q)}} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\Big]\Big)$
3. $\tilde{h}_{tr}^{(q+1)} \propto \alpha_{k\ell r}^{(q)}(t; \boldsymbol{\xi}_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{(q)^T} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)$

where :

- $\tilde{z}_{ik} = \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$,
- $\tilde{w}_{j\ell} = \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$,
- $\tilde{h}_{tr} = \mathbb{P}(h_{tr} = 1|z_i, w_j, y_{ij}(t), x_{ij}(t))$

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules :

3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules :
3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

   The regression parameters updates consist in analytic WLS problems :
4. $\boldsymbol{\beta}_{k\ell r}^{(q+1)} = \left[\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \mathbf{X}_{ij}\right]^{-1} \sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \boldsymbol{y}_{ij}$
5. $\sigma_{k\ell r}^{2(q+1)} = \frac{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \|\sqrt{\boldsymbol{\Lambda}_{ijkr}^{(q)}}(\boldsymbol{y}_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_{kr}^{(q+1)})\|^2}{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \ \mathrm{trace}(\boldsymbol{\Lambda}_{ijkr}^{(q)})}$ where $\mathbf{X}_{ij}$ is the design matrix for the ith curve, $\boldsymbol{\Lambda}_{ijkr}^{(q)}$ is the diagonal matrix whose diagonal elements are the posterior segment memberships $\{\tilde{h}_{ijtr}^{(q)}; t = 1, \ldots, T_{ij}\}$.

$\hookrightarrow$ It is also possible to use the Classification EM (CEM) approximation of EM [Celeux and Govaert, 1992].

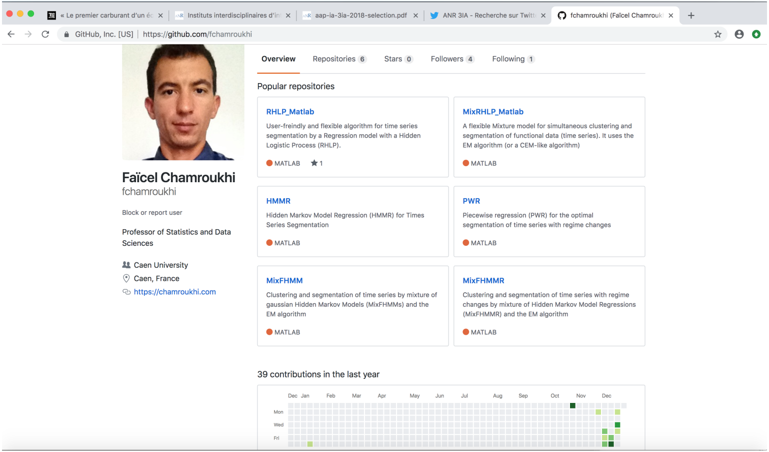
Parameter estimation by an SEM algorithm : SEM-FLBM

- $\hookrightarrow$ The SEM algorithm [Celeux and Diebolt, 1985] allows to overcome some drawbacks of the variational-EM algorithm, including its sensitivity to starting values ; SEM does not use an approximation.
- Eg. SEM for latent block models for categorical data [Keribin et al., 2012, 2014]
- The formulas of VEM-FLBM and SEM-FLBM are essentially the same, except that we incorporate a stochastic step consisting of sampling binary indicator variables z_{ik}, $w_{j\ell}$ and h_{tr} according to $\tilde{z}_{ik}$, $\tilde{w}_{j\ell}$ and $\tilde{h}_{tr}$.

Source codes are/will be made available on github

Matlab/R/Python

https://github.com/fchamroukhi



Data science, Big-Data, AI

The way of the future !

Eg. In France : Interdisciplinary Institutes of Artificial Intelligence (3IA) :

AI Strategy

6/11/2018

Interdisciplinary Institutes of Artificial Intelligence (3IA): the four selected projects

The results of the 3IA (Interdisciplinary Institute of Artificial Intelligence) call for expressions of interest were made public on 6 November 2018 by Frédérique Vidal, French Minister of Higher Education, Research and Innovation, and Mounir Mahjoubi, French Secretary of State for Digital Affairs. The projects of the Grenoble (MIAI@Grenoble-Alpes), Nice (3IA Côte d'Azur), Paris (PRAIRIE) and Toulouse (ANITI) sites have been selected. Inria participates in three of the four successful projects.

At the heart of the national AI strategy

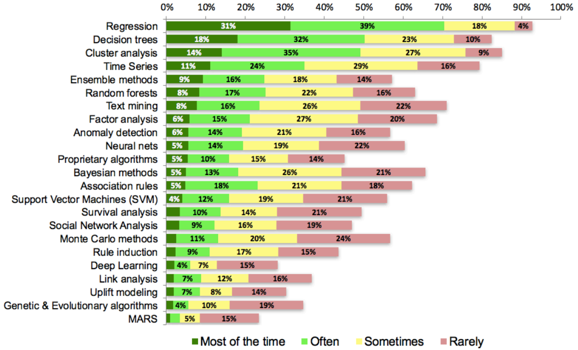
Each of the Inria research centres took part in formulating 3IA Institute projects within the framework of the considerable mobilisation of the regional academic and industry ecosystems: they are part of the dynamic stimulated by the national plan on artificial intelligence announced by the French president following the report by Cédric Villani.

Data science models/algorithms

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

Our Core Algorithms Remain the Same

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.



Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

2016 Rexer Analytics 4

References I

Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3) :803–821, 1993.

Y. Ben Slimen, S. Allio, and J. Jacques. Model-Based Co-clustering for Functional Data. HAL preprint hal-01422756, December 2016. URL `https://hal.inria.fr/hal-01422756`.

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.

C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Analysis and Classification*, 5(4) :281–300, 2011.

C. Bouveyron, L. Bozzi, J. Jacques, and F.-X. Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society, Series C*, 2018.

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

G. Celeux and J. Diebolt. The SEM algorithm a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1) :73–82, 1985.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14 :315–332, 1992.

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.

References II

F. Chamroukhi and C. Biernacki. Model-Based Co-Clustering of Multivariate Functional Data. In *ISI 2017 - 61st World Statistics Congress*, Marrakech, Morocco, Jul 2017. URL https://hal.archives-ouvertes.fr/hal-01653782.

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. URL https://chamroukhi.com/papers/chamroukhi_neucomp_2010.pdf.

Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, To appear, 2018. URL https://chamroukhi.com/papers/MBCC-FDA.pdf. arXiv :1803.00276v2.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B*, 39(1) :1–38, 1977.

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. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2) :463 – 473, 2003. Biometrics.

G. Govaert and M. Nadif. Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5) : 415–422, 2006.

G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6) :3233 –3245, 2008.

G. Govaert and M. Nadif. *Co-Clustering*. Computer engineering series. Wiley-ISTE, November 2013. 256 pages.

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.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 : 417–441, 1933.

References III

Julien Jacques and Cristian Preda. Functional data clustering : A survey. *Adv. Data Anal. Classif.*, 8(3) :231–255, September 2014. ISSN 1862-5347. doi: 10.1007/s11634-013-0158-y. URL `http://dx.doi.org/10.1007/s11634-013-0158-y`.

G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B*, 63 :533–550, 2001.

G. M. James and C. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98 (462), 2003.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, 2012.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014. ISSN 0960-3174. doi: 10.1007/s11222-014-9472-2. URL `http://dx.doi.org/10.1007/s11222-014-9472-2`.

A. Lomet. *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne, 2012.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

G. J. McLachlan and D. Peel. *Finite mixture models*. New York : Wiley, 2000.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, June 2005.

S. Roweis. EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Proceedings of the 11th Conference on Advances in Neural Information Processing Systems (NIPS)*, volume 10. MIT Press, 1998.

A. Samé, F. Chamroukhi, G. Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, pages 1–21, 2011. ISSN 1862-5347.

C. Spearman. General intelligence, objectively determined and measured. *American Journal of psychology*, 15 :201–293, 1904.

References IV

L. L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, 1947.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Comuting Research Group, Aston University, 1997.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61 : 611–622, 1999.

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.

Thank you for your attention !