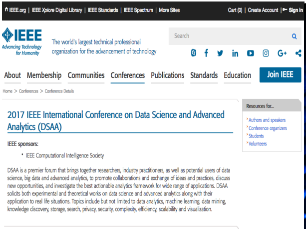
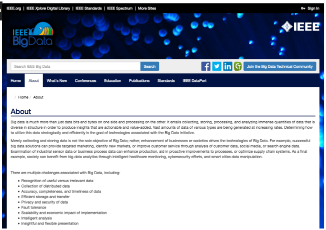
Learning from Heterogeneous & Non-Stationary Time-Series Data

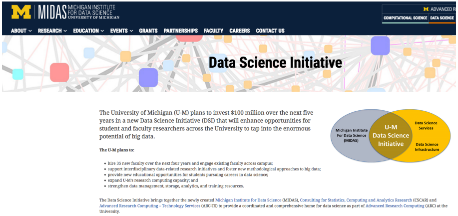
Faicel Chamroukhi

https://chamroukhi.com/

- 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

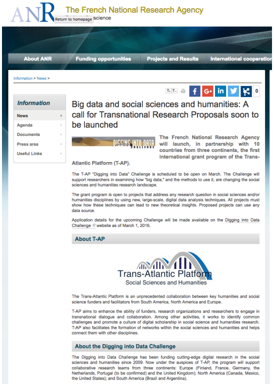


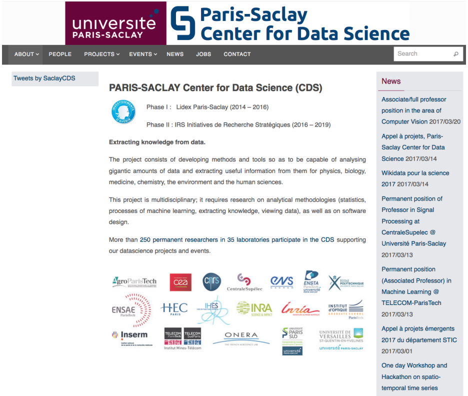





Le CNRS | Annuaires | Mots-Clefs CNRS | Autres sites
Mission interdisciplinarité
Centre national de la recherche scientifique
La Mission pour l'interdisciplinarité
Actualités interdisciplinaires
Les défis
Actions de site
PEPS en réseau
Plateforme réseaux
Colloques et manifestations
Espace communication
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 Mastodons
Défi Mastodons - Les Big Data en recherche - AAP 2017
Grandes masses de données scientifiques - MASTODONS
AAP 2016 - La qualité des données dans les Big Data : résultats


MaDICS
Masses de Données, Informations et Connaissances en Sciences
Big Data - Data Science
Accueil
Actions
Manifestations
Réseaux
Actualités
Offres d'emploi
Calendrier
Nous contacter
Intranet
Présentation MaDICS
Se Connecter
Username
Password
Login →
Devenir membre
Mot de passe oublié
MaDICS On Twitter
Tweets by @GdR_MaDICS
Prochainement


ANR The French National Research Agency
Return to homepage 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
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 that 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

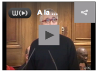
Publié dans Colloques, conférences et débats

Conférence-débat de l'Académie des sciences

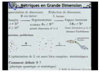
Nous vivons dans une "société de l'information" dont les avancées scientifiques et techniques rapides, associées au développement d'usages nouveaux, conduisent à produire des quantités toujours plus gigantesques de données numériques. Cette situation d'abondance ouvre des perspectives nouvelles tant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

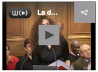
À la découverte des connaissances massives de la Toile
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences



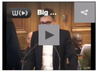
Des mathématiques pour l'analyse de données massives
Stéphane Mallat, professeur à l'École normale supérieure, Paris



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

Outline

Unsupervised Learning

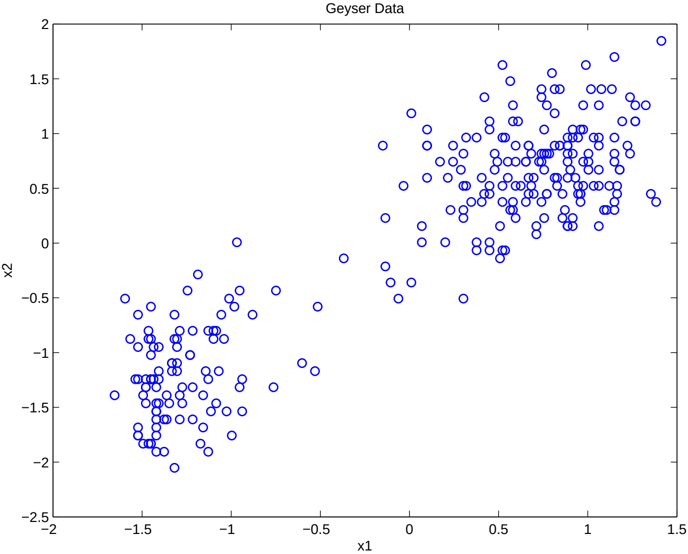
1 Introduction
- Statistics and Data Science
- Unsupervised Learning

2 Latent data models for temporal data segmentation

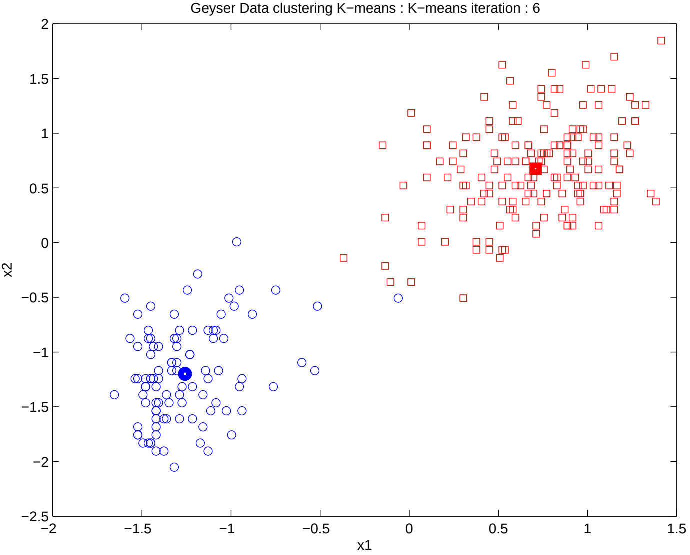
3 Mixture models for functional data analysis

4 Mixture-of-Experts for fitting complex non-normal distributions

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.
- Observed data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbb{R}^d$ with unknown cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$ $(z_i \in 1, \ldots, K)$
- Each of the K clusters 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.
- Observed data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbb{R}^d$ with unknown cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$ $(z_i \in 1, \ldots, K)$
- Each of the K clusters is represented by its mean (cluster centroid) $\boldsymbol{\mu}_k$ in $\mathbb{R}^d$.

K-means [MacQueen, 1967]

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

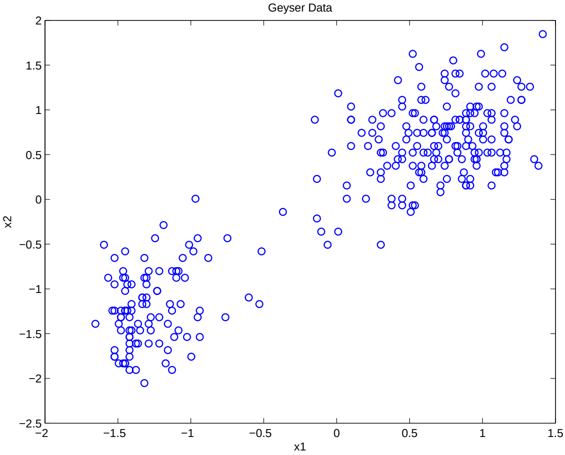
objective function : $\mathcal{J}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}_{z_i}\|^2$

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

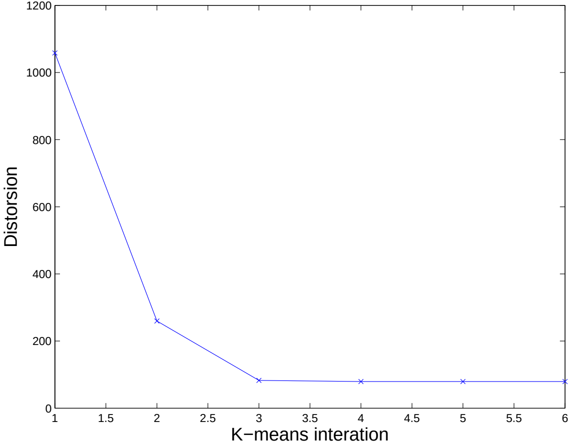
1. **Assignment step** : $z_i^{(t)} = \arg \min_{z \in \mathcal{Z}} \|\mathbf{x}_i - \boldsymbol{\mu}_z\|^2$
2. **Relocation step** : $\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} z_{ik}^{(t)}}$,

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

Example



Example



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

K-means

How to measure uncertainty ?

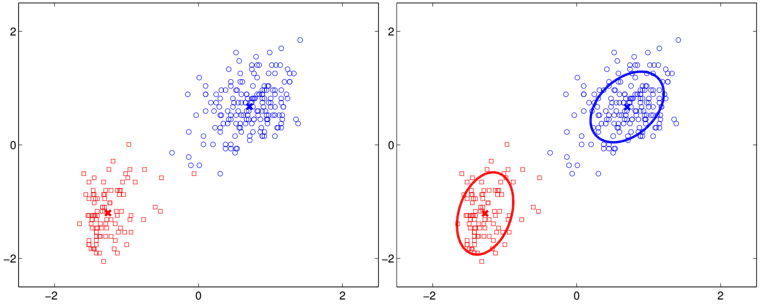


FIGURE – K-means partition (left) vs GMM-EM partition (right)

Scientific context

- The data are assumed to represent samples from random variables with unknown probability distributions
- The area of statistical learning and analysis of complex data.
- **Data :** Complex data $\hookrightarrow$ *heterogeneous, temporal/dynamical, high-dimensional/functional, incomplete,...*
- **Objective :** Transform the data into knowledge :
  $\hookrightarrow$ Reconstruct hidden structure/information, groups/hierarchy of groups, summarizing prototypes, underlying dynamical processes, etc

Modeling framework

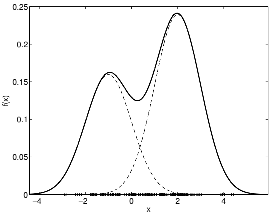
- Latent variable models : $f(x|\boldsymbol{\theta}) = \int_z f(x, z|\boldsymbol{\theta})\mathrm{d}z$
  Generative formulation :
$$\begin{aligned} z &\sim q(z|\boldsymbol{\theta}) \\ x|z &\sim f(x|z, \boldsymbol{\theta}) \end{aligned}$$

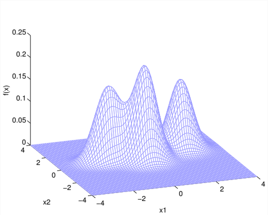
$\hookrightarrow$ Mixture models : $f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{P}(z = k) f(x|z = k, \boldsymbol{\theta}_k)$ and extensions

Mixture models [McLachlan and Peel., 2000]

Mixture modeling framework

- Mixture density : $f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(x|\boldsymbol{\theta}_k)$





- Generative model

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

$\hookrightarrow$ learn $\boldsymbol{\theta}$ from the data

Model-Based Clustering

Clustering based on finite mixture models [McLachlan, 1982, McLachlan and Basford, 1988, Banfield and Raftery, 1993, McLachlan and Peel., 2000]
eg. ; Gaussian mixture models (GMMs) :

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

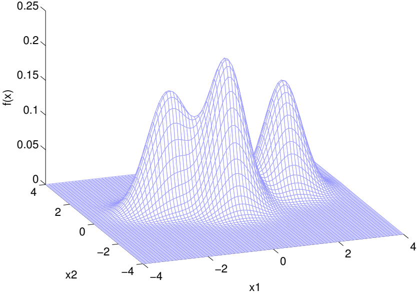


FIGURE – An example of a three-component Gaussian mixture density in $\mathbb{R}^2$.

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

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

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

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

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

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

The EM algorithm [Dempster et al., 1977, McLachlan and Krishnan, 2008]

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

complete log-likelihood : $\log L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log\left[\pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]$ 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

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

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

The EM algorithm [Dempster et al., 1977, McLachlan and Krishnan, 2008]

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

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

Clustering

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

EM for Gaussian mixture models

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

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

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

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

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

Example

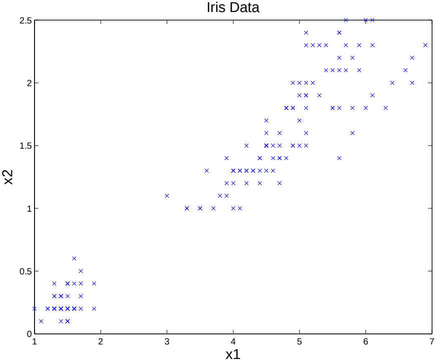
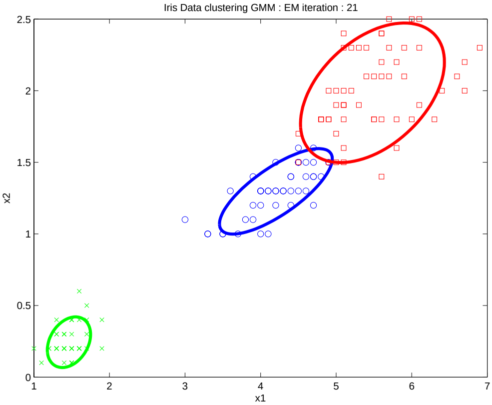


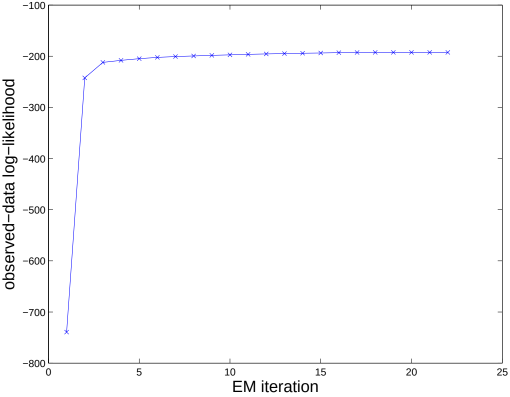
FIGURE – A three-class example of a real data set : Iris data of Fisher.

Example



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

Example


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

Mixtures in a high-dimensional setting

- Parsimonious GMMs [Banfield and Raftery, 1993, Celeux and Govaert, 1995] :
  - Eigenvalue decomposition of the covariance matrices : $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$
  - λ_k the volume of the kth cluster (the amount of space of the cluster).
  - $\mathbf{D}_k = (\mathbf{v}_{k1}, \ldots \mathbf{v}_{kp})$ orthogonal matrix of eigenvectors $\mathbf{v}$ of $\mathbf{\Sigma}_k$: determines the orientation of the cluster.
  - $\mathbf{A}_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kp})/|\mathbf{\Sigma}_k|^{1/p}$ a normalized diagonal matrix (its determinant is 1) of the eigenvalues of $\mathbf{\Sigma}_k$ arranged in a decreasing order. This matrix is associated with the shape of the cluster.

Mixtures in a high-dimensional setting

- Parsimonious GMMs [Banfield and Raftery, 1993, Celeux and Govaert, 1995] :
  - Eigenvalue decomposition of the covariance matrices : $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$
  - λ_k the volume of the kth cluster (the amount of space of the cluster).
  - $\mathbf{D}_k = (\mathbf{v}_{k1}, \ldots \mathbf{v}_{kp})$ orthogonal matrix of eigenvectors $\mathbf{v}$ of $\mathbf{\Sigma}_k$: determines the orientation of the cluster.
  - $\mathbf{A}_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kp})/|\mathbf{\Sigma}_k|^{1/p}$ a normalized diagonal matrix (its determinant is 1) of the eigenvalues of $\mathbf{\Sigma}_k$ arranged in a decreasing order. This matrix is associated with the shape of the cluster.

for $p > n$:

- use regularization (LASSO etc) of the log-likelihood
- Mixtures of Factor Analyzers [McLachlan et al., 2003] (or extensions MCFA, MCUFSA..)
  $\mathbf{\Sigma}_k = \mathbf{B}_k \mathbf{B}_k^T + \mathbf{\Lambda}_k$:
  $\mathbf{B}_k$ is a $p \times q$ (with $q < p$) matrix and $\mathbf{\Lambda}_k$ is a diagonal matrix.
  $\hookrightarrow \left(\mathbf{B}_k \mathbf{B}_k^T + \mathbf{\Lambda}_k\right)^{-1}$ and $|\mathbf{B}_k \mathbf{B}_k^T + \mathbf{\Lambda}_k|$ are calculated in a q-dimensional space !

How many clusters in the data ?

- 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.

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

  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) [Akaike, 1974] :

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

- Bayesian Information Criterion (BIC) [Schwarz, 1978] :

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

- Integrated Classification Likelihood (ICL) [Biernacki et al., 2000] :

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

where $\log L_c(\widehat{\boldsymbol{\theta}}_m)$ 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

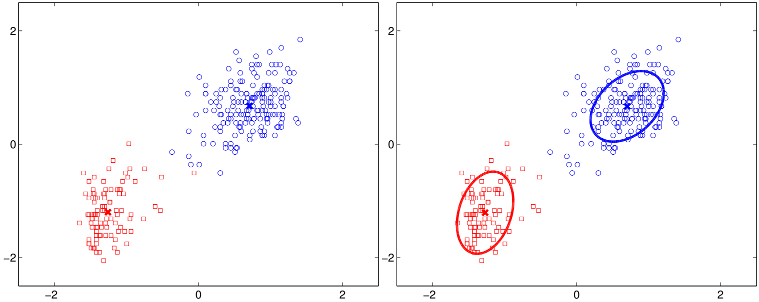


FIGURE – Clustering results obtained with K-means algorithm (left) with $K = 2$ and the EM algorithm (right). The cluster centers are shown by the red and blue crosses and the ellipses are the contours of the Gaussian component densities at level 0.4 estimated by EM. The number of clusters for EM have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

Examples

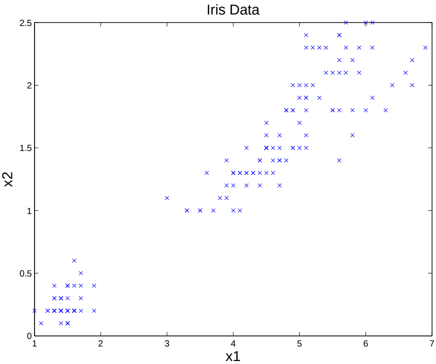


FIGURE – A three-class example of a real data set : Iris data of Fisher.

Examples

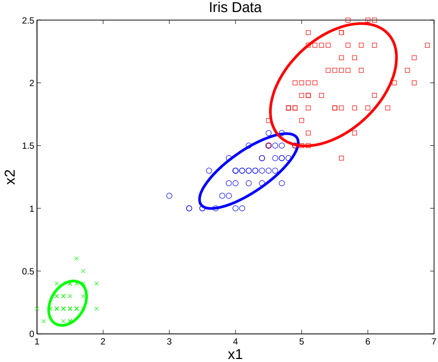
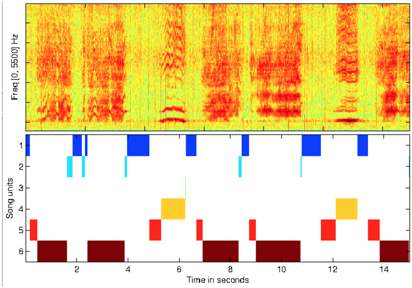
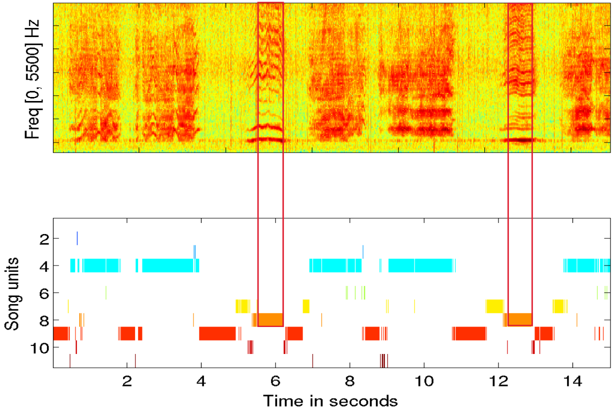


FIGURE – Iris data : Clustering results with EM for a GMM and AIC.

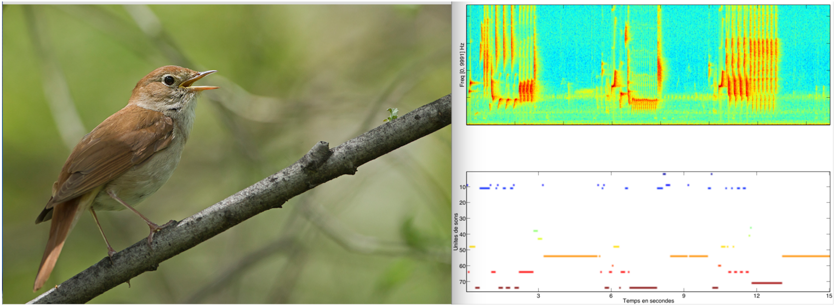
Unsupervised Sparse Signal Decomposition



Unsupervised Sparse Signal Decomposition



Unsupervised Sparse Signal Decomposition



Outline

Outline

1 Introduction

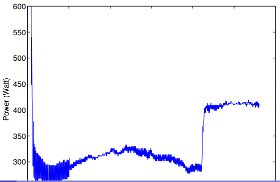
2 Latent data models for temporal data segmentation

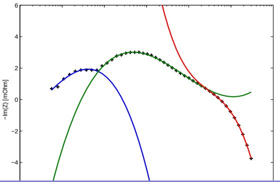
- Piecewise Regression
- Regression with hidden logistic process
- Multiple hidden process regression
- **SaMUraiS** : Open-Source Software

3 Mixture models for functional data analysis

4 Mixture-of-Experts for fitting complex non-normal distributions

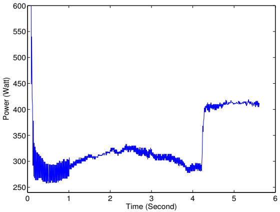
Temporal data with regime changes

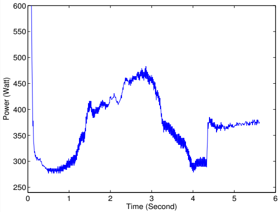




Temporal data

Temporal data with regime changes





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

Objectives

Temporal data modeling and segmentation

Latent data models for temporal data segmentation

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

Times series segmentation context

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

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

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

Piecewise regression [McGee & Carleton 70], Chamroukhi et al. [2009]

- The data : $((t_1, y_1), \ldots, (t_n, y_n))$ where y_i is the observation et time t_i
- The piecewise polynomial regression model is defined as :

$$\forall i = 1, \ldots, n, \quad x_i = \begin{cases} \boldsymbol{\beta}_1^T \boldsymbol{x}_i + \varepsilon_{i1} & \text{if } i \in I_1 \\ \boldsymbol{\beta}_2^T \boldsymbol{x}_i + \varepsilon_{i2} & \text{if } i \in I_2 \\ \vdots & \\ \boldsymbol{\beta}_K^T \boldsymbol{x}_i + \varepsilon_{iK} & \text{if } i \in I_K \end{cases},$$

  - $I_k =]\gamma_k..\gamma_{k+1}]$: indexes of elements of segment k with ($\gamma_1 = 0$ and $\gamma_{K+1} = n$).
  - $\boldsymbol{x}_i = (1, t_i, \ldots, t_i^p)^T$: time-dependent covariates vector
  - $\boldsymbol{\beta}_k \in \mathbb{R}^{p+1}$: regression coefficients vector for the k^{th} segment
  - $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma_k^2)$: independent additive Gaussian noise on the segment k.

The model parameters

$(\boldsymbol{\psi}, \boldsymbol{\gamma})$ with $\boldsymbol{\psi} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+1})^T$.

Parameter estimation piecewise regression

Maximize the likelihood of $(\boldsymbol{\psi}, \boldsymbol{\gamma})$ or equivalently minimize, with respect to $(\boldsymbol{\psi}, \boldsymbol{\gamma})$:

$$J(\boldsymbol{\psi}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{i \in I_k} \left[\log \sigma_k^2 + \frac{(y_i - \boldsymbol{\beta}_k^T \boldsymbol{x}_i)^2}{\sigma_k^2} \right].$$

- Global optimization using by dynamic programming [Bellman 61 ; Stone 61 ; Lechevallier 90, C. 2009] since the criterion J is additive on k

Time series approximation and segmentation

- $\hat{y}_i = \sum_{k=1}^{K} \hat{z}_{ik} \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i \quad ; \quad \forall i = 1, \ldots, n$
- $\hat{z}_{ik} = 1$ if $i \in \hat{I}_k = (\hat{\gamma}_k, \hat{\gamma}_{k+1}]$ (y_i belongs to the k^{th} segment) and $\hat{z}_{ik} = 0$ otherwise

- Using dynamic programming can be computationally expensive
- Provides a hard partition $\Rightarrow$ adapted for regimes with abrupt changes

Regression with hidden logistic process

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

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

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

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

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

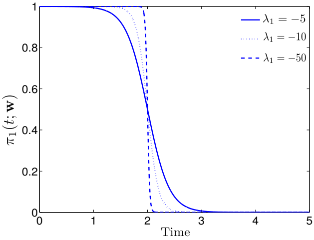
$$f(y_i | t_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)$$

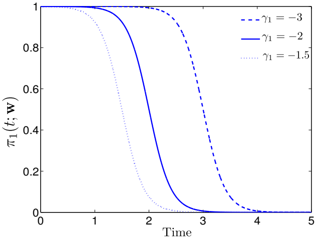
- Both the mixing proportions and the component parameters are time-varying
- Parameter vector of the model :$\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_K^T, \sigma_1^2, \dots, \sigma_K^2)^T$

Illustration

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

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



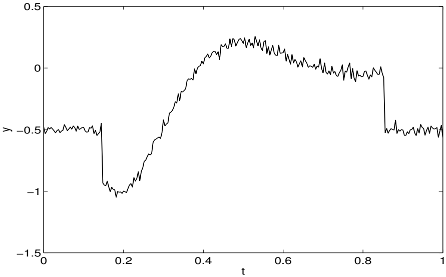


$\Rightarrow$ The parameter $\lambda_k = w_{k1}$ controls the quality of transitions between regimes

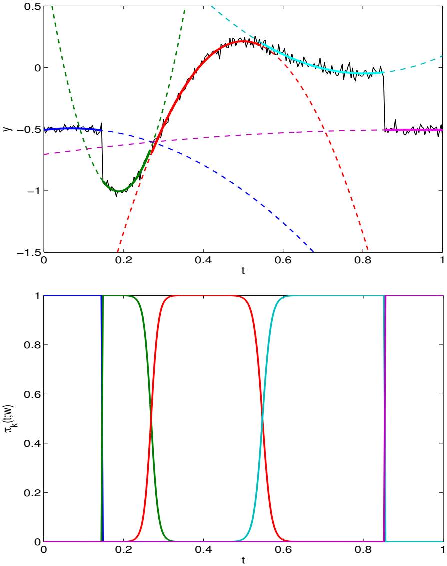
$\Rightarrow$ The parameter $\gamma_k = w_{k0}/w_{k1}$ is related to the transition time point

- Ensure time series segmentation into contiguous segments

Illustration



Illustration



$K = 5$ polynomial components of degree $p = 2$

Parameter estimation : MLE via EM : EM-RHLP

- Parameter vector : $\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T$
- Maximize the observed-data log-likelihood :

$$\log L(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}\big(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\big)$$

- Complete-data log-likelihood

$$\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log[\pi_k(t_i; \mathbf{w}) \mathcal{N}\big(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\big)]$$

$Z_{ik} = 1$ if $Z_i = k$ (i.e., when y_i belongs to the kth component)

- The Q-function

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) &= \mathbb{E}\left[\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) | \boldsymbol{y}, \mathbf{t}; \boldsymbol{\theta}^{(q)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \left[\log \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)\right] \end{aligned}$$

EM-RHLP

- **E-Step** : compute the posterior component memberships :

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|y_i, t_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t_i; \mathbf{w}^{(q)})\mathcal{N}(y_i; \boldsymbol{\beta}_k^{T(q)}\boldsymbol{x}_i, \sigma_k^{2(q)})}{\sum_{\ell=1}^{K} \pi_\ell(t_i; \mathbf{w}^{(q)})\mathcal{N}(y_i; \boldsymbol{\beta}_\ell^{T(q)}\boldsymbol{x}_i, \sigma_\ell^{2(q)})} \cdot$$

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

$$\boldsymbol{\beta}_k^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} y_i \boldsymbol{x}_i$$ weighted polynomial regression

$$\sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{\beta}_k^{T(q+1)} \boldsymbol{x}_i)^2$$

$$\mathbf{w}^{(q+1)} = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(t_i; \mathbf{w})$$ weighted logistic regression

EM-RHLP algorithm

M-Step : Weighted multi-class logistic regression

$$\mathbf{w}^{(q+1)} = \arg\max_{\mathbf{w}} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(t_i;\mathbf{w})$$

- A convex optimization problem
- Solved with a multi-class Iteratively Reweighted Least Squares (IRLS) algorithm (Newton-Raphson)

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \left[\frac{\partial^2 Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial\mathbf{w}\partial\mathbf{w}^T}\right]^{-1}_{\mathbf{w}=\mathbf{w}^{(l)}} \frac{\partial Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial\mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(l)}}$$

- Analytic calculation of the Hessian and the gradient
- EM-RHLP algorithm complexity : $\mathcal{O}(I_{\mathsf{EM}} I_{\mathsf{IRLS}} K^3 p^3 n)$ (more advantageous than dynamic programming).

Time series approximation and segmentation

1. Approximation : a prototype mean curve

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

$\hookrightarrow$ A smooth and flexible approximation thanks to the the logistic weights

$\hookrightarrow$ 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]

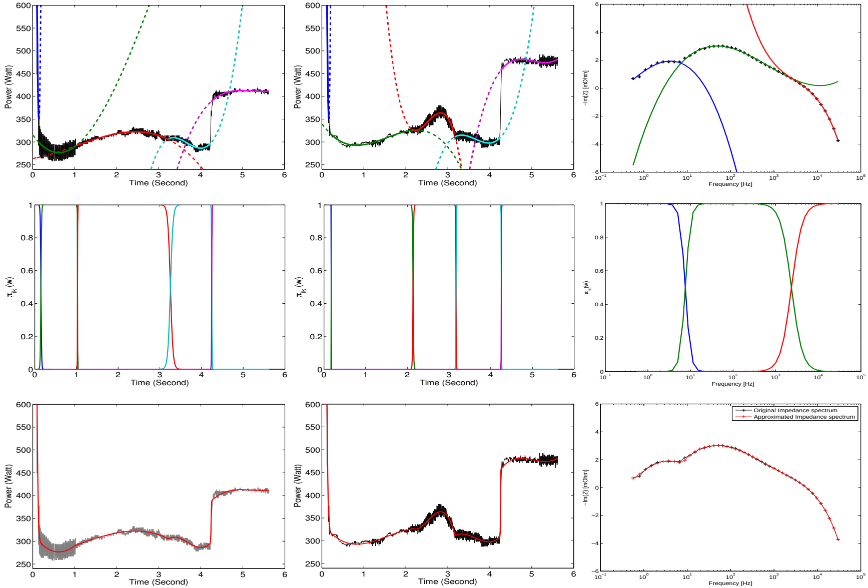
2. Curve segmentation :

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

3. **Model selection** Application of BIC, ICL

BIC$(K,p) = \log L(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$; ICL$(K,p) = \log L_c(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$ where $\nu_{\boldsymbol{\theta}} = K(p+4) - 2$.

Application to real data



Joint segmentation of multivariate time series

Multiple hidden process regression

- Data : $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ a time series of n multidimensional observations $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at instants $\mathbf{t} = (t_1, \ldots, t_n)$.
- Model

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

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

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

↪ Multiple regression with hidden logistic process : Multiple RHLP [6]

↪ Multiple Hidden Markov model regression (MHMMR) [7]

Multiple hidden Markov model regression

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

  $\hookrightarrow$ Solve multiple regression problems

Application to human activity time series

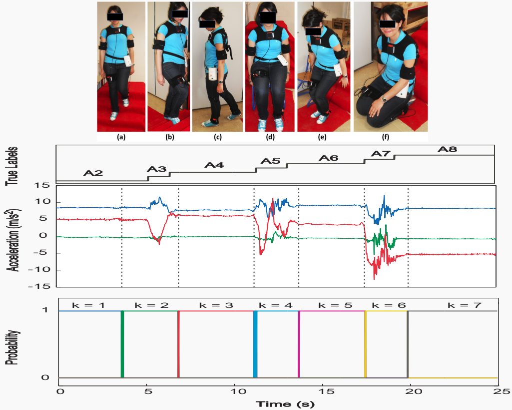


FIGURE – MHMMR Segmentation of acceleration data issued from three body-worn sensors

Multiple regression with hidden logistic process

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

  ↪ Solve multiple regression problems

Application to human activity time series

Problem : Activity recognition from multivariate acceleration time series

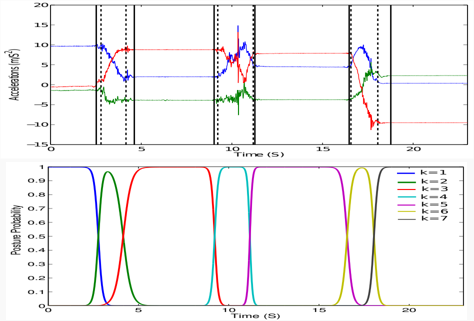


FIGURE – MRHLP segmentation of acceleration data issued from three body-worn sensors

SaMUraiS : open source software for time-series

SaMUraiS : **S**t**A**tistical **M**odels for the **U**nsupe**R**vised segment**A**t**I**on of time-**S**eries [1]

R software Matlab software

1. credit : pictures above created via dreamscopeapp.com

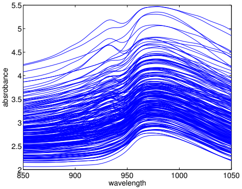
SaMUraiS : open source software for time-series

Available algorithms and Packages

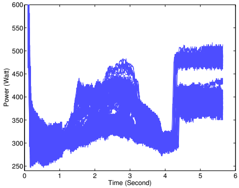
- **RHLP** : Regression with Hidden Logistic Process ▸ R software ▸ Matlab software
- **HMMR** : Hidden Markov Model Regression ▸ R software ▸ Matlab software
- **PWR** : Piece-Wise Regression ▸ R software ▸ Matlab software
- **MRHLP** : Multivariate RHLP ▸ R software ▸ Matlab software
- **MHMMR** : Multivariate HMMR ▸ R software ▸ Matlab software

Outline

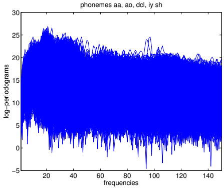
Functional data are increasingly frequent



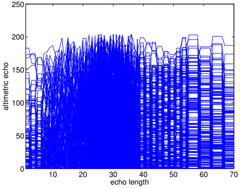
Tecator data



Railway switch curves



Phonemes curves



Satellite waveforms

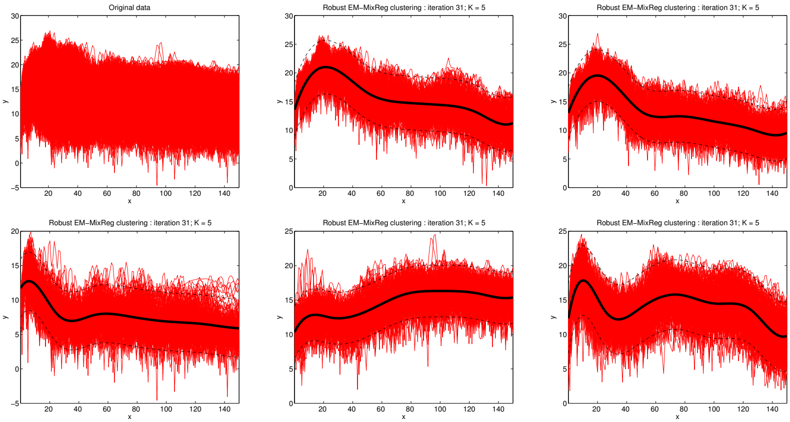
Statistical analysis of functional data

A broad literature :
[James and Hastie, 2001, James and Sugar, 2003]
[Ramsay and Silverman, 2005]
[Ferraty and Vieu, 2006]
[Ramsay et al., 2011]
[Bouveyron and Jacques, 2011]
[Samé et al., 2011]
[Delaigle et al., 2012]
[Jacques and Preda, 2014]
[Bouveyron et al., 2018]
[Qiao et al., 2018]
A review can be found in [Chamroukhi and Nguyen, 2018] ▸ pdf available here

- Functional regression
- Functional classification
- Functional clustering, including model-based
- Functional graphical models
- ...

Clustering of functional data

Phonemes data set [2] : $n = 1000$ log-periodograms for $m = 150$ frequencies



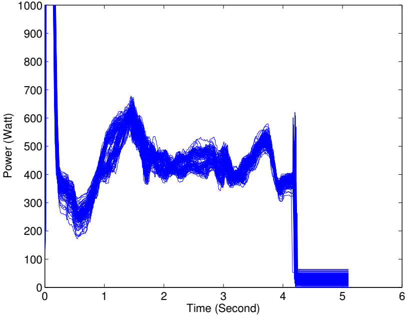
2. Data from `http://www.math.univ-toulouse.fr/staph/npfda/`, used in Ferraty and Vieu [2003]

Clustering of functional data

Clustering real curves of high-speed railway-switch operations

Data : $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters : operating state without/with possible defect

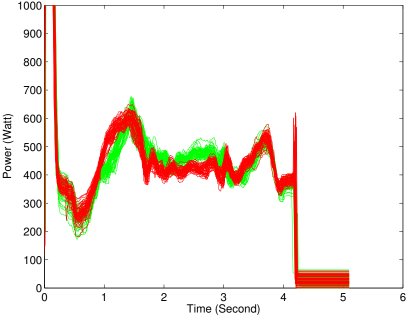


Clustering switch operations

Clustering real curves of high-speed railway-switch operations

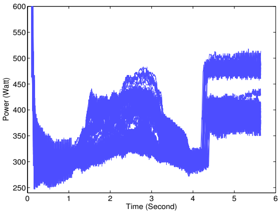
Data : $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters : operating state without/with possible defect

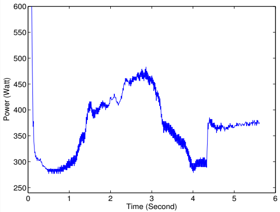


Functional data analysis context

Unsupervised analysis of heterogeneous curves with regime changes



Railway switch curves



An individual curve

Objectives

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

Outline

Functional data analysis context

Data

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

Objectives : exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes : [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 process (RHLP)
3. Regression model with Markov process (HMMR)

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 vector $\boldsymbol{\Psi} = (\alpha_1, \ldots, \alpha_{K-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T, \boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_K^T)^T$ with $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_{k1}^T, \ldots, \boldsymbol{\beta}_{kR_k}^T, \sigma_{k1}^2, \ldots, \sigma_{kR_k}^2)^T$ and $\boldsymbol{\xi}_k = (\xi_{k1}, \ldots, \xi_{k,R_k+1})^T$

Parameter estimation

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

EM-PWRM

- Maximize the observed-data log-likelihood :

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2\right)$$

- The complete-data log-likelihood

$$\log L_c(\boldsymbol{\Psi}, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} Z_{ik} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} Z_{ik} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2)$$

- The conditional expected complete-data log-likelihood

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(q)} \log \alpha_k + \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} \tau_{ik}^{(q)} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2)$$

EM-PWRM algorithm

E-step : Compute the $Q-$function

$\hookrightarrow$ Compute the posterior probability that the ith curve belongs to component k :

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(q)}) = \frac{\alpha_k^{(q)} f_k\left(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)}\right)}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} f_{k'}\left(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}_{k'}^{(q)}\right)}$$

M-step : Compute the update $\mathbf{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$

- $\alpha_k^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n}, \quad (k = 1, \ldots, K)$
- maximization w.r.t the piecewise regression parameters $\{\boldsymbol{\xi}_{kr}, \boldsymbol{\beta}_{kr}, \sigma_{kr}^2\}$ $\hookrightarrow$ a weighted piecewise regression problem $\hookrightarrow$ dynamic programming :

$$\boldsymbol{\beta}_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_{ir}^T \mathbf{X}_{ir}\Big]^{-1} \sum_{i=1}^{n} \mathbf{X}_{ir} \boldsymbol{y}_{ir}$$

$$\sigma_{kr}^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j \in I_{kr}^{(q)}} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \|\boldsymbol{y}_{ir} - \mathbf{X}_{ir} \boldsymbol{\beta}_{kr}^{(q+1)}\|^2$$

$\boldsymbol{y}_{ir}$ are the observations of segment r of the ith curve and $\mathbf{X}_{ir}$ its design matrix

Maximum classification likelihood estimation : CEM-PWRM

- Maximize the complete-data log-likelihood w.r.t $(\boldsymbol{\Psi}, \mathbf{z})$ simultaneously
- C-step : Bayes' optimal allocation rule : $\hat{z}_i = \arg\max_{1 \leq k \leq K} \tau_{ik}(\hat{\boldsymbol{\Psi}})$

CEM-PWRM is equivalent to the K-means-like algorithm of Hébrail et al. [2010] :

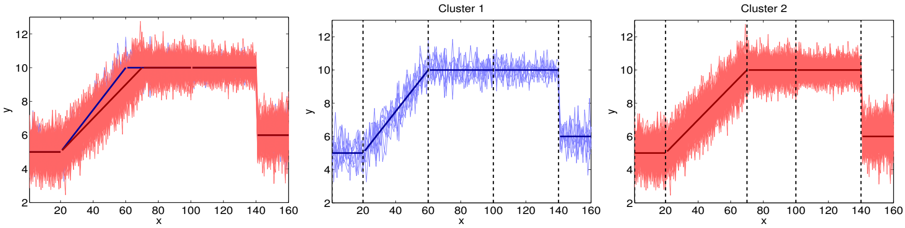
$$\log L_c(\mathbf{z}, \boldsymbol{\Psi}) \propto \mathcal{J}\big(\mathbf{z}, \{\mu_{kr}, I_{kr}\}\big) = \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{i|Z_i=k} \sum_{j \in I_{kr}} \big(y_{ij} - \mu_{kr}\big)^2$$

if the following conditions hold :

- $\alpha_k = \frac{1}{K}\ \forall K$ (identical mixing proportions) ;
- $\sigma^2_{kr} = \sigma^2\ \forall r$ and $\forall k$; (isotropic and homoskedastic model) ;
- μ_{kr} : piecewise *constant* regime approximation

- Curve clustering : $\hat{z}_i = \arg\max_k \tau_{ik}(\hat{\boldsymbol{\Psi}})$ with $\tau_{ik}(\hat{\boldsymbol{\Psi}}) = \mathbb{P}(Z_i|\boldsymbol{x}_i, \boldsymbol{y}_i; \hat{\boldsymbol{\Psi}})$
- Model selection : Application of BIC, ICL
- Complexity in $\mathcal{O}(I_{\mathsf{EM}} K R n m^2 p^3)$: Significant computational load for large

Simulation results



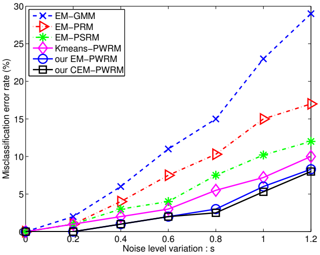
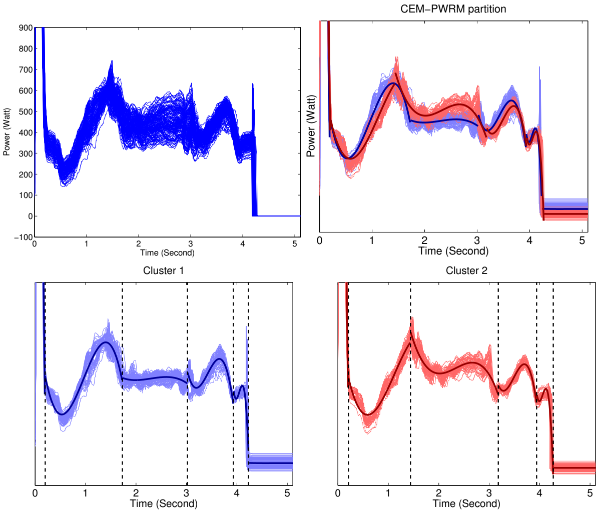


FIGURE – Misclassification error rate versus the noise level variation.

Application to switch operation curves

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

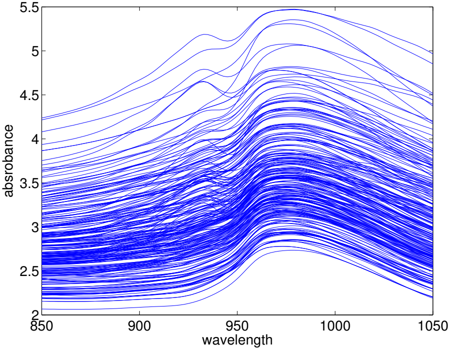
Each curve is composed of $R = 6$ electromechanical phases (regimes)



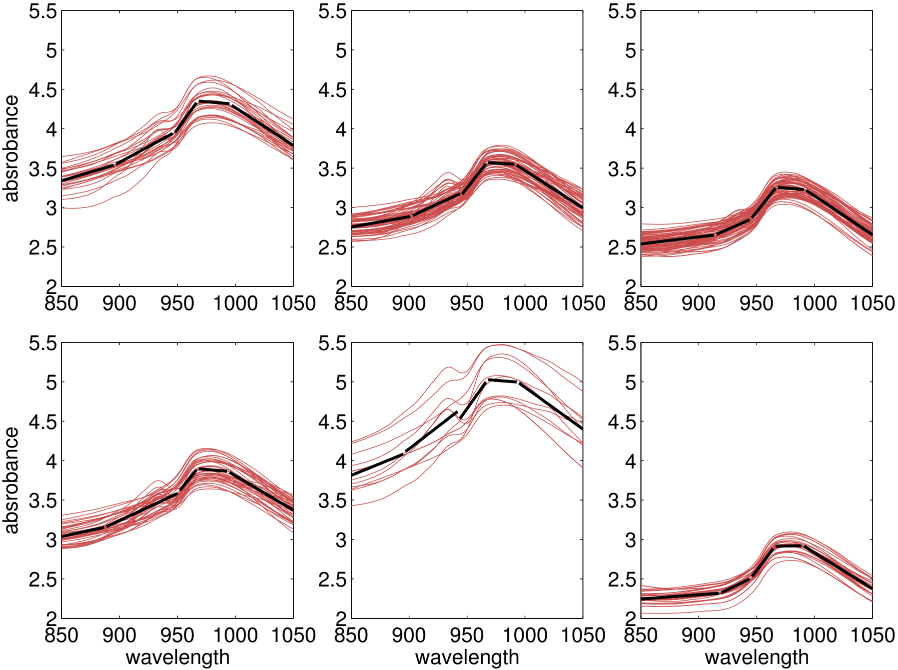
Application to Tecator data

The Tecator data set [3] contains $n = 240$ spectra with $m = 100$ observations for each spectrum

Data considered in the same setting as in Hébrail et al. [2010] (six clusters, each cluster is approximated by five linear segments $(R = 5, p = 1)$)



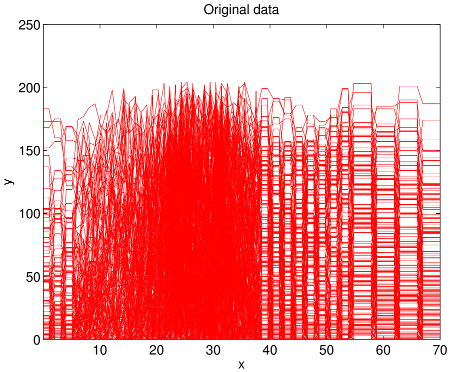
3. Tecator data are available at `http://lib.stat.cmu.edu/datasets/tecator`.



absrobance
absrobance
wavelength
wavelength
wavelength

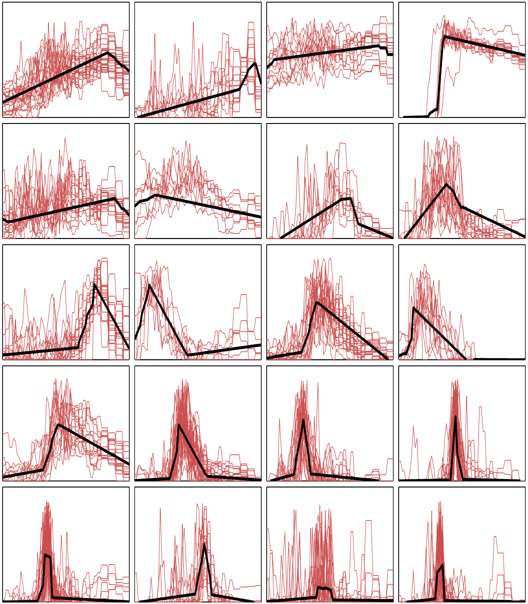
Topex/Poseidon satellite data

The Topex/Poseidon radar satellite data [4] 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].



4. Satellite data are available at
`http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html`.

CEM-PWRM clustering



Summary

- Probabilistic approach to the simultaneous curve clustering and optimal segmentation
- Two algorithms : EM-PWRM and CEM-PWRM
- CEM-PWRM is a probabilistic-based version of the K-means-like algorithm Hébrail et al. [2010]

- If the aim is density estimation, the EM version is suggested (CEM provides biased estimators but is well-tailored to the segmentation/clustering end)
- For continuous functions the PWRM in its current formulation, may lead to discontinuities between segments for the piecewise approximation.
- This may be avoided by posterior interpolation as in Hébrail et al. [2010].
- May lead to significant computational load especially for large time series. However, for quite reasonable dimensions, the algorithms remain usable

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

MLE estimation via the EM algorithm

- The observed-data log-likelihood

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\big(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\big)$$

- The complete-data log-likelihood :

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} Z_{ik} H_{ijr} \log \left[\pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\right)\right]$$

- The conditional expected complete-data log-likelihood

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) &= \mathbb{E}\left[\log L_c(\boldsymbol{\Psi}) \middle| \mathcal{D}; \boldsymbol{\Psi}^{(q)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i,j} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \tau_{ik}^{(q)} \gamma_{ijr}^{(q)} \log \left[\pi_{kr}(x_j; \mathbf{w}_k) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2\right)\right]. \end{aligned}$$

EM-MixRHLP algorithm

E-step

- The posterior cluster memberships :

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)}) = \frac{\alpha_k^{(q)} f(\boldsymbol{y}_i|Z_i = k, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)})}{\sum_{k\prime=1}^{K} \alpha_{k'}^{(q)} f(\boldsymbol{y}_i|Z_i = k\prime, \boldsymbol{x}_i; \boldsymbol{\Psi}_{k\prime}^{(q)})}$$

- the posterior regime memberships :

$$\gamma_{ijr}^{(q)} = \mathbb{P}(H_{ij} = r|Z_i = k, y_{ij}, t_j; \boldsymbol{\Psi}_k^{(q)}) = \frac{\pi_{kr}(x_j; \mathbf{w}_k^{(q)})\mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^{T(q)}\boldsymbol{x}_j, \sigma_{kr}^{2(q)})}{\sum_{r\prime=1}^{R_k} \pi_{kr\prime}(x_j; \mathbf{w}_k^{(q)})\mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr\prime}^{T(q)}\boldsymbol{x}_j, \sigma_{kr\prime}^{2(q)})}$$

Computed directly (i.e, without a forward-backward recursion as in the Markovian model).

M-step of the EM-MixRHLP

M-step : calculate the update $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$.

- Mixing proportions update : standard

$$\alpha_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(q)}, \quad (k = 1, \ldots, K).$$

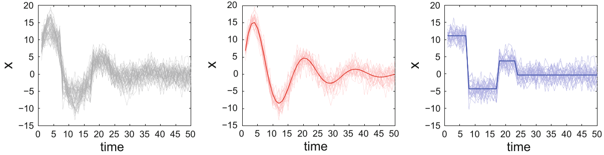
- Regression parameters update : Analytic weighted least-squares problems

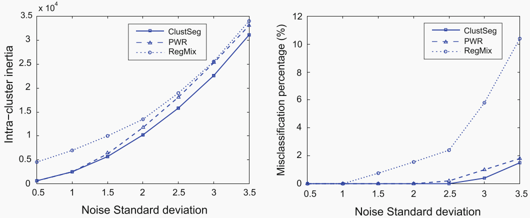
$$\begin{aligned}
\boldsymbol{\beta}_{kr}^{(q+1)} &= \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \mathbf{X}_i\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \boldsymbol{y}_i, \\
{\sigma_{kr}^2}^{(q+1)} &= \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \| \sqrt{\mathbf{W}_{ikr}^{(q)}} (\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{\beta}_{kr}^{(q+1)}) \|^2}{\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathrm{trace}(\mathbf{W}_{ikr}^{(q)})},
\end{aligned}$$

where $\mathbf{W}_{ikr}^{(q)} = \mathrm{diag}(\gamma_{ijr}^{(q)}; j = 1, \ldots, m_i)$.

- Maximization w.r.t the logistic processes' parameters $\{\mathbf{w}_k\}$: solving multinomial logistic regression problems $\Rightarrow$ IRLS
- $\hookrightarrow$ EM-MixRHLP has complexity in $\mathcal{O}(I_{\mathsf{EM}} I_{\mathsf{IRLS}} K R^3 n m p^3)$ (K-means like algo. for PWR is in $\mathcal{O}(I_{\mathsf{KM}} K R n m^2 p^3)$ $\hookrightarrow$ computationally attractive for large m with moderate value of R.

EM-MixRHLP clustering of simulated data

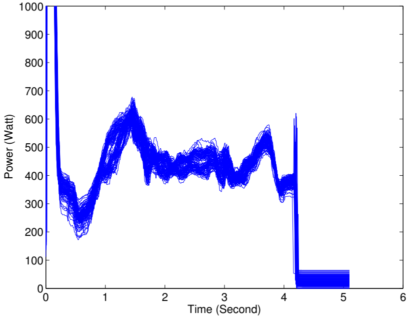




Clustering switch operations

Clustering real curves of switch operations The data set contains 115 curves of $R = 6$ operations electromechanical process

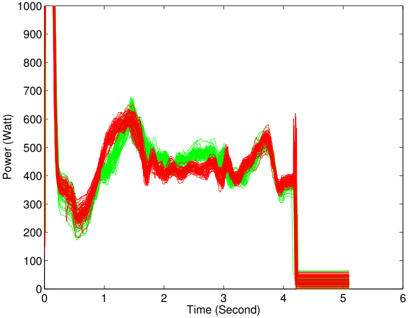
$K = 2$ clusters : operating state without/with possible defect



Clustering switch operations

Clustering real curves of switch operations The data set contains 115 curves of $R = 6$ operations electromechanical process

$K = 2$ clusters : operating state without/with possible defect



Functional discriminant analysis

Supervised classification context

- Data : a training set of labeled functions $((\boldsymbol{x}_1, \boldsymbol{y}_1, c_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n, c_n))$ where $c_i \in \{1, \ldots, G\}$ is the class label of the ith curve
- Problem : predict the class label c_i for a new unlabeled function $(\boldsymbol{x}_i, \boldsymbol{y}_i)$

Tool : Discriminant analysis

Use the Bayes' allocation rule

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

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

- Homogeneous classes : Functional Linear Discriminant Analysis [8]
- Dispersed classes : Functional Mixture Discriminant Analysis [5]

Summary

- A full generative model for curve clustering and segmentation
- The segmentation is smoothly controlled by logistic functions
- An alternative to the previously described mixture of piecewise regressions
- more advantageous compared to approaches involving dynamic programming namely when using piecewise regression especially for large samples.
- Could be extended to the multivariate case without a major effort

FLaMingoS : open source softw. for functional data

FLaMingoS : **F**unctional **L**atent dat**A** **M**odels for cluster**ING** heterogene**O**us time-**S**eries

R software Matlab software

FLaMingoS : open source softw. for functional data

Available algorithms and Packages

- **mixRHLP** : Mixture of Regressions with Hidden Logistic Processes ▸ R software ▸ Matlab software
- **mixHMM** : Mixtures of Hidden Markov Models ▸ R software ▸ Matlab software
- **mixHMMR** : Mixtures of Hidden Markov Model Regressions ▸ R software ▸ Matlab software
- **PWRM** : Piece-Wise Regression Mixture ▸ R software ▸ Matlab software

Coming soon : Learning of regression mixtures with unknown number of components

unsupPRM ▸ R software ▸ Matlab software
unsupSRM ▸ R software ▸ Matlab software
unsupBSRM ▸ R software ▸ Matlab software

Outline

Mixture-of-Experts (MoE) modeling framework

- Data : an observed i.i.d sample of the pair $(\boldsymbol{X}, Y)$ where the response $Y \in \mathbb{R}$ for the vector of predictors $\boldsymbol{X} \in \mathbb{R}^p$ is governed by a hidden categorical variable Z

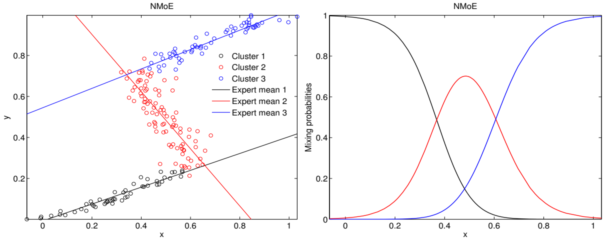
  $z_i \in [K]$ is the expert label for $(\boldsymbol{X}_i, Y_i)$
- Mixture of experts (MoE) [Jacobs et al., 1991, Jordan and Jacobs, 1994] :

$$f(y|\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K} \underbrace{\pi_k(\boldsymbol{x};\mathbf{w})}_{\text{Gating network}} \underbrace{f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)}_{\text{Expert Network}}$$

- Gating network (e.g softmax) : $\pi_k(\boldsymbol{x};\mathbf{w}) = \frac{\exp(w_{k0}+\boldsymbol{w}_k^T\boldsymbol{x})}{1+\sum_{\ell=1}^{K-1}\exp(w_{\ell 0}+\boldsymbol{w}_\ell^T\boldsymbol{x})}$
- Experts network (e.g Gaussian regressors) : $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k) = \phi\big(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big)$ with parametric (non-)linear regression functions $\mu(\boldsymbol{x};\boldsymbol{\beta}_k)$
- parameter vector $\boldsymbol{\Psi} = (\mathbf{w}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$

  $\hookrightarrow$ For a review, see Nguyen and Chamroukhi [2018] ▸ pdf available here

Illustration



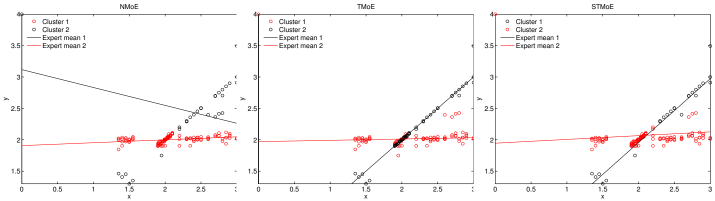


FIGURE – Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

For a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations, the use of normal experts may be unsuitable and can unduly affect the fit

Objectives

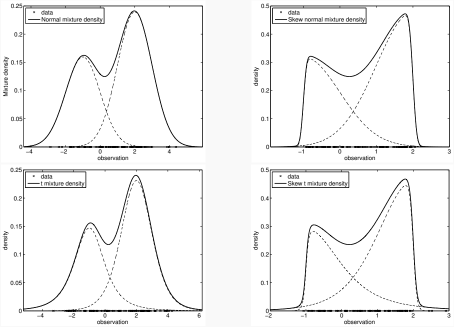
- Overcome these imitations of MoE modeling with the normal distribution.
- We proposed three non-normal derivations including two robust mixture of experts (MoE) models. $\hookrightarrow$ suitable to accommodate data which exhibit additional features such as skewness, heavy-tails and which may be affected by atypical data [Chamroukhi, 2017, 2016a,b]

Non-normal mixtures of experts

Non-normal mixtures of experts (NNMoE)

1. the t MoE (TMoE) (Robustness, heavy tails) [11]
2. the skew-normal MoE (SNMoE) (skewness) [14]
3. the skew-t MoE (STMoE) (skewness, robustness, heavy tails) [15]

Non-normal mixtures



$\pi_k = [0.4, 0.6]$, $\mu_k = [-1, 2]$; $\sigma_k = [1, 1]$; $\nu_k = [3, 7]$; $\lambda_k = [14, -12]$;

The skew-normal mixture of experts model

- The SNMoE is defined as

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

where each expert component k has indeed a skew-normal distribution, whose density is defined by (1).

- *The skew-normal distribution* [Azzalini, 1985, 1986] with location $\mu \in \mathbb{R}$, scale $\sigma^2 \in (0, \infty)$ and skewness parameter $\lambda \in \mathbb{R}$ has density

$$\mathsf{SN}(y; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi(\frac{y - \mu}{\sigma}) \Phi\left(\lambda(\frac{y - \mu}{\sigma})\right)$$

$\phi(.)$ and $\Phi(.)$ denote the pdf and the cdf of the standard normal distribution.

- The parameter vector is $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ with $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$ the parameter vector for the kth skewed-normal expert component.
- It is obvious to see that if the skewness parameter $\lambda_k = 0$ for each k, the SNMoE model reduces to the NMoE model.

The skew-normal mixture of experts model

- **Stochastic representation of the SNMoE :** A random variable Y_i is said to follow the SNMoE model if it has the following representation :

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \delta_{z_i}\sigma_{z_i}|U_i| + \sqrt{1-\delta^2_{z_i}}\,\sigma_{z_i}E_i.$$

  where U and E be independent random variables following the standard normal distribution $\mathsf{N}(0,1)$ with pdf $\phi(.)$, $|U|$ denotes the magnitude of U and $\delta_{z_i} = \frac{\lambda_{z_i}}{\sqrt{1+\lambda^2_{z_i}}}$ where $Z_i \in \{1,\ldots,K\}$ is a categorical variable following the multinomial distribution

$$Z_i|\boldsymbol{r}_i \sim \mathsf{Mult}(1; \pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha}))$$

  where each $\pi_{z_i}(\boldsymbol{r}_i;\boldsymbol{\alpha}) = \mathbb{P}(Z_i = z_i|\boldsymbol{r}_i)$ is given by the logistic function.

- **Hierarchical representation of the SNMoE**

$$\begin{aligned}
Y_i|u_i, Z_{ik}=1, \boldsymbol{x}_i &\sim \mathsf{N}\Big(\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k) + \delta_k|u_i|, (1-\delta_k^2)\sigma_k^2\Big),\\
U_i|Z_{ik}=1 &\sim \mathsf{N}(0,\sigma_k^2),\\
\boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \mathsf{Mult}\left(1;\pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha})\right)
\end{aligned}$$

  where Z_{ik} are the binary latent component-indicators such that $Z_{ik}=1$ iff $Z_i = k$, $\boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{iK})$ and $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$

Maximum likelihood parameter estimation

- Given an observed i.i.d sample of n observations $\{(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)\}_{i=1}^n$, the parameter vector $\boldsymbol{\Psi}$ of the SNMoE model can be estimated by maximizing the observed-data log-likelihood :

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \mathsf{SN}\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\right).$$

- $\Rightarrow$ A dedicated Expectation Conditional Maximization (ECM) algorithm
- The ECM algorithm [Meng and Rubin, 1993] is an EM variant that mainly aims at addressing the optimization problem in the M-step of the EM algorithm. In ECM, the M-step is performed by several conditional maximization (CM) steps by dividing the parameter space into sub-spaces. The parameter vector updates are then performed sequentially, one coordinate block after another in each sub-space.
- This is also the generative process for sampling data according to the SNMoE model

MLE via the ECM algorithm

- The complete-data log-likelihood of $\boldsymbol{\Psi}$, where the complete-data are $\{y_i, z_i, u_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n$, is given by :

$$\log L_c(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \log L_c(\boldsymbol{\Psi}_k),$$

with

$$\log L_c(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$\log L_c(\boldsymbol{\Psi}_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ u_i}{(1-\delta_k^2)\sigma_k} - \frac{u_i^2}{2(1-\delta_k^2)\sigma_k^2} \Big],$$

where $d_{ik} = \frac{y_i - \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k)}{\sigma_k}$.

ECM for the SNMoE : E-Step

E-Step calculates the Q-function

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)})=\mathbb{E}\left[\log L_c(\boldsymbol{\Psi})|\{y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\}_{i=1}^n;\boldsymbol{\Psi}^{(m)}\right]=Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})+\sum_{k=1}^K Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$$

with

$$\begin{aligned}
Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^n\sum_{k=1}^K \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),\\
Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^n \tau_{ik}^{(m)}\Bigg[-\log(2\pi)-\log(\sigma_k^2)-\frac{1}{2}\log(1-\delta_k^2)\\
&\quad +\frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k}-\frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2}-\frac{d_{ik}^2}{2(1-\delta_k^2)}\Bigg]
\end{aligned}$$

where the required conditional expectations (analytic) are given by :

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

CM-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}} Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$. does not exist in closed form (Unlike in skew-normal (regression) mixtures)

The Iteratively Reweighted Least Squares (IRLS) algorithm :

$$\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l)} - \left[\frac{\partial^2 Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\right]^{-1}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \frac{\partial Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}}$$

Then, for $k = 1 \ldots, K$,

CM-Step 2 Calculate $\boldsymbol{\beta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$

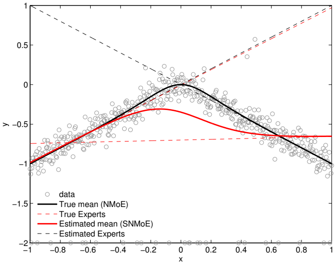
$$\boldsymbol{\beta}_k^{(m+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(y_i - \delta_k^{(m)} e_{1,ik}^{(m)}\right) \boldsymbol{x}_i.$$

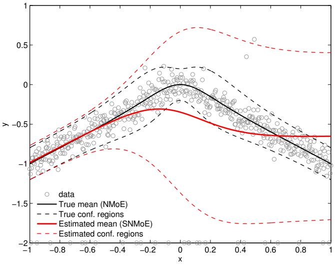
CM-Step 3 : Calculate $\sigma_k^{2\,(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$

$$\sigma_k^{2\,(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left[\left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)} e_{1,ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i) + e_{2,ik}^{(m)}\right]}{2\left(1 - \delta_k^{2\,(m)}\right) \sum_{i=1}^{n} \tau_{ik}^{(m)}}.$$

CM-Step 4 Calculate $\lambda_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$: Solution of :
$\sigma_k^{2\,(m+1)} \delta_k (1 - \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} + (1 + \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i)\, e_{1,ik}^{(m)}$
$- \delta_k \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[e_{2,ik}^{(m)} + \left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i\right)^2\right] = 0 \cdot$ root finding (Brent's method [Brent, 1973]).

.





- SNMoE model is tailored to model the skewness in the data, it may be not adapted to handle data containing groups or a group with heavy-tailed distribution.
- The SNMoE as NMoE may thus be affected by outliers.
- $\Rightarrow$ Handle the problem of sensitivity of normal mixture of experts to outliers and heavy tails.
  $\hookrightarrow$ robust mixture of experts modeling using the t distribution.

The t mixture of experts model

- The proposed t mixture of experts model extends the t mixture model, first proposed by Mclachlan and Peel [1998], Peel and Mclachlan [2000] for multivariate data, as well as the regression mixture model using the t-distribution as in [Bai et al., 2012, Wei, 2012, Ingrassia et al., 2012] to the MoE framework.
- A K-component TMoE model is defined by :

$$f(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\Psi}) \quad = \quad \sum_{k=1}^{K}\pi_k(\boldsymbol{r};\boldsymbol{\alpha})\ t_{\nu_k}\left(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2,\nu_k\right).$$

  The t-distribution with location $\mu \in \mathbb{R}$, scale $\sigma^2 \in (0,\infty)$ and degrees of freedom $\nu \in (0,\infty)$ has the probability density function

$$t_{\nu_k}(y;\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{d_y^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

  where $d_y^2 = \left(\frac{y-\mu}{\sigma}\right)^2$ denotes the squared Mahalanobis distance
- The parameter vector of the TMoE model is given by $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_{K-1}^T,\boldsymbol{\Psi}_1^T,\ldots,\boldsymbol{\Psi}_K^T)^T$ where $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T,\sigma_k^2,\nu_k)^T$
- When the robustness parameter $\nu_k \to \infty$ for each experts k, the TMoE model approaches the NMoE model

The t mixture of experts model

- **Stochastic representation for the TMoE** Let $E \sim \phi(.)$. Suppose that, conditional on the hidden variable $Z_i = z_i$, a random variable W_i is distributed as $\text{Gamma}(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2})$. Then, given the covariates $(\boldsymbol{x}_i, \boldsymbol{r}_i)$, a random variable Y_i is said to follow the TMoE model if

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_{z_i}}},$$

where the categorical variable $Z_i | \boldsymbol{r}_i$ is multinomial

- **Hierarchical representation of the TMoE model**

$$\begin{aligned}
Y_i | w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \mathsf{N}\left(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \frac{\sigma_k^2}{w_i}\right), \\
W_i | Z_{ik} = 1 &\sim \text{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\
\boldsymbol{Z}_i | \boldsymbol{r}_i &\sim \text{Mult}\left(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\right).
\end{aligned}$$

- This hierarchical representation involves the hidden variables Z_i and W_i facilitates the ML inference of model parameters $\boldsymbol{\Psi}$ via E(C)M.

MLE of the TMoE model

- Given an i.i.d sample of n observations, $\boldsymbol{\Psi}$ can be estimated by maximizing the observed-data log-likelihood :

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) t\nu_k \left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right).$$

- $\Rightarrow$ EM algorithm and then describe an ECM extension
- The complete data consist of the responses $(y_1, \ldots, y_n)$ and their corresponding predictors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$, as well as the latent variables $(w_1, \ldots, w_n)$ (in the hierarchical representation) and the latent labels $(z_1, \ldots, z_n)$.

MLE of the TMoE model

- $\Rightarrow$ The complete-data log-likelihood of $\boldsymbol{\Psi}$ is given by :

$$\log L_c(\boldsymbol{\Psi}) \quad = \quad \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \left[\log L_{2c}(\boldsymbol{\Psi}_k) + \log L_{3c}(\nu_k) \right],$$

where

$$\log L_{1c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$\log L_{2c}(\boldsymbol{\Psi}_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2} w_i d_{ik}^2 \Big],$$

$$\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\log \Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right) \log\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2} - 1\right) \log(w_i) - \left(\frac{\nu_k}{2}\right) w_i \Big].$$

MLE of the TMoE model : E-Step

E-Step Calculate the Q-function :

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K}\Big[Q_2(\boldsymbol{\theta}_k,\boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k,\boldsymbol{\Psi}^{(m)})\Big],$$

where $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T,\sigma_k^2)^T$ and

$$Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)}\log\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),$$

$$Q_2(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}\Big[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{1}{2}\,w_{ik}^{(m)}d_{ik}^2\Big].$$

$$Q_3(\nu_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}\Big[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right)\,w_{ik}^{(m)} + \left(\frac{\nu_k}{2}-1\right)e_{1,ik}^{(m)}\Big]$$

$\rightarrow$ requires the following conditional expectations (analytic) :

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

MLE of the TMoE model : M-Step

M-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\alpha}$. $\Rightarrow$ Iteratively via IRLS (92) as for the mixture of SNMoE.

M-Step 2 Calculate $\boldsymbol{\theta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\theta}_k$

$$\begin{aligned}
\boldsymbol{\beta}_k^{(m+1)} &= \Big[\sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} y_i \boldsymbol{x}_i, \\
\sigma_k^{2\,(m+1)} &= \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i\right)^2 .
\end{aligned}$$

M-Step 3 Calculate $\nu_k^{(m+1)}$ by maximizing $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$ w.r.t ν_k
$\Rightarrow$ iteratively solve the following equation in ν_k :

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \big(\log(w_{ik}^{(m)}) - w_{ik}^{(m)}\big)}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} + \psi\Big(\frac{\nu_k^{(m)} + 1}{2}\Big) - \log\Big(\frac{\nu_k^{(m)} + 1}{2}\Big) = 0.$$

This scalar non-linear equation can be solved with a root finding algorithm, such as Brent's method [Brent, 1973].

The skew t mixture of experts (STMoE) model

- A K-component mixture of skew t experts (STMoE) is defined by :

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

- kth expert : has a skew t distribution [Azzalini and Capitanio, 2003] :

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

  The skew t mixture of experts (STMoE) model extends the univariate skew t mixture model Lin et al. [2007], to the MoE framework.

Model characteristics

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

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

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

$\hookrightarrow$ The STMoE is flexible as it generalizes the previously described models

Representation of the STMoE model

- **Stochastic representation** Suppose that conditional on a Multinomial categorical variable Z_i, E_i and W_i are independent univariate random variables such that $E_i \sim \mathsf{SN}(\lambda_{z_i})$ and $W_i \sim \mathsf{Gamma}(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2})$, and $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$ are given covariates. A variable Y_i having the following representation :

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_i}}$$

  is said to follow the STMoE distribution

- **Hierarchical representation**

$$\begin{aligned}
Y_i | u_i, w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \mathsf{N}\Big(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k) + \delta_k |u_i|, \frac{1-\delta_k^2}{w_i}\sigma_k^2\Big),\\
U_i | w_i, Z_{ik} = 1 &\sim \mathsf{N}\Big(0, \frac{\sigma_k^2}{w_i}\Big),\\
W_i | Z_{ik} = 1 &\sim \mathsf{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)\\
\boldsymbol{Z}_i | \boldsymbol{r}_i &\sim \mathsf{Mult}\big(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\big).
\end{aligned}$$

  The variables U_i and W_i are hidden in this hierarchical representation

MLE via the ECM algorithm

- Maximize the observed-data log-likelihood :

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha})\mathsf{ST}(y;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k),\sigma_k^2,\lambda_k,\nu_k)\,.$$

- $\Rightarrow$ This is performed iteratively by a dedicated ECM algorithm.
- The complete-data log-likelihood :

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K}\big[\log L_{2c}(\boldsymbol{\theta}_k) + \log L_{3c}(\nu_k)\big];\;\; \boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T,\sigma_k^2,\lambda_k)^T$$

$$\begin{aligned}
\log L_{1c}(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),\\
\log L_{2c}(\boldsymbol{\theta}_k) &= \sum_{i=1}^{n} Z_{ik}\Big[-\log(2\pi)-\log(\sigma_k^2)-\frac{1}{2}\log(1-\delta_k^2)-\frac{w_i\ d_{ik}^2}{2(1-\delta_k^2)}+\frac{w_i\ u_i\ \delta_k\ d_{ik}}{(1-\delta_k^2)\sigma_k}-\frac{w_i\ u_i^2}{2(1-\delta_k^2)\sigma_k^2}\Big]\\
\log L_{3c}(\nu_k) &= \sum_{i=1}^{n} Z_{ik}\Big[-\log\Gamma\left(\frac{\nu_k}{2}\right)+\left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right)+\left(\frac{\nu_k}{2}\right)\log(w_i)-\left(\frac{\nu_k}{2}\right)w_i\Big]\,.
\end{aligned}$$

MLE via the ECM algorithm : E-Step

- **E-Step** Calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood , given the observed data $\{y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^{n}$ and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$ given by :

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K} \Big[Q_2(\boldsymbol{\theta}_k, \boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k, \boldsymbol{\Psi}^{(m)}) \Big],$$

where

$$Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \Bigg[-\log(2\pi\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{w_{ik}^{(m)}\ d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k} - \frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2} \Bigg],$$

$$Q_3(\nu_k; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right)\ w_{ik}^{(m)} + \left(\frac{\nu_k}{2}\right) e_{3,ik}^{(m)} \right].$$

MLE via the ECM algorithm : E-Step

- $\Rightarrow$ The E-Step requires the following conditional expectations :

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

- These conditional expectations are calculated analytically except $e_{3,ik}^{(m)}$ for which I adopted a one-step-late (OSL) approach as in Lee and McLachlan [2014], rather than using a Monte Carlo approximation as in Lin et al. [2007].
- I also mention that, for the multivariate skew t mixture models, recently Lee and McLachlan [2015] presented a series-based truncation approach, which exploits an exact representation of this conditional expectation and which can also be used here.

MLE via the ECM algorithm : M-Step

- **CM-Step 1** update the mixing parameters $\boldsymbol{\alpha}^{(m+1)}$ by maximizing the function $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$ by using IRLS. Then, for $k = 1\ldots,K$,
- **CM-Step 2** Update the regression params $(\boldsymbol{\beta}_k^{T(m+1)}, \sigma_k^{2(m+1)})$:

$$\boldsymbol{\beta}_k^{(m+1)}=\Big[\sum_{i=1}^{n}\tau_{ik}^{(q)}w_{ik}^{(m)}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}\sum_{i=1}^{n}\tau_{ik}^{(q)}\left(w_{ik}^{(m)}y_i-\boldsymbol{e}_{1,ik}^{(m)}\delta_k^{(m+1)}\right)\boldsymbol{x}_i,$$

$$\sigma_k^{2(m+1)}=\frac{\sum_{i=1}^{n}\tau_{ik}^{(m)}\left[w_{ik}^{(m)}\left(\boldsymbol{y}_i-\boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i\right)^2-2\delta_k^{(m+1)}\boldsymbol{e}_{1,ik}^{(m)}(y_i-\boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i)+\boldsymbol{e}_{2,ik}^{(m)}\right]}{2\left(1-\delta_k^{2(m)}\right)\sum_{i=1}^{n}\tau_{ik}^{(m)}}$$

- **CM-Step 3** Update the skewness parameters λ_k by solving the following equation :

$$\delta_k(1-\delta_k^2)\sum_{i=1}^{n}\tau_{ik}^{(m)}+(1+\delta_k^2)\sum_{i=1}^{n}\tau_{ik}^{(m)}\frac{d_{ik}^{(m+1)}e_{1,ik}^{(m)}}{\sigma_k^{(m+1)}}-\delta_k\sum_{i=1}^{n}\tau_{ik}^{(m)}\Big[w_{ik}^{(m)}d_{ik}^{2\,(m+1)}+\frac{e_{2,ik}^{(m)}}{\sigma_k^{2\,(m+1)}}\Big]=0\,.$$

- **CM-Step 4** Update the degree of freedom ν_k by solving of the following equation :

$$-\psi\left(\frac{\nu_k}{2}\right)+\log\left(\frac{\nu_k}{2}\right)+1+\frac{\sum_{i=1}^{n}\tau_{ik}^{(m)}\left(\boldsymbol{e}_{3,ik}^{(m)}-w_{ik}^{(m)}\right)}{\sum_{i=1}^{n}\tau_{ik}^{(m)}}=0.$$

Prediction, clustering and model selection

- **Prediction** Predicted response : $\hat{y} = \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})$ with

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

$$\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}_n) \big[(\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x}))^2 + \mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x}) \big] - \big[\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})\big]$$

  where $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x})$ and $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x})$ are respectively the component-specific (expert) means and variances.

- **Clustering of regression data** Calculate the cluster label as

$$\hat{z}_i = \arg\max_{k=1}^{K} \mathbb{E}[Z_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}] = \arg\max_{k=1}^{K} \frac{\pi_k(\boldsymbol{r}; \hat{\boldsymbol{\Psi}}) f_k\Big(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}\Big)}{\sum_{k'=1}^{K} \pi_{k'}(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}) f_{k'}\Big(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}_{k'}\Big)}$$

- **Model selection** The value of (K, p) can be computed by using BIC, ICL
  Number of free parameters :
  $\eta_{\boldsymbol{\Psi}} = K(p + 4) - 2$ for the NMoE model,
  $\eta_{\boldsymbol{\Psi}} = K(p + 5) - 2$ for both the SNMoE and the TMoE models,
  $\eta_{\boldsymbol{\Psi}} = K(p + 6) - 2$ for the STMoE model.

Illustation on Bishop's data set

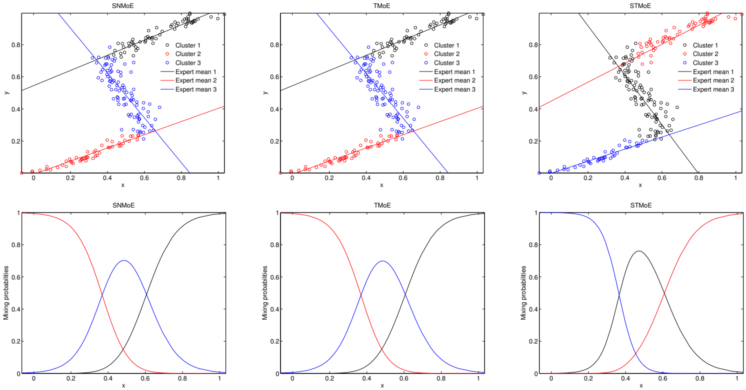


FIGURE – Fitting the the non-normal mixture of experts models (SNMoE, TNMoE, STMoE) to the toy data set analyzed in Bishop and Svensén [2003] : $n = 250$ values of input variables x_i generated uniformly in $(0, 1)$ and output variables y_i generated as $y_i = x_i + 0.3\sin(2\pi x_i) + \epsilon_i$, with ϵ_i drawn from a zero mean Normal distribution with standard deviation 0.05.

Robustness of the NNMoE

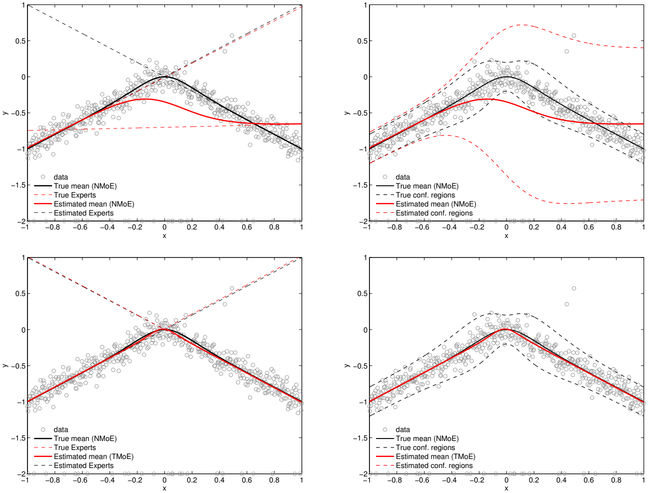


FIGURE – Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (bottom).

Robustness of the NNMoE

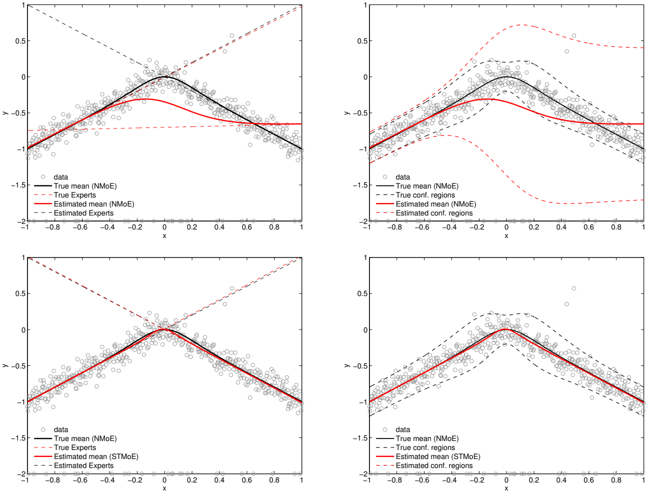


FIGURE – Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), STMoE fit (bottom).

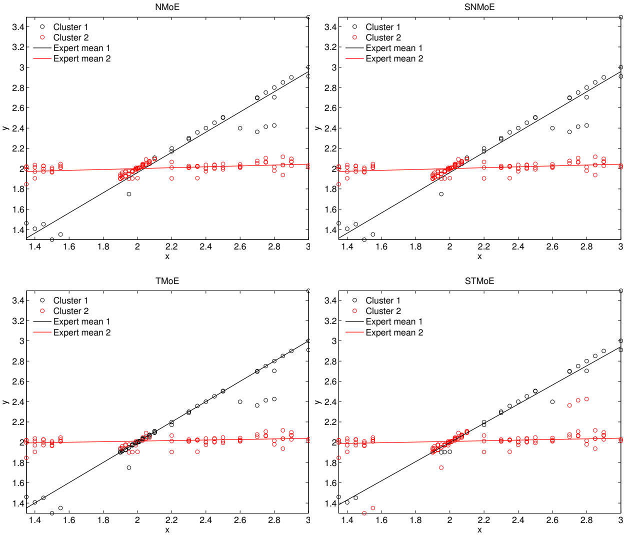


FIGURE – Fitting the MoE models to the tone data set studied by Bai et al. [2012] and Song et al. [2014] by using robust regression mixture models based on, respectively, the t distribution and the Laplace distribution : $n = 150$ pairs of "tuned" predictors (x), and their corresponding "strech ratio" responses (y).

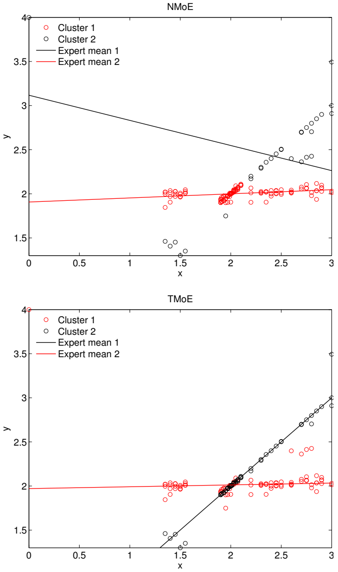


FIGURE – Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

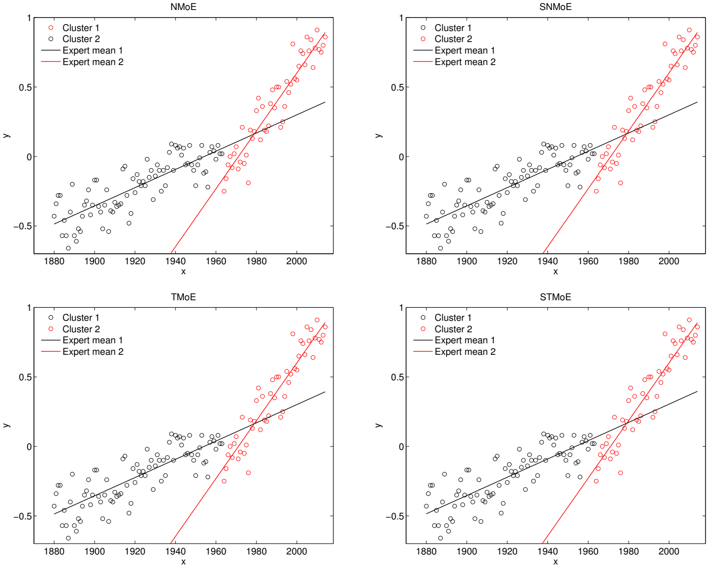
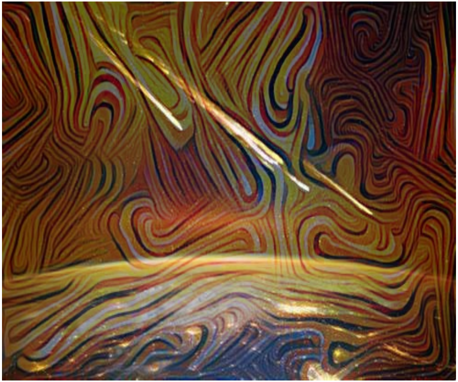


FIGURE – Fitting the MoLE models to the temperature anomalies data set.

MEteorits : open-source soft. for mixtures-of-experts



MEteorits : **M**ixtures-of-**E**xper**T**s mod**E**ling for c**O**mplex and non-no**R**mal d**I**s**T**ribution**S**

R software Matlab software

MEteorits : open-source soft. for mixtures-of-experts

Available algorithms and Packages

- **NMoE** : Normal Mixture-of-Experts R software Matlab software
- **SNMoE** : Skew-Normal Mixture-of-Experts R software Matlab software
- **tMoE** : Robust modeling of mixture-of-experts using the t-distribution R software Matlab software
- **StMoE** : Skew-t Mixture-of-Experts R software Matlab software

High-dimensional Mixtures-of-Experts [Huynh and Chamroukhi, 2019]

Estimation and Feature Selection in Mixtures of Generalized Linear Experts Models

- **prEMME** : proximal Newton EM for estimation and feature selection in high-dimensional Mixtures-of-Experts [Huynh and Chamroukhi, 2019] R software
- Expert models : Poisson R software Logistic R software Gaussian R software

Coming soon : MoE for functional data (functional predictors)

- **FunME** Matlab software

MoE models in high-dimension

Maximum Likelihood Estimation via EM [Dempster et al., 1977, Jacobs et al., 1991]

- MLE : $\boldsymbol{\Psi}$ is commonly estimated by maximizing the observed-data log-likelihood : $\widehat{\boldsymbol{\Psi}}_n \in \arg\max_{\boldsymbol{\Psi}\in\Theta} L(\boldsymbol{\Psi})$ with $L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}_k)$

Regularized MLE of the MoE [Khalili, 2010] [Huynh and Chamroukhi, 2019]

$\boldsymbol{\Psi}$ is estimated by maximizing a penalized observed-data log-likelihood :

$$\widehat{\boldsymbol{\Psi}}_n \in \arg\max_{\boldsymbol{\Psi}\in\Theta} L(\boldsymbol{\Psi}) - \mathrm{Pen}_\lambda(\boldsymbol{\Psi})$$

- $\hookrightarrow$ $\mathrm{Pen}_\lambda(\boldsymbol{\Psi})$ LASSO penalties for experts and the gating network
- encourages sparse solutions
- parameter estimation and selection problem

High-dimensional Mixtures-of-Experts

Estimation and Feature Selection in Mixtures of Generalized Linear Experts Models

- **prEMME** : proximal Newton EM for estimation and feature selection in high-dimensional Mixtures-of-Experts ▸ R software
- Poisson ▸ R software Logistic ▸ R software Gaussian ▸ R software

FunME : Functional Mixtures-of-Experts

Ongoing work : MoE with functional predictors/responses

- Let $\{X_i(\cdot), Y_i\}_{i=1}^{n}$, be a random i.i.d sample where $Y_i \in \mathbb{R}$ is the response and $X_i(t); t \in \mathcal{T} \subset \mathbb{R}$ is a functional predictor, for example the time in time series.
- The input $X(\cdot)$ is a function (eg. data continuously recorded for some time period) eg. $\boldsymbol{X}(\cdot)$ are data continuously recorded from multiple subject' sensors

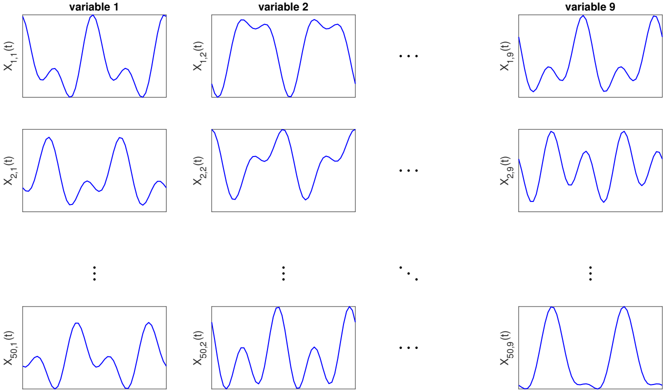


FIGURE – Functional predictors $X_{ij}(t)$ $t \in \mathcal{T}$, $i = 1, \cdots, n$ and $j = 1, \ldots, p$.

References I

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723, 1974.

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

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

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B*, 65 :367–389, 2003.

Xiuqin Bai, Weixin Yao, and John E. Boyer. Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, 56(7) :2347 – 2359, 2012.

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

C. Biernacki, G. Celeux, and G Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :719–725, 2000.

C. Bishop and M. Svensén. Bayesian hierarchical mixtures of experts. In *In Uncertainty in Artificial Intelligence*, 2003.

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.

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

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

F. Chamroukhi. Skew-normal mixture of experts. 2016a. URL `https://chamroukhi.com/papers/Chamroukhi-SNMoE.pdf`.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 79 :20–36, 2016b. URL `https://chamroukhi.com/papers/TMoE.pdf`.

References II

F. Chamroukhi. Skew t mixture of experts. *Neurocomputing - Elsevier*, 266 :390–408, 2017. URL https://chamroukhi.com/papers/STMoE.pdf.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A regression model with a hidden logistic process for feature extraction from time series. In *International Joint Conference on Neural Networks (IJCNN)*, 2009.

Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, Dec 2018. URL https://chamroukhi.com/papers/MBCC-FDA.pdf. DOI : 10.1002/widm.1298.

A. Delaigle, P. Hall, and N. Bathia. Componentwise classification and clustering of functional data. *Biometrika*, 99(2) :299–313, 2012.

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.

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

F. Ferraty and P. Vieu. *Nonparametric functional data analysis : theory and practice*. Springer series in statistics, 2006. ISBN 0-387-30369-3.

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.

T. Huynh and F. Chamroukhi. Estimation and feature selection in mixtures of generalized linear experts models. *arXiv :1810.12161*, July 2019. URL https://chamroukhi.com/papers/prEMME.pdf.

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

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1) : 79–87, 1991.

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.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6 :181–214, 1994.

A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4) : 519–539, 2010.

Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of multivariate skew t-distributions : some recent and new results. *Statistics and Computing*, 24(2) :181–202, 2014.

Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of canonical fundamental skew *t*-distributions. *Statistics and Computing (To appear)*, 2015. doi: 10.1007/s11222-015-9545-x.

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

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. The classification and mixture maximum likelihood approaches to cluster analysis. In P.R. Krishnaiah and L. Kanal, editors, *In Handbook of Statistics, Vol. 2*, pages 199–208. Amsterdam : North-Holland, 1982.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. New York : Wiley, second edition, 2008.

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

G. J. McLachlan, D. Peel, and R. W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4) :379–388, January 2003. URL http://www.sciencedirect.com/science/article/B6V8V-472JRC1-12/1/4d40244841bb6f7c8c454ca92e6cc347.

References IV

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

G.J. McLachlan and K.E. Basford. *Mixture Models : Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

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

Hien D. Nguyen and Faicel Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling : An overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, pages e1246–n/a, Feb 2018. ISSN 1942-4795. doi: 10.1002/widm.1246. URL `http://dx.doi.org/10.1002/widm.1246`.

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

Xinghao Qiao, Shaojun Guo, and Gareth M. James. Functional graphical models. *Journal of the American Statistical Association*, 0(0) :1–12, 2018. doi: 10.1080/01621459.2017.1390466.

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

J.O. Ramsay, T.O. Ramsay, and L.M. Sangalli. Spatial functional data analysis. In F. Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, pages 269–275. Springer, 2011.

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.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

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

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

Thank you for your attention !