Université de Toulon

Habilitation à Diriger les Recherches (HDR)

Discipline : Informatique et Mathématiques appliquées

présentée par

Faicel CHAMROUKHI

Maître de Conférences, UMR CNRS LSIS

Statistical learning of latent data models for complex data analysis

Soutenue publiquement le 07 décembre 2015

JURY

Geoffrey McLachlan	Professor, University of Queensland Australian Academy of Science Fellow	Rapporteur
Christophe Ambroise	Professeur, Université d'Evry	Rapporteur
Younès Bennani	Professeur, Université Paris Nord	Rapporteur
Stéphane Derrode	Professeur, Ecole Centrale de Lyon	Rapporteur
Mohamed Nadif	Professeur, Université Paris 5	Examinateur
Christophe Biernacki	Professeur, Université Lille 1, INRIA	Examinateur
Hervé Glotin	Professeur, Université de Toulon	Examinateur

Acknowledgements

First, I would like to address all my thanks and express my gratitude to Professor Geoff McLachlan, Professor Christophe Ambroise, Professor Younès Bennani and Professor Stéphane Derrode for having given me the honor of reviewing my habilitation and for the quality of their reports. A very special mention to Geoff; I was also very honored by your visit this year and greatly appreciated all the time you have spent in Toulon. You are inspiring me a lot to continue in this way in science.

I would also like to address my very special thanks to Professor Christophe Biernacki, Professor Mohamed Nadif and Professor Hervé Glotin for accepting to be part of my habilitation committee. A special warm mention to Professor Hervé Glotin for these past four years of collaboration in Toulon.

I would also like to address my special warm thanks to my colleagues of the DYNI team and the computer science department with whom it is a great pleasure to work. I also extend my warm thanks to all my colleagues of the LSIS lab and the faculty of science, with whom I have worked during these past four years, since I have been recruited in Toulon.

This year I am in CNRS research leave and I would like to warmly thank my new colleagues at the lab of mathematics of Lille 1 and at INRIA-Modal for their warm welcome. A particular mention to Professor Christophe Biernacki for having given me the honor to join his team and for his support.

I am also grateful to all the colleagues with whom I have collaborated during my research since the beginning of my PhD and I would like to address my thanks and express my gratitude to all of them at the Heudiasyc lab of UTC Compiègne, the Grettia group of IFSTTAR-Marne, the LiSSi lab of Paris 12, the LIPN lab of Paris 13, the LIPADE lab of Paris 5. A particular mention to my PhD advisors Doctor Allou Samé, Doctor Patrice Aknin, and Professor Gérard Govaert.

I have been a teaching assistant at Université Paris 13 during my first four academic years and I take this opportunity to warmly thank all my colleagues at the computer science department with whom I have collaborated.

I am also grateful to my former PhD and MSc students Dr. Marius Bartcus, Dr. Dorra Trabelsi, Dr. Rakia Jaziri, MSc. Ahmed Hosni, MSc. Céline Rabouy, MSc. Ahmad Tay and MSc. Hiba Badri. Thanks to all of you for your collaboration and I wish you all the success in your career and in your life.

Finally, I address my thanks to my friends and my family.

This manuscript has been finished in august, at Hyères.

Faicel Chamroukhi
Lille, November 26, 2015

Contents

1 Introduction **1**
- 1.1 Contributions during my thesis (2007-2010) . . . 2
- 1.2 Contributions after my thesis (2011-2015) . . . 2
  - 1.2.1 Latent data models for non-stationary multivariate temporal data . . . 2
  - 1.2.2 Functional data analysis . . . 2
  - 1.2.3 Bayesian regularization of mixtures for functional data analysis . . . 3
  - 1.2.4 Bayesian non-parametric parsimonious mixtures for multivariate data . . . 3
  - 1.2.5 Non-normal mixtures of experts . . . 4
  - 1.2.6 Applications . . . 4

2 Latent data models for temporal data segmentation **5**
- 2.1 Introduction . . . 7
  - 2.1.1 Personal contribution . . . 8
  - 2.1.2 Problem statement . . . 9
- 2.2 Regression with hidden logistic process . . . 9
  - 2.2.1 The model . . . 9
  - 2.2.2 Maximum likelihood estimation via a dedicated EM . . . 10
  - 2.2.3 Experiments . . . 12
  - 2.2.4 Conclusion . . . 13
  - 2.2.5 Multiple hidden process regression for joint segmentation of multivariate time series 13
- 2.3 Multiple hidden logistic process regression . . . 13
  - 2.3.1 The model . . . 14
  - 2.3.2 Maximum likelihood estimation via a dedicated EM . . . 14
  - 2.3.3 Application on human activity time series . . . 15
  - 2.3.4 Conclusion . . . 15
- 2.4 Multiple hidden Markov model regression . . . 16
  - 2.4.1 The model . . . 17
  - 2.4.2 Maximum likelihood estimation via a dedicated EM . . . 17
  - 2.4.3 Application on human activity time series . . . 17
  - 2.4.4 Conclusion . . . 18

3 Latent data models for functional data analysis **19**
- 3.1 Introduction . . . 21
  - 3.1.1 Personal contribution . . . 22
  - 3.1.2 Mixture modeling framework for functional data . . . 22
- 3.2 Mixture of piecewise regressions . . . 23
  - 3.2.1 The model . . . 23
  - 3.2.2 Maximum likelihood estimation via a dedicated EM . . . 24
  - 3.2.3 Maximum classification likelihood estimation via a dedicated CEM . . . 25
  - 3.2.4 Experiments . . . 26
  - 3.2.5 Conclusion . . . 28
- 3.3 Mixture of hidden Markov model regressions . . . 30
  - 3.3.1 The model . . . 30
  - 3.3.2 Maximum likelihood estimation via a dedicated EM . . . 31

3.3.3 Experiments . . . 32
3.3.4 Conclusion . . . 33
3.4 Mixture of hidden logistic process regressions . . . 34
3.4.1 The model . . . 34
3.4.2 Maximum likelihood estimation via a dedicated EM algorithm . . . 35
3.4.3 Experiments . . . 36
3.4.4 Conclusion . . . 37
3.5 Functional discriminant analysis . . . 37
3.5.1 Functional linear discriminant analysis . . . 38
3.5.2 Functional mixture discriminant analysis . . . 38
3.5.3 Experiments . . . 39
3.5.4 Conclusion . . . 40

4 Bayesian regularization of mixtures for functional data **43**
4.1 Introduction . . . 45
4.1.1 Personal contribution . . . 46
4.1.2 Regression mixtures . . . 46
4.2 Regularized regression mixtures for functional data . . . 47
4.2.1 Introduction . . . 47
4.2.2 Regularized maximum likelihood estimation via a robust EM-like algorithm . . . 49
4.2.3 Experiments . . . 51
4.2.4 Conclusion . . . 53
4.3 Bayesian mixtures of spatial spline regressions . . . 54
4.3.1 Bayesian inference by Markov Chain Monte Carlo (MCMC) sampling . . . 54
4.3.2 Mixtures of spatial spline regressions with mixed-effects . . . 55
4.3.3 Bayesian spatial spline regression with mixed-effects . . . 57
4.3.4 Bayesian mixture of spatial spline regressions with mixed-effects . . . 59
4.3.5 Experiments . . . 61
4.3.6 Conclusion . . . 62

5 Bayesian non-parametric parsimonious mixtures for multivariate data **63**
5.1 Introduction . . . 65
5.1.1 Personal contribution . . . 67
5.2 Finite mixture model model-based clustering . . . 67
5.2.1 Bayesian model-based clustering . . . 68
5.2.2 Parsimonious Gaussian mixture models . . . 68
5.3 Dirichlet Process Parsimonious Mixtures . . . 69
5.3.1 Dirichlet Process Parsimonious Mixtures . . . 69
5.3.2 Chinese Restaurant Process parsimonious mixtures . . . 71
5.3.3 Bayesian inference via Gibbs sampling . . . 72
5.3.4 Bayesian model comparison via Bayes factors . . . 73
5.3.5 Experiments . . . 74
5.4 Conclusion . . . 77

6 Non-normal mixtures of experts **79**
6.1 Introduction . . . 81
6.1.1 Personal contribution . . . 82
6.1.2 Mixture of experts for continuous data . . . 83
6.1.3 The normal mixture of experts model and its MLE . . . 83
6.2 The skew-normal mixture of experts model . . . 84
6.2.1 The model . . . 84
6.2.2 Maximum likelihood estimation via the ECM algorithm . . . 85
6.3 The t mixture of experts model . . . 88
6.3.1 The model . . . 88
6.3.2 Maximum likelihood estimation . . . 89
6.3.3 MLE via the EM algorithm . . . 89

6.3.4 MLE via the ECM algorithm . . . 91
6.4 The skew t mixture of experts model . . . 91
6.4.1 The model . . . 91
6.4.2 Identifiability of the STMoE model . . . 92
6.4.3 Maximum likelihood estimation via the ECM algorithm . . . 92
6.5 Prediction, clustering and model selection with the non-normal MoE . . . 94
6.6 Experiments . . . 96
6.7 Conclusion . . . 98

7 Conclusion and perspectives **99**
7.1 Conclusion . . . 99
7.2 Perspectives . . . 99
7.2.1 Advanced mixtures for complex data (My ongoing CNRS research leave project) . . . 99
7.2.2 LEarning from biG cOmplex FunctIonal daTa - LegoFit (2015 - an ANR proposal) 100
7.2.3 Non-normal mixture modeling . . . 100
7.2.4 Feature selection in model-based clustering . . . 101
7.2.5 Bayesian latent variable models for sparse representations . . . 101
7.2.6 Unsupervised learning of feature hierarchies: Deep learning . . . 101

8 Personal bibliography **103**
8.1 Monograph and editorials . . . 103
8.2 Journal papers . . . 103
8.2.1 Publications (9) . . . 103
8.2.2 Submitted papers (6) . . . 104
8.2.3 Papers in preparation (2) . . . 104
8.3 International conference papers . . . 104
8.4 Invited talks in international conferences . . . 106
8.5 Francophone conferences . . . 106
8.6 Theses . . . 106
8.7 Award and distinctions . . . 106
8.8 Invited and contributed seminars . . . 107

Bibliography **120**

Chapter 1

Introduction

The problem of complex data analysis is a central topic of modern statistical science as well as computer and information sciences, and is connected to both theoretical and applied parts of these sciences. The analysis of complex data in general implies the development of statistical models and autonomous algorithms that aim at acquiring knowledge from raw data for analysis, interpretation and to make accurate decisions and predictions for future data. Such analysis by learning models from raw data requires, from a theoretical point of view, models which rely on well-established statistical background, as well as, from a practical point of view, the derivation of efficient algorithmic tools to address problems regarding the data complexity, including heterogeneity, missing information, high dimensionality, dynamical structure, and big volume. To ensure such reliability of models and algorithms for the analysis, it is important to understand the processes generating the data. From a statistical learning prospective, this in general arises in generative learning approaches. Generative model-based approaches are indeed well-established statistical models that explicit the processes generating the data and for which the computational part regarding the development of dedicated efficient inference algorithms has took and is still taking a lot of investigations in the computer science field particularly machine learning field as well as in statistics. They are well-suitable in many contexts, in particular the unsupervised context when the supervision (e.g., expert information required for the analysis in the large sense) is missing, hidden or difficult to obtain, and are useful for many applications, including clustering and classification of heterogeneous data. Latent data models, including (Bayesian) mixture model-based approaches and their Markovian extensions, are one of the most popular and successful generative unsupervised learning approaches. They are very used in particular in cluster analysis for automatically finding clusters, in discriminant analysis when the classes are dispersed, as well as in non-linear regression when the response exhibits a non-stationary behavior conditional on predictors and in segmentation for data arranged in sequences, etc. Fitting such models is the core of the analysis task and has lead to an important research to derive efficient algorithmic tools such as the expectation-maximization (EM) algorithms or Markov Chain Monte Carlo (MCMC) sampling techniques in the Bayesian framework. Thanks to their flexibility and their sound statistical background, these successful latent data models, in general used in multivariate analysis in which the analyzed data are composed of individuals described by vectors, have took and are still taking a growing investigation for adapting them to the framework of functional data analysis, in which the individuals are describing functions (e.g., curves, surfaces), rather than simple vectors. In many areas of application, including signal and image processing, functional imaging, handwritten text recognition, bio-informatics, etc., the analyzed data are indeed often available in the form of discretized values of functions or curves (e.g., time series, waveforms) and surfaces (e.g., 2D-images, spatio-temporal data) which makes them very structured, and for which classical multivariate analyses are not suitable. This "functional" aspect of the data adds additional difficulties compared to the case of a classical multivariate (non-functional) analysis.

These modeling questions and the methodological issues, as well as their related practical and computational issues, are in the core of my research. This manuscript synthesizes the research activities I conducted on the subject after my PhD thesis, defended in December 2010 at Université de Technologie de Compiègne. My research activities are at the interface between applied mathematics (statistics) and computer science, with a special interest to statistical signal processing, and primary lie into the multidisciplinary area of statistical learning and analysis of complex data. The data complexity aspect refers to temporal non-stationary data, high dimensional multivariate and functional data, spatial structured data with possibly random effects, and non-normal, skewed and noisy (with atypical observations) data. They

are more precisely structured around the following five thematics of research, summarized after a brief account on my thesis research. My publications are given in my personal bibliography provided in Chapter 8, as well as in the long French version of my CV: `http://chamroukhi.univ-tln.fr/CV/FChamroukhi_CV_fr.pdf` or its short English version `http://chamroukhi.univ-tln.fr/CV/FChamroukhi_CV_en.pdf`, which contain my other academic activities.

1.1 Contributions during my thesis (2007-2010)

Curve modeling and classification using hidden process regression

My thesis research has mainly resulted in a contribution to the approximation and segmentation of temporal non-stationary data. This contribution addressed two main issues of the state of the art approaches related to the subject, which provide non-smooth approximations, or smooth but very costly in optimization: It provides smooth approximations thanks to the regression model with a hidden logistic process (RHLP) which allows to control the smoothness in the underlying process governing the data, with a very reasonable complexity thanks to efficient maximum-likelihood fitting via the expectation-maximization (EM) algorithm. I also showed that this model presents a well-principled alternative to address the classical problem of nonlinear regression by showing that the resulting RHLP regression function is asymptotically identical to the one obtained by the classical nonlinear regression model. The developments have also been of a great interest and successfully applied to, in particular, the diagnosis and the survey of the french high-speed railway system in a collaboration with the SNCF.

1.2 Contributions after my thesis (2011-2015)

1.2.1 Latent data models for non-stationary multivariate temporal data

This first theme of research after my thesis focuses on the modeling and joint segmentation of non-stationary multivariate time series which present regime changes. The main part of this research, initiated in 2010, was conducted in the framework of the PhD thesis of Dorra Trabelsi defended in 2013 at Paris 12 University - LISSI Lab, that I co-supervised with Pr. Latifa Oukhellou, Pr. Yacine Amirat and Dr. Samer Mohammed. Motivated by a problem of recognition of human activities from acceleration data issued from on-body sensors in the framework of assistive robotics, I reformulated the problem, generally addressed from a supervised prospective, into the methodological one of joint segmentation of multiple time series with hidden process regression. I proposed a new unsupervised generative modeling based on two latent data models: The multiple hidden Markov model regression (MHMMR) which naturally allows to recover the underlying states (i.e. activities) governing the data via the Markov chain, and provides a better fit for this structured data thanks to the conditional regression density for each state, compared to standard Markovian modeling and other unsupervised and supervised techniques. Furthermore, I proposed the multiple regression model with hidden logistic process (MRHLP), which tackles the problem from the same generative point of view, but with better theoretical modeling capabilities as in the univariate RHLP, that is, particularly the possibility of better and explicitly addressing possible smoothness in the time series, and with no restrictions on the state modeling, that is, a state residence time which is not necessarily Geometrically distributed as in the Markovian case. For the inference of the proposed models, I developed dedicated EM algorithms for MLE which were successfully applied to the real-world problem of human activity recognition and provided very interesting results.

1.2.2 Functional data analysis

Beyond the standard learning problem for the analysis of data which are univariate or multivariate variables, in this reach direction, that I mainly conducted since the end of my thesis and pursued, under additional modeling considerations, until now, focuses on functional data analysis (FDA). The key tenet of the modeling and analyses I perform in this research is that the unit of information (individual) is a function (e.g., curves, time series, surfaces). This is a learning problem for structured data analysis in which I essentially seek to provide solutions to the problems of classification, supervised or not (e.g., segmentation, clustering) of such complex data. I considered namely complex functional data arising in time series presenting regime changes and organized in groups. The considered functional data exhibit a hidden complex structure in two respects, that is, on the one hand, the dynamical behavior within the

individuals generated by regime changes, and, on the other hand, a grouping aspect between individuals, that is, a clustering property. Moreover, I also considered modeling data presenting random effects. I proposed latent data models, particularly, dedicated functional mixture models to accommodate the complexity of the data density. The models rely on hierarchical mixtures with regression components governed by hidden processes and were applied to the classification and the segmentation of functional data. More specifically, the developed models are: the mixture of piecewise regressions (MixPWR), the mixture of hidden Markov model regressions (MixHMMR), and the mixture of regressions with hidden logistic process (MixRHLP), in which the classification is performed directly in the space of curves. I also showed that the probabilistic formulation of such functional probabilistic models naturally optimized with EM type algorithms generalizes alternatives based on the of optimization of distortion criteria (i.e. K-means based functional methods). The models firstly applied in clustering, thanks to their generative formulation, are naturally used in discriminant analysis for functional data, particularly when the classes are dispersed, by proposing functional (mixture) discriminant analyses (FMDA) where the classification is directly performed in the space of curves and accounts for the structure of the curves.

1.2.3 Bayesian regularization of mixtures for functional data analysis

In this third axis that I initiated in 2013, I consider the subject of fitting latent data models for FDA and I aim at answering the two following main questions: i) how to automatically and simultaneously fit the model and its structure, and ii) how to overcome issues encountered in the ML point estimation framework, regarding possible singularities and degeneracies of the ML estimator. More precisely, in a first stage, I was interested in constructing fully unsupervised regression mixture models for univariate functional data, that is, where both the number of mixture components and the parameters are unknown and have to be inferred from the data. I handled this problem by proposing a penalized maximum likelihood estimation framework carried out via a robust-EM like algorithm which namely encourages sparse structures. Both the parameters and the model structure are simultaneously inferred from the data as the learning proceeds. As such, the proposal constitutes an alternative to conventional approaches on mixture-model based functional data clustering where the model selection is performed in a two-steps strategy by selecting a model from several pre-estimated model candidates. The algorithm effectiveness has been shown on application on functional data classification from various fields.

Then, since 2014, I considered the problem of fitting latent data models for FDA, treated in the two previous directions for some of the models considered here; The angle of the approach is different though, since here I am placed only in the Bayesian framework, that is, maximum a posteriori estimation (MAP) via Bayesian sampling techniques, in particular Markov Chain Monte Carlo to simulate directly under the posterior distribution of the parameters. Furthermore, here I considered the problem of spatial functional data analysis where the data are surfaces, rather than univariate or multivariate curves as before. I introduced a Bayesian spatial spline regression model with mixed-effects (BSSR) for modeling spatial functional data. The model accommodates both common mean behavior for the data through a fixed-effects part, and variability inter-individuals thanks to a random-effects part. Then, in order to model populations of spatial functional data issued from heterogeneous groups, I proposed a Bayesian mixture of spatial spline regressions with mixed-effects (BMSSR) for density estimation and model-based surface clustering. The two models, through their Bayesian formulation, allow to integrate possible prior knowledge on the data structure and constitute a Bayesian alternative to recent mixture of spatial spline regressions model estimated in a maximum likelihood framework via the EM algorithm. The Bayesian model inference is performed by Gibbs sampling and is applied on surface approximation as well as on a problem of handwritten digit recognition using the MNIST data set. The obtained results highlight the potential benefit of the proposed Bayesian approaches.

1.2.4 Bayesian non-parametric parsimonious mixtures for multivariate data

I initiated this research direction in 2012 with the beginning of the PhD thesis of Marius Bartcus and for whom I was the principal supervisor, in collaboration with Pr. Hervé Glotin. The PhD defense is scheduled for October 26th, 2015. In this research theme, I investigated the problem of fitting mixtures and model-based clustering under a Bayesian point of view where the aim is to deal with limitations of the previously and classically studied MLE based approaches. I study the case of Bayesian mixture fitting by investigating two sub-axis. The first one corresponds to the investigation of Bayesian finite

mixtures and their inference using mainly MCMC sampling, with a particular focus on finite parsimonious mixtures which offer great modeling flexibilities. The second one, however, addresses the problem from a non-parametric perspective by investigating the Dirichlet process mixture derivation for Bayesian mixtures, which can be interpreted as infinite mixture models. Then, I introduced Bayesian non-parametric parsimonious mixture models by relying on general flexible priors, that is, Dirichlet processes, or by equivalence the Chinese Restaurant Process. The developed DPPM models provide a flexible framework for modeling different data structures as well as a well-principled alternative to tackle the problem of model selection. The derived Gibbs sampling techniques to infer the models and the Bayes Factors used for model selection and comparison perform well on several real data and particularly have shown very interesting results in a challenging problem of unsupervised bioacoustic signals decomposition.

1.2.5 Non-normal mixtures of experts

The previously developed models in my research, as well as those classically used in learning for the analysis of continuous data, are usually or at least very often based on the normal hypothesis regarding the distribution of the data or a group of the data. In this research direction that I initiated very recently in May 2015, I attempt to overcome the (well-known) limitations of modeling with the normal distribution in terms of non suitability to heavy-tailed data, skewed data, and data possibly affected by outliers. I particularly focused on the problem of mixture modeling in such situations including for model-based clustering and for fitting non-linear regression. I investigated the framework of non-normal mixture models for density estimation and clustering of regression data, particularly mixture of experts (MoE), a popular framework for modeling heterogeneity in data in the computer science field in particular in machine learning, as well as in statistics. I proposed three non-normal and robust derivations for standard normal of mixture of experts models which can deal with possibly skewed, heavy-tailed data and are robust to atypical data. The proposed models are the skew-normal MoE (SNMoE), and the robust t MoE (TMoE) and skew-t MoE (STMoE). I developed dedicated expectation-maximization (EM) type algorithms for ML fitting of the models. The obtained results on simulations and real-world data show very interesting results and confirm that the proposed MoE models are well-suited and robust generalizations of the standard normal MoE.

1.2.6 Applications

The contributed models has lead to the development of numerous codes and were applied at least in the following structuring project applications

SwitchRdf project (2007-2010) This project has constituted the applicative context of my PhD research and was in collaboration between IFSTTAR and Heudyasic Lab where I did my PhD, and the SNCF (the French railway company). The objectives were the diagnosis and the monitoring of the high-speed railway switches based on temporal data of switch operations.

Human Activity recognition (2010-2013) This application relates the problem of human activity recognition, which is central for understanding and predicting the human behavior, in particular in a prospective of assistive services to humans, such as health monitoring, well being, security, etc. It was conducted with the LISSI Lab/Universié Paris-Est Créteil (Paris 12) mainly in framework of the PhD thesis of Dorra Trabelsi. The aim was the analysis of multidimensional raw acceleration data measured using body-worn sensors.

Bio-acoustic signals decomposition (2012-2015) SABIOD - Scaled Acoustic BIODiversity is CNRS MASTODONS project coordinated by LSIS, started in 2012 and calls for learning techniques to analyze bioacoustic data, mainly songs of marine mammals (e.g. whales). The idea was to apply unsupervised Bayesian learning techniques for the decomposition of whale song signals to discover possible call units which can be considered as a kind of whale alphabet of the whale.

Other projects I was also member and served for other projects in our team, that is, the ANR (french research council) Project COGNILEGO : from Pixels to Semantics (2011-2014) and the DGA-RAPID Project PHRASE: Augmented Reality and Autonomous Perception (2012-2015). I also Initiated the ANR proposal "LegoFit" proposed to ANR in collaboration with Paris 13, Paris 5, IFSTTAR and AIRBUS (see the perspectives section).

More additional information is available in my CV (in French) `http://chamroukhi.univ-tln.fr/CV/FChamroukhi_CV_fr.pdf` or in English (short version) `http://chamroukhi.univ-tln.fr/CV/FChamroukhi_CV_en.pdf`.

Chapter 2

Latent data models for temporal data segmentation

Contents

2.1 Introduction 7
2.1.1 Personal contribution 8
2.1.2 Problem statement 9
2.2 Regression with hidden logistic process 9
2.2.1 The model 9
2.2.2 Maximum likelihood estimation via a dedicated EM 10
2.2.3 Experiments 12
2.2.4 Conclusion 13
2.2.5 Multiple hidden process regression for joint segmentation of multivariate time series 13
2.3 Multiple hidden logistic process regression 13
2.3.1 The model 14
2.3.2 Maximum likelihood estimation via a dedicated EM 14
2.3.3 Application on human activity time series 15
2.3.4 Conclusion 15
2.4 Multiple hidden Markov model regression 16
2.4.1 The model 17
2.4.2 Maximum likelihood estimation via a dedicated EM 17
2.4.3 Application on human activity time series 17
2.4.4 Conclusion 18

Related PhD thesis:

[1] D. Trabelsi. *Contribution à la reconnaissance non-intrusive d'activités humaines.* Ph.D. thesis, Université Paris-Est Créteil, Laboratoire Images, Signaux et Systèmes Intelligents (LiSSi), June 2013

Related journal papers:

[1] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009d. URL `http://chamroukhi.univ-tln.fr/papers/Chamroukhi_Neural_Networks_2009.pdf`

[2] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_neucomp_2010.pdf`

[3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011c. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_same_govaert_aknin_rnti.pdf`

[4] F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013b.pdf`

[5] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on Hidden Markov Model Regression. *IEEE Transactions on Automation Science and Engineering*, 3(10):829–335, 2013. URL `http://arxiv.org/pdf/1312.6965.pdf`

[6] F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 2015. URL `http://chamroukhi.univ-tln.fr/papers/Sensors-2015.pdf`. submitted

Some related conference papers:

[1] F. Chamroukhi and H. Glotin. Mixture model-based functional discriminant analysis for curve classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, June 2012b

[2] F. Chamroukhi, H. Glotin, and C. Rabouy. Functional Mixture Discriminant Analysis with hidden process regression for curve classification. In *Proceedings of XXth European Symposium on Artificial Neural Networks ESANN*, pages 281–286, Bruges, Belgium, April 2012b

[3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Classification automatique de données temporelles en classes ordonnées. In *Actes des 44 ème Journées de Statistique*, Bruxelles, Belgique, Mai 2012c

[4] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Supervised and unsupervised classification approaches for human activity recognition using body-mounted sensors. In *Proceedings of the XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 417–422, Bruges, Belgium, April 2012

[5] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Activity Recognition Using Hidden Markov Models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, september 2011

This first research theme concerns the modeling and segmentation of complex temporal data, univariate and multivariate, and directly follows some of work I developed during my PhD thesis. This research axis can be organized into two sub-axes which are developed in what follows. The first deals with latent process regression models for univariate time series and mainly resulted into 3 methodological articles [J-1][J-2] [J-3] . The second is dedicated to latent data models for dealing with the joint segmentation of multivariate time series. This main part initiated in 2010 was conducted in the framework of the PhD thesis of Dorra Trabelsi defended in 2013 at Paris 12 University - LISSI Lab, and that I co-supervised with Pr. Latifa Oukhellou, Pr. Yacine Amirat and Dr. Samer Mohammed. This work resulted into 2 methodological articles [J-7][J-6] and the preprint [J-15] is issued from this work.

2.1 Introduction

In many application domains of data analysis, the data to be analyzed are presented as time series (may be called in other communities signals, curves, etc). Time series analysis is a popular problem with a broad literature, and is studied by several scientific communities, including statistics, (statistical) signal processing, economics as well as statistical learning in pattern recognition. In this study, we particularly focus on complex non-stationary time series that present non-linearities through various regime changes. When analyzing such data, very often of large size, it is often necessary to approximate them in order to extract relevant knowledge such as a relevant feature representation, a simplified model resuming the data for prediction, a segmentation for further categorization, etc. In such a context of time series with regime changes, the problem of time series analysis is in general reformulated into a time series segmentation problem possibly via models, parametric or not. The general problem of time series segmentation has taken great interest from different communities, including statistics, detection, signal processing, and machine learning. Earlier contributions in the subject were taken from a statistical point of view, one can cite for example the following references on the subject, among many others (McGee and Carleton, 1970; Rabiner and Juang, 1986; Brailovsky and Kempner, 1992; Basseville and Nikiforov, 1993; Eamonn Keogh and Pazzani, 1993; Fridman, 1993; Fearnhead, 2006; Fearnhead and Liu, 2007; Dobigeon et al., 2007). The problem can be stated as a detection problem via hypothesis testing as in Basseville and Nikiforov (1993). This in general requires a detection threshold to reject the null hypothesis. In addition, the hypothesis testing is often used in a binary setting, the problem of multiple hypothesis testing, which is the case in this multi-class segmentation problem, is not very common while one can take the hypotheses to be tested in pairs independently. The time series modeling for segmentation can also be handled via piecewise regression (PWR) which goes back to McGee and Carleton (1970), and which partitions a time series into several regimes (segments) where each regime is assumed to be a noisy constant function, or polynomial function as in Brailovsky and Kempner (1992). Each regime being characterized by its mean and possibly its own noise variance in the case of a heteroskedastic PWR model, and activated in a time range defined by its boundary transition points. In such a model, which is perhaps the most frequently used one in time series segmentation, at least from a statistical inference point of view, the problem of time series segmentation becomes the one of estimating the regression parameters and the change points. Thanks to the additivity of the error criterion, the usual tool for parameter estimation of the PWR model is dynamic programming (Bellman, 1961; Stone, 1961) which provides an optimal segmentation. However, it is known that the dynamic programming may be computationally expensive especially for time series with a large number of observations. The other alternative, which I mainly consider in this analysis, is in the framework of statistical learning of latent data models. Recall that latent data models which go back to Spearman (1904) with factor analysis, are statistical models that aim at representing the distribution of the observed data in terms of a number of latent (hidden, unknown, missing) variables. Mixture models (Titterington et al., 1985; McLachlan and Peel., 2000; Frühwirth-Schnatter, 2006) and hidden Markov models (HMMs) (Rabiner and Juang, 1986; Rabiner, 1989; Frühwirth-Schnatter, 2006) are two well-known widely used examples of such models. In deed, in this framework of regime changing time series, it is natural to think that the observed time series is generated by an underlying stochastic process, with several possibly parametric states. The problem of time series modeling and segmentation therefore becomes the one of recovering the underling process and inferring the statistical parameters of each of its states. The classical model in that case is the well-known hidden Markov model (HMM) (Rabiner and Juang, 1986; Rabiner, 1989) which assumes that the observed time series is generated by a hidden state sequence following a Markov chain, and for a time series segmentation problem, each regime

might be associated with a state. A classical assumption for such analysis is to consider that conditional on each state, the observation has a Gaussian density with constant or vectorial mean. Another way for time series with more structured regimes is to use a HMM regression (HMMR) (Fridman, 1993) which is a formulation of the standard HMM into an regression context. The parameters of such HMM-based models are estimated using the well-adapted expectation-maximization algorithm (Dempster et al., 1977) known as the Baum-Welch algorithm (Baum et al., 1970) in the context of HMMs. The problem of the analysis of time series with multiple change points based on Markov processes has also been considered as from a Bayesian prospective by using MCMC sampling as in Fearnhead (2006) an sequential MCMC for online change point detection as in Fearnhead and Liu (2007). Another Bayesian sampling approach is the one based on hierarchical Bayesian model for a joint segmentation of multidimensional astronomical time series (Dobigeon et al., 2007). These approaches, while at the time when I developed the proposed method were out of the scope, since I mainly focused on the frequentist paradigm of inference, as they use MCMC sampling, this can be limiting in terms of computational time compared to optimization methods I developed using deterministic EM algorithms. In addition, the previously described approaches particularly concern abrupt change point detection. Hence, if we are placed in the situation where we have nonlinear regimes their changes may be smooth and/or abrupt, in such a context, such models, particularly the piecewise regression model and the HMM based model, are not very appropriate to in particular provide regular approximation of the time series. In such a context, the aim of the analysis might be two-fold, that is *i)* to build a dedicated generative model, possibly parametric, to capture the dynamical behavior of the data, that is, mainly through detecting the temporal regime locations, while *ii)* providing a precise approximation to preserve a relevant data representation.

So in summary I focused on the approaches that attempt to address these issues by naturally thinking that a curve or a time series is generated by a process with several states. Conventional solutions to find these states generally are subject to limitations in the control of the transitions between these states, leading to a non-smooth approximation. One can force the resulting approximation to be regular, but this leads to combinatorial optimization problems for the optimal choice of the positions of the regime transitions. Relaxing the regularity conditions leads to efficient dynamic programming algorithms, but also to non smooth approximations. In what I proposed, I particularly relied on generative parametric latent data modeling, given the flexibility and easy interpretation of the generative framework which helps understanding the underlying processes generating the data. In addition, a generative model is in general directly usable for classification and clustering perspectives.

2.1.1 Personal contribution

My personal contribution consists in the study of generic generative regression models for both curve approximation and segmentation. I also studied the implementation of efficient estimation algorithms and applied them on real data and evaluated them with the alternatives. By tackling the problem from a statistical generative modeling prospective, the regression model that incorporates a discrete hidden logistic process (RHLP) presented in [J-1] for example, addresses these two issues regarding accurate regular curve approximation and segmentation. The RHLP which represents a dynamical mixture model, allows for activating, simultaneously and preferentially, time-varying polynomial regression models with both smooth and abrupt regime changes. Then, as showed in [J-3], the RHLP model, by producing smooth approximation, can be an alternative to solve the classical problem of nonlinear regression. I proposed two variants of efficient model inference algorithms by monotonically maximizing the observed-data (respectively complete-data) likelihood with a dedicated EM algorithm as in [J-1] (respectively classification EM (CEM)) algorithm [C-1]. The proposal provides results clearly better than those provided by standard methods, by considering real-world data issued from mainly diagnosis of complex systems [J-1] [C-13] [C-14] [C-16] [C-17] [C-18] and energy application for fuel cell life time estimation [C-15].

Then, I studied the problem of modeling and joint segmentation of multivariate temporal data and proposed two multiple hidden process regression models. I extended the previously developed univariate RHLP model to the multivariate case to develop a multiple RHLP (MRHLP) model as in [J-6]. I also investigated the use of HMM in such a regression context, and proposed, when the main aim is the segmentation, the multiple hidden Markov model regression (MHMMR) as in [J-7] which is better adapted to these structured time series thanks to the conditional regression density, compared to namely standard Markovian modeling. Both the MRHLP and the MHMMR models are naturally tailored to recover the underlying states governing these non-stationary time series via efficient EM algorithms.

The proposed models were successfully applied to a real-world problem of human activity recognition in an assistive robotics context and both provide better results compared to standard unsupervised and supervised techniques for activity recognition as shown in [J-6] [J-7] [J-15] [C-10] [C-12].

The remainder of this chapter is organized as follows. In section 2.2.1, I first present the first theme regarding the approximation and the segmentation of univariate time series, mainly by developing the regression model with hidden logistic process (RHLP). Then, I present the second theme which focuses on the on modeling and joint segmentation of multiple time series with regime changes by developing two latent data models, that is, the multiple regression model with hidden logistic process (MRHLP) presented in Section 2.3 and the multiple hidden Markov model regression (MHMMR) presented in 2.4.

2.1.2 Problem statement

The developed models are based on (multiple) regression with hidden processes. The aim of regression is to explore the relationship of an observed random variable Y given a covariate vector $\boldsymbol{X} \in \mathbb{R}^p$ via conditional density functions for $Y|\boldsymbol{X} = \boldsymbol{x}$ of the form $f(y|\boldsymbol{x})$, rather than only exploring the unconditional distribution of Y. For time series, the independent vector $\boldsymbol{x}$ in general relates the sampling time t, which we will consider hereafter. We are interested in parametric (non-)linear regression functions $f(y|\boldsymbol{x}) = \mu(\boldsymbol{x};\boldsymbol{\beta})$. Let $\boldsymbol{y} = (y_1,\ldots,y_n)$ be a time series composed of n univariate observations $y_i \in \mathbb{R}$ $(i = 1,\ldots,n)$ observed at the time points $\mathbf{t} = (t_1,\ldots,t_n)$.

2.2 Regression with hidden logistic process

2.2.1 The model

The regression model with a hidden logistic process (RHLP) assumes that the observed time series is governed by a K-"state" hidden process $\mathbf{z} = (z_1,\ldots,z_n)$ with the categorical random variable $z_i \in \{1,\ldots,K\}$ representing the unknown (hidden) label of the state (or class), in this case associated with a regime, generating the ith observation y_i. Each regime Z_i is modeled by a Gaussian polynomial regression model. Thus, the ith observation of the time series is modeled as

$$y_i = \boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i + \sigma_{z_i}\epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0,1), \quad (i = 1,\ldots,n) \tag{2.1}$$

where $\boldsymbol{\beta}_{z_i} \in \mathbb{R}^{p+1}$ is the regression coefficient vector of the polynomial regression model characterizing regime Z_i, p being the polynomial degree, $\boldsymbol{x}_i = (1, t_i, \ldots, t_i^p)^T \in \mathbb{R}^{p+1}$ is the time dependent covariate vector at time t_i, ϵ_i are independent standard zero-mean Gaussian variables representing an additive noise and $\sigma_{z_i}^2$ the associated noise variance. Under the RHLP, the process $\boldsymbol{Z} = (Z_1,\ldots,Z_m)$ is assumed to be logistic, that is, the hidden variable Z_i that allows for the switching from one regression model to another, given t_i, is generated independently according to the multinomial distribution $\mathcal{M}(1, \pi_1(t_i;\mathbf{w}),\ldots,\pi_K(t_i;\mathbf{w}))$, where the component probabilities have a logistic distribution:

$$\pi_k(t_i;\mathbf{w}) = \mathbb{P}(Z_i = k|t_i;\mathbf{w}) = \frac{\exp\left(\boldsymbol{w}_k^T \boldsymbol{v}_i\right)}{\sum_{\ell=1}^K \exp\left(\boldsymbol{w}_\ell^T \boldsymbol{v}_i\right)}, \tag{2.2}$$

where $\boldsymbol{v}_i = (1, t_i, , \ldots, t_i^u)^T \in \mathbb{R}^{u+1}$ is time-dependent covariate vector, $\boldsymbol{w}_k \in \mathbb{R}^{u+1}$ is the coefficients vector associated with $\boldsymbol{v}_i$ and $\mathbf{w} = (\boldsymbol{w}_1^T,\ldots,\boldsymbol{w}_{K-1}^T)^T \in \mathbb{R}^{(K-1)\times(u+1)}$ is the parameter vector of the logistic model, with $\boldsymbol{w}_K$ being the null vector. This modeling with the logistic distribution allows activating simultaneously and preferentially several regimes during time, and hence switching from one regime to another during time. The relevance and flexibility of the logistic distribution for modeling the dynamical behavior within a time series through accurately capturing smooth/abrupt etc regime transitions is discussed in Section 4.1 of [J-1]. Particularly, if the goal is to segment the curves into contiguous segments, one just needs to use linear logistic functions, that is by setting the value u of $\boldsymbol{w}_k$ to 1, which leads to linear logistic regression, what will be assumed hereafter.

From the above, the observation y_i at each time point t_i is distributed according to the following dynamical conditional mixture density:

$$f(y_i|t_i;\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(t_i;\mathbf{w})\mathcal{N}\big(y_i;\boldsymbol{\beta}_k^T\boldsymbol{x}_i,\sigma_k^2\big), \tag{2.3}$$

where $\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2)^T$ is the unknown parameter vector to be estimated.

In the RHLP model (2.3), both the mixing proportions and the component parameters are time-varying, contrary to for example standard switching regression models or mixture of regression models (Quandt, 1972; Quandt and Ramsey, 1978; Veaux, 1989). It can be seen as a mixture of experts (ME) (Jordan and Jacobs, 1994) where the logistic weights are time-dependent, that is, the particular covariate variable used for the mixing proportions represents the time.

2.2.2 Maximum likelihood estimation via a dedicated EM

The parameter vector $\boldsymbol{\theta}$ is estimated by monotonically maximizing the observed-data likelihood. Under the RHLP model, the maximization of the log-likelihood of $\boldsymbol{\theta}$, which is given by:

$$\log L(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}\big(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\big) \tag{2.4}$$

can not be performed in a closed form since the data are incomplete, that is, the labels $(z_1, \ldots, z_m)$ indicating from which component each observation of the time series is originated from, are unknown. However, in this latent data framework, the expectation-maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) is particularly adapted to achieve this task. By artificially completing the observed data with the indicator variables z_{ik} such that $z_{ik} = 1$ if $z_i = k$ (i.e., when y_i is generated by the kth regression model), and 0 otherwise, the EM algorithm monotonically maximizes $\log L(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t})$ (2.4) iteratively by alternating between the two following steps until convergence (see for example [J-1] for more details).

The E-Step computes the expected complete-data log-likelihood (also often called the Q-function), given the observations $(\mathbf{t}, \boldsymbol{y})$ and the current parameter value $\boldsymbol{\theta}^{(q)}$, q being the current iteration:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathbb{E}\left[\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) | \boldsymbol{y}, \mathbf{t}; \boldsymbol{\theta}^{(q)}\right] = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \left[\log \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)\right], \tag{2.5}$$

which simply requires the computation of the posterior probability $\tau_{ik}^{(q)}$ that y_i $(i = 1, \ldots, m)$ originates from component k $(k = 1, \ldots, K)$:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | y_i, t_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i; \boldsymbol{\beta}_k^{T(q)} \boldsymbol{x}_i, \sigma_k^{2(q)})}{\sum_{\ell=1}^{K} \pi_\ell(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i; \boldsymbol{\beta}_\ell^{T(q)} \boldsymbol{x}_i, \sigma_\ell^{2(q)})}. \tag{2.6}$$

The M-Step computes the parameter vector update $\boldsymbol{\theta}^{(q+1)}$ by maximizing the expected complete-data log-likelihood, that is, $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ where $\boldsymbol{\Theta}$ is the parameter space. The maximization of the Q-function w.r.t the regression coefficient vector $\boldsymbol{\beta}_k$ for each component k consists in analytically solving a weighted least-squares problem and the one w.r.t σ_k^2 is a weighted variant of the problem of estimating the variance of an univariate Gaussian density. In both cases the weights are the posterior membership probabilities $\tau_{ik}^{(q)}$ and the updates of these two parameters are given by

$$\boldsymbol{\beta}_k^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} y_i \boldsymbol{x}_i, \tag{2.7}$$

$$\sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{\beta}_k^{T(q+1)} \boldsymbol{x}_i)^2. \tag{2.8}$$

The maximization of the Q-function with respect to $\mathbf{w}$ is a multinomial logistic regression problem weighted by the $\tau_{ik}^{(q)}$'s and however cannot be solved in a closed form. It is solved with a multi-class Iteratively Reweighted Least Squares (IRLS) algorithm (Green, 1984; Chen et al., 1999a), see for example [C-14], which is equivalent to the Newton-Raphson algorithm, and in which the parameter $\mathbf{w}$ is updated as follows:

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \Big[\frac{\partial^2 Q_{\mathbf{w}}(\mathbf{w}, \boldsymbol{\theta}^{(q)})}{\partial \mathbf{w} \partial \mathbf{w}^T}\Big]_{\mathbf{w}=\mathbf{w}^{(l)}}^{-1} \frac{\partial Q_{\mathbf{w}}(\mathbf{w}, \boldsymbol{\theta}^{(q)})}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(l)}} \tag{2.9}$$

where $\frac{\partial^2 Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial \mathbf{w}\partial \mathbf{w}^T}$ and $\frac{\partial Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})}{\partial \mathbf{w}}$ are respectively the Hessian matrix and the gradient of $Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})$. The parameter update $\mathbf{w}^{(q+1)}$ is taken at convergence of the IRLS algorithm (2.9). Note that one can limit the number of iterations of the IRLS procedure of the EM algorithm. This version consists in only increasing the criterion $Q_{\mathbf{w}}(\mathbf{w},\mathbf{w}^{(q)})$ at each EM iteration rather than maximizing it. One can, for example, limit the number of IRLS iterations up to a single iteration. This scheme yields to a Generalized EM (GEM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) which has the same convergence properties as the EM algorithm. However, in practice, the IRLS is very fast and takes only up to forty iterations for the first three or four EM iterations and then only up to three or four iterations.

The time complexity of the E-step of this EM algorithm is of $\mathcal{O}(Kn)$. The calculation of the regression coefficients in the M-step requires the computation and the inversion of the square matrix $\mathbf{X}^T\mathbf{X}$ which is of dimension $p+1$, and a multiplication by the observation sequence of length n which has a time complexity of $\mathcal{O}(p^3 n)$. In addition, each IRLS loop requires an inversion of the Hessian matrix which is of dimension $(u+1)\times(K-1)$. Therefore for a small u (here we used $u=1$), the complexity of the IRLS loop is approximatively of $\mathcal{O}(I_{\text{IRLS}}K^2)$ where I_{IRLS} is the average number of iterations required by the internal IRLS algorithm. Therefore the proposed algorithm is performed with a time complexity of $\mathcal{O}(I_{\text{EM}}I_{\text{IRLS}}K^3p^3n)$, where I_{EM} is the number of iterations of the EM algorithm.

Time series approximation and segmentation Given the MLE $\hat{\theta}$, the time series approximation under the RHLP model i by its mean

$$\mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] \quad = \quad \int_{\mathbb{R}} y_i \; p(y_i|t_i;\hat{\boldsymbol{\theta}})\mathrm{d}y_i = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i \;\; (i=1,\ldots,n) \tag{2.10}$$

which results in a smooth and flexible approximation thanks to the flexibility of the logistic weights as illustrated namely in [J-1]. Thus, as proved in [J-3], if we consider the classical nonlinear regression model $y_i = f(t_i;\boldsymbol{\theta}) + \epsilon_i$ (Antoniadis et al., 1992), and, in order to cover a broad range of non-linear regression functions that are easily parameterizable, regression functions which can be written in the form of a finite sum of weighted polynomials with logistic weights $f(t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\boldsymbol{\beta}_k^T\boldsymbol{x}_i$, thanks to the desirable asymptotic properties the MLE estimator of $\boldsymbol{\theta}$, the regression function (2.10) estimated under the RHLP model is asymptotically identical to the one obtained from the classical nonlinear regression model. This strengthens us in the fact that the proposed model can be a good well-principled alternative to solve the nonlinear regression problem, if a suitable algorithm for parameter estimation is available, which is the case of the EM algorithm here. On the other hand, a time series segmentation can be obtained by computing the estimated label $\hat{z}_i$ of the polynomial regime generating y_i according to the following rule:

$$\hat{z}_i = \arg\max_{1\leq k\leq K} \pi_k(t_i;\hat{\mathbf{w}}), \quad (i=1,\ldots,n). \tag{2.11}$$

We show that applying this rule guarantees the curves are segmented into contiguous segments if the probabilities π_k are computed with a dimension $u=1$ of $\boldsymbol{w}_k$ $(k=1,\ldots,K)$. The separation between the polynomial regimes is linear in t in this case.

Model selection In a general application of the proposed model, the optimal values of (K,p,q) can be computed by using the Bayesian Information Criterion (BIC) (Schwarz, 1978) which is in our case defined by $\text{BIC}(K,p,u) = \log L(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}}\log(n)}{2}$ where $\nu_{\boldsymbol{\theta}} = K(p+u+3)-(u+1)$ is the number of free parameters of the model and $\log L(\hat{\boldsymbol{\theta}})$ is the observed-data log-likelihood obtained at convergence of the EM algorithm. Note that in the case of contiguous segmentation $(u=1)$ which we adopt here, the dimension of the parameter space reduces to $\nu_{\boldsymbol{\theta}} = K(p+4)-2$.

In this particular regression model, the variable z_i controls the switching from one regression model to another of K regression models at each time t_i. Therefore, unlike basic polynomial regression models which can be seen as stationary models as they assume uniform regression parameters over time, the proposed dynamical regression model allows for polynomial coefficients to vary over time by switching from one regression model to another. This modeling is therefore beneficial for capturing non-stationary behavior involved by regime changes for a curve.

2.2.3 Experiments

In [J-1][C-14][C-16], I evaluated the RHLP approach on simulated data and real data and compared it to alternatives, including piecewise regression and univariate HMM regression. Two evaluation criteria were used. the mean squared error between the true simulated mean curve (before noise is added) and the estimated mean curve, this criterion is used to assess the models with regard to the true curve approximation. The second criterion is the segmentation error rate between the true simulated and the estimated partitions and is used to assess the models with regard to time series segmentation.
Three experiments were performed: the first aims to observe the effect of the smoothness level of transitions on estimation quality, by varying the smoothness level of transitions from abrupt to very smooth changes, for different data situations. The second aims to observe the effect of the sample size n on estimation quality (to observe the convergence of the MLE of the model), and the third aims to observe the effect of the noise level σ. In terms of modeling and segmentation, while the results are closely similar when the transitions are abrupt, the proposed approach clearly outperforms the alternatives for smooth transitions in all situations. The proposed approach performs the time series segmentation and approximation better than the piecewise regression and the HMRM approaches. We also observed that the approximation error and the segmentation error rate decrease when the sample size n increases for the proposed model which provides more accurate results than the piecewise and the HMRM approaches. When the noise level increases, the RHLP provides more stable results than to the two other alternatives. In terms of computing time, as shown for example in [C-14], the proposed EM algorithm for the RHLP is clearly faster compared to using dynamic programming for piecewise regression, slightly faster that the EM for HMM regression, and faster than both when derived into a CEM version as in [C-1]. The proposed RHLP were also applied to the approximation of real time series issued from a complex system diagnostic application [C-18][C-17] [C-14][J-1]. The data are signals of the power consumed during high-speed railway switch operations, each operation signal is composed of five successive movements, each of them is associated with a regime in the time series model. The provided results are very satisfactory in terms of both segmentation and approximation, the model retrieves the actual phases precisely (see an illustrative example[1] in Figure 2.1). This result has also been reveled relevant for signal representation for a classification perspective. In deed, for example in [J-1], we considered the RHLP model in a classification context to diagnosis the railway switch by predicting the class label of a new measured signal based on a probabilistic discrimination model trained on a labeled training set of curves.

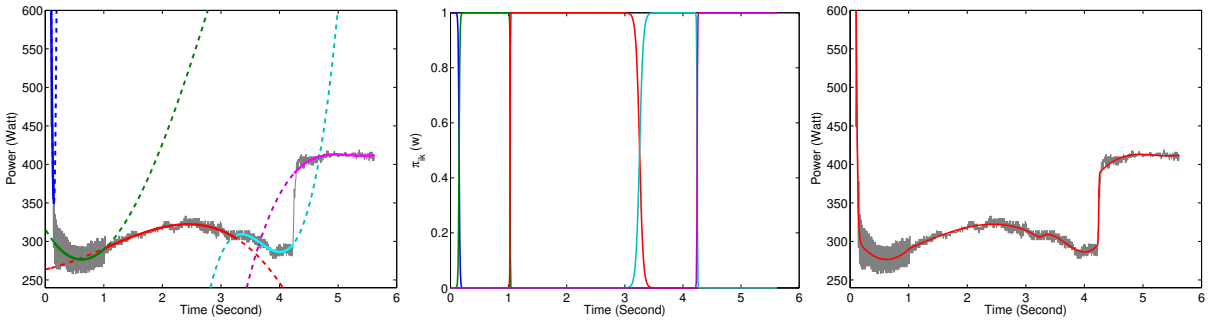

Figure 2.1: Results obtained with the proposed RHLP on a real switch operation time series: The signal and the polynomial regimes (left), the corresponding estimated logistic proportions (middle) and the obtained mean curve (right).

The RHLP was used first to extract features from the data, each feature vector being the MLE of the RHLP model from a given signal. Then, the estimated features wer used to train a mixture discriminant analysis (MDA) Hastie and Tibshirani (1996) classifier, that is, by using a Gaussian mixture as conditional density. The Bayes rule is finally applied for class prediction (by covering three classes: without defect, with minor defect, and with critical defect). The results in terms of correct classification rates confirm

[1] **Please note that, in this experiment, as well as along the whole manuscript, I'm trying to summarize at best the (numerous) experiments and just provide some graphical illustrations, at some places where I think this may help better understanding. However, the complete results, for all the considered data sets are available in the references as cited in the manuscript.**

that the RHLP can be a good representation of such time series to be used as input for external classifiers (the gain is about 8% when using PWR and 3% when using HMM regression).

In [J-3], I considered the RHLP as an alternative to the estimation of the classical non linear regression model by covering a broad range of non-linear regression functions that are easily parameterizable, and which can be written in the form of a finite sum of weighted polynomials with logistic weights. The provided results, compared to other alternatives, confirmed this claim.

2.2.4 Conclusion

In conclusion, the RHLP, thanks to its generative modeling, is naturally tailored to deal with the problem of modeling regime changing time series, and, is particularly particularly useful for situations with smooth regime transitions. In addition to accurate time series approximation and segmentation, the RHLP can be used as a signal representation method in a context of time series classification by using classical classification approaches such as MDA as experimented in [J-1].

2.2.5 Multiple hidden process regression for joint segmentation of multivariate time series

It is natural to extend the RHLP model as well as its alternative, in particular the Markovian model and the piecewise regression model, to the segmentation of multivariate time series. Indeed, it is natural to think that the univariate components of the multivariate time series are simultaneously governed by a hidden process and thus the problem of segmentation becomes the one of recovering the hidden process. This what we did respectively in [J-6] and [J-7], respectively, where we focused on the extension of the two generative models, that is, the RHLP model and the hidden Markov model regression model, which are described in the two following sections.

In this multivariate case, let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be a time series of n multidimensional observations $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at the time points $\mathbf{t} = (t_1, \ldots, t_n)$. Then, the multiple regression with hidden process model is formulated as a set of several polynomial regression models (RHLP or HMMR) for univariate time series and is stated as follows:

$$\begin{aligned} y_i^{(1)} &= \boldsymbol{\beta}_{z_i}^{(1)T}\boldsymbol{x}_i + \sigma_{z_i}^{(1)}\epsilon_i \\ y_i^{(2)} &= \boldsymbol{\beta}_{z_i}^{(2)T}\boldsymbol{x}_i + \sigma_{z_i}^{(2)}\epsilon_i \\ &\vdots \quad\quad \vdots \\ y_i^{(d)} &= \boldsymbol{\beta}_{z_i}^{(d)T}\boldsymbol{x}_i + \sigma_{z_i}^{(d)}\epsilon_i \end{aligned} \tag{2.12}$$

which can be written in a matrix form as

$$\boldsymbol{y}_i = \mathbf{B}_{z_i}^T\boldsymbol{x}_i + \mathbf{e}_i \quad ; \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{z_i}), \quad (i = 1, \ldots, n) \tag{2.13}$$

where $\mathbf{B}_k = \left[\boldsymbol{\beta}_k^{(1)}, \ldots, \boldsymbol{\beta}_k^{(d)}\right]$ is a $(p+1)\times d$ dimensional matrix of the multiple regression model parameters associated with the regime (class) $Z_i = k$ and $\boldsymbol{\Sigma}_{z_i}$ its corresponding $d \times d$ covariance matrix. with $\boldsymbol{x}_i = (1, t_i, t_i^2 \ldots, t_i^p)^T$ with p is the degree of the polynomial model associated with the class $z_i = k$. The latent process $\mathbf{z}$ that simultaneously governs all the univariate time series components controls the switching from one regime to another during time and allows therefore for a joint segmentation of the time series. We investigated the case where this process is logistic as it will be presented in Section 2.3, as well as the alternative case in which the hidden process is assumed to be a Markov chain as it will be presented in Section 2.4.

2.3 Multiple hidden logistic process regression

The Multiple hidden logistic process regression (MRHLP) model proposed in [J-6] assumes that the observed multivariate time series is governed by hidden states following a logistic process and conditional on each state, the observed data has a Gaussian multiple regression model. More specifically, the proposed approach is based on a specific multiple regression model incorporating a hidden discrete logistic process

which governs the switching from one regime to another over time. The model is learned in an unsupervised context by maximizing the observed-data log-likelihood via a dedicated expectation-maximization (EM) algorithm. The proposed approach extends the regression model with a hidden logistic process (RHLP) [J-1] which is concerned with univariate time series, to the multivariate case. Note that the general model formulation may include the possibility to train the polynomial dynamics with different orders rather than assuming a common order for all the polynomials. In this way, the model offers more flexibility allowing the capture of nonlinearities generated by the different regimes.

2.3.1 The model

Under the MRHLP model, the probability distribution of the process $\mathbf{z} = (z_1, \ldots, z_n)$, that allows for the switching from one regression model to another is assumed to be logistic and hence is well-adapted for capturing both abrupt and/or smooth regime changes thanks to the flexibility of the logistic distribution as illustrated in [J-1]. The observation $\boldsymbol{y}_i$ at each time point t_i is therefore distributed according to the following conditional normal mixture density:

$$f(\boldsymbol{y}_i|t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\mathcal{N}\big(\boldsymbol{y}_i;\mathbf{B}_k^T\boldsymbol{x}_i,\boldsymbol{\Sigma}_k\big), \tag{2.14}$$

where $\boldsymbol{\theta} = (\mathbf{w}, \mathrm{vec}(\mathbf{B}_1)^T, \ldots, \mathrm{vec}(\mathbf{B}_K)^T, \mathrm{vech}(\boldsymbol{\Sigma}_1)^T, \ldots, \mathrm{vech}(\boldsymbol{\Sigma}_K)^T)^T$ is the unknown parameter vector to be estimated where vec is the vectorization operator and vech is the half-vectorization operator which produces the lower triangular portion of the symmetric matrix it operates on.

2.3.2 Maximum likelihood estimation via a dedicated EM

The parameter vector $\boldsymbol{\theta}$ of the MRHLP model is estimated by maximizing the observed-data log-likelihood of $\boldsymbol{\theta}$:

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\mathcal{N}\big(\boldsymbol{y}_i;\mathbf{B}_k^T\boldsymbol{x}_i,\boldsymbol{\Sigma}_k\big). \tag{2.15}$$

via the EM algorithm developed in [J-6] which alternates between the two following steps until convergence:

The E-Step computes the expected complete-data log-likelihood which is given by:

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log\left[\pi_k(t_i;\mathbf{w})\mathcal{N}\left(\boldsymbol{y}_i;\mathbf{B}_k^T\boldsymbol{x}_i,\boldsymbol{\Sigma}_k\right)\right], \tag{2.16}$$

and consists in only computing the posterior component membership probabilities given by:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i,t_i;\boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t_i;\mathbf{w}^{(q)})\mathcal{N}(\boldsymbol{y}_i;\mathbf{B}_k^{T(q)}\boldsymbol{x}_i,\boldsymbol{\Sigma}_k^{(q)})}{\sum_{\ell=1}^{K}\pi_\ell(t_i;\mathbf{w}^{(q)})\mathcal{N}(\boldsymbol{y}_i;\mathbf{B}_\ell^{T(q)}\boldsymbol{x}_i,\boldsymbol{\Sigma}_\ell^{(q)})}\cdot \tag{2.17}$$

The M-Step computes the parameter vector update $\boldsymbol{\theta}^{(q+1)}$ by maximizing the Q-function w.r.t $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)})$. In this case the regression model parameters update correspond to analytic solutions of weighted multiple regression problems where the weights are the posterior probabilities $\tau_{ik}^{(q)}$ and are given by:

$$\mathbf{B}_k^{(q+1)} = \Big[\sum_{i=1}^{n}\tau_{ik}^{(q)}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}\sum_{i=1}^{n}\tau_{ik}^{(q)}\boldsymbol{x}_i\boldsymbol{y}_i^T \tag{2.18}$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ik}^{(q)}}\sum_{i=1}^{n}\tau_{ik}^{(q)}(\boldsymbol{y}_i - \mathbf{B}_k^{T(q+1)}\boldsymbol{x}_i)^T(\boldsymbol{y}_i - \mathbf{B}_k^{T(q+1)}\boldsymbol{x}_i)\cdot \tag{2.19}$$

The maximization of $Q_{\mathbf{w}}(\mathbf{w},\boldsymbol{\theta}^{(q)})$ with respect to $\mathbf{w}$ is a multinomial logistic regression problem weighted by $\tau_{ik}^{(q)}$ which we solve with a multi-class IRLS.

The complexity of the E-step of this EM algorithm is of $\mathcal{O}(Kn)$. The calculation of the regression coefficients in the M-step requires the computation and the inversion of the $d \times (p+1)$ square matrix, and a multiplication by the observation sequence of length n which has a time complexity of $\mathcal{O}(d^3p^3n)$. In addition, each IRLS loop requires an inversion of the Hessian matrix which is of dimension $2 \times (K-1)$. The complexity of the IRLS loop is then approximatively of $\mathcal{O}(I_{\text{IRLS}}K^2)$ where I_{IRLS} is the average number of iterations required by the internal IRLS algorithm. The proposed algorithm has therefore a time complexity of $\mathcal{O}(I_{\text{EM}}I_{\text{IRLS}}K^3d^3p^3n)$, where I_{EM} is the number of iterations of the EM algorithm. So this attractive for a reasonable number of regimes and dimensions and large number of individuals.

Once the model parameters are estimated by the EM algorithm, the time series segmentation can be obtained by computing the estimated label $\hat{z}_i$ of the polynomial regime generating each measurement $\boldsymbol{y}_i$ according to the rule (2.11).

The model selection, mainly related to choosing the optimal value of K can be performed from the data by using for example the BIC defined as $\text{BIC}(K,p,u) = \log L(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$ where $\nu_{\boldsymbol{\theta}} = K(d(p+1)+ d \times (d+1)/2 + 2) - 2$ is the number of free parameters of the model and $\log L(\hat{\boldsymbol{\theta}})$ is the observed-data log-likelihood obtained at convergence of the EM algorithm.

2.3.3 Application on human activity time series

In [J-6](Trabelsi, 2013)[C-10], the MRHLP model was applied on real-world problem of segmentation of human activity time series for activity recognition in a context of assistive robotics. The experiments were conducted at the LISSI Lab/University of Paris-Est Créteil (UPEC) by six different healthy subjects of different ages and have consisted in collecting raw acceleration data over time when from three body-worn sensor units, each unit being a tri-axial accelerometer. The 9-dimensional acceleration time series recorded overtime for each activity, present regime changes, each regime is associated to an activity. The MRHLP was used to jointly segment the time series in order to recover the activities in an unsupervised way. Twelve activities including dynamical and transitory activities have been studied. The model performances have been compared to those obtained with alternative unsupervised and supervised activity recognition approaches. The evaluation criterion is the error segmentation between the obtained segmentation and the ground truth (the data were labeled by an independent operator). The estimated probabilities of the logistic process that govern the switching from one activity to another over time correspond to a quite accurate segmentation of the acceleration time series (see an example in Figure 2.2). Moreover, the flexibility of the logistic process allows to get smooth probabilities in particular for the transitions. On the other hand, for the standard HMM, which represents a standard temporal segmentation approach, we clearly observed segmentation errors in the transitory phases and even when the person maintains the same activity. Furthermore, comparison with other classical supervised techniques, including MLP and SVM, showed that the MRHLP performs better thanks to its time-varying parameters formulation. While in some situations the KNN can provide slightly better results, it may require an important computational load for each classification (about 5 seconds for a single time series) while for the MRHLP one, the Bayes assignment rule is instantaneous once the model is trained. While the HMM model is also a dynamic model for time series modeling, it was observed that it does not outperform the proposed MRHLP approach. These misclassification errors can be attributed to the fact that the transitions here are similar the problem of class overlap in the case of multidimensional data classification problem. However, it was seen that, unlike the proposed model, for the HMM approach, confusions can occur even within a homogeneous part of the class that is not necessarily near the transition.

2.3.4 Conclusion

In summary, the proposed MRHLP model provides a well established statistical latent data model with very encouraging performance for automatic segmentation of human activity time series. The model formulation explicits the switching from one regime to another during time through a flexible logistic process which is also particularly well adapted for abrupt or smooth transitions. Furthermore, the expectation-maximization algorithm offers a stable efficient optimization tool to learn the model. The proposed MRHLP approach is applied on a real-world activity recognition problem based on multidimensional acceleration time series measured using body-worn accelerometers. The approach has shown very encouraging results compared to alternative models for activity recognition. Future work will also concern the use of other non-linear models to describe each activity signal rather than polynomial bases.

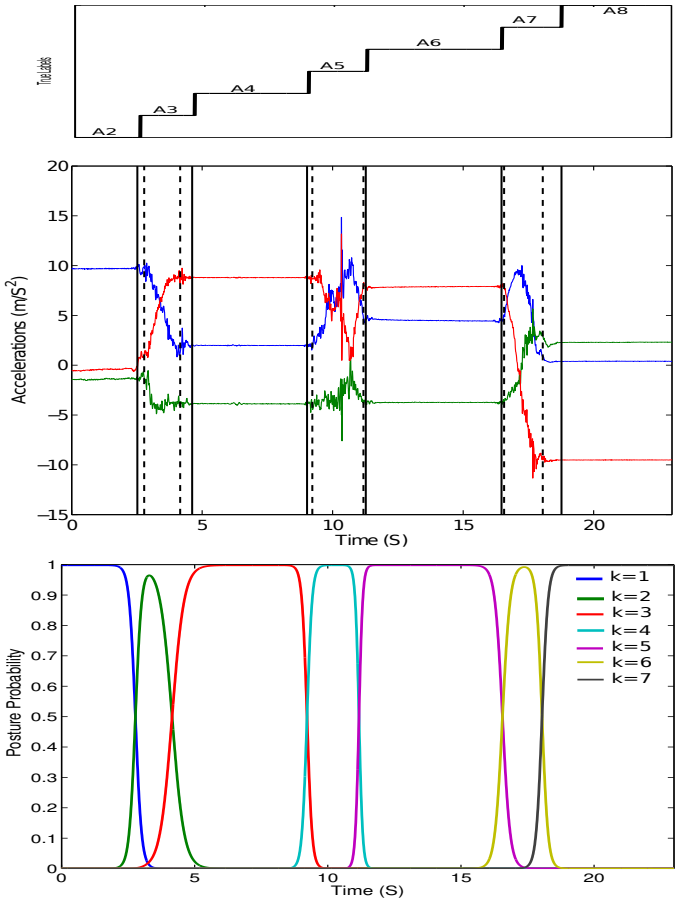

Figure 2.2: MRHLP segmentation results for the scenario: Standing A_2 (k=1) - Sitting down A_3 (k=2)- Sitting A_4 (k=3) - From sitting to sitting on the ground A_5 (k=4) - Sitting on the ground A_6 (k=5) - Lying down A_7 (k=6) - Lying A_8 (k=7) with (top) the true labels, (middle) the time series and the actual segments in bold line and the estimated segments in dotted line, and (bottom) the logistic probabilities.

This may improve in particular the representation of each activity. Then, another extension may consist int integrating the model into a Bayesian non-parametric framework which will be useful for any kind of complex activities and in which the number of activities might be selected as the learning proceeds.

2.4 Multiple hidden Markov model regression

The Multiple hidden Markov model regression (MHMMR) model is a HMM model in a multiple regression setting which allows to better handle the dynamical structure of the time series through the underlying Markov chain as well as the structure within each state thanks to the multiple regression model. More specifically, it is an extension of the standard hidden Markov model (HMM) Rabiner and Juang (1986); Rabiner (1989) to regression analysis as in univariate hidden Markov model regression Fridman (1993), particularly for regression on multivariate data.

Each regime is described by a regression model rather than a simple constant mean over time, while preserving the Markov process modelling for the sequence of unknown (hidden) regimes. Indeed, standard HMM-based approaches use in general standard Gaussian density as emission density. However, in the HMM regression context, each observation is assumed to be a noisy polynomial function to better model very structured observations for each state. The approach we propose further extends the HMM model to a multiple regression setting.

2.4.1 The model

The proposed multiple HMM regression (MHMMR) model is formulated by (2.12) or in a matrix form by (2.13) and assumes that the hidden sequence $\mathbf{z} = (z_1, \ldots, z_n)$ is a homogeneous Markov chain of first order parameterized by the initial state distribution $\boldsymbol{\pi}$ and the transition matrix $\mathbf{A}$. IEach regime is represented by a multiple regression model and the switching from one regime to another is thus governed by the hidden Markov chain. The model is therefore fully parameterized by the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \text{vec}(\mathbf{A})^T, \text{vec}(\mathbf{B}_1)^T, \ldots, \text{vec}(\mathbf{B}_K)^T, \text{vech}(\boldsymbol{\Sigma}_1)^T, \ldots, \text{vech}(\boldsymbol{\Sigma}_K)^T)^T$.

2.4.2 Maximum likelihood estimation via a dedicated EM

The parameter vector $\boldsymbol{\theta}$ of the MHMMR model is estimated by maximizing the observed-data log-likelihood, which is in this case given by:

$$\log L(\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} p(z_1; \pi) \prod_{i=2}^{n} p(z_i | z_{i-1}; A) \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{y}_i; \mathbf{B}_{z_i}^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_{z_i}) \cdot \tag{2.20}$$

Since this log-likelihood cannot be maximized directly in this incomplete-data framework (the segment labels are missing), this can be performed by using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008), known as the Baum-Welch algorithm in the HMM context (Baum et al., 1970; Rabiner and Juang, 1986; Rabiner, 1989). The EM algorithm for the MHMMR model alternates between the two following steps:

The E-step computes the conditional expectation of the complete-data log-likelihood given the observed data $\mathbf{Y}$, the time points $\mathbf{t}$ and a current parameter estimation denoted by $\boldsymbol{\theta}^{(q)}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathbb{E}\Big[\log p(\mathbf{Y}, \mathbf{z} | \mathbf{t}; \boldsymbol{\theta}) | \mathbf{Y}, \mathbf{t}; \boldsymbol{\theta}^{(q)} \Big] \cdot \tag{2.21}$$

It can be easily shown that this step only requires the calculation of the posterior probability $\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | \mathbf{Y}, \mathbf{t}; \boldsymbol{\theta}^{(q)})$ $(i = 1, \ldots, n; k = 1, \ldots, K)$ that $\boldsymbol{y}_i$ originates from the kth polynomial regression component given the whole observation sequence and the current parameter estimation $\boldsymbol{\theta}^{(q)}$, as well as the joint posterior probability of the state k at time i and the state ℓ at time $i-1$ given the whole observation sequence and $\boldsymbol{\theta}^{(q)}$, that is $\xi_{i\ell k}^{(q)} = \mathbb{P}(Z_i = k, Z_{i-1} = \ell | \mathbf{Y}, \mathbf{t}; \boldsymbol{\theta}^{(q)})$ $(i = 2, \ldots, n; k, \ell = 1, \ldots, K)$.

These posterior probabilities are computed by the forward-backward procedures in the same way as for a standard HMM (Rabiner and Juang, 1986; Rabiner, 1989).

The M-step updates the value of the parameter vector $\boldsymbol{\theta}$ by computing $\boldsymbol{\theta}^{(q+1)}$ that maximizes the conditional expectation (2.21) with respect to $\boldsymbol{\theta}$. It can be shown that this maximization leads to the following updating rules. The updates of the parameters governing the hidden Markov chain $\mathbf{z}$ are the ones of a standard HMM and are given by:

$$\pi_k^{(q+1)} = \tau_{1k}^{(q)}; \quad \mathbf{A}_{\ell k}^{(q+1)} = \frac{\sum_{i=2}^{n} \xi_{ik\ell}^{(q)}}{\sum_{i=2}^{n} \tau_{ik}^{(q)}}$$

The updates of the regression parameters consist in analytically solving K weighted multiple polynomial regressions and the analytic update of each of the covariance matrices corresponds to a weighted variant of the estimation of the covariance of a multivariate Gaussian with polynomial mean. They are respectively given by (2.18) and (2.19) as in the case of the previously presented MRHLP model.

The most likely sequence of activities is then estimated using the Viterbi algorithm (Viterbi, 1967) which is performed in a complexity of $\mathcal{O}(nK^2)$.

2.4.3 Application on human activity time series

Similarly like with the previously devoloped MRHLP mode, the MHMMR model was applied to an activity recognition problem based on multivariate acceleration time series. See for example [J-7](Trabelsi, 2013). For the particular problem of human activity recognition, the standard HMM is used for example

in (Lin and Kulić, 2011) wand thus was also applied for comparison, as well GMMs or other supervised techniques such as MLP or SVM. Note that in the generative models, including the proposed MHMMR approach, the raw acceleration data are directly used without a prior feature extraction step as for example in (Altun et al., 2010), Ravi et al. (2005) and Yang and Hsu (2010) in such application context. The MHMMR gives 91.4% as a mean correct classification rate averaged over all observations. It highlights the potential benefit of the proposed approach in terms of automatic segmentation and classification of human activity. Both the transitions and the stationary activities are well identified. It significantly outperforms the alternative standard unsupervised classification approaches (k-Means, the GMM and the standard HMM which only provide 60%, 72% and 84% of correct classification, respectively). Notice that, the GMM and K-means approaches are not well suitable for this kind of longitudinal data. The model also outperforms the majority of the considered supervised classification approaches (such as random forests, SVM, MLP wich provide respectively correct classification of 93.5%, 88.1% and 83.1% respectively). While the k-NN might provide slightly better results, it might need a strong data storage especially for large sequences, contrary to the classification with the MHMMR, which is direct. In addition, the K-NN is supervised and can not be applied in an unsupervised context. In summary, compared to standard supervised classification techniques (using class labels), these results are very encouraging since the proposed approach performs in an unsupervised way. Se an example in Figure 2.3.

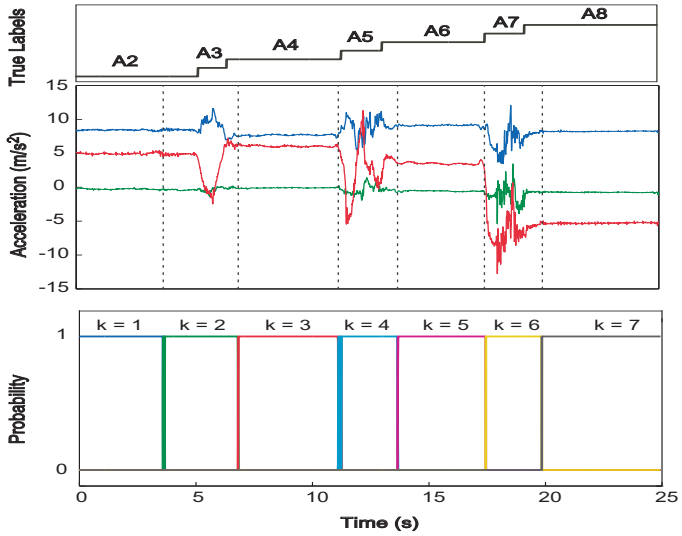

Figure 2.3: MHMRM segmentation for the sequence (Standing A2 - Sitting down A3 - Sitting A4 - From sitting to sitting on the ground A5 - Sitting on the ground A6 - Lying down A7- Lying A8) for the seven classes k=(1,. . . , 7)

2.4.4 Conclusion

In this section, we presented a statistical approach based on hidden Markov models in a multiple regression context for the joint segmentation of multivariate time series. The main advantage of the proposed approach comes from the fact that the statistical model explains the regime changes over time through the hidden Markov chain, each regime being interpreted as a regime (a segment) and its internal structure is accommodated by a multiple regression model. The parameter estimates are computed by maximizing the log-likelihood by using a dedicated EM algorithm. Application on real time series of human activities based upon the use of raw accelerometer data acquired from body mounted inertial sensors in a health-monitoring context, and the comparison with well-known unsupervised and supervised classification approaches demonstrated the effectiveness of the model. This work can be extended in several directions, namely by integrating the model into a Bayesian context to better control the model complexity via choosing suitable prior distributions on the model parameters. Then, and perhaps more interestingly, another step to explore is to built a fully Bayesian non-parametric model which will be useful for any kind of complex activities and in which the number of activities might be inferred from the data.

Chapter 3

Latent data models for functional data analysis

Contents

3.1 Introduction 21
3.1.1 Personal contribution 22
3.1.2 Mixture modeling framework for functional data 22
3.2 Mixture of piecewise regressions 23
3.2.1 The model 23
3.2.2 Maximum likelihood estimation via a dedicated EM 24
3.2.3 Maximum classification likelihood estimation via a dedicated CEM 25
3.2.4 Experiments 26
3.2.5 Conclusion 28
3.3 Mixture of hidden Markov model regressions 30
3.3.1 The model 30
3.3.2 Maximum likelihood estimation via a dedicated EM 31
3.3.3 Experiments 32
3.3.4 Conclusion 33
3.4 Mixture of hidden logistic process regressions 34
3.4.1 The model 34
3.4.2 Maximum likelihood estimation via a dedicated EM algorithm 35
3.4.3 Experiments 36
3.4.4 Conclusion 37
3.5 Functional discriminant analysis 37
3.5.1 Functional linear discriminant analysis 38
3.5.2 Functional mixture discriminant analysis 38
3.5.3 Experiments 39
3.5.4 Conclusion 40

Related journal papers:

[1] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_neucomp_2010.pdf`

[2] A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011. URL `http://chamroukhi.univ-tln.fr/papers/adac-2011.pdf`

[3] F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013a.pdf`

[4] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015d. URL `http://arxiv.org/pdf/1312.6974v2.pdf`. Accepted

[5] F. Chamroukhi. Mixture of hidden Markov model regressions for functional data clustering and segmentation. *Neural Networks*, 2015b. In preparation

Some related conference papers:

[1] F. Chamroukhi and H. Glotin. Mixture model-based functional discriminant analysis for curve classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Brisbane, Australia, June 2012a

[2] F. Chamroukhi, H. Glotin, and C. Rabouy. Functional Mixture Discriminant Analysis with hidden process regression for curve classification. In *Proceedings of XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 281–286, Bruges, Belgium, April 2012a

[3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Classification automatique de données temporelles en classes ordonnées. In *Actes des 44 ème Journées de Statistique*, Bruxelles, Belgique, Mai 2012c

[4] F. Chamroukhi, A. Samé, P. Aknin, and G. Govaert. Model-based clustering with Hidden Markov Model regression for time series with regime changes. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 2814–2821, Jul-Aug 2011a

This research direction that I initiated at the end of my PhD thesis in 2010 with the development of a functional classification models for temporal data, and pursued, under additional modeling considerations, until now, focuses on functional data analysis (FDA) where the individuals are entire functions. This theme is structured into two sub-axes. In the first, I seek to provide solutions to the problem of unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes. This has lead mainly to three contributions: [J-4], [J-9], and [C-11] [J-16]. In the second sub-axis, which concerns the supervised case, that is, discrimination of functional data, I am interested in building discriminant analyses that handle the problem of classification of functional data that might be organized in homogeneous or heterogeneous groups and further exhibit a non-stationary behavior due to regime changes. This has lead to mainly two following contributions [J-2][J-5].

3.1 Introduction

Most statistical analyses involve vectorial data where the observations are finite dimensional vectors. However, in many application domains, such as diagnosis of complex systems [J-2][J-4], electrical engineering (Hébrail et al., 2010), speech recognition (e.g. the phoneme data studied in (Delaigle et al., 2012)), medicine for example with the study of human growth curves (Liu and Yang, 2009), bioinformatics (Gui and Li, 2003), spectroscopy (Ferraty and Vieu, 2002), etc, the individuals are described by entire functions (.i.e curves) rather than finite dimensional vectors. The most frequent case is that in which the studied individuals have a temporal variability. The functional representations are also present in other cases which are not necessarily temporal, for example, spectroscopy in which samples are charactered by their spectrum, a function that, at a wavelength associates a measurement of interest such as the absorbance (e.g. near infrared (NIR) absorbance spectra[1], see Ferraty and Vieu (2002)). This "functional" aspect of the data adds additional difficulties in the analysis compared to the case of a classical multivariate (non functional) analysis, which ignores the structure of individuals, and there is therefore a need to formulate "functional" models that explicitly integrate the functional form of the data, rather than directly and simply considering them as vectors to apply classical multivariate analysis methods, which is of course possible, but is subject to a strong loss of information. The paradigm of analyzing such data is known as functional data analysis (FDA) (Ramsay and Silverman, 2005, 2002; Ferraty and Vieu, 2006). The key tenet of FDA is to treat the data not just as multivariate observations but as (discretized) values of possibly smooth functions. FDA is indeed the paradigm of data analysis in which the individuals are functions (e.g., curves or surfaces) rather than vectors of reduced dimension and the statistical approaches for FDA allow to fully exploit the structure of the data.

The goals of FDA, like in multivariate data analysis, may be exploratory for example clustering or segmentation prospectives when the curves arise from sub-populations, or when each individual functional itself presents a heterogeneity aspect say for example a non-stationary temporal curve, or decisional for example to make prediction on future data, that is, regression when predicting continuous responses or classification when predicting categorical ones. Additional background on FDA, examples and analysis techniques can be found for example in Ramsay and Silverman (2005). In this FDA field, I considered the problems of functional data clustering, segmentation and classification. The methods on which I focus here rely on generative functional regression models which are based on the finite mixture formulation with tailored component densities. Finite mixture models (McLachlan and Peel., 2000; Frühwirth-Schnatter, 2006; Titterington et al., 1985), known in multivariate analysis by their well-established theoretical background, flexibility, easy interpretation and associated efficient estimation tools in many problems particularly in cluster and discriminant analyses, say the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008), are taking a growing investigation for adapting them to the framework of FDA. See for example (Devijver, 2014; Jacques and Preda, 2014; Bouveyron and Jacques, 2011; Chamroukhi, 2010a; Liu and Yang, 2009; Gaffney and Smyth, 2004; Gaffney, 2004; James and Sugar, 2003; James and Hastie, 2001). Indeed, when the data are curves, which are in general very structured, relying on standard multivariate mixture analysis may lead to unsatisfactory results in terms of modeling and classification accuracy. The classic mixture methods for data analysis, in particular the multivariate (Gaussian) mixture model might be used but ignore the functional structure of the data since they simply assume them as vectors in $\mathbb{R}^m$, which leads to rough approximations and strong loss of information. Addressing the problem from a functional data analysis prospective, that is, formulating functional mixture models, al-

[1] For example the tecator data available at `http://lib.stat.cmu.edu/datasets/tecator`.

lows to better handle the issue as shown for example in [J-1][J-2][J-4](Gaffney and Smyth, 1999; Gaffney, 2004; Gaffney and Smyth, 2004; Liu and Yang, 2009). In this case of model-based curve clustering, one can distinguish the regression mixture approaches (Gaffney and Smyth, 1999; Gaffney, 2004), including polynomial regression and spline regression, or random effects polynomial regression as in Gaffney and Smyth (2004) or (B-)spline regression as in Liu and Yang (2009). When clustering sparsely sampled curves, one may use the mixture approach based on splines in (James and Sugar, 2003). In (Devijver, 2014) and Giacofci et al. (2013), the clustering is performed by spanning the data on wavelet basis instead of (B-)spline ones. Another alternative, which concerns mixture-model based clustering of multivariate functional data, is the one in which the clustering is performed in the space of reduced functional principal components (Jacques and Preda, 2014). One can also mention the K-means based clustering for functional data by using B-spline bases (Abraham et al., 2003) or wavelets as in Antoniadis et al. (2013). ARMA mixtures have also been introduced in Xiong and Yeung (2004) for time series clustering. Beyond these (semi-)parametric approaches, one can also cite non-parametric statistical methods (Ferraty and Vieu, 2003) using kernel density estimators (Delaigle et al., 2012), or those using mixture of Gaussian processes regression (Shi et al., 2005; Shi and Wang, 2008; Shi and Choi, 2011) of hierarchical Gaussian process mixtures for regression (Shi and Choi, 2011; Shi et al., 2005).

In functional data discrimination, the generative approaches for functional data related to this work are essentially based on functional linear discriminant analysis using splines, including B-splines as in James and Hastie (2001), of mixture discriminant analysis (Hastie and Tibshirani, 1996) in the context of functional data by relying on B-spline bases as in Gui and Li (2003). Delaigle et al. (2012) have also addressed the functional data discrimination problem from an non-parametric prospective using a kernel based method.

3.1.1 Personal contribution

My personal contribution to FDA consists in studying latent data models, particularly the finite mixture modeling in the framework of functional data and proposing models to deal with the problem of i) functional data clustering [J-4][J-5][J-9][J-16][C-11] and ii) the one of functional data discrimination [J-2][J-5], particularly when the data present a complex structure due to regime changes. More specifically, I proposed mixture-model based cluster and discriminant analyzes based on latent processes. The heterogeneity of a population of functions arising in several sub-populations is naturally accommodated by a mixture distribution, and the dynamic behavior within each subpopulation, generated by a non-stationary process typically governed by a regime change, is captured via a dedicated latent process. Here the latent process is modeled by either a Markov chain or a logistic process, or as a deterministic piecewise segmental process. I first investigated the use of a mixture model with piecewise regression components (PWRM) for simultaneous clustering and segmentation of univariate regime changing functions [J-9]. Then, I formulated the problem from a full generative prospective by proposing the mixture of hidden Markov model regressions (MixHMMR) [C-11][J-16] and the mixture of regressions with hidden logistic processes (MixRHLP) [J-4][J-5] which offers additional attractive features including the possibility to deal with possible smooth dynamics within the curves. I also investigated discriminant analyzes for homogeneous ones [J-2] as well as for heterogeneous curves [J-5]. For each model, I also studied and evaluated the algorithmic tools, which mainly consist in EM algorithms, on real data from different application domains.

The remainder of this chapter is organized as follows. After giving the general modeling framework in subsection 3.1.2, I derive the proposed finite mixtures for simultaneous functional data clustering and segmentation. The PWRM model is presented in Section 3.2. Then, Section 3.3 presents the MixHMMR model and Section 3.4 is dedicated to the MixRHLP. Finally, In Section 3.5, I derive the proposed functional discriminant analyses, in particular, the functional mixture discriminant analysis with hidden process regression (FMDA).

3.1.2 Mixture modeling framework for functional data

The considered modeling framework for the analysis is the one of the finite mixture model. The finite mixture model decomposes the density of the observed data as a convex sum of a finite number of component densities. In multivariate analysis, the most frequently used model is the finite Gaussian

mixture model (GMM) in which each mixture component is a multivariate Gaussian. In what follows, I will specify the global mixture distribution whose components are tailored to functional data modeling. Let $\boldsymbol{D} = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ be a set of n independent individuals describing functions (e.g curves) where each individual $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ consists of m_i observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ regularly observed at the independent covariates, for example the time in time series, $(x_{i1}, \ldots, x_{im_i})$ with $x_{i1} < \ldots < x_{im_i}$. The mixture model for such functional data, which will referred to hereafter as the "functional mixture model", assumes that the observed pairs $(\boldsymbol{x}, \boldsymbol{y})$ are generated from K tailored functional components, more particularly, tailored regressors explaining the response $\boldsymbol{y}$ by the covariate $\boldsymbol{x}$, and are governed by a hidden categorical random variable Z indicating from which component each function is generated. Thus, the functional mixture model can be defined as:

$$f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Psi}_k) \tag{3.1}$$

where the α_k's defined by $\alpha_k = \mathbb{P}(Z = k)$ are the mixing proportions such that $\alpha_k > 0$ for each k and $\sum_{k=1}^{K} \alpha_k = 1$, $\boldsymbol{\Psi}_k$ $(k = 1, \ldots, K)$ is the parameter vector of the kth component density and $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ is the parameter vector of the functional mixture model.

In mixture modeling for FDA, the component densities $f_k(y|\boldsymbol{x})$ may be the ones of polynomial (B-)spline regression, regression using wavelet bases etc or Gaussian process regressions. As we focus on functions arising in curves with regimes changes, possibly smooth, these regression mixture models, as mentioned before, however do not address the problem of regime changes. For the (hierarchical) mixture of Gaussian processes for functional regression (Shi and Choi, 2011) which might be used in this context as a non parametric alternative, they are more tailored to approximate smooth functions, and would not be suited to deal with possible abrupt regime changes. For curves with regime changes, they only provide a non-linear smooth approximation, without account for the segmentation.

In the models I present, the mixture component density $f_k(y|\boldsymbol{x})$ is itself assumed to exhibit a complex structure arising in sub-components, each one is associated with a regime. In what follows, we investigate mainly three choices for this component specific density, that is, first a piecewise regression density (PWR), then a hidden Markov regression (HMMR) density and finally a regression model with hidden logistic process (RHLP) density.

3.2 Mixture of piecewise regressions

The idea described here and proposed in [J-9] is in the same spirit of the one proposed byHugueney et al. (2009); Hébrail et al. (2010) for curve clustering and optimal segmentation based on a piecewise regression model that allows for fitting several constant (or polynomial) models to each cluster of functional data with regime changes. However, unlike the distance-based approach, which uses a K-means-like algorithm (Hugueney et al., 2009; Hébrail et al., 2010), the proposed model provides a general probabilistic framework to address the problem. Indeed, in the proposed approach, the piecewise regression model is included into a mixture framework to generalize the deterministic K-means like approach. As a result, both fuzzy clustering and hard clustering are possible. I also provide two algorithms for learning the model parameters. The first one is a dedicated EM algorithm to find a fuzzy partition of the data and an optimal segmentation by maximizing the observed-data log-likelihood. The EM version being the natural way to the maximum likelihood estimation of a mixture model, including the proposed piecewise regression mixture. The second algorithm consists in maximizing a specific classification likelihood criterion by using a dedicated CEM algorithm in which the curves are partitioned in a hard way and optimally segmented simultaneously as the learning proceeds. The K-means-like algorithm of Hébrail et al. (2010) is shown to be a particular case of the proposed CEM algorithm if some constraints are imposed on the piecewise regression mixture.

3.2.1 The model

The piecewise regression mixture model (PWRM) assumes that each curve $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $(i = 1, \ldots, n)$ is generated by a piecewise regression model among K models, with a prior probability α_k, that is, each

component density in (3.1) is the one of a piecewise regression model, defined by:

$$f_k(\boldsymbol{y}_i|Z_i = k, \boldsymbol{x}_i; \boldsymbol{\Psi}_k) = \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2) \tag{3.2}$$

where $I_{kr} = (\xi_{kr}, \xi_{k,r+1}]$ represents the element indices of segment (regime) r ($r = 1, \ldots, R_k$) for component (cluster) k, R_k being the corresponding number of segments, $\boldsymbol{\beta}_{kr}$ is the vector of its polynomial coefficients and σ_{kr}^2 the associated Gaussian noise variance. Thus, the PWRM density if defined by:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2), \tag{3.3}$$

where the 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 $\boldsymbol{\xi}_k = (\xi_{k1}, \ldots, \xi_{k,R_k+1})^T$ are respectively the vector of all the polynomial coefficients and noise variances, and the vector of transition points which correspond to the segmentation of cluster k. The proposed mixture model is therefore suitable for clustering and optimal segmentation of complex shaped curves. More specifically, by integrating the piecewise polynomial regression into the mixture framework, the resulting model is able to approximate curves issued from different groups. Furthermore, the problem of regime changes within each cluster of curves is addressed as well thanks to the optimal segmentation provided by dynamic programming for each piecewise regression component. These two simultaneous outputs are clearly not provided by the standard generative curve clustering approaches namely the regression mixtures. On the other hand, the PWRM is a probabilistic model and as it will be shown in the following, generalizes the deterministic K-means-like algorithm.
I derived two approaches for learning the model parameters. The former is an estimation approach and consists in maximizing the observed-data likelihood via a dedicated EM algorithm. A fuzzy partition of the curves in K clusters is then obtained at convergence by maximizing the posterior component probabilities. The latter however focuses on the classification and optimizes a specific classification likelihood criterion through a dedicated CEM algorithm. The optimal curve segmentation is performed by using dynamic programming. In the classification approach, both the curve clustering and the optimal segmentation are performed simultaneously as the CEM learning proceeds. I showed that the classification approach using the PWRM model with the CEM algorithm is the probabilistic version that generalizes the deterministic K-means-like algorithm proposed in Hébrail et al. (2010).

3.2.2 Maximum likelihood estimation via a dedicated EM

In this estimation (maximum likelihood) approach, the parameter estimation is performed by monotonically maximizing the observed-data log-likelihood

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2\right) \tag{3.4}$$

iteratively via the EM algorithm [J-9]. In this EM framework, the complete-data log-likelihood which will be denoted by $\log L_c(\boldsymbol{\Psi}, \mathbf{z})$, and which represents the log-likelihood of the parameter vector given the observed data, completed by the unknown variables representing the component labels $\mathbf{Z} = (Z_1, \ldots, Z_n)$ with $Z_i \in \{1, \ldots, K\}$ the label of the ith individual, is given by:

$$\log L_c(\boldsymbol{\Psi}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \alpha_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{j \in I_{kr}} Z_{ik} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2) \tag{3.5}$$

where Z_{ik} is an indicator binary-valued variable such that $Z_{ik} = 1$ iff $Z_i = k$ (i.e., if and only if the ith curve is generated from component k). The EM algorithm for the PWRM model (EM-PWRM) alternates between the two following steps until convergence (e.g., when there is no longer change in the relative variation of the log-likelihood):

The E-step computes the expected complete-data log-likelihood given the observed curves $\boldsymbol{D} = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ and the current value of the parameter vector $\boldsymbol{\Psi}^{(q)}$:

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \mathbb{E}\big[\log L_c(\boldsymbol{\Psi}; \boldsymbol{D}, \mathbf{z}) | \boldsymbol{D}; \boldsymbol{\Psi}^{(q)}\big] = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k}\sum_{j \in I_{kr}} \tau_{ik}^{(q)} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_{ij}, \sigma_{kr}^2) \tag{3.6}$$

where

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(q)}) = \frac{\alpha_k^{(q)} f_k\big(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)}\big)}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} f_{k'}\big(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}_{k'}^{(q)}\big)} \tag{3.7}$$

is the posterior probability that the curve $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ belongs to cluster k. This step therefore only requires the computation of the posterior component membership probabilities $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$ for each of the K components.

The M-step computes the parameter vector update $\boldsymbol{\Psi}^{(q+1)}$ by maximizing the Q-function with respect to $\boldsymbol{\Psi}$, that is: $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$. The mixing proportions are updated as in standard mixtures and their updates are given by:

$$\alpha_k^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n}, \quad (k = 1, \ldots, K). \tag{3.8}$$

The maximization of the Q-function w.r.t $\boldsymbol{\Psi}_k$, that is, w.r.t the piecewise segmentation $\{I_{kr}\}$ of component (cluster) k and the corresponding piecewise regression representation through $\{\boldsymbol{\beta}_{kr}, \sigma_{kr}^2\}$, $(r = 1, \ldots, R_k)$, corresponds to a weighted version of the piecewise regression problem for a set of homogeneous as described in [J-9], with the weights being the posterior component membership probabilities $\tau_{ik}^{(q)}$. The maximization simply consists in solving a weighted piecewise regression problem where the optimal segmentation of each cluster k, represented by the parameters $\{\boldsymbol{\xi}_{kr}\}$ is performed by running a dynamic programming procedure Finally, the regression parameters are updated as:

$$\boldsymbol{\beta}_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_{ir}^T \mathbf{X}_{ir}\Big]^{-1} \sum_{i=1}^{n} \mathbf{X}_{ir} \boldsymbol{y}_{ir} \tag{3.9}$$

$$\sigma_{kr}^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j \in I_{kr}^{(q)}} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \|\boldsymbol{y}_{ir} - \mathbf{X}_{ir} \boldsymbol{\beta}_{kr}^{(q+1)}\|^2 \tag{3.10}$$

where $\boldsymbol{y}_{ir}$ is the segment (regime) r of the ith curve, that is the observations $\{y_{ij} | j \in I_{kr}\}$ and $\mathbf{X}_{ir}$ its associated design matrix with rows $\{\boldsymbol{x}_{ij} | j \in I_{kr}\}$.

Thus, the proposed EM algorithm for the PWRM model provides a fuzzy partition of the curves into K clusters through the posterior cluster probabilities τ_{ik}, each fuzzy cluster is optimally segmented into regimes with indices $\{I_{kr}\}$. At convergence of the EM algorithm, a hard partition of the curves can then be deduced by assigning each curve to the cluster that maximizes the posterior probability (3.7), that is:

$$\hat{z}_i = \arg \max_{1 \leq k \leq K} \tau_{ik}(\hat{\boldsymbol{\Psi}}), \quad (i = 1, \ldots, n) \tag{3.11}$$

where $\hat{z}_i$ denotes the estimated cluster label for the ith curve.

3.2.3 Maximum classification likelihood estimation via a dedicated CEM

As mentioned before, in addition to the MLE approach via EM, here I present another scheme to achieve both the model estimation (including the segmentation) and the clustering. It consists in a maximum classification likelihood approach which uses the Classification EM (CEM) algorithm. The CEM algorithm (see for example (Celeux and Govaert, 1992)) is the same as the so-called classification maximum likelihood approach as described earlier in McLachlan (1982) and dates back to the work of Scott and Symons (1971). The CEM algorithm was initially proposed for model-based clustering of multivariate data. We adopt it here in order to perform model-based curve clustering with the proposed PWRM model. The resulting CEM simultaneously estimates both the PWRM parameters and the class labels by maximizing the complete-data log-likelihood (3.5) w.r.t both the model parameters $\boldsymbol{\Psi}$ and the partition represented by the vector of cluster labels $\mathbf{z}$, in an iterative manner, by alternating between the two following steps:

Step 1 update the cluster labels for the current model defined by $\boldsymbol{\Psi}^{(q)}$ by maximizing the complete-data log-likelihood (3.5) w.r.t the cluster labels $\mathbf{z}$, that is: $\mathbf{z}^{(q+1)} = \arg\max_{\mathbf{z}} \log L_c(\boldsymbol{\Psi}^{(q)}, \mathbf{z})$.

Step 2 Given the estimated partition defined by $\mathbf{z}^{(q+1)}$, update the model parameters by maximizing (3.5) w.r.t to the PWRM parameters $\boldsymbol{\Psi}$: $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} \log L_c(\boldsymbol{\Psi}, \mathbf{z}^{(q+1)})$. Equivalently, the CEM algorithm consists in integrating a classification step (C-step) between the E- and the M- steps of the EM algorithm presented previously. The C-step computes a hard partition of the n curves into K clusters by applying the Bayes' optimal allocation rule (3.11). The only difference between this CEM algorithm and the previously derived EM one is that the posterior probabilities τ_{ik} in the case of the EM-PWRM algorithm are replaced by the cluster label indicators Z_{ik} in the CEM-PWRM algorithm; The curves being assigned in a hard way rather than in a soft way. By doing so, the CEM monotonically maximizes the complete-data log-likelihood (3.5).

Another attractive feature of the proposed PWRM model is that when it is estimated by the CEM algorithm, as shown in [J-9], it is equivalent to a probabilistic generalization for the K-means-like algorithm of Hébrail et al. (2010). Indeed, maximizing the complete-data log-likelihood (3.5) optimized by the proposed CEM algorithm for the PWRM model, is equivalent to minimizing the following distortion criterion w.r.t the cluster labels $\mathbf{z}$, the segments indices I_{kr} and the segments constant means μ_{kr}, which is exactly the criterion optimized by the K-means-like algorithm:

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

if the following constraints are imposed:

- $\alpha_k = \frac{1}{K}\ \forall K$ (identical mixing proportions);
- $\sigma_{kr}^2 = \sigma^2\ \forall r = 1, \ldots, R_k$ and $\forall k = 1, \ldots, K$; (isotropic and homoskedastic model);
- piecewise constant approximation of each segment rather than a polynomial approximation.

The proposed CEM algorithm for piecewise polynomial regression mixture is therefore the probabilistic version for hard curve clustering and optimal segmentation of the K-means-like algorithm.

Model selection The problem of model selection here is equivalent to the one of choosing the optimal number of mixture components K, the number of regimes R and the polynomial degree p. The optimal value of the triplet (K, R, p) can be computed by using some model selection criteria such as the BIC (Schwarz, 1978), the ICL (Biernacki et al., 2000), etc. Let us recall that BIC is a penalized log-likelihood criterion which can be defined as a function to be maximized as: $\text{BIC}(K, R, p) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\nu_{\boldsymbol{\Psi}} \log(n)}{2}$, while ICL consists in a penalized complete-data log-likelihood and can be expressed as: $\text{ICL}(K, R, p) = \log L_c(\hat{\boldsymbol{\Psi}}) - \frac{\nu_{\boldsymbol{\Psi}} \log(n)}{2}$, where $\log L(\hat{\boldsymbol{\Psi}})$ and $\log L_c(\hat{\boldsymbol{\Psi}})$ are respectively the incomplete (observed) data log-likelihood and the complete data log-likelihood, obtained at convergence of the (C)EM algorithm, $\nu_{\boldsymbol{\Psi}} = \sum_{k=1}^{K} R_k(p+3) - 1$ is the number of free parameters of the model and n is the sample size. The number of free model parameters includes $K-1$ mixing proportions, $\sum_{k=1}^{K} R_k(p+1)$ polynomial coefficients, $\sum_{k=1}^{K} R_k$ noise variances and $\sum_{k=1}^{K}(R_k - 1)$ transition points.

3.2.4 Experiments

The performance of the PWRM with both the EM and CEM algorithms is studied in [J-9] by comparing it to the polynomial regression mixture models (PRM) (Gaffney, 2004), the standard polynomial spline regression mixture model (PSRM) (Gaffney, 2004; Gui and Li, 2003; Liu and Yang, 2009) and the piecewise regression model used with the K-means-like algorithm (Hébrail et al., 2010). I also included comparisons with standard model-based clustering of multivariate data including the GMM. The algorithms have been evaluated in terms of curve classification and approximation accuracy. The used evaluation criteria are the classification error rate between the true partition (when it is available) and the estimated partition, and the intra-cluster inertia.

In the simulation studies, in summary, for some situations, when all the algorithms retrieve the actual partition, which is possible for obvious partitions, in terms of curves approximation, we clearly saw that, on the one hand, the standard model-based clustering using the GMM is not adapted as it does not take into account the functional structure of the curves and therefore does not account for the smoothness,

they rather compute an overall mean curve. On the other hand, the proposed probabilistic model, when trained with the EM algorithm (EM-PWRM) or with the CEM algorithm (CEM-PWRM), as well as the K-means-like one of Hébrail et al. (2010), as expected, provide the quasi identical results in terms of clustering and segmentation. This is attributed to the fact that the K-means PWRM approach is a particular case of the proposed probabilistic approach. The best curves approximation are however those provided by the PWRM models. The GMM mean curves are simply over all means, and the PRM and the PSRM models, as they are based on continuous curve prototypes, do not account for the segmentation, unlike the PWRM models which are well adapted to perform simultaneous curve clustering and segmentation. When we varied the noise level, for a small noise level variation, the results are very similar. However, as the noise level increases, the misclassification error rate increases faster for the other models compared to the proposed PWRM model. The EM and the CEM algorithm for the proposed approach provide very similar results with a slight advantage for the CEM version, which can be attributed to the fact that CEM is by construction tailored to the classification end. When the proposed PWRM approach is used, the misclassification error can be improved by 4% compared to the K-means like approach, about 7% compared to both the PRM and the PSRM, an more that 15% compared to the standard multivariate GMM. In addition, when the data have non-uniform mixing proportions, the K-means based approach can fail namely in terms of segmentation. This is attributed to the fact that the K-means-like approach for PWRM is constrained as it assumes the same proportion for each cluster, and does not sufficiently take into account the heteroskedasticity within each cluster compared to the proposed general probabilistic PWRM model. For the model selection, the ICL was used on simulated data. We remarked that when using the proposed EM-PWRM and CEM-PWRM approaches, the actual model may be selected up to more than 10% cases compared to when using the K-means-like algorithm for piecewise regression. The number of regimes was underestimated with only around 10% for the proposed EM and CEM algorithms, while the number of clusters is correctly estimated. However, the K-means-like approach overestimates the number of clusters in 12% of cases. These results highlight an advantage of the fully probabilistic approach compared to the one based on the K-means-like approach.

Application to real curves In [J-9] the model was also applied on real curves issued from three different data sets, other railway switch curves (different from those used in the previous chapter), the Tecator curves, and the Topex/Poseidon satellite data as studied in Hébrail et al. (2010). The actual partitions for these data are however unknown and we used the intra-class inertia as well as a qualitative assessment of the results. The first studied curves are the railway switch curves issued from a railway diagnosis application of the railway switch. Roughly, the railway switch is the component that enables (high speed) trains to be guided from one track to another at a railway junction, and is controlled by an electrical motor. The considered curves are the signals of the consumed power during the switch operations. These curves present several changes in regime due to successive mechanical motions involved in each switch operation. A preliminary data preprocessing task is to automatically identify homogeneous groups (typically, curves without defect and curves with possible defect (we assumed $K = 2$). The used database is composed of $n = 146$ real curves of $m = 511$ observations. The number of regression components was set to $R = 6$ in accordance with the number of electromechanical phases of these switch operations and the degree of the polynomial regression p was set to 3 which is appropriate for the different regimes in the curves. The obtained results show that, for the CEM-PWRM approach, the curves the two obtained clusters do not have the same characteristics with quite clearly different shapes and may correspond to two different states of the switch mechanism. According to the experts, this can be attributed to a default in the measurement process, rather than a default of the switch itself. The device used for measuring the power would have been used slightly differently for this cluster of curves. The intra-class inertia results are also significantly better compared to the standard alternatives. This confirms that the piecewise regression mixture model has an advantage for giving homogeneous and well approximated clusters from curves of regime changes.

The second data set is the Tecator data[1] which consist of near infrared (NIR) absorbance spectra of 240 meat samples. The NIR spectra are recorded on a Tecator Infratec food and feed Analyzer working in the wavelength range $850 - 1050$ nm. The full Tecator data set contains $n = 240$ spectra with $m = 100$ for each spectrum. This data set has been considered in Hébrail et al. (2010) and in our experiment we consider the same setting, that the data set is summarized with six clusters ($K = 6$), each cluster being composed of five linear regimes (segments) ($R = 5, p = 1$). The retrieved clusters are informative (see

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

Fig. 3.1) in the sense that the shapes of the clusters are clearly different, and the piecewise approximation is in concordance with the shape of each cluster. On the other hand, the obtained result is very close to the one obtained by Hébrail et al. (2010) by using the K-means-like approach. This is not surprising and confirms that the proposed CEM-PWRM algorithm is a probabilistic alternative for the K-means-like approach.

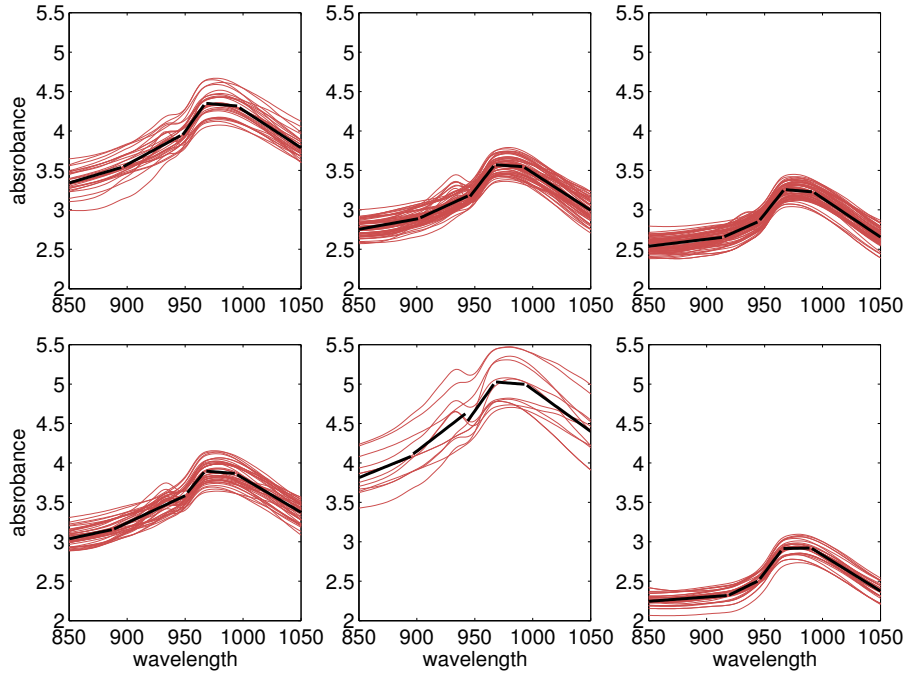

Figure 3.1: Clusters and the corresponding piecewise prototypes for each cluster obtained with the CEM-PWRM algorithm for the Tecator data set.

The third data set is the Topex/Poseidon radar satellite data[1] which were registered by the satellite Topex/Poseidon around an area of 25 kilometers upon the Amazon River and contain $n = 472$ 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 used in Hébrail et al. (2010). We note that, in our approach, we directly apply the proposed CEM-PWRM algorithm to raw the satellite data without a preprocessing step. However, in Hébrail et al. (2010), the authors used a two-fold scheme. They first perform a topographic clustering step using the Self Organizing Map (SOM), and then apply their K-means-like approach to the results of the SOM. The proposed CEM-PWRM algorithm for the satellite data provide clearly informative clustering and segmentation which reflect the general behavior of the hidden structure of this data set (see Fig. 3.2). The structure is indeed more clear with the mean curves of the clusters (prototypes) than with the raw curves. The piecewise approximation thus helps to better understand the structure of each cluster of curves from the obtained partition, and to more easily infer the general behavior of the data set. On the other hand, the result is similar to the one found in Hébrail et al. (2010). Most of the profiles are present in the two results. There is a slight difference that can be attributed to the fact that the result in Hébrail et al. (2010) is provided from a two-stage scheme which includes and additional pre-clustering step using the SOM, instead of directly applying the piecewise regression model to the raw data.

3.2.5 Conclusion

Here I introduced a new probabilistic approach based on a piecewise polynomial regression mixture (PWRM) for simultaneous clustering and optimal segmentation of curves with regime changes. I provided two algorithms to learn the model. The first (EM-PWRM) consists of using the EM algorithm to maximize the observed data log-likelihood and the latter (CEM-PWRM) is a CEM algorithm to maximize

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

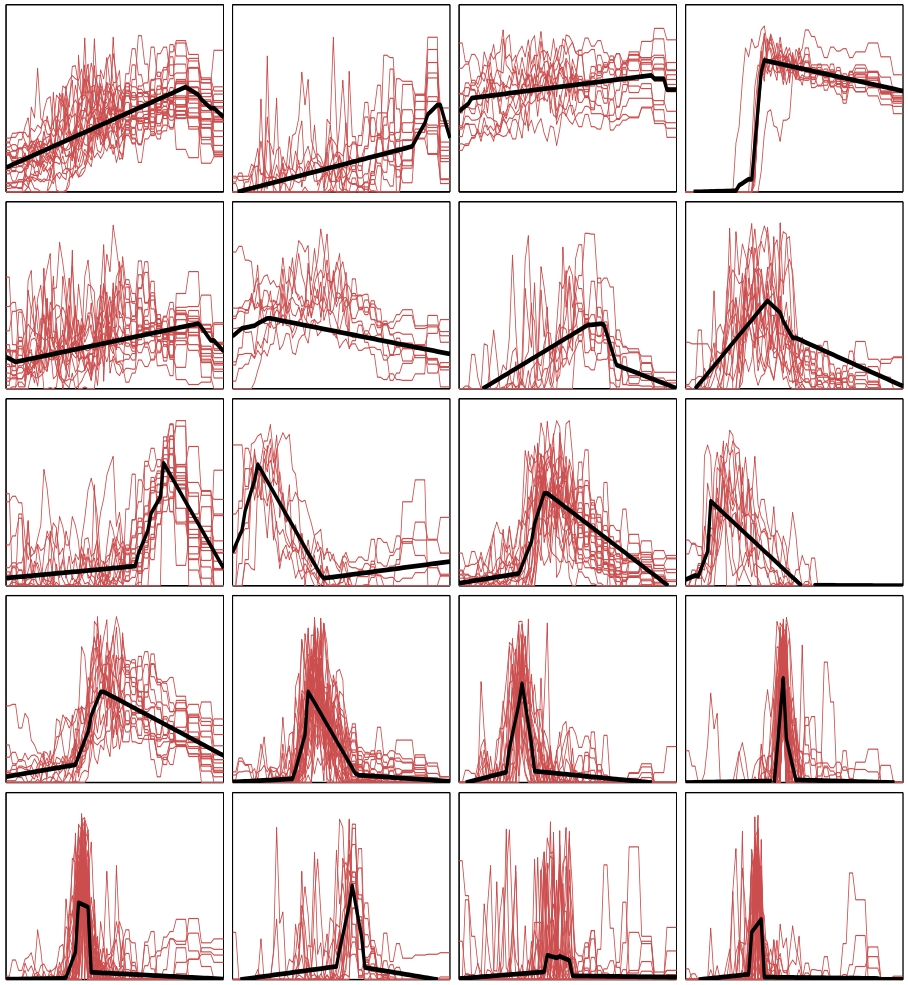

Figure 3.2: Clusters and the corresponding piecewise prototypes for each cluster obtained with the CEM-PWRM algorithm for the satellite data set.

the complete-data log-likelihood. I showed that the CEM-PWRM algorithm is a general probabilistic-based version of possible K-means-like algorithm. However, it is worth to mention that if the aim is density estimation, the EM version is suggested since the CEM provides biased estimators but is well-tailored to the segmentation/clustering end. The obtained results demonstrated the benefit of the proposed approach in terms of both curve clustering and piecewise approximation and segmentation of the regimes of each cluster. In particular, the comparisons with the K-means-like algorithm approach confirm that the proposed CEM-PWRM is an interesting probabilistic alternative. We note that in some practical situations involving continuous functions the proposed piecewise regression mixture, in its current formulation, may lead to discontinuities between segments for the piecewise approximation. This may be avoided by slightly modifying the algorithm by adding an interpolation step as performed in Hébrail et al. (2010). We also note that in this work we are interested in piecewise regimes which dot not overlap; only the clusters can overlap. However, one way to address the regime overlap is to use more segments so that a regime that overlaps (for example it occurs in two different time ranges) can be treated as two sub-regimes. These two reconstructed non-overlapping regimes would have very close characteristics so that as to correspond to a single overlapping regime. In terms of computing time, I mention that in some situations, especially for large sample sizes and large value of the number segments, the algorithms may lead to significant computational load. However, for quite reasonable dimensions, the algorithms remain usable without significant difficulty.

3.3 Mixture of hidden Markov model regressions

The mixture of piecewise regressions presented previously can be seen as not being completely generative, since the transition points, while assumed unknown and determined automatically from the data, are not governed by a probability distribution. This however achieves the clustering and segmentations tasks and was useful at least to show that K-means based alternatives may be particular cases of such model with even so a probabilistic dimension thanks to its mixture formulation. The aim now is to build a full generative model. It is natural to think, as previously for the univariate case, that for each group the regimes governing the observed curves follow a discrete hidden process, typically a hidden Markov chain. By doing so, it is assumed that, within each cluster k, the observed curve is governed by a hidden process which enables for switching from one state to another among R_k states following a homogeneous Markov chain, which leads to the mixture of hidden Markov models introduced by Smyth (1996). Two different approaches can be adopted for estimating this mixture of HMMs. The first one is the K-means-like approach for hard clustering used in Smyth (1996) and in which the optimized function is the complete-data log-likelihood. The resulting clustering scheme consists of assigning sequences to clusters at each iteration and using only the sequences assigned to a cluster for re-estimation of the HMM parameters related to that cluster. The second one is the soft clustering approach described in Alon et al. (2003) where the model parameters are estimated in a maximum likelihood framework by the EM algorithm. The model I propose here can be seen as an extension of the model-clustering approach using mixture of standard HMMs introduced by Smyth (1996), where each HMM state has a conditional Gaussian density with simple scalar mean, by considering polynomial regressors, and by performing a MLE using EM, rather that K-means. In addition, the use of polynomial regime modeling rather than simple constant means should be indeed more suitable for fitting the non-linear regimes governing the time series, and the EM fitting should better capture the uncertainty regarding the curve assignments thanks to the fuzzy posterior component memberships. This results into the mixture of hidden Markov model regressions (MiXHMMR) [C-11][J-16].

3.3.1 The model

The proposed mixture of HMM regressions (MixHMMR) assumes that each curve is issued from one of K-component mixture where, conditional on each component k ($k = 1, \ldots, K$), the curve is distributed according to an R_k-state hidden Markov model regression. That is, given the label $Z_i = k$ of the component generating the ith curve, and given the state $H_{ij} = r$ ($r = 1, \ldots, R_k$), the jth observation y_{ij} (e.g., the one observed at time t_j in the case of temporal data) is generated according to a Gaussian polynomial regression model with regression coefficient vector $\boldsymbol{\beta}_{kr}$ and noise variance σ^2_{kr}:

$$y_{ij} = \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j + \sigma_{kr}\epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, 1) \tag{3.12}$$

where $\boldsymbol{x}_j$ is a covariate vector, the ϵ_{ij} are independent random variables distributed according to a standard zero-mean unit-variance Gaussian distribution and the hidden state sequence $\boldsymbol{H}_i = (H_{i1}, \ldots, H_{im})$ for each mixture component k is assumed to be Markov chain with initial state distribution $\boldsymbol{\pi}_k$ with components $\pi_{kr} = \mathbb{P}(H_{i1} = r|Z_i = k)$ ($r = 1, \ldots, R_k$) and transition matrix $\mathbf{A}_k$ whose general term is $A_{k\ell r} = \mathbb{P}(H_{ij} = r|H_{i,j-1} = \ell, Z_i = k)$. Thus, the change from one regime to another is governed by the hidden Markov Chain. Note that if the time series we aim to model consist of successive contiguous regimes, one may use a left-right model (Rabiner and Juang, 1986; Rabiner, 1989) by imposing order constraints on the hidden states via constraints on the transition probabilities. From (3.12), it follows that the response $\boldsymbol{y}_i$ for the predictor (bsx_i), conditional on each mixture component $Z_i = k$ is therefore distributed according to a HMM regression distribution, defined by:

$$f_k(\boldsymbol{y}_i|Z_i = k, \boldsymbol{x}_i; \boldsymbol{\Psi}_k) = \sum_{\boldsymbol{H}_i} \mathbb{P}(H_{i1}; \boldsymbol{\pi}_k) \prod_{j=2}^{m_i} \mathbb{P}(H_{ij}|H_{i,j-1}; \boldsymbol{A}_k) \times \prod_{j=1}^{m_i} \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kh_{ij}}^T \boldsymbol{x}_j, \sigma^2_{kh_{ij}}) \tag{3.13}$$

with parameter vector $\boldsymbol{\Psi}_k = (\boldsymbol{\pi}_k^T, \mathrm{vec}(\boldsymbol{A}_k)^T, \boldsymbol{\beta}_{k1}^T, \ldots, \boldsymbol{\beta}_{kR}^T, \sigma^2_{k1}, \ldots, \sigma^2_{kR})^T$. Finally, the distribution of a curve $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is defined by the following MixHMMR density:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{y}_i|Z_i = k, \boldsymbol{x}_i; \boldsymbol{\Psi}_k) \tag{3.14}$$

described by the parameter vector $\boldsymbol{\Psi} = (\alpha_1, \ldots, \alpha_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$.

3.3.2 Maximum likelihood estimation via a dedicated EM

The MixHMMR parameter vector $\boldsymbol{\Psi}$ is estimated by monotonically maximizing the observed-data log-likelihood

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \sum_{\boldsymbol{H}_i} \mathbb{P}(H_{i1}; \boldsymbol{\pi}_k) \prod_{j=2}^{m_i} \mathbb{P}(H_{ij}|H_{i,j-1}; \boldsymbol{A}_k) \times \prod_{j=1}^{m_i} \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kh_{ij}}^T \boldsymbol{x}_j, \sigma_{kh_{ij}}^2) \tag{3.15}$$

by using a dedicated EM algorithm as devoloped in [C-11][J-16]. By introducing the two following indicator binary variables for indicating the cluster memberships and the regime memberships for a given cluster, that is, $Z_{ik} = 1$ if $Z_i = k$ (i.e., $\boldsymbol{y}_i$ belongs to cluster k) and $Z_{ik} = 0$ otherwise, and $H_{ijr} = 1$ if $H_{ij} = r$ (i.e., the ith time series $\boldsymbol{y}_i$ belongs to cluster k and its jth observation y_{ij} at time t_j belongs to regime r) and $H_{ijr} = 0$ otherwise, the complete-data likelihood of $\boldsymbol{\Psi}$ can be written as:

$$\log L_c(\boldsymbol{\Psi}) = \sum_{k=1}^{K} \Big[\sum_{i} Z_{ik} \log \alpha_k + \sum_{i,r} Z_{ik} H_{i1r} \log \pi_{kr} + \sum_{i,j=2,r,\ell} Z_{ik} H_{ijr} H_{i(j-1)\ell} \log A_{k\ell r} + \sum_{i,j,r} Z_{ik} H_{ijr} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2) \Big]. \tag{3.16}$$

The EM algorithm for the MixHMMR model starts from an initial parameter $\boldsymbol{\Psi}^{(0)}$ and alternates between the two following steps until convergence:

The E-Step computes the expected complete-data log-likelihood given the responses $\boldsymbol{y}$, the covariates $\boldsymbol{x}$ and the current value of the parameter vector $\boldsymbol{\Psi}^{(q)}$: $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \mathbb{E}\big[\log L_c(\boldsymbol{\Psi})|\{\boldsymbol{y}, \boldsymbol{x}\}; \boldsymbol{\Psi}^{(q)}\big]$ which is given by:

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \sum_{k,i} \tau_{ik}^{(q)} \log \alpha_k + \sum_{k} \Big[\sum_{r,i} \tau_{ik}^{(q)} \big[\gamma_{i1r}^{(q)} \log \pi_{kr} + \sum_{j=2,\ell} \xi_{ij\ell r}^{(q)} \log A_{k\ell r}\big] + \sum_{r,i,j}^{m_i} \tau_{ik}^{(q)} \gamma_{ijr}^{(q)} \log \mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2) \Big] \tag{3.17}$$

and therefore only requires the computation of the posterior probabilities $\tau_{ik}^{(q)}$, $\gamma_{ijr}^{(q)}$ and $\xi_{ij\ell r}^{(q)}$ defined as:

- $\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(q)})$ is the posterior probability that the ith curve belongs to the kth mixture component;
- $\gamma_{ijr}^{(q)} = \mathbb{P}(H_{ij} = r|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)})$ is the posterior probability of the rth polynomial regime in the mixture component (cluster) k;
- $\xi_{ij\ell r}^{(q)} = \mathbb{P}(H_{ij} = r, H_{i(j-1)} = \ell|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}_k^{(q)})$ is the joint posterior probability of having the regime r at time t_j and the regime ℓ at time t_{j-1} in cluster k.

The E-step probabilities $\gamma_{ijr}^{(q)}$ and $\xi_{ij\ell r}^{(q)}$ for each time series $\boldsymbol{y}_i$ $(i = 1, \ldots, n)$ are computed recursively by using the forward-backward algorithm (see [C-11]Rabiner and Juang (1986); Rabiner (1989)). The posterior cluster probabilities $\tau_{ik}^{(q)}$ are given by:

$$\tau_{ik}^{(q)} = \frac{\alpha_k^{(q)} f_k(\boldsymbol{y}_i|\mathbf{t}; \boldsymbol{\Psi}_k^{(q)})}{\sum_{k'=1}^{K} \alpha_{k'}^{(q)} f_{k'}(\boldsymbol{y}_i|\mathbf{t}; \boldsymbol{\Psi}_{k'}^{(q)})}, \tag{3.18}$$

where the conditional probability distribution $f_k(\boldsymbol{y}_i|\mathbf{t}; \boldsymbol{\Psi}_k^{(q)})$ is the one of an HMM regression likelihood (given by (3.13) and is obtained after the forward procedure like in the standard HMM.

The M-Step computes the parameter vector update $\boldsymbol{\Psi}^{(q+1)}$ by maximizing the expected complete-data log-likelihood, that is $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$. The maximization w.r.t the mixing proportions is the one of a standard mixture model and the updates are given by:

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

The maximization w.r.t the Markov chain parameters $(\boldsymbol{\pi}_k, \mathbf{A}_k)$ correspond to a weighted version of updating the parameters of the Markov chain in a standard HMM where the weights in this case are the posterior component membership probabilities $\tau_{ik}^{(q)}$. The updates are given by:

$$\begin{aligned}\pi_{kr}^{(q+1)} &= \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \gamma_{i1r}^{(q)}}{\sum_{i=1}^{n} \tau_{ik}^{(q)}}, \\ A_{k\ell r}^{(q+1)}) &= \frac{\sum_{i=1}^{n} \sum_{j=2}^{m_i} \tau_{ik}^{(q)} \xi_{ij\ell r}^{(q)}}{\sum_{i=1}^{n} \sum_{j=2}^{m_i} \tau_{ik}^{(q)} \gamma_{ijr}^{(q)}}.\end{aligned}$$

Finally, the maximization w.r.t the regression parameters $\boldsymbol{\beta}_{kr}$ consists in analytically solving weighted least-squares problems and the one w.r.t the noise variances σ_{kr}^2 consists in a weighted variant of the problem of estimating the variance of a univariate Gaussian density. The weights consist in both the posterior cluster probabilities τ_{ik} and the posterior regime probabilities $\gamma_{ijr}^{(q)}$ for each cluster k. The parameter updates are given by:

$$\boldsymbol{\beta}_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \mathbf{X}_i\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{W}_{ikr}^{(q)} \boldsymbol{y}_i, \tag{3.19}$$

$$\sigma_{kr}^{2(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \| \sqrt{\mathbf{W}_{ikr}^{(q)}} (\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{\beta}_{kr}^{(q+1)}) \|^2}{\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathrm{trace}(\mathbf{W}_{ikr}^{(q)})}, \tag{3.20}$$

where $\mathbf{W}_{ikr}^{(q)}$ is an m_i by m_i diagonal matrix whose diagonal elements are the weights $\{\gamma_{ijr}^{(q)}; j = 1, \ldots, m_i\}$. It can be seen that here, the parameters for each regime are updated from the whole curve weighted by the posterior regime memberships $\{\gamma_{ijr}\}$, while in the previously presented piecewise regression model, they are only updated from the observations assigned to that regime, that is, in a hard way. This may better take into account possible uncertainty regarding whether the regime change in abrupt or not.

Complexity of the algorithm The proposed EM algorithm includes forward-backward procedures at the E-step to compute the joint posterior probabilities for the HMM states and the conditional distribution (the HMM likelihood) for each time series. The complexity of the Forward-Backward procedure is the one of a standard R state HMM for univariate n curves of size m. The complexity of this step is of $\mathcal{O}(R^2nm)$ per iteration. In addition, in this regression context, the calculation of the regression coefficients for each regime and for each cluster in the M-step of the EM algorithm requires an inversion of a $(p+1) \times (p+1)$ matrix and n multiplications associated with each curve of length m, which is done with a complexity of $\mathcal{O}(p^3nmRK)$. The proposed EM algorithm has therefore a time complexity of $\mathcal{O}(I_{\mathrm{EM}} K R^2 p^3 nm)$ where I_{EM} is the number of EM iterations, K being the number of clusters.

curves approximation, segmentation and model selection Once the model parameters are estimated, a mean curve can be estimated for each cluster by relying on the so-called smoothed signal in the context of HMMs, that is, the curve wighted by the posterior regime membership probabilities. An approximated cluster "centroid" can be computed as a weighted empirical mean of its smoothed curves as in [C-11]. For the segmentation, the most likely sequence hidden states can be inferred given the observations and the estimated model, by using the Viterbi decoder (Viterbi, 1967) which is performed in a complexity of $\mathcal{O}(mR^2)$.

The model selection task consists in estimating the optimal values of the triplet (K, R, p). This can be performed by maximizing for example the BIC Schwarz (1978) (which was used here) defined by: $\mathrm{BIC}(K, R, p) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\nu(K,R,p)}{2} \log(n)$, where $\hat{\boldsymbol{\Psi}}$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\Psi}$ provided by the EM algorithm, $\nu(K, R, p) = KR(R + p + 2) - 1$ is the number of free parameters of the MixHMMR model. For a left-right model, the number of free parameters reduces to $\nu(K, R, p) = KR(\frac{R+1}{2} + p + 2) - 1$

3.3.3 Experiments

The performance of the developed MixHMMR model was studied in [C-11][J-7] by comparing it to the regression mixture model, the standard mixture of HMMs, as well as two standard multidimensional data clustering algorithms: the EM for Gaussian mixtures and K-means.

Simulation results The evaluation criteria are used in the simulations are the misclassification error rate between the true simulated partition and the estimated partition and the intra-cluster inertia. From the obtained results, it was clearly observed that the proposed approach provides more accurate classification results and smaller intra-class inertias compared to the considered alternatives. For example, the MixHMMR provides a clustering error 3% less than the standard mixture of HMMs, which is the most competitive model, and more than 10% compared to standard multivariate clustering alternatives. Applying the MixHMMR for clustering time series with regime changes also provided accurate results in terms of clustering and approximation of each cluster of time series. This is attributed to the fact that the proposed MixHMMR model, thanks to its flexible generative formulation, addresses better both the problem of time series heterogeneities by the mixture formulation and the dynamical aspect within each homogeneous set of time series by the underlying Markov chain. It was also observed that the standard EM for GMM and standard K-means are not well suitable for this kind of functional data.

Clustering the real time series of switch operations The model was also applied in [C-11][J-16] to a real problem of clustering time series issued from a railway diagnosis application. The aim is to discover non-normal time series for a diagnosis prospective. The data set contains 115 curves, each of them results from a process of $R = 6$ operations electromechanical process. We used the model with cubic polynomials (which was enough to approximate each regime) and applied it with $K = 2$ clusters in order to separate curves that would correspond to a defective operating state and curves corresponding to a normal operating state. Since the true class labels are unknown, we only considered the intra-class inertias as wall as a graphical inspection by observing the obtained partitions and each cluster approximation. The algorithm provided a partition of curves where the cluster shapes are clearly different (see Figure 3.3) and might correspond to two different states of the switch mechanism. According to the experts, one cluster could correspond to a default in the measurement process. These results are also in concordance with those obtained by the previously introduced piecewise regression mixture model; The partitions are quasi-identical.

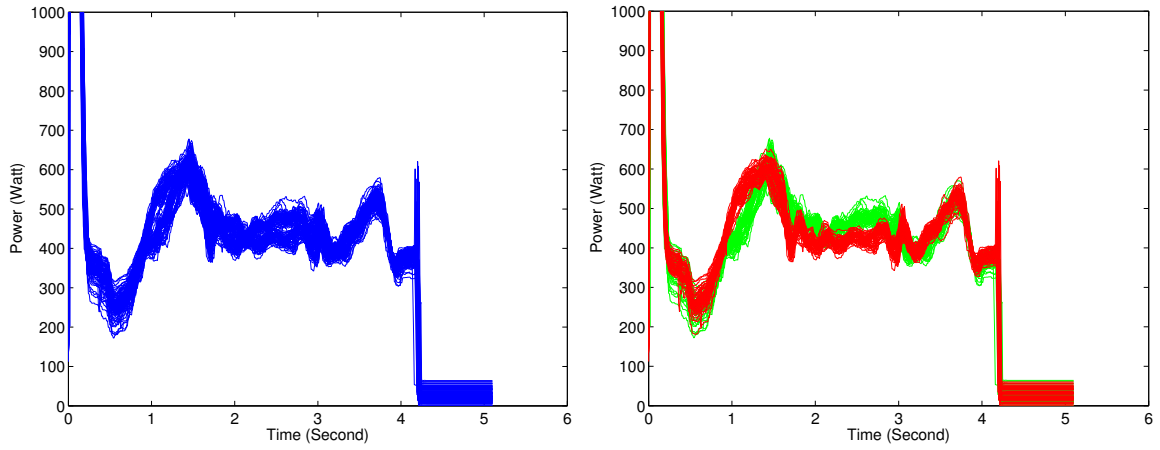

Figure 3.3: Clustering results for switch operation time series obtained with the MixHMMR model.

3.3.4 Conclusion

The introduced mixture of polynomial regression models governed by hidden Markov chains is particularly appropriate for clustering curves with various changes in regime and rely on a suitable generative formulation. The experimental results demonstrated the benefit of the proposed approach as compared to existing alternative methods, including the regression mixture model and the standard mixture of hidden Markov models. It also represents a full generative alternative to the previously described mixture of piecewise regressions. Also while the model in its current version only concerns univariate time series, I think that its extension to the multivariate case could be done without a major effort. while only the EM version is derived here, however, its extension to derive a CEM variant, for example to privilege the classification rather than the density estimation, is obvious.

Also while the model is a full generative one, one disadvantage is that as each hidden regime sequence is a Markov chain, the regime residence time is geometrically distributed, which is not adapted especially for long duration regimes, which might be the case for regimes of the analyzed functional data. However, I notice that this issue is more pronounced for the standard mixture of HMMs. In the proposed MixHMMR model, the fact that the conditional distribution rely on polynomial regressors, contribute to stabilize this effect by providing well-structured regimes even when those are activated for a long time period. For modeling different state length distributions, one might also use a non-homogeneous Markov chain as (Diebold et al., 1994; Hughes et al., 1999), that is, a Markov chain with time-dependent transition probabilities. The model proposed in the next section addresses the problem by using a logistic process rather than a Markov chain which as it was seen in the previous chapter, provides a modeling with better flexibility.

3.4 Mixture of hidden logistic process regressions

We saw in Section 3.2 that a first natural idea to cluster and segment complex functional data arising in curves with regime changes is to use piecewise regression integrated into a mixture formulation. This model however does not define a probability distribution over the change points and in practice may be time consuming especially for large time series. A first full generative alternative is to use mixtures of HMMs or the one more adapted for structured regimes in time series, that is, the proposed mixture of HMM regressions, seen in the previous section. However, if we look at how are we dealing with the quality of regime changes, that is, particularly regarding their smoothness, it appears that for the piecewise approach, it handles only abrupt changes, and for the HMM-based approach, while the posterior regime probabilities can be seen as fuzzy partition for the regimes and hence in some sense accomodate smoothness, there is no however explicit formulation regarding the nature of transition points and the smoothness of the resulting estimated functions. On the other hand, the regime residence time is necessarily geometrically distributed in these HMM-based models which might result in the fact that a transition may occur even within structured observations of the same regime. This what we saw in some obtained results in Section 2.4 when applying the HMM models, especially the standard HMM. However, using polynomial regressors for the state conditional density is a quite sufficient way to stabilize this behavior. The modeling can however be further improved by adopting a process that explicitly takes into account the smoothness of transitions in the process governing the regime changes.

Here, we attempt to overcome this by using a logistic process rather than a Markov process. The resulting model is a mixture of regressions with hidden logistic processes (MixRHLP) [J-4][J-5].

3.4.1 The model

In the proposed mixture of regression models with hidden logistic processes (MixRHLP) [J-4][J-5], each of the functional mixture components (3.1) is an RHLP [J-1][J-2]. That is, as seen in Chapter 2, the conditional distribution of a curve is defined by an RHLP:

$$f(\boldsymbol{y}_i|Z_i=k,\boldsymbol{x}_i;\boldsymbol{\Psi}_k)=\prod_{j=1}^{m_i}\sum_{r=1}^{R_k}\pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\big(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\big) \tag{3.21}$$

whose parameter vector is $\boldsymbol{\Psi}_k=(\mathbf{w}_k^T,\boldsymbol{\beta}_{k1}^T,\ldots,\boldsymbol{\beta}_{kR_k}^T,\sigma_{k1}^2,\ldots,\sigma_{kR_k}^2)^T$ and where the distribution of the discrete variable H_{ij} governing the hidden regimes is assumed to be logistic, that is, in this segmental case,

$$\pi_{kr}(x_j;\mathbf{w}_k)=\mathbb{P}(H_{ij}=r|Z_i=k,x_j;\mathbf{w}_k)=\frac{\exp\left(w_{kr0}+w_{kr1}x_j\right)}{\sum_{r'=1}^{R_k}\exp\left(w_{kr'0}+w_{kr'1}x_j\right)}, \tag{3.22}$$

whose parameter vector is $\mathbf{w}_k=(\boldsymbol{w}_{k1}^T,\ldots,\boldsymbol{w}_{kR_k-1}^T)^T$ where $\boldsymbol{w}_{kr}=(w_{kr0},w_{kr1})^T$ being the 2-dimensional coefficient vector for the rth logistic component with $\boldsymbol{w}_{kR_k}$ being the null vector. This choice is due to the flexibility of the logistic function in both determining the regime transition points and accurately modeling abrupt and/or smooth regimes changes. Indeed, as shown in [J-1][J-2], the logistic function (3.22) parameters (w_{kr0},w_{kr1}) control the regime transition points and the quality of regime (smooth or abrupt). Remark that here we used a linear logistic function for contiguous regime segmentation.

The resulting distribution of a curve is given by the following MixRHLP density:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\big(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\big) \tag{3.23}$$

with parameter vector $\boldsymbol{\Psi} = (\alpha_1,\ldots,\alpha_{K-1},\boldsymbol{\Psi}_1^T,\ldots,\boldsymbol{\Psi}_K^T)^T$. Notice that the key difference between the proposed MixRHLP and the standard regression mixture model is that the proposed model uses a generative hidden process regression model (RHLP) for each component rather than polynomial or spline components; The RHLP is itself based on a dynamic mixture formulation. Thus, the proposed approach is more adapted for accomodating the regime changes within curves during time.

3.4.2 Maximum likelihood estimation via a dedicated EM algorithm

The unknown parameter vector $\boldsymbol{\Psi}$ is estimated from an independent sample of unlabeled curves $\boldsymbol{D} = ((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ by monotonically maximizing the following observed-data log-likelihood

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

via a dedicated EM algorithm. The EM scheme requires the definition of the complete-data log-likelihood. The complete-data log-likelihood for the proposed MixRHLP model, given the observed data which we denote by $\boldsymbol{D}$, the hidden component labels $\mathbf{Z}$, and the hidden process $\{\mathbf{H}_k\}$ for each of the K components, is given by:

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log\alpha_k + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K}\sum_{r=1}^{R_k} Z_{ik}H_{ijr}\log\Big[\pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\Big(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\Big)\Big]. \tag{3.24}$$

The EM algorithm for the MixRHLP model (EM-MixRHLP) starts with an initial parameter $\boldsymbol{\Psi}^{(0)}$ and alternates between the two following steps until convergence:

The E-step computes the expected complete-data log-likelihood, given the observations $\boldsymbol{D}$, and the current parameter estimation $\boldsymbol{\Psi}^{(q)}$ and is given by:

$$\begin{aligned} Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) &= \mathbb{E}\Big[\log L_c(\boldsymbol{\Psi})\big|\mathcal{D};\boldsymbol{\Psi}^{(q)}\Big] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m_i}\sum_{r=1}^{R_k}\tau_{ik}^{(q)}\gamma_{ijr}^{(q)}\log\Big[\pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\Big(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\Big)\Big]. \end{aligned} \tag{3.25}$$

As shown in the expression of $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$, this step simply requires the calculation of each of the posterior component probabilities, that is, the probability that the ith observed curve originates from component k which is given by applying Bayes' theorem:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i,\boldsymbol{x}_i;\boldsymbol{\Psi}_k^{(q)}) = \frac{\alpha_k^{(q)} f(\boldsymbol{y}_i|Z_i=k,\boldsymbol{x}_i;\boldsymbol{\Psi}_k^{(q)})}{\sum_{k\prime=1}^{K}\alpha_{k'}^{(q)} f(\boldsymbol{y}_i|Z_i=k\prime,\boldsymbol{x}_i;\boldsymbol{\Psi}_{k\prime}^{(q)})} \tag{3.26}$$

where the conditional densities are given by (3.21), and the posterior regime probabilities given a mixture component, that is, the probability that the observation y_{ij}, for example at time x_j in a temporal context, originates from the rth regime of component k, which is given by applying the Bayes' theorem:

$$\gamma_{ijr}^{(q)} = \mathbb{P}(H_{ij}=r|Z_i=k,y_{ij},t_j;\boldsymbol{\Psi}_k^{(q)}) = \frac{\pi_{kr}(x_j;\mathbf{w}_k^{(q)})\mathcal{N}(y_{ij};\boldsymbol{\beta}_{kr}^{T(q)}\boldsymbol{x}_j,\sigma_{kr}^{2(q)})}{\sum_{r\prime=1}^{R_k}\pi_{kr\prime}(x_j;\mathbf{w}_k^{(q)})\mathcal{N}(y_{ij};\boldsymbol{\beta}_{kr\prime}^{T(q)}\boldsymbol{x}_j,\sigma_{kr\prime}^{2(q)})}. \tag{3.27}$$

It can be seen that here the posterior regime probabilities are computed directly without need of a forward-backward recursion as in the Markovian model.

The M-step updates the value of the parameter vector $\boldsymbol{\Psi}$ by maximizing the Q-function (3.25) w.r.r $\boldsymbol{\Psi}$, that is: $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$. The mixing proportions updates are given as in the case of standard mixtures by:

$$\alpha_k^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(q)}, \quad (k = 1, \ldots, K). \tag{3.28}$$

The maximization w.r.t the regression parameters consists in separate analytic solutions of weighted least-squares problems where the weights are the product of the posterior probability $\gamma_{ik}^{(q)}$ of component k and the posterior probability $\gamma_{ijr}^{(q)}$ of regime r. Thus, the updating formula for the regression coefficients and the variances are respectively given by (3.19) and (3.20). These updates are indeed the same those of the MixHMMR model, the only difference in that posterior cluster and regime memberships are calculated in a different way because of the different modeling for the hidden categorical variable H representing the regime. It follows a Markov chain for the MixHMMR model and a logistic process for the MixRHLP model.
Finally, the maximization w.r.t the logistic processes' parameters $\{\mathbf{w}_k\}$ consists in solving multinomial logistic regression problems weighted by the posterior probabilities $\tau_{ik}^{(q)}\gamma_{ijr}^{(q)}$ which we solve with a multi-class IRLS algorithm (see for example [C-14] for more detail on IRLS). The parameter update $\mathbf{w}_k^{(q+1)}$ is then taken at convergence of the IRLS algorithm.

Algorithmic complexity The algorithmic complexity of the proposed EM algorithm depends on the computation costs of the E- and M- steps. The complexity of the E-step is $\mathcal{O}(KRnmp)$, which mainly comprises the calculation of the logistic probabilities π_{kr} and the normal densities $\mathcal{N}(y_{ij}; \boldsymbol{\beta}_{kr}^T \boldsymbol{x}_j, \sigma_{kr}^2)$ for all k, r, i, j . For each k and r, the regression coefficients update requires the computation and inversion of a $(p+1) \times (p+1)$ matrix which can be done in $\mathcal{O}(nmp^3)$, and the variance update is computed in $\mathcal{O}(nmp)$. Each iteration of the IRLS algorithm requires, in the case of contiguous segmentation, a $2(R-1) \times 2(R-1)$ Hessian matrix to be computed and inverted, which is done in $\mathcal{O}(R^3nm)$. From the computation costs of the regression coefficients, the variances and the logistic functions coefficients, it can be deduced that the M-step has complexity $\mathcal{O}(KRnmp^3)$. Consequently, the computational complexity of the proposed EM algorithm is $\mathcal{O}(I_{\text{EM}} I_{\text{IRLS}} KR^3nmp^3)$, where I_{EM} is the number of iterations of the EM algorithm and I_{IRLS} is the maximum number of iterations of the inner IRLS loops. Compared to other clustering and segmentation algorithms such as the K-means type algorithm based on piecewise polynomial regression (Hébrail et al., 2010), whose complexity is $\mathcal{O}(I_{\text{KM}} KRnm^2p^3)$ where I_{KM} is the number of iterations of the algorithm, our EM algorithm is computationally attractive for large values of m and small values of R.

Curve approximation, segmentation and model selection Concerning the curves approximation, each cluster k is summarized by approximating it by a single "mean" curve, which we denote by $\hat{\boldsymbol{y}}_k$. Each point $\hat{y}_{kj}$ of this mean curve is defined by the conditional expectation $\hat{y}_{kj} = \mathbb{E}[y_{ij}|Z_i = k, t_j; \hat{\boldsymbol{\Psi}}_k]$ given by: $\hat{y}_{kj} = \sum_{r=1}^{R_k} \pi_{kr}(x_j; \hat{\mathbf{w}}_k)\hat{\boldsymbol{\beta}}_{kr}^T \boldsymbol{x}_j$ which is a sum of polynomials weighted by the logistic probabilities π_{kr} that model the regime variability over time and which constitutes a smooth flexible approximation.

The number of mixture components K, the number regimes R_k and the polynomial degree p can be estimated by maximizing some information criteria such the BIC. The number of free parameters of the MixRHLP model $\nu_{\boldsymbol{\Psi}} = K - 1 + \sum_{k=1}^{K} \nu_{\boldsymbol{\Psi}_k}$ with $\nu_{\boldsymbol{\Psi}_k} = (p+4)R_k - 2$ represents the number of free parameters of each RHLP component.

3.4.3 Experiments

In [J-4], the clustering accuracy of the proposed algorithm was evaluated using experiments carried out on simulated time series and real-world time series issued from a railway application. The obtained results are compared with those provided by the standard mixture of regressions and the K-means-like clustering approach based on piecewise regression Hébrail et al. (2010). To measure the clustering accuracy, two criteria were used: the misclassification percentage between the true partition and the estimated partition, and the intra-cluster inertia.

Simulation results The results in terms of misclassification error and intra-cluster inertia have shown that the proposed EM-MixRHLP algorithm outperforms the EM when used with regression mixtures. Although the misclassification percentages of the two approaches are close in particular in some situations, particularly for a small noise variance, the intra-cluster inertia differs from about 10^4. The misclassification provided by the regression mixture EM algorithm more rapidly increases with the noise variance level, compared to the proposed EM-MixRHLP approach. When the noise variance increases, the intra-cluster inertia obtained by the two approaches naturally increases, but the increase is less pronounced for the proposed approach compared to the regression mixture alternative. In addition, the obtained results showed that, as expected, contrary to the proposed model, the regression mixture model cannot accurately model time series which are subject to changes in regime. For model selection using BIC, the overall performance of the proposed algorithm is better than that of the regression mixture EM algorithm and the K-means like approach.

Experiments using real railway time series We used 140 times series issued from a railway diagnosis application. The specificity of the time series to be analyzed in this context as mentioned before is that they are subject to various changes in regime as a result of the mechanical movements involved in a switching operation. We accomplished this clustering task using our EM-MixRHLP algorithm, designed for estimating the parameters of a mixture of hidden process regression models. We compared the proposed EM algorithm to the regression mixture EM algorithm and the K-means like algorithm for piecewise regression. The obtained results show that the proposed regression approach provides the smallest intra-cluster error and misclassification rate.

3.4.4 Conclusion

In this section I presented a new mixture model-based approach for clustering and segmentation of univariate functional data with changes in regime. This approach involves modeling each cluster using a particular regression model whose polynomial coefficients vary across the range of the inputs, typically during time, according to a discrete hidden process. The transition between regimes is smoothly controlled by logistic functions. The model parameters are estimated by maximum likelihood method via a dedicated EM algorithm. The proposed approach can also be regarded as a clustering approach which operates by finding groups of time series having common changes in regime. The BIC is used to determine the numbers of clusters and segments, as well as the regression order. We note that these computations of the BIC for selecting three values (K, R, p) can be computationally more expensive compared the ones in classical model selection namely for standard mixture where only the number of clusters has to be selected. However, we notice that for small values of the dimensions to be selected, the computational cost is around few minutes and is not dramatically high, compared to approaches involving dynamic programming namely when using piecewise regression especially for large samples. The experimental results, both from simulated time series and from a real-world application, show that the proposed approach is an efficient means for clustering univariate time series with changes in regime. A CEM derivation of the current version is direct and obvious, and consists in assigning the curves in a hard way during the EM iterations, rather than in a soft way as what is done now via the posterior cluster memberships. One can further extend this to the regimes, by assigning the observations to the regimes also in a hard way, especially in the case where there are only abrupt change points in order to promote the segmentation. Then, in the framework of model selection, in a such extension, as well as for the current version of the model, it would be interesting to derive an ICL type criterion (Biernacki et al., 2000) which is known to be suited to the clustering and segmentation objectives.

3.5 Functional discriminant analysis

The previous sections were dedicated to cluster analysis of functional data where the aim was to explore a functional data set to automatically determine groupings of individuals described only by observations from that functions, that is, where the group labels indicating from which group each individual is issued are unknown. Here, I investigate the problem of prediction for functional data, specifically, the one of predicting the group label C_i of new observed unlabeled individual $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ describing a function, based on a training set of triplets of labeled individuals $\boldsymbol{D} = ((\boldsymbol{x}_1, y_1, c_1), \ldots, (\boldsymbol{x}_n, y_n, c_n))$ where $c_i \in \{1, \ldots, G\}$

is the class label of the ith individual. I focused on probabilistic discriminant analysis in which the discrimination task consists in estimating the class conditional density $f(\boldsymbol{y}_i|C_i, \boldsymbol{x}_i; \boldsymbol{\Psi}_g)$ and the prior class probabilities $\mathbb{P}(C_i)$ from the training set, and predicting the class label C_i for new data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ by using the Bayes' optimal allocation rule:

$$\hat{c}_i = \arg \max_{1 \leq g \leq G} \frac{w_g f(\boldsymbol{y}_i|C_i = g, \boldsymbol{x}_i; \boldsymbol{\Psi}_g)}{\sum_{g'=1}^{G} w_{g'} f(\boldsymbol{y}_i|C_i = g', \boldsymbol{x}_i; \boldsymbol{\Psi}_{g'})}, \tag{3.29}$$

where $w_g = \mathbb{P}(C_i = g)$ is the proportion of group g in the training set and $\boldsymbol{\Psi}_g$ the parameter vector of the conditional density. As in discriminant analysis for multivariate data, in this functional data discrimination context, one can rely on discriminant analyses by adopting dedicated conditional densities accounting for the functional aspect of the data. Two different ways are possible to accomplish the discriminant task, depending on how to model this conditional density.

3.5.1 Functional linear discriminant analysis

The first one consists in functional linear discriminant analysis (FLDA), firstly proposed in James and Hastie (2001) for irregularly sampled curves, and arises when we model each class conditional density with a single component model, for example a polynomial, spline or a B-spline regression model, that is in (3.29) $f(\boldsymbol{y}_i|C_i = g, \boldsymbol{x}_i; \boldsymbol{\Psi}_g) = \mathcal{N}(\boldsymbol{x}_i; \mathbf{X}_i\boldsymbol{\beta}_g, \sigma_g^2\mathbf{I}_m)$ with $\mathbf{X}_i$ is the design matrix of the chosen regression type and $\boldsymbol{\Psi}_g = (\boldsymbol{\beta}_g^T, \sigma_g^2)^T$ the parameter vector of class g. However, for curves with regime changes, these models are not adapted. In [J-2], the proposed FLDA with hidden process regression, in which each class is modeled with the regression model with a hidden logistic process (RHLP) (as presented in Section 2.2.1) accounts for regime changes through the tailored the class-specific density given by:

$$f(\boldsymbol{y}_i|C_i = g, \boldsymbol{x}_i; \boldsymbol{\Psi}_g) = \prod_{j=1}^{m_i} \sum_{r=1}^{R_g} \pi_{gr}(t_j; \mathbf{w}_g) \mathcal{N}\left(y_{ij}; \boldsymbol{\beta}_{gr}^T \boldsymbol{x}_j, \sigma_{gr}^2\right) \tag{3.30}$$

where $\boldsymbol{\Psi}_g = (\mathbf{w}_g^T, \boldsymbol{\beta}_{g1}^T, \ldots, \boldsymbol{\beta}_{gR_g}^T, \sigma_{g1}^2, \ldots, \sigma_{gR_g}^2)^T$ is the parameter vector of class g. In this FLDA context, each class estimation itself involves an unsupervised task regarding the hidden regimes, which is performed by the EM algorithm as described in [J-2]. However, the FLDA approaches are more adapted to homogeneous classes of curves and are not adapted to deal with dispersed classes, that is, when each class is itself composed of several sub-populations of curves.

3.5.2 Functional mixture discriminant analysis

The more flexible way in such a context of heterogeneous classes of functions is to rely on the idea of mixture discriminant analysis (MDA) for dispersed groups, introduced by Hastie and Tibshirani (1996) for multivariate data discrimination. Indeed, while the global discrimination task is supervised, in some situations, it may include an unsupervised task which in general relates clustering possibly dispersed classes into homogeneous sub-classes. In many areas of application of classification, a class may itself be composed of several unknown (unobserved) sub-classes. For example, in handwritten digit recognition, there are several characteristic ways to write a digit, and therefore a creation of several sub-classes within the class of a digit itself, which may be modeled using a mixture density as in Hastie and Tibshirani (1996). In complex systems diagnosis application, for example when we have to decide between two classes, say without or wit defect, one would have only the class labels indicating just either with or without defect, however no labels according to how a defect would happen, namely the type of defect, the degree of defect, etc. Another example is the one of gene function classification based on time course gene expression data. As stated in Gui and Li (2003) when considering the complexity of the gene functions, one functional class may include genes which involve one or more biological profiles. Describing each class as a combination of sub-classes is therefore necessary to provide realistic class representation, rather than providing a rough representation through a simple class conditional density. Here I consider the classification of functional data, particularly curves with regime changes, into classes arising in sub-populations. It is therefore assumed that each class g ($g = 1, \ldots, G$) has a complex shape arising in K_g homogeneous sub-classes. Furthermore, each sub-class k ($k = 1, \ldots, K_g$) of class g is itself governed by R_{gk} unknown regimes ($r = 1, \ldots, R_{gk}$). In such a context, the global discrimination task includes a two-level unsupervised task.

The first one is the one that attempts to automatically cluster possibly dispersed classes into several homogeneous clusters (i.e., sub-classes), and the second aims at automatically determining the regime locations of each sub-class, which is a segmentation task. A first idea on functional mixture discriminant analysis, motivated by the complexity of the time course gene expression functional data was proposed by Gui and Li (2003) and is based on B-spline regression mixtures. However, using polynomial or spline regressions for class representation, as studied for example in [J-2] is more adapted for smooth and stationary curves. In case of curves exhibiting a dynamical behavior through abrupt changes, one may relax the spline regularity constraints, which leads to the previously developed MixPWR model (see Section 3.2). Thus, in such context the generative functional mixture models presented previously can also be used as class conditional densities, that is, the MiHMMR, and the MixRHLP presented respectively in Sections 3.3 and 3.4. Here I only focus on the use of the mixture of RHLP since it is also dedicated to clustering and is flexible and explicitly integrates the smooth and/or abrupt regime changes via the logistic process. This leads to functional mixture discriminant analysis (FMDA) with hidden logistic process regression [J-5][C-8][C-10], in which the class conditional density for a function is given by a MixRHLP (3.23):

$$\begin{aligned} f(\boldsymbol{y}_i|C_i=g,\boldsymbol{x}_i;\boldsymbol{\Psi}_g) &= \sum_{k=1}^{K_g}\mathbb{P}(Z_i=k|C_i=g)f(\boldsymbol{y}_i|C_i=g,Z_i=k,\boldsymbol{x}_i;\boldsymbol{\Psi}_{gk}) \\ &= \sum_{k=1}^{K_g}\alpha_{gk}\prod_{j=1}^{m}\sum_{r=1}^{R_{gk}}\pi_{gkr}(x_j;\mathbf{w}_{gk})\mathcal{N}\big(y_{ij};\boldsymbol{\beta}_{gkr}^T\boldsymbol{x}_j,\sigma_{gkr}^2\big) \end{aligned} \tag{3.31}$$

where $\boldsymbol{\Psi}_g = (\alpha_{g1},\ldots,\alpha_{gK_g},\boldsymbol{\Psi}_{g1}^T,\ldots,\boldsymbol{\Psi}_{gK_g}^T)^T$ is the parameter vector for class g, with $\alpha_{gk} = \mathbb{P}(Z_i = k|C_i = g)$ is the proportion of component k of the mixture for group g and $\boldsymbol{\Psi}_{gk}$ the parameter vector of its RHLP component density. Then, once we have an estimate $\hat{\boldsymbol{\Psi}}_g$ of the parameter vector of the functional mixture density MixRHLP (provided by the EM algorithm described in the previous section) for each class, a new curve $(\boldsymbol{y}_i,\boldsymbol{x}_i)$ is then assigned to the class maximizing the posterior probability, that is, the Bayes' optimal allocation rule, using Equation (3.29).

3.5.3 Experiments

The proposed FMDA approach was evaluated in [J-5] on simulated data and real-world data issued from a railway diagnosis application. We performed comparisons with alternative functional discriminant analysis approaches using polynomial regression (FLDA-PR) or a spline regression (FLDA-SR) model (James and Hastie, 2001), and the FLDA one that uses a single RHLP model per class (FLDA-RHLP) as in [J-2]. I also performed comparisons with alternative FMDA approaches that use polynomial regression mixtures (FMDA-PRM), and spline regression mixtures (FMDA-SRM) as in Gui and Li (2003). Two evaluation criteria were used: the misclassification error rate computed by a 5-fold cross-validation procedure, which evaluates the discrimination performance, and the mean squared error between the observed curves and the estimated mean curves, which is equivalent to the intra-class inertia, and evaluates the performance of the approaches with respect to the curves modeling and approximation.

Simulation results The obtained results have shown that the proposed FMDA-MixRHLP approach accurately decomposes complex shaped classes into homogeneous sub-classes of curves and account for underlying hidden regimes for each sub-class. Furthermore, the flexibility of the logistic process used to model the hidden regimes allows for accurately approximating both abrupt and/or smooth regime changes within each sub-class. We also notice that the FLDA approach with spline or polynomial regression, provide rough approximations in the case of non-smooth regime changes compared to alternatives. The FLDA with RHLP accounts better for the regime changes, however, not surprising, for complex classes having sub-classes, it provides unsatisfactory results. This is confirmed on the obtained intra-class inertia results. Indeed, the smallest intra-class inertia is obtained for the proposed FMDA-MixRHLP approach which outperforms the alternative FMDA based on polynomial regression mixtures (FMDA-PRM) and spline regression mixtures (FMDA-SRM). This performance is attributed to the flexibility of the MixRHLP model thanks to the logistic process which is well adapted for modeling the regime changes. Also in terms of curve classification, the FMDA approaches provide better results compared to FLDA approaches. This is due to the fact that using a single model for complex-shaped classes (i.e., when using

FLDA approaches) is not adapted as it does not take into account the class dispersion when modeling the class conditional density. On the other hand, the proposed FMDA-MixRHLP approach provides a better modeling which results into a more accurate class prediction. The percentage of choosing the best model is also satisfactory and high (around 91%).

Experiments on real data Here the used data are issued from a railway diagnosis application as studied in [J-1][J-2][J-4]. The used data are the curves of the instantaneous electrical power consumed during the switch actuation period. The used database is composed of 120 labeled real switch operation curves. Each curve consists of 564 points. Two classes were considered where the first one is composed by the curves with no defect and with a minor defect and the second one contains curves without defect. The goal is therefore to provide an accurate automatic modeling especially for the first class which is henceforth dispersed into two sub-classes. The proposed method ensure both an accurate decomposition of the complex shaped class into sub-classes and at the same time, a good approximation of the underlying regimes within each homogeneous set of curves. The logistic process probabilities are close to 1 when the regression model seems to be the best fit for the curves and vary over time according to the smoothness degree of regime transition. Figure 3.4 shows modeling results provided by the proposed FMDA-MixRHLP for each of the two classes. We see that the proposed method ensure both an accurate decomposition of the complex shaped class into sub-classes and at the same time, a good approximation of the underlying regimes within each homogeneous set of curves. This also illustrates the clustering and segmentation using the MixRHLP presented in the previous section. The obtained classification results, while they were similar for the FMDA approaches, the difference in terms of curves modeling (approximation) is significant, for which the proposed FMDA-MixRHLP approach clearly outperforms the alternative ones. This is attributed to the fact that the use of polynomial regression mixtures for FMDA-PRM or spline regression mixtures (FMDA-SRM) does not fit at best the regime changes compared to the proposed model. The proposed approach provides the better results, but also has more parameters to estimate compared to the alternatives. But note that, for this real data, in terms of required computational effort to train each of the compared methods, the FLDA approaches are faster than the FMDA ones. In FLDA, both the polynomial regression and the spline regression approaches are analytic and does not require a numerical optimization scheme. The FLDA-RHLP is based on an EM algorithm which, therefore performs in an iterative way, but the learning scheme is quite fast and the computing time is in mean around one minute for the described real data, and is less demanding compared to the alternative piecewise regression using dynamic programming. On the other hand, the alternative FMDA approaches, that is the regression mixture and the spline regression mixture-based approaches still more fast and their EM algorithm requires only few seconds to converge. However, these approaches are clearly not adapted for the regime changes problem; to do that, one needs to built a piecewise regression-based model which requires dynamic programming and therefore may need a significant computational time especially for large curves, and is mainly adapted to abrupt regime changes. The training procedure for the proposed MixFRHLP-FMDA approach is not dramatically time consuming, the training for the data of class 1 (which is the more complex class), requires a mean computational time of around up to few minutes on a Matlab software using a laptop CPU of 2Ghz and 8GB of memory.

3.5.4 Conclusion

The presented mixture model-based approach for functional data discrimination includes unsupervised tasks that relate clustering dispersed classes and determining possible underlying unknown regimes for each sub-class. It is therefore suggested for the classification of curves organized in sub-groups and presenting a non-stationary behaviour arising in regime changes. Furthermore, the proposed functional discriminant analysis approach, as it uses a hidden logistic process regression model for each class, is particularly adapted for modeling abrupt and smooth regime changes. Each class is trained in an unsupervised way by a dedicated EM algorithm and a model selection using the BIC may be suggested as it provides satisfactory results.

Another possible direction is to train the MixRHLP of each class by maximizing the classification likelihood criterion, in which one will mainly be interested into classification, rather than maximizing a likelihood criterion as in this approach where we mainly focus on model estimation. This is direct and will rely on the CEM algorithm (Celeux and Govaert, 1992). In such context, the model selection relying on ICL (Biernacki et al., 2000) could also be used.

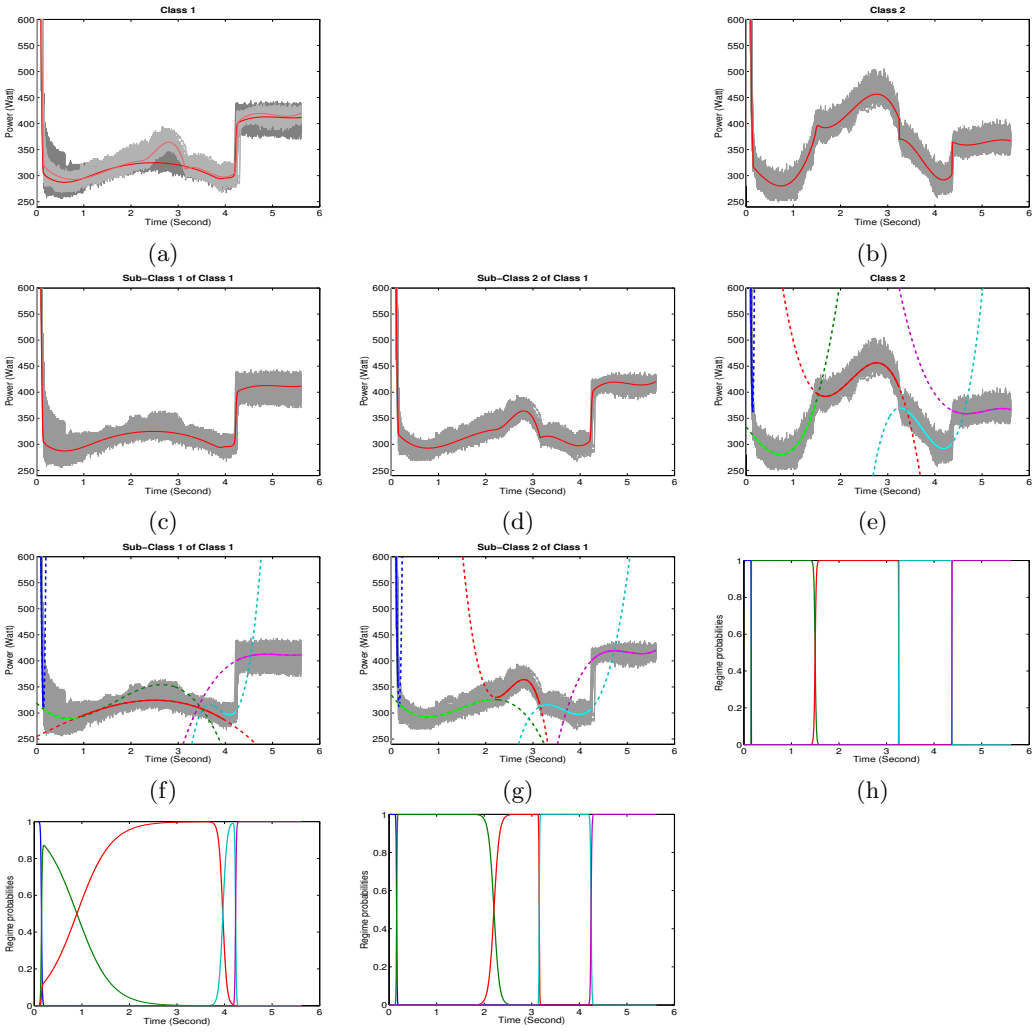

Figure 3.4: Results obtained with the proposed FMDA-MiXRHLP for the real switch operation curves. The estimated clusters (sub-classes) for class 1 and the corresponding mean curves (a); Then, we show separately each sub-class of class 1 with the estimated mean curve presented in a bold line (c,d), the polynomial regressors (degree $p = 3$) (f,g) and the corresponding logistic proportions that govern the hidden processes (i,j). Similarly, for class 2, we show the estimated mean curve in bold line (b), the polynomial regressors (e) and the corresponding logistic proportions (h).

Chapter 4

Bayesian regularization of mixtures for functional data

Contents

4.1 Introduction 45
4.1.1 Personal contribution 46
4.1.2 Regression mixtures 46
4.2 Regularized regression mixtures for functional data 47
4.2.1 Introduction 47
4.2.2 Regularized maximum likelihood estimation via a robust EM-like algorithm . . 49
4.2.3 Experiments 51
4.2.4 Conclusion 53
4.3 Bayesian mixtures of spatial spline regressions 54
4.3.1 Bayesian inference by Markov Chain Monte Carlo (MCMC) sampling 54
4.3.2 Mixtures of spatial spline regressions with mixed-effects 55
4.3.3 Bayesian spatial spline regression with mixed-effects 57
4.3.4 Bayesian mixture of spatial spline regressions with mixed-effects 59
4.3.5 Experiments 61
4.3.6 Conclusion 62

Related journal papers:

[1] F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015e. doi: 10.1080/00949655.2015.1109096. URL `http://chamroukhi.univ-tln.fr/papers/Chamroukhi-JSCS-2015.pdf`. Published online: 05 Nov 2015

[2] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, 2015a. URL `http://arxiv.org/pdf/1508.00635.pdf`

Related conference papers:

[1] F. Chamroukhi. Robust EM algorithm for model-based curve clustering. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Dallas, Texas, August 2013

[2] F. Chamroukhi. Model-based cluster and discriminant analysis for functional data. ERCIM 2014 : The 7th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2014a. Pisa, Italy

This research direction I initiated in 2013 is two-fold. First, I attempt to learn regression mixture models from univariate functional data, in a full unsupervised way, so that to provide an alternative to standard EM fitting which uses external information criteria for model selection. I am therefore interested in constructing on what can be seen as a non-parametric approach to simultaneously learn the model structure characterized by the number of mixture components, the model density, and the data partition. This is performed by regularizing the standard MLE of the regression mixtures and has lead to the following contribution [J-8][C-5].
On the other hand, I investigate regression mixtures, with mixed effects, the angle of approach and the type of data are different tough, compared to the previously studied mixtures with ML fitting, since here I am placed fully in the Bayesian inference framework by using Markov Chain Monte Carlo sampling. In addition, I consider mixture models dedicated to spatial functional data. The pre-publication [J-11] is issued from this work.

4.1 Introduction

In the previous Chapter, I investigated the problem of functional data analysis and I proposed latent data models, particularly functional mixture models, for such analysis which involve the construction of maximum likelihood estimators. Among the previously discussed models, there is the regression mixtures and their use in model-based cluster and discriminant analyses of functional data. The maximum likelihood estimation of the mixture density is mainly performed by using the EM algorithm thanks to its good desirable properties of stability and reliable convergence. In this chapter, I focus on regression mixtures and their use in model-based functional data clustering, particularly for univariate smooth functions and for spatial functional data (2D surfaces).

First, I revisit these mixture models and their estimation from another prospective by considering regularized MLE rather than standard MLE. This particularly attempts to address the issue of the ML fitting with the EM which requires careful initialization, and the one of model selection, from another point of view, say regularization. Indeed, it is well-known that the initialization is crucial for EM. The EM algorithm also requires the number of mixture component to be given a priori. The problem of selecting the number of mixture components in this case can be addressed by using, in an afterward step, some model selection criteria (e.g. AIC, BIC, and ICL as seen before) to choose one from a set of pre-estimated candidate models. Here I propose a penalized MLE approach carried out via a robust EM-like algorithm which simultaneously infers the model parameters, the model structure and the partition [J-8][C-5], and in which the initialization is simple.

On the other hand, these regression models seen until now have been constructed by relying of deterministic parameters which account for fixed effects that model the mean behavior of a population of homogeneous curves. However, in some situations, it is necessary to take into account possible random effects governing the inter-individual behavior. This is in general achieved by random effects regression or mixed effects regressions (**?**), that is, a regression model accounting for fixed effects, to which is added a random effects part. In a model-based clustering context, this is achieved by deriving mixtures of these models, for example the mixture of linear mixed models (Celeux et al., 2005). Despite the growing investigation for adapting multivariate mixture to the framework of FDA, for example as in (Devijver, 2014; Jacques and Preda, 2014; Bouveyron and Jacques, 2011; Chamroukhi, 2010a; Liu and Yang, 2009; Gaffney and Smyth, 2004; Gaffney, 2004; James and Sugar, 2003; James and Hastie, 2001), the most investigated type of data however is univariate or multivariate functions. The problem of learning from spatial functional data, that is, surfaces, is still less well studied. For example, one can cite the following quite recent approaches on the subject (Malfait and Ramsay, 2003; Ramsay et al., 2011; Sangalli et al., 2013; Nguyen et al., 2014). In particular, the very recent approach proposed by Nguyen et al. (2014) for clustering and classification of surfaces is based on the regression spatial spline regression as in Sangalli et al. (2013) in a mixture of linear mixed-effects model framework as in Celeux et al. (2005). The model estimation tool is the usual maximum likelihood estimation (MLE) by using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008). While the MLE via the EM algorithm is the standard way to fit finite mixture-based models, a common alternative is the Bayesian inference, that is, the maximum a posteriori (MAP) estimation. It is promoted to avoid singularities and degeneracies of the MLE as highlighted namely in Stephens (1997); Snoussi and Mohammad-Djafari (2001, 2005); Fraley and Raftery (2005) and Fraley and Raftery (2007) by regularizing it through a prior distribution over the model.

The MAP estimator is in general constructed by using Markov Chain Monte Carlo (MCMC) sampling, such as the Gibbs sampler (e.g., see Neal (1993); Raftery and Lewis (1992); Bensmail et al. (1997); Marin et al. (2005); Robert and Casella (2011)). For the Bayesian analysis of regression data, Lenk and DeSarbo (2000) introduced a Bayesian inference for finite mixtures of generalized linear models with random effects. Int their mixture model, each component is a regression model with a random-effects parts and the model is dedicated to multivariate regression data.

So in the second axis of this part of my research I first introduce the Bayesian spatial spline regression with mixed-effects (BSSR) for fitting a population of homogeneous surfaces. Then, I introduce the Bayesian mixtures of SSR (BMSSR) for fitting populations of heterogeneous surfaces organized in groups. The BSSR model is applied in surface approximation and the BMSSR model is applied in model-based surface clustering by considering real-world handwritten digits from the MNIST data set (LeCun et al., 1998).

4.1.1 Personal contribution

My personal contribution in this reach theme is two-fold. First, I proposed in [J-8][C-5][C-1] a new fully unsupervised learning algorithm to fit regression mixture models with unknown number of components. The developed approach consists in a penalized maximum likelihood estimation carried out by a robust EM-like algorithm. I derive it for polynomial, spline, and B-spline regression mixtures. *i*) it simultaneously infers the model parameters and the optimal number of the regression mixture components from the data as the learning proceeds, rather than in a two-fold scheme as in standard model-based clustering using afterward model selection criteria, and *ii*) its initialization is simple unlike the standard EM for regression mixtures which requires careful initialization. I validated the proposed algorithm on simulations and, the obtained results on real-world data covering three different application area, that is, phoneme recognition, clustering gene expression time course data for bio-informatics and clustering radar waveform data, confirm its benefit for practical applications.

Second, in [J-11], I investigated the problem of regression models with mixed effects and their use in FDA, particularly in model-based clustering of spatial functional data. I first introduced a Bayesian spatial spline regression model with mixed-effects (BSSR) for modeling spatial function data. The BSSR model is based on Nodal basis functions for spatial regression and accommodates both common mean behavior for the data through a fixed-effects part, and variability inter-individuals thanks to a random-effects part. Then, in order to model populations of spatial functional data issued from heterogeneous groups, I introduced a Bayesian mixture of spatial spline regressions with mixed-effects (BMSSR) used for density estimation and model-based surface clustering. The models, through their Bayesian formulation, allow to integrate possible prior knowledge on the data structure and constitute a good alternative to recent mixture of spatial spline regressions model estimated in a maximum likelihood framework via the EM algorithm. I derived MCMC sampling technique to infer each of the two model and applied them on simulated surfaces and a real problem of handwritten digit recognition using the MNIST data set. The obtained results highlight the potential benefit of the proposed Bayesian approaches for modeling surfaces possibly dispersed in particular in clusters.

The remainder of the chapter is organized as follows. After giving a brief background on regression mixtures and their use in model-based curve clustering in Section 4.1.2, I present in section 4.2, the proposed regularization of the mixture model and the fully unsupervised EM-like algorithm for fitting the resulting model. An experimental study is performed on numerous simulations and real-world data sets to apply and assess the proposed approach. Then, Section 4.3.2 describes recent related work on mixture of spatial spline regressions, and some formulation necessary to derive the proposed BSSR model, which I present in Section 4.3.3 where I also present its inference technique using Gibbs sampling. Then, in Section 4.3.4 I present the Bayesian mixture model for spatial functional data, that is, the BMSSR model, and show how to apply it in model-based clustering of surfaces. A Gibbs sampler is derived to estimate the BMSSR model parameters.

4.1.2 Regression mixtures

Modeling with regression mixtures is an important topic in the general family of mixture models. The finite regression mixture model (Quandt, 1972; Quandt and Ramsey, 1978; Veaux, 1989; Jones and

McLachlan, 1992; Gaffney and Smyth, 1999; Viele and Tong, 2002; Faria and Soromenho, 2010; Chamroukhi, 2010a; Young and Hunter, 2010; Hunter and Young, 2012) provides a way to model data arising from a number of unknown classes of conditionally dependent observed data. Let us denote by $\boldsymbol{D} = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ an observed independently and identically distributed (i.i.d) sample where each individual is a couple of a response $\boldsymbol{y}_i$ and its corresponding covariate $\boldsymbol{x}_i$. For example, in the case of temporal curves, the response consists of m_i observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ (regularly) observed at the inputs $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im_i})$ for all $i = 1, \ldots, n$ (e.g., x may represent the sampling time in a temporal context). The finite regression mixture model assumes that each individual $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is drawn from a mixture density of K (possibly unknown) components, whose mixing proportions are $(\pi_1, \ldots, \pi_K)$ where $\pi_k = \mathbb{P}(Z_i = k)$ is the prior probability of component k, $Z_i \in \{1, \ldots, K\}$ being the hidden variable representing the class label of the ith individual. A common way to model the conditional dependence in the observed data is to use regression functions. The regression mixture model assumes that each mixture component k is a conditional component density $f_k(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ of a regression model with parameters $\boldsymbol{\theta}_k$. This includes polynomial, spline, and B-spline regression mixtures, see for example DeSarbo and Cron (1988); Jones and McLachlan (1992); Gaffney (2004). These three models are considered here and the global Gaussian regression mixture is defined by the following conditional mixture density:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_{m_i}) \tag{4.1}$$

where $\mathbf{X}_i$ is the regression matrix constructed according to the chosen bases for the model, that is polynomial, spline, etc, $\boldsymbol{\beta}_k$ is the vector of regression coefficients for component k, and σ_k^2 is the noise variance with $\mathbf{I}_{m_i}$ denotes the $m_i \times m_i$ identity matrix. The regression matrix construction depends on the chosen type of regression, it may be Vandermonde for polynomial regression or a spline regression matrix for splines (Deboor, 1978)(Ruppert and Carroll, 2003) which are widely used for function approximation based on constrained piecewise polynomials, or the one of B-splines, which allow for more efficient computations compared to splines Ruppert and Carroll (2003)

The regression mixture model parameter vector is given by $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ where $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k^2)$ represents the parameter vector of component k composed of the regression coefficients vector and the noise variance. The use of regression mixtures for density estimation as well as for cluster and discriminant analyses, requires the estimation the mixture parameters. The problem of fitting regression mixture models is a widely studied problem in statistics, machine learning and data analysis, particularly for cluster analysis. It is usually performed by maximum likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_{m_i}) \tag{4.2}$$

by using the EM algorithm (Jones and McLachlan, 1992; Dempster et al., 1977; Gaffney and Smyth, 1999; Gaffney, 2004; McLachlan and Krishnan, 2008).

4.2 Regularized regression mixtures for functional data

4.2.1 Introduction

It is however well-known that the initialization is crucial for EM. If the initialization is not appropriately performed, the EM algorithm may lead to unsatisfactory results see for example Biernacki et al. (2003); Reddy et al. (2008); Yang et al. (2012). Thus, these regression mixture models when trained with the standard EM algorithm are sensitive to initialization since it might yield poor estimations if the regression mixture parameters are not initialized properly. The EM initialization in general can be performed from a randomly chosen partition of the data or by computing a partition from another clustering algorithm such as K-means, Classification EM (Celeux and Diebolt, 1985), Stochastic EM (Celeux and Govaert, 1992), etc or by initializing EM with a few number of iterations of EM itself. Several works have been proposed in the literature in order to overcome this drawback and making the EM algorithm for Gaussian mixtures robust with regard initialization, see for example Biernacki et al. (2003); Reddy et al. (2008); Yang et al. (2012). Further details about choosing starting values for the EM algorithm for Gaussian mixtures can be found for example in Biernacki et al. (2003). In addition to the sensitivity regarding the initialization, the EM algorithm requires the number of mixture components (clusters in a clustering context) to be

known. While the number of components can be chosen by some model selection criteria such as the BIC Schwarz (1978), the AIC Akaike (1974) or the ICL Biernacki et al. (2000), or resampling methods such as bootstrap McLachlan (1978), this requires performing an afterward model selection procedure, to choose one from a set of pre-estimated candidate models. Some authors have considered this issue in order to estimate the unknown number of mixture components in Gaussian mixture models, for example by an adapted EM as in Figueiredo and Jain (2000) and Yang et al. (2012) or from a Bayesian prospective Richardson and Green (1997) by reversible jump MCMC. However, in general, these two issues have been considered each separately. Among the approaches that consider the problem of robustness with regard to initial values and the one of estimating the number of mixture components, in the same algorithm, one can cite the EM algorithm proposed by Figueiredo and Jain (2000). This EM algorithm is capable of selecting the number of components and attempts to be not sensitive with regard to initial values by optimizing a minimum message length (MML) criterion, which is a penalized log-likelihood, rather than the observed-data log-likelihood. It starts by fitting a mixture model with a large number of clusters and discards invalid clusters as the learning proceeds. The degree of validity of each cluster is measured through the penalization term which includes the mixing proportions to know if the cluster is small or not to be discarded, and therefore to reduce the number of clusters. More recently, in Yang et al. (2012), the authors developed a robust EM-like algorithm for model-based clustering of multivariate data using Gaussian mixture models that simultaneously addresses the problem of initialization and the one of estimation of the number of mixture components. This algorithm overcomes some initialization drawback of the EM algorithm proposed in Figueiredo and Jain (2000). As shown in Yang et al. (2012), this problem regarding initialization can become more serious especially for a data set with a large number of clusters. However, these presented model-based clustering approaches, including those in Yang et al. (2012) and Figueiredo and Jain (2000), are concerned with vectorial data where the observations are assumed to be vectors of reduced dimension. When the data are rather curves or functions, they are not adapted. Indeed, when the data are functional described by individuals presented as curves or surfaces they are in general very structured and approaches relying on standard multivariate mixture analysis may therefore lead to unsatisfactory results in terms of modeling and classification accuracy since in that case we ignore the structure of the individuals[J-1][J-2][J-4][J-5]. However, addressing the problem from a functional data analysis prospective, that is formulating "functional" mixture models, allows to overcome these limitations, e.g., as in [J-1][J-2][J-4]Gaffney (2004). So here we attempt to overcome the limitations of the EM algorithm in the case of regression mixtures and their use in model-based curve clustering by regularizing the model and proposing an EM-like algorithm for the inference, which is robust with regard initialization and automatically estimates the optimal number of clusters as the learning proceeds.

The presented approach as developed in [J-8][C-5] is in the same spirit of the EM-like algorithm presented in Yang et al. (2012), but by extending the idea to the case of functional data (curve) clustering, rather than multivariate data clustering. This leads to fitting regression mixture models (including splines or B-splines) of the form (4.1). For estimating the regression mixture model (4.1), rather than maximizing the standard observed-data log-likelihood (4.2), we attempt to maximize a penalized version of it. The penalized log-likelihood function we propose to maximize is thus constructed by penalizing the observed-data log-likelihood by a regularization term related as we will see to the model complexity, and is defined by:

$$\mathcal{J}(\lambda, \boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) - \lambda H(\mathbf{Z}), \quad \lambda \geq 0 \tag{4.3}$$

where $\log L(\boldsymbol{\theta})$ is the observed-data log-likelihood maximized by the standard EM algorithm for regression mixtures (see Eq. (4.2)) and $\lambda \geq 0$ is a parameter that controls the complexity of the fitted model. This penalized log-likelihood function allows to control the complexity of the model fit through the roughness penalty $H(\mathbf{Z})$ accounting for the model complexity. As the model complexity is related to particularly the number of mixture components and therefore the structure of the hidden variables Z_i (recall that Z_i represents the class label of the ith curve), we chose to use the entropy of the hidden variable Z_i as penalty. The penalized log-likelihood criterion is therefore derived as follows. The (differential) entropy of Z_i is defined by:

$$H(Z_i) = -\sum_{k=1}^{K} \mathbb{P}(Z_i = k) \log \mathbb{P}(Z_i = k) = -\sum_{k=1}^{K} \pi_k \log \pi_k. \tag{4.4}$$

By assuming that the variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ are i.i.d, which is in general the assumption in clustering using mixtures where the cluster labels are assumed to be distributed according to a Multinomial

distribution, the whole entropy for $\mathbf{Z}$ is therefore additive and we have

$$H(\mathbf{Z}) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k \log \pi_k, \tag{4.5}$$

which leads to the following penalized log-likelihood criterion:

$$\mathcal{J}(\lambda, \boldsymbol{\theta}) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_{m_i}) + \lambda n \sum_{k=1}^{K}\pi_k \log \pi_k. \tag{4.6}$$

This penalized log-likelihood function (4.6) we attempt to optimize allows to control the complexity of the model fit through the roughness penalty $\lambda n \sum_{k=1}^{K}\pi_k \log \pi_k$. The entropy term $-n\sum_{k=1}^{K}\pi_k \log \pi_k$ in the penalty measures the complexity of a fitted model for K clusters. When the entropy is large, the fitted model is rougher, and when it is small, the fitted model is smoother. The non-negative smoothing parameter λ is for establishing a trade-off between closeness of fit to the data and a smooth fit. As λ decreases, the fitted model tends to be less complex, and we get a smoother fit. However, when λ increases, the result is a rougher fit.

The next section presents the proposed robust EM-like algorithm to maximize the penalized observed-data log-likelihood $\mathcal{J}(\lambda, \boldsymbol{\theta})$ for regression mixture density estimation and model-based curve clustering.

4.2.2 Regularized maximum likelihood estimation via a robust EM-like algorithm

Given an i.i.d sample of n curves $\boldsymbol{D} = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$, the penalized log-likelihood (4.6) is iteratively maximized by using the following robust EM-like algorithm. Before giving the EM steps, we give the penalized complete-data log-likelihood, on which the algorithm formulation is based. The complete-data log-likelihood, in this penalized case, is given by:

$$\mathcal{J}_c(\lambda, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log\big[\pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_{m_i})\big] + \lambda n \sum_{k=1}^{K}\pi_k \log \pi_k \tag{4.7}$$

where Z_{ik} is an indicator binary-valued variable such that $Z_{ik} = 1$ if $Z_i = k$ (i.e., if the ith curve $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is generated from the kth regression mixture component) and $Z_{ik} = 0$ otherwise. After starting with an initial solution (see section 4.2.2 for the initialization strategy and stopping rule), the proposed algorithm alternates between the two following steps until convergence.

E-step This step computes the expectation of the penalized complete-data log-likelihood (4.7), given the observed data $\boldsymbol{D}$ and a current parameter vector $\boldsymbol{\theta}^{(q)}$:

$$Q(\lambda, \boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = \mathbb{E}\big[\mathcal{J}_c(\lambda, \boldsymbol{\theta})|\boldsymbol{D}; \boldsymbol{\theta}^{(q)}\big] = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\big[\pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}_{m_i})\big] + \lambda n \sum_{k=1}^{K}\pi_k \log \pi_k \tag{4.8}$$

where

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)}\mathcal{N}\big(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k^{T(q)}, \sigma_k^{2(q)}\mathbf{I}_{m_i}\big)}{\sum_{h=1}^{K}\pi_h^{(q)}\mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_h^{(q)}, \sigma_h^{2(q)}\mathbf{I}_{m_i})} \tag{4.9}$$

is the posterior probability that the curve $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is generated by the kth cluster. This step therefore only requires the computation of the posterior component memberships $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$ for each of the K components.

M-step This step updates the value of the parameter vector $\boldsymbol{\theta}$ by maximizing the Q-function (4.8) with respect to $\boldsymbol{\theta}$, that is by computing the parameter vector update $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} Q(\lambda, \boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$. The mixing proportions updates are given by (see for example Appendix B in [J-8] for more calculation details):

$$\pi_k^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ik}^{(q)} + \lambda\pi_k^{(q)}\Big(\log\pi_k^{(q)} - \sum_{h=1}^{K}\pi_h^{(q)}\log\pi_h^{(q)}\Big). \tag{4.10}$$

We can remark here that the update of the mixing proportions (4.10) is close to the standard EM update $\left(\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}^{(q)}\right)$ for very small value of λ. However, for a large value of λ, the penalization term will play its role in order to make clusters competitive and thus allows for discarding invalid clusters and enhancing actual clusters. Indeed, in the updating formula (4.10), we can see that for cluster k if

$$\log \pi_k^{(q)} - \sum_{h=1}^{K} \pi_h^{(q)} \log \pi_h^{(q)} > 0, \tag{4.11}$$

that is, for the (logarithm of the) current proportion $\log \pi_k^{(q)}$, the entropy of the hidden variables is decreasing, and therefore the model complexity tends to be stable, the cluster k has therefore to be enhanced. This explicitly results in the fact that the update of the kth mixing proportion $\pi_k^{(q+1)}$ in (4.10) will increase. On the other hand, if (4.11) is less than 0, the cluster is not informative its proportion will decrease. Furthermore, the penalization coefficient λ can be set in an adaptive way (see [J-8]) in such a way to be large for enhancing competition when the proportions are not increasing enough from one iteration to another. In that case, the robust algorithm plays its role for estimating the number of clusters (which is decreasing in such a situation, by discarding small invalid clusters). In practice a cluster k can be discarded if its proportion is not significant, e.g. less than $\frac{1}{n}$, that is $\pi_k^{(q)} < \frac{1}{n}$. On the other hand, λ has to become small when the proportions are sufficiently increasing as the learning proceeds and the partition can therefore be considered as stable. In this case, the robust EM-like algorithm tends to have the same behavior as the standard EM. The regularization coefficient λ is also set in $[0, 1]$ to prevent very large values.

Then, the regression parameters $(\boldsymbol{\beta}_k, \sigma_k^2)$ are updated by analytically solving weighted least-squares problems where the weights are the posterior probabilities $\tau_{ik}^{(q)}$ and the updates are given by:

$$\boldsymbol{\beta}_k^{(q+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{X}_i\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \boldsymbol{y}_i, \tag{4.12}$$

$$\sigma_k^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)} m_i} \sum_{i=1}^{n} \tau_{ik}^{(q)} \|\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{\beta}_k^{(q+1)}\|^2, \tag{4.13}$$

where the posterior cluster probabilities $\tau_{ik}^{(q)}$ given by (4.9) are computed using the updated mixing proportions derived in (4.10).

Finally, once the model parameters have been estimated, a fuzzy partition of the data into K clusters, represented by the estimated posterior cluster probabilities $\hat{\tau}_{ik}$, is obtained. A hard partition can also be computed according to the Bayes' optimal allocation rule, that is, by assigning each curve to the component having the highest posterior probability (4.9).

Initialization strategy and stopping rule The initial number of clusters is $K^{(0)} = n$, n being the total number of curves and the initial mixing proportions are $\pi_k^{(0)} = \frac{1}{K^{(0)}}$, $(k = 1, \ldots, K^{(0)})$. Then, to initialize the regression parameters $\boldsymbol{\beta}_k$ and the noise variances σ_k^2 $(k = 1, \ldots, K^{(0)})$, we fitted a polynomial regression model on each curve k, $(k = 1, \ldots, K^{(0)})$; The initial values of the regression parameters are thus given by $\boldsymbol{\beta}_k^{(0)} = \left(\mathbf{X}_k^T \mathbf{X}_k\right)^{-1} \mathbf{X}_k \boldsymbol{y}_k$ and the noise variance can be deduced as $\sigma_k^{2(0)} = \frac{1}{m_k} \|\boldsymbol{y}_k - \mathbf{X}_k \boldsymbol{\beta}_k^{(0)}\|^2$. To avoid singularities at the starting point, we set $\sigma_k^{2(0)}$ as a middle value in the following sorted range $\|\boldsymbol{y}_k - \mathbf{X}_k \boldsymbol{\beta}_k^{(0)}\|^2$ for $k = 1, \ldots, n$. The algorithm is stopped when the maximum variation of the estimated regression parameters between two iterations $\max_{1 \leq k \leq K^{(q)}} \|\boldsymbol{\beta}_k^{(q+1)} - \boldsymbol{\beta}_k^{(q)}\|$ was less than a fixed threshold υ (e.g., 10^{-6}).

Choosing the order of regression and spline knots number and locations For a general use of the proposed algorithm for the polynomial regression mixture, the order of regression can be chosen by cross-validation techniques as in Gaffney (2004). In our experiments, we report the results corresponding to the degree for which the polynomial regression mixture provides the best fit. However, in some situations, the PRM model may be too simple to capture the full structure of the data, in particular for curves with high non-linearity or with regime changes, even if it can be seen as providing a useful first-order approximation of the data structure. The (B-)spline regression models in such case are more adapted. For these models, one may need to choose the spline order as well as the number of knots and

their locations. For the order of regression in (B-)splines, we notice that, in practice, the most widely used orders are $M = 1$,2, and 4 (Hastie et al., 2010). For smooth function approximation, cubic (B-)splines, which correspond to a (B-)spline of order 4 and thus with twice continuous derivatives, are sufficient to approximate smooth functions. When the data present irregularity, such as a kind of piecewise non continuous functions, a linear spline (of order 2) is more adapted. This was namely used for the satellite data set. The order 1 can be chosen for piecewise constant data. Concerning the choice of the number of knots and their locations, a common choice is to place a number of knots uniformly spaced across the range of x. In general more knots are needed for functions with high non-linearity or regime changes. One can also use automatic techniques for the selection of the number of knots and their locations as reported in Gaffney (2004). For example, this can be performed by using cross validation as in Ruppert and Carroll (2003). In Kooperberg and Stone (1991), the knots are placed at selected order statistics of the sample data and the number of knots is determined including by minimizing a variant of AIC. The general goal is to use a sufficient number of knots to fit the data while at the same time to avoid over-fitting and to not make the computation excessive. The current algorithm can be easily extended to handle this type of automatic selection of spline knots placement, but as the unsupervised clustering problem itself requires much attention and is difficult, it is wise to fix the number and location of knots. In this proposal knot sequences uniformly spaced across the range of x are used. The studied problems are not very sensitive to the number and location of knots; Few number of equispaced knots (less than ten for the data studied here) are sufficient to provide a reasonable fit of the data.

4.2.3 Experiments

The proposed unsupervised algorithm was evaluated in [J-8][C-5] for the three regression mixture models, that is, polynomial, spline, and B-spline regression mixtures, respectively abbreviated as PRM, SRM, and bSRM by performing numerous experiments carried on simulations, the Breiman wavefrom Benchmark (Breiman et al., 1984) and three real-world data sets covering three different application area: phoneme recognition in speech recognition, clustering gene expression time course data for bio-informatics and clustering radar waveform data. The evaluation is performed in terms of estimating the actual partition by considering the estimated number of clusters and the clustering accuracy (misclassification error) when the true partition is known.

Simulation results In summary, the number of clusters is correctly estimated by the proposed algorithm for three models. The spline regression models provide slightly better results in terms of clusters approximation than the polynomial regression mixture. On the other hand, the regression mixture models with the proposed EM-like algorithm outperform the standard K-means and EM-GMM clustering algorithms.

Different simulations scenarios were also designed to assess the behavior of the proposed approach in terms of the number of observations, the dimension of each observation, as well as the number of clusters in the data, the cluster and the cluster proportions. Simulations $S1$ were designed to assess the capacity of the proposed approach to retrieve partitions with a small number of clusters while simulations $S2$ were designed to retrieve partitions with a large number of clusters. Bot include well separated clusters as well as poorly separated clusters. The true partition is correctly estimated in most cases. The clusters which are not well separated (merged) are also retrieved with success. The model indeed takes into account mixture components with different noise variances (heteroskedastic model) which allow to recover merged functions with only different noise variances.

Phonemes data The phonemes data set used in Ferraty and Vieu (2003)[1] is a part of the original one available at `http://www-stat.stanford.edu/ElemStatLearn` and was described and used namely in Hastie et al. (1995). The application context related to this data set is a phoneme classification problem. The phonemes data correspond to log-periodograms y constructed from recordings available at different equispaced frequencies x for different phonemes. The data set contains five classes corresponding to the following five phonemes: "sh" as in "she", "dcl" as in "dark", "iy" as in "she", "aa" as in "dark", and "ao" as in "water". For each phoneme we have 400 log-periodograms at a 16-kHz sampling rate. We only retain the first 150 frequencies from each subject as in Ferraty and Vieu (2003). This data set has

[1] Data from `http://www.math.univ-toulouse.fr/staph/npfda/`

been considered in a phoneme discrimination problem as in Hastie et al. (1995) and Ferraty and Vieu (2003) where the aim is to predict the phoneme class for a new log-periodogram. Here we reformulate the problem into a clustering problem where the aim is to automatically group the phonemes data into classes. We therefore assume that the cluster labels are missing. We also assume that the number of clusters is unknown. Thus, the proposed algorithm will be assessed in terms of estimating both the actual partition and the optimal number of clusters from the data. The number of phoneme classes (five) is correctly estimated by the three models. The spline regression mixture (SRM) results are closely similar to those provided by the bSRM model. The spline regression models provide better results in terms of classification error (14.2 %) and clusters approximation than the polynomial regression mixture. In functional data modeling, splines are indeed more adapted than simple polynomial modeling. The number of clusters decreases very rapidly from 1000 to 51 for the polynomial regression mixture model, and to 44 for the spline and B-spline regression mixture models. The grand majority of invalid clusters is discarded at the beginning of the learning process. Then, the number of clusters gradually decreases from one iteration to another for the three models and the algorithm converges toward a partition with the actual number of clusters for the three models after at most 43 iterations. Figure 4.1 shows the used 1000 phonemes log-periodograms (upper-left) and the clustering partition obtained by the proposed unsupervised algorithm with the B-spline regression mixture (bSRM).

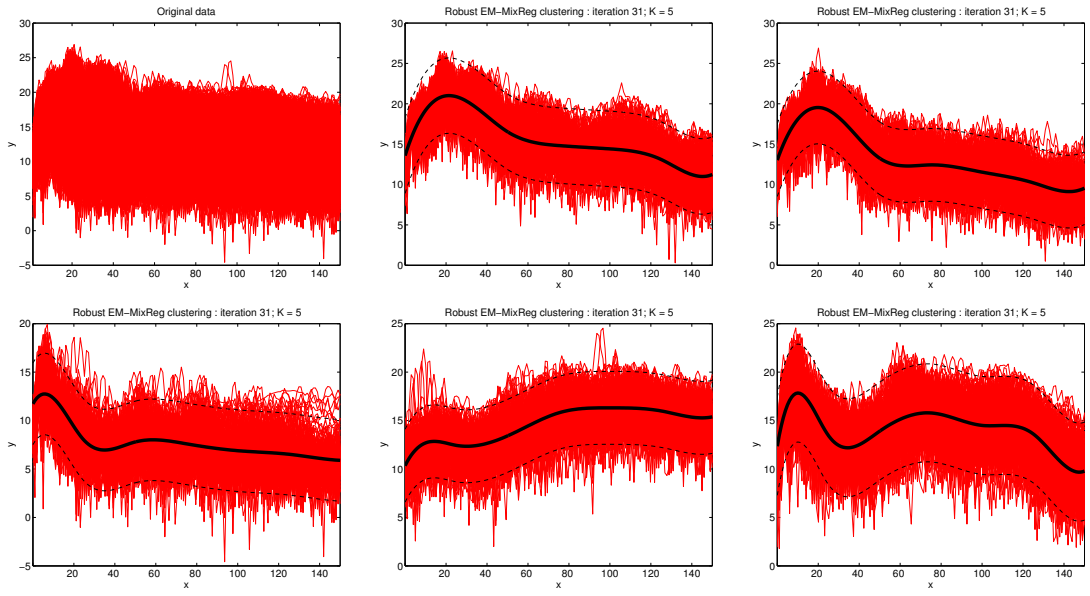

Figure 4.1: Phonemes data and clustering results obtained by the proposed robust EM-like algorithm and the bSRM model with a cubic B-spline of seven knots for the phonemes data. The five sub-figures correspond to the automatically retrieved clusters which correspond to the phonemes “ao”, “aa”, “yi”, “dcl”, “sh”.

Yeast cell cycle data In this experiment, we consider the yeast cell cycle data set Cho et al. (1998). The original yeast cell cycle data represent the fluctuation of expression levels of approximately 6000 genes over 17 time points corresponding to two cell cycles Cho et al. (1998). This data set has been used to demonstrate the effectiveness of clustering techniques for time course Gene expression data in bio-informatics such as model-based clustering as in Yeung et al. (2001). We used the standardized subset constructed by Yeung et al. (2001) available in `http://faculty.washington.edu/kayee/model/`[1]. This data set referred to as the subset of the 5-phase criterion in Yeung et al. (2001) contains $n = 384$ gene expression levels over $m = 17$ time points. The usefulness of the cluster analysis in this case is therefore to automatically reconstruct this five class partition. Both the PRM and the SRM models provide similar partitions with four clusters with two clusters which are merged into one cluster. Note that some model selection criteria in Yeung et al. (2001) also provide four clusters in some situations. However, the bSRM

[1]The complete data are from `http://genome-www.stanford.edu/cellcycle/`.

model correctly infers the actual number of clusters. The Rand index (RI)[1] for the obtained partition equals 0.7914 which indicates that the partition is quite well defined.

Topex/Poseidon data set The last considered real data set is the Topex/Poseidon radar satellite data set[2] namely used in Dabo-Niang et al. (2007) and Hébrail et al. (2010). This data set was registered by the satellite Topex/Poseidon around an area of 25 kilometers upon the Amazon River. The data contain $n = 472$ waveforms of the measured echoes, sampled at $m = 70$ number of echoes. The actual number of clusters and the actual partition are unknown for this data set. The provided solution for the polynomial regression mixture (PRM) is rather an overall rough approximation and provides three clusters. The polynomial fitting for this type of curves is not adapted. This is because the curves present in particular peaks and transitions. The solutions provided by the proposed algorithm with the spline regression mixture (SRM) and the B-spline regression mixture (bSRM) are very close and are more informative about the underlying structure of this data set. We used a linear (B-)spline for this data set in order to allow piecewise linear function approximation and thus to better recover the possible peaks and transitions in the curves. As a result, both the SRM and the bSRM provide a five class partition. The partitions are quasi-identical and contain clearly informative clusters with different shapes of waves that summarize the general underlying structure governing this dataset. In addition, the found number of clusters (five) also equals the one found by Dabo-Niang et al. (2007) by using another hierarchical nonparametric kernel-based unsupervised classification technique. The mean curves for the five groups provided by the proposed approach for both the SRM and the bSRM are similar to those in Dabo-Niang et al. (2007). On the other hand, this result is similar to the one found in Hébrail et al. (2010); Most of the profiles are indeed present in the two results. There is a slight difference which can be attributed to the fact that the results in Hébrail et al. (2010) are provided from a two-stage scheme which includes an additional pre-clustering step using the Self Organizing Map (SOM), rather by directly applying the piecewise regression model to the raw data. We also notice that, in the study of Hébrail et al. (2010), the number of clusters was set to twenty and the clustering procedure was two-fold. The authors first performed a topographic clustering step using the SOM, and then applied a K-means-like approach to the results of the SOM. However, in our approach, we directly apply the proposed algorithm to the raw satellite data without a preprocessing step. In addition, the number of clusters is automatically inferred from the data. The found five clusters here do summarize the general behavior of the twenty clusters in Hébrail et al. (2010) which can be summarized in clusters with one narrow shifted peak, less narrow peak, two large peaks, and finally a cluster containing curves with sharp increase followed by a slow decrease.

For this dataset, the algorithm converged after at most 35 iterations. After starting with $n = 472$ clusters, the number of clusters rapidly decreases to 59 for the PRM and to 95 for both the SRM and the bSRM models. Then it gradually decreases until the number of clusters is stabilized. The variation of the value of the objective function during the iterations of the algorithm also shows that it becomes horizontal at convergence which corresponds to the stabilization of the partition.

4.2.4 Conclusion

Here I presented a new robust EM-like algorithm for fitting regression mixtures and model-based curve clustering. It optimizes a penalized observed-data log-likelihood and overcomes both the problem of sensitivity to initialization and determining the optimal number of clusters for standard EM for regression mixtures. Note that the proposed algorithm, as it proceeds to the estimation of the number of components, does not guarantee the ascent property of the objective function, and, thus, is not a true EM algorithm. Note that even if this property is not established, in practice the algorithm works very well and does converged towards very satisfactory solutions for the several data on which it was applied. Compared to standard EM fitting, this constitutes an interesting fully unsupervised alternative that simultaneously infers the model and its optimal number of components. The experimental results on simulated data and real-world data demonstrate the benefit of the proposed approach for applications in curve clustering. The obtained clustering results are quite precise and the number of clusters was always correctly selected. For the phonemes data and the yeast cell cycle data, the polynomial degree with the best solution was

[1] The Rand Index measures the similarity between two data clusterings. It has a value between 0 and 1, with 0 indicating that the two partitions do not agree on any pair of observations and 1 indicating that the data clusters are exactly the same. For more details on the RI, see Rand (1971).

[2] Available at `http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html`.

retained. However, for a more general use in functional data clustering and approximation, the splines are clearly more adapted. In practice, for the spline and B-spline regression mixtures, we used cubic (B-)splines because cubic splines, which correspond to a spline of order 4 which are are sufficient to approximate smooth functions. However, when the data present irregularity, such as a kind of piecewise non continuous functions, which is the case of the Topex/Poseidon satellite data, we use a linear (B-)spline approximation. We also note that the algorithm is fast for the three models. It converged after a few number of iterations, and took at most less than 45 seconds for the phonemes data. For the other data, it took only few seconds. This makes it useful for real practical situations.

Here, I considered the problem of unsupervised fitting of regression mixtures with unknown number of components. One interesting future direction is to extend the proposed approach to the problem of fitting hidden process regression (e.g. those seen in Chapter 2) or mixture of experts Jacobs et al. (1991) and hierarchical mixture of experts Jordan and Jacobs (1994) with unknown number of experts. A further challenging extension might consist in extending this approach to the unsupervised simultaneous clustering and segmentation of functional data, say the models seen in Chapter 3.

4.3 Bayesian mixtures of spatial spline regressions

The previous section was dedicated to regression mixtures for univariate functional data with a kind of regularization. In this section, I investigate regression mixtures, but with three additional features: the first relates regression mixtures extended by including random effects, the second one relates further formulating the model for spatial functional data, and the third one is the full Bayesian inference of the proposed models. This sub-axis therefore relates the framework of Bayesian regression mixture modeling for spatial functional data where the data are surfaces. I present a probabilistic Bayesian formulation to model spatial functional data by extending the approaches of Nguyen et al. (2014) and apply the proposal to surface approximation and clustering. The model is also related to the random-effects mixture model of Lenk and DeSarbo (2000) in which I explicitly add mixed-effects and derive it for spatial functional data by using the Nodal basis functions (NBFs). The NBFs (Malfait and Ramsay, 2003) used in Ramsay et al. (2011), Sangalli et al. (2013), and Nguyen et al. (2014), represent an extension of the univariate B-spline bases to bivariate surfaces. I thus first introduce a Bayesian spatial spline regression model with mixed-effects (BSSR) for fitting a population of homogeneous surfaces. The BSSR model accommodates both common mean behavior for the data through a fixed-effects part, and variability inter-individuals thanks to a random-effects part. Then, in order to model populations of spatial functional data issued from heterogeneous groups, I integrate the BSSR model into a mixture framework. The resulting model is a Bayesian mixture of spatial spline regressions with mixed-effects (BMSSR) used for density estimation and model-based surface clustering. The models, through their Bayesian formulation, allow to integrate possible prior knowledge on the data structure and constitute a good alternative to the recent mixture of spatial spline regressions model of Nguyen et al. (2014) estimated in a maximum likelihood framework via the expectation-maximization (EM) algorithm. The inference of the proposed Bayesian modeling is performed by Markov Chain Monte Carlo (MCMC) sampling and I derive two Gibbs samplers to infer the BSSR and the BMSSR models. The BSSR model is first applied in surface approximation. Then, the BMSSR model is applied in model-based surface clustering by considering the real-world handwritten digits from the MNIST data set (LeCun et al., 1998). The obtained results highlight the potential benefit of the proposed Bayesian approaches for modeling surfaces possibly dispersed in particular in clusters.

4.3.1 Bayesian inference by Markov Chain Monte Carlo (MCMC) sampling

In this section I open a parenthesis to introduce the principle of Bayesian inference using MCMC sampling and its use for latent data models. I will use $p(.)$ as a generic notation for a density function. In the Bayesian inference framework, the estimation of the parameter vector $\boldsymbol{\Psi}$ of a model is performed by maximizing the posterior distribution $p(\boldsymbol{\Psi}|\mathbf{X})$ for a given prior distribution $p(\boldsymbol{\Psi})$ and a likelihood function $L(\mathbf{X}|\boldsymbol{\Psi})$. By using Bayes' theorem, the posterior distribution of the model parameters is defined, up to a constant, by

$$p(\boldsymbol{\Psi}|\mathbf{X}) \propto p(\boldsymbol{\Psi})L(\mathbf{X}|\boldsymbol{\Psi})\cdot \tag{4.14}$$

Often this posterior distribution is difficult to calculate directly. In such situations, we use techniques to simulate realizations from this distribution. These techniques, known as sampling techniques, are

grouped under the generic name of Markov Chain Monte Carlo (MCMC) (see for example Gilks et al. (1996); Robert and Casella (1999); Neal (1993)).

Markov Chain Monte Carlo The general principle of MCMC algorithms is to construct, from a target distribution $p(y)$, an ergodic Markov chain $(Y^{(1)}, \ldots, Y^{(M)})$ with stationary distribution equal to the target distribution, if we can sample from the conditional distributions $p(y_i|\boldsymbol{y}_{\setminus i})$ (i.e. as in Gibbs sampling) or more generally when we can calculate $\frac{p(y_i)}{p(y_j)}$ (i.e., as in Metropolis-Hasting). This is particularly useful when we can't directly sample from the target distribution $p(y)$, say as in inference in latent data models, particularly in mixture models. The, $Y^{(M)}$, for a sufficiently large value of M, can be considered as an approximate sample from the target distribution $p(y)$ (convergence in law). This principle of MCMC can also be used to approximate the expectation of any function $g(Y)$ by the ergodic mean

$$\mathbb{E}[g(Y)] = \lim_{x \to +\infty} \frac{1}{M} \sum_{t=1}^{M} g(y^{(t)}) \tag{4.15}$$

and hence in practice the expectation can be approximated by

$$\mathbb{E}[g(Y)] \approx \frac{1}{M - M_0} \sum_{t=M_0+1}^{M} g(y^{(t)}) \tag{4.16}$$

that is, after removing M_0 burn-in samples.

One of the most used MCMC algorithms is the Gibbs sampler, which will be considered frequently in the manuscript. The first form of Gibbs sampler goes back to Geman and Geman (1984) and was proposed in a framework of Bayesian image restoration. A very close to it was introduced by Tanner and Wong (1987) under the name of "data augmentation" for missing data data problems, and also presented in Gelfand and Smith (1990) and Diebolt and Robert (1994). The Gibbs sampler simulates successively realizations from the distribution of y_i from $\boldsymbol{y}$, conditional on the other components, that is:

$$y_i^{(t+1)} \sim p(y_i|y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_{i+1}^{(t)}, \ldots, y_n^{(t)})$$

and we then cycle iteratively until we have a sufficiently large number of samples. Of course one issue is how to determine the sufficient number of samples.

In Bayesian inference, the target distribution is the posterior distribution of the model parameters to be estimated $\boldsymbol{\Psi}$, that is, $p(\boldsymbol{\Psi}|\mathbf{X})$ given in (4.14). The sampling hence consists in drawing $(\boldsymbol{\Psi}^{(1)}, \ldots, \boldsymbol{\Psi}^{(M)})$ from the Markov chain to approximate the posterior.

However, in the latent data models, for example in mixture models, the unknown parameters are augmented by the hidden components labels $\mathbf{z}$ and thus the target distribution in this case corresponds to the posterior joint distribution of the model parameters $\boldsymbol{\Psi}$ and the component indicators $\mathbf{z}$. The sampling then consists in alternating between generating the missing labels $\mathbf{z}$ given the observations and the current parameter vector, that is, according to $p(\mathbf{z}|\mathbf{X}, \boldsymbol{\Psi}^{(t)})$, and the parameter vector $\boldsymbol{\Psi}$ given the observations and the current component labels, that is, according to $p(\boldsymbol{\Psi}|\mathbf{X}, \mathbf{z}^{(t)})$, to finally produce a Markov chain on the model parameters and another one on the missing labels. The posterior inference of mixtures with MCMC goes back to the first works of Tanner and Wong (1987) and Gelfand and Smith (1990). Other key initial papers on Bayesian inference of mixtures using MCMC include Diebolt and Robert (1994); Escobar and West (1994) as well as some more recent papers in the broad literature such as Richardson and Green (1997); Bensmail et al. (1997); Stephens (2000a); Celeux et al. (2000).
In the next sections, I introduce the two Bayesian models and their Bayesian inference using MCMC (Gibbs) sampling.

4.3.2 Mixtures of spatial spline regressions with mixed-effects

Before introducing the proposed Bayesian modeling, this section is dedicated to related work on mixture of spatial spline regressions (SSR) with mixed-effects (MSSR), introduced by Ng and McLachlan (2014), since the key difference between the two approaches resides in the added prior distributions on the model parameters and the resulting posterior inference. I first describe the regression model with linear mixed-effects and its mixture formulation, in the general case, and then describe the models for spatial regression data.

Regression with mixed-effects

The mixed-effects regression models (see for example Laird and Ware (1982), Verbeke and Lesaffre (1996) and Xu and Hedeker (2001)), are appropriate when the standard regression model (with fixed-effects) can not sufficiently explain the data. For example, when representing dependent data arising from related individuals or when data are gathered over time on the same individuals. In that case, the mixed-effects regression model is more appropriate as it includes both fixed-effects and random-effects terms. In the linear mixed-effects regression model, the $m_i \times 1$ response $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^T$ is modeled as:

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{T}_i\mathbf{b}_i + \mathbf{e}_i \tag{4.17}$$

where the $p \times 1$ vector $\boldsymbol{\beta}$ is the usual unknown fixed-effects regression coefficients vector describing the population mean, $\mathbf{b}_i$ is a $q \times 1$ vector of unknown subject-specific regression coefficients corresponding to individual effects, independently and identically distributed (i.i.d) according to the normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{R}_i)$ and independent from the $m_i \times 1$ error terms $\mathbf{e}_i$ which are distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\mathbf{X}_i$ and $\mathbf{T}_i$ are respectively $m_i \times p$ and $m_i \times q$ known covariate matrices. A common choice for the noise covariance-matrix is to take a diagonal matrix $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{m_i}$ where $\mathbf{I}_{m_i}$ denotes the $m_i \times m_i$ identity matrix. Thus, under this model, the joint distribution of the observations $\boldsymbol{y}_i$ and the random effects $\mathbf{b}_i$ is the following joint multivariate normal distribution (see for example Xu and Hedeker (2001)):

$$\begin{bmatrix} \boldsymbol{y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X}_i\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{I}_{m_i} + \mathbf{T}_i\mathbf{R}_i\mathbf{T}_i^T & \mathbf{T}_i\mathbf{R}_i \\ \mathbf{R}_i\mathbf{X}_i^T & \mathbf{R}_i \end{bmatrix} \right). \tag{4.18}$$

Then, from (4.18) it follows that the observations $\boldsymbol{y}_i$ are marginally distributed according to the following normal distribution (see Verbeke and Lesaffre (1996) and Xu and Hedeker (2001)):

$$f(\boldsymbol{y}_i|\mathbf{X}_i, \mathbf{T}_i; \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\mu}_i, \sigma^2\mathbf{I}_{m_i} + \mathbf{T}_i\mathbf{R}_i\mathbf{T}_i^T). \tag{4.19}$$

Mixture of regressions with mixed-effects

The regression model with mixed-effects (4.17) can be integrated into a finite mixture framework to deal with regression data arising from a finite number of groups. The resulting mixture of regressions model with linear mixed-effects (Verbeke and Lesaffre, 1996; Xu and Hedeker, 2001; Celeux et al., 2005; Ng et al., 2006) is a mixture model where every component k ($k = 1, \ldots, K$) is a regression model with mixed-effects given by (4.17), K being the number of mixture components. Thus, the observation $\boldsymbol{y}_i$ conditionally on each component k is modeled as:

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{T}_i\mathbf{b}_{ik} + \mathbf{e}_{ik} \tag{4.20}$$

where $\boldsymbol{\beta}_k$, $\mathbf{b}_{ik}$ and $\mathbf{e}_{ik}$ are respectively the fixed-effects regression coefficients, the random-effects regression coefficients for individual i, and the error terms, for component k. The random-effect coefficients $\mathbf{b}_{ik}$ are i.i.d according to $\mathcal{N}(\boldsymbol{\mu}_{ki}, \mathbf{R}_{ki})$ and are independent from the error terms $\mathbf{e}_{ik}$ which follow the distribution $\mathcal{N}(\mathbf{0}, \sigma_k^2\mathbf{I}_{m_i})$. Let Z_i denotes the categorical random variable representing the component memebership for the ith observation. Thus, conditional on the component $Z_i = k$, the observation $\boldsymbol{y}_i$ and the random effects $\mathbf{b}_i$ have the following joint multivariate normal distribution:

$$\left.\begin{bmatrix} \boldsymbol{y}_i \\ \mathbf{b}_i \end{bmatrix}\right|_{Z_i=k} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X}_i\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\mu}_k \\ \boldsymbol{\mu}_k \end{bmatrix}, \begin{bmatrix} \sigma_k^2\mathbf{I}_{m_i} + \mathbf{T}_i\mathbf{R}_{ki}\mathbf{T}_i^T & \mathbf{T}_i\mathbf{R}_{ki} \\ \mathbf{R}_{ki}\mathbf{X}_i^T & \mathbf{R}_{ki} \end{bmatrix} \right) \tag{4.21}$$

and thus the observations $\boldsymbol{y}_i$ are marginally distributed according to the following normal distribution :

$$f(\boldsymbol{y}_i|\mathbf{X}_i, \mathbf{T}_i, Z_i = k; \boldsymbol{\Psi}_k) = \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{T}_i\boldsymbol{\mu}_{ki}, \mathbf{T}_i\mathbf{R}_{ki}\mathbf{T}_i^T + \sigma_k^2\mathbf{I}_{m_i}). \tag{4.22}$$

The unknown parameter vector of this component-specific density is given by: $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \boldsymbol{\mu}_{k1}^T, \ldots, \boldsymbol{\mu}_{kn}^T, \text{vech}(\mathbf{R}_{k1})^T, \ldots, \text{vech}(\mathbf{R}_{kn})^T)^T$. Thus, the marginal distribution of $\boldsymbol{y}_i$ unconditional on component memberships is given by the following mixture distribution:

$$f(\boldsymbol{y}_i|\mathbf{X}_i, \mathbf{T}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{T}_i\boldsymbol{\mu}_{ki}, \mathbf{T}_i\mathbf{R}_{ki}\mathbf{T}_i^T + \sigma_k^2\mathbf{I}_{m_i}) \tag{4.23}$$

where the π_k's are the usual mixing proportions. The unknown mixture model parameters given by the parameter vector $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ where $\boldsymbol{\Psi}_k$ is the parameter vector of component k, are usually estimated, given an i.i.d sample of n observations, by maximizing the observed-data log-likelihood

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{T}_i\boldsymbol{\mu}_{ki}, \mathbf{T}_i\mathbf{R}_{ki}\mathbf{T}_i^T + \sigma_k^2\mathbf{I}_{m_i}) \tag{4.24}$$

via the EM algorithm as in (Verbeke and Lesaffre, 1996; Xu and Hedeker, 2001; Celeux et al., 2005; Ng et al., 2006).

Mixtures of spatial spline regressions with mixed-effects

For spatial regression data, Nguyen et al. (2014) introduced the spatial spline regression with liner mixed-effects (SSR). The model is given by (4.17) where the covariate matrices, which are assumed to be identical in Nguyen et al. (2014), that is, $\mathbf{T}_i = \mathbf{X}_i$ and denoted by $\mathbf{S}_i$, in this spatial case, represent a spatial structure and are calculated from the Nodal Basis Functions (NBF) (Malfait and Ramsay, 2003). Note that in what follows I will denote the number of columns of $\mathbf{S}_i$ by d. The NBF idea is an extension of the B-spline bases used in general for univariate or multivariate functions, to bivariate surfaces and was first introduced by Malfait and Ramsay (2003) and then used namely in Ramsay et al. (2011) and Sangalli et al. (2013) for surfaces. As in Nguyen et al. (2014), it is assumed that the random-effects are centered with isotropic covariance matrix common to all the individuals, that is $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \xi^2\mathbf{I}_{m_i})$. Thus, from (4.22) it follows that under the spatial spline regression model with linear mixed-effects, the density of the observation $\boldsymbol{y}_i$ is given by

$$f(\boldsymbol{y}_i|\mathbf{S}_i; \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{y}_i; \mathbf{S}_i\boldsymbol{\beta}, \xi^2\mathbf{S}_i\mathbf{S}_i^T + \sigma^2\mathbf{I}_{m_i}). \tag{4.25}$$

Then, under the mixture of spatial spline regression models with linear mixed-effects, the density of a surface is given by:

$$f(\boldsymbol{y}_i|\mathbf{S}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{S}_i\boldsymbol{\beta}_k, \xi_k^2\mathbf{S}_i\mathbf{S}_i^T + \sigma_k^2\mathbf{I}_{m_i}) \tag{4.26}$$

where $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \sigma_1^2, \ldots, \sigma_K^2, \xi_1^2, \ldots, \xi_K^2)^T$ is the model parameter vector. Both of models are fitted by using the EM algorithm (Nguyen et al., 2014). In particular, for the mixture of spatial spline regressions, the EM algorithm maximizes the following observed-data log-likelihood:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_i; \mathbf{S}_i\boldsymbol{\beta}_k, \xi_k^2\mathbf{S}_i\mathbf{S}_i^T + \sigma_k^2\mathbf{I}_{m_i}). \tag{4.27}$$

More details on the EM developments for the two models can be found in detail in Nguyen et al. (2014). Note that Nguyen et al. (2014) assumed a common noise variance σ^2 for all the mixture components in (4.26) and hence in (4.27).

4.3.3 Bayesian spatial spline regression with mixed-effects

I introduce a Bayesian probabilistic approach to the spatial spline regression model with mixed-effects presented in Nguyen et al. (2014) in a maximum likelihood context. The proposed model is thus the Bayesian spatial spline regression with linear mixed-effects (BSSR) model. I first present the model, the parameter distributions and then derive the Gibbs sampler for parameter estimation.

The model

The Bayesian spatial spline regression with mixed-effects (BSSR) model is defined by:

$$\boldsymbol{y}_i = \mathbf{S}_i(\boldsymbol{\beta} + \mathbf{b}_i) + \mathbf{e}_i \tag{4.28}$$

where the model parameters in this Bayesian framework are assumed to be random variables with specified prior distributions, and the spatial covariates matrix $\mathbf{S}_i$ is computed from the Nodal basis functions.

Introduced by Malfait and Ramsay (2003), the idea of Nodal basis functions (NBFs) extends the use of B-splines for univariate function approximation (Ramsay and Silverman, 2005), to the approximation of surfaces. For a fixed number of basis functions d, defined on a regular grid with regularly spaced points $c(l)$ $(l = 1, \ldots, d)$ of the domain we are working on, with d defined as $d = d_1 d_2$ where d_1 and d_2 are respectively the columns and rows number of nodes, the ith surface can be approximated using piecewise linear Lagrangian triangular finite element NBFs constructed as in Sangalli et al. (2013) and Nguyen et al. (2014) (see also [J-11]). Thus, this construction leads to the following $m_i \times d$ spatial covariates matrix:

$$\mathbf{S}_i = \begin{pmatrix} s(\boldsymbol{x}_1; \mathbf{c}_1) & s(\boldsymbol{x}_1; \mathbf{c}_2) & \cdots & s(\boldsymbol{x}_1; \mathbf{c}_d) \\ s(\boldsymbol{x}_2; \mathbf{c}_1) & s(\boldsymbol{x}_2; \mathbf{c}_2) & \cdots & s(\boldsymbol{x}_2; \mathbf{c}_d) \\ \vdots & \vdots & \ddots & \vdots \\ s(\boldsymbol{x}_{m_i}; \mathbf{c}_1) & s(\boldsymbol{x}_{m_i}; \mathbf{c}_2) & \cdots & s(\boldsymbol{x}_{m_i}; \mathbf{c}_d) \end{pmatrix} \tag{4.29}$$

where $s(\boldsymbol{x}; \mathbf{c})$ is a shortened notation of the NBF $s(\boldsymbol{x}, \mathbf{c}, \delta_1, \delta_2)$ with $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2})$ the two spatial coordinates of y_{ij} and $\mathbf{c} = (c_1, c_2)$ is a node center parameter, δ_1 and δ_1 being respectively the vertical and horizontal shape parameters representing the distances between two consecutive centers. An example of a NBF function defined on the rectangular domain $(x_1, x_2) \in [-1, 1] \times [-1, 1]$ with a single node $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$ is presented in the Figure 4.2.

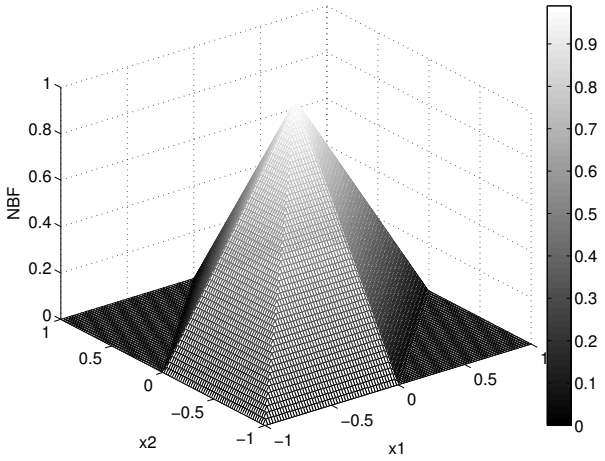

Figure 4.2: Nodal basis function $s(\boldsymbol{x}, \mathbf{c}, \delta_1, \delta_2)$, where $\mathbf{c} = (0, 0)$ and $\delta_1 = \delta_2 = 1$.

The model parameters of the proposed Bayesian model, which are given by the parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \sigma^2, \mathbf{b}_1, \ldots, \mathbf{b}_n, \xi^2)^T$ are assumed to be unknown random variables with the following prior distributions. I use conjugate priors for ease of calculation as those mostly used priors in the literature for example as in Diebolt and Robert (1994), Richardson and Green (1997) ,and Stephens (2000a). The used priors for the parameters are as follows:

$$\begin{array}{lcl} \boldsymbol{\beta} & \sim & \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \mathbf{b}_i | \xi^2 & \sim & \mathcal{N}(\mathbf{0}_d, \xi^2 \mathbf{I}_d) \\ \xi^2 & \sim & \mathcal{IG}(a_0, b_0) \\ \sigma^2 & \sim & \mathcal{IG}(g_0, h_0) \end{array} \tag{4.30}$$

where $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ are the hyper-parameters of the normal prior over the fixed-effects coefficients, ξ^2 is the variance of the normal distribution over the random-effect coefficients, a_0 and b_0 (respectively g_0 and h_0) are respectively the shape and scale parameters of the Inverse Gamma ($\mathcal{IG}$) prior over the variance ξ^2 (respectively σ^2).

Bayesian inference using Gibbs sampling

I use MCMC sampling for the Bayesian inference of the BSSR model. As seen before, MCMC sampling is indeed one of the most commonly used inference techniques in Bayesian analysis of mixtures, in particular

the Gibbs sampler (e.g see Diebolt and Robert (1994)). The Gibbs sampler is implemented by deriving the full conditional posterior distributions of the model parameters. Due to the chosen conjugate hierarchical prior (4.30) presented in the previous section, the full conditional posterior distributions can then be calculated analytically (see [J-11] for more detail). Applying the Bayes theorem to the joint distribution leads to the following posterior distributions used in the Gibbs sampler (see [J-11] for details). In what follows the notation |... is used to denote a conditioning of the parameter in question on all the other parameters and the observed data.

$$\boldsymbol{\beta}|... \sim \mathcal{N}(\boldsymbol{\nu}_0, \mathbf{V}_0) \quad \text{with} \quad \mathbf{V}_0^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i, \quad \boldsymbol{\nu}_0 = \mathbf{V}_0\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{S}_i\mathbf{b}_i) - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu_0}\right) \tag{4.31}$$

$$\mathbf{b}_i|... \sim \mathcal{N}(\boldsymbol{\nu}_1, \mathbf{V}_1) \quad \text{with} \quad \mathbf{V}_1^{-1} = \frac{1}{\sigma^2}\mathbf{S}_i^T\mathbf{S}_i + \frac{1}{\xi^2}, \quad \boldsymbol{\nu}_1 = \mathbf{V}_1\big(\frac{1}{\sigma^2}\mathbf{S}_i^T(\boldsymbol{y}_i - \mathbf{S}_i\boldsymbol{\beta})\big), \tag{4.32}$$

$$\sigma^2|... \sim \mathcal{IG}(g_1, h_1) \quad \text{with} \quad g_1 = g_0 + \frac{n}{2}, \quad h_1 = h_0 + \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{S}_i\boldsymbol{\beta} - \mathbf{S}_i\mathbf{b}_i)^T(\boldsymbol{y}_i - \mathbf{S}_i\boldsymbol{\beta} - \mathbf{S}_i\mathbf{b}_i)}{2}, \tag{4.33}$$

$$\xi^2|... \sim \mathcal{IG}(a_1, b_1) \quad \text{with} \quad a_1 = a_0 + \frac{n}{2}, \quad b_1 = b_0 + \frac{\sum_{i=1}^{n}\mathbf{b}_i^T\mathbf{b}_i}{2}. \tag{4.34}$$

The Gibbs sampler for the BSSR model then cycles by sampling from each of the above posterior distributions until a sufficiently large number of samples is reached.

4.3.4 Bayesian mixture of spatial spline regressions with mixed-effects

The BSSR model presented previously is dedicated to learn from a single surface or a set of homogeneous surfaces. However, when the data exhibit a grouping aspect, this may be restrictive, and its extension to accommodate clustered data is needed. I therefore integrate the BSSR model into a mixture framework. This is mainly motivated by a clustering prospective. The resulting model is therefore a Bayesian mixture of spatial spline regression with mixed-effects (BMSSR) and is described in the following section.

The model

Consider that there are K sub-populations within the n surfaces, that is the responses $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and their corresponding spatial covariates $(\mathbf{S}_1, \ldots, \mathbf{S}_n)$. The proposed BMSSR model has the following stochastic representation. Conditional on component k, the individual $\boldsymbol{y}_i$ given $\mathbf{S}_i$ is modeled by a BSSR model as:

$$\boldsymbol{y}_i = \mathbf{S}_i(\boldsymbol{\beta}_k + \mathbf{b}_{ik}) + \mathbf{e}_{ik}. \tag{4.35}$$

Thus, a K component Bayesian mixture of spatial spline regression models with mixed-effects (BMSSR) has the following density:

$$f(\boldsymbol{y}_i|\mathbf{S}_i;\boldsymbol{\Psi}) = \sum_{k=1}^{K}\pi_k\,\mathcal{N}\left(\boldsymbol{y}_i;\mathbf{S}_i(\boldsymbol{\beta}_k + \mathbf{b}_{ik}), \sigma_k^2\mathbf{I}_{m_i}\right) \tag{4.36}$$

where $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \mathbf{B}_1^T, \ldots, \mathbf{B}_K^T, \sigma_1^2, \ldots, \sigma_K^2, \xi_1^2, \ldots, \xi_K^2)^T$ is the parameter vector of the model, $\mathbf{B}_k = (\mathbf{b}_{1k}^T, \ldots, \mathbf{b}_{nk}^T)^T$ being the vector of the random-effect coefficients of the kth BSSR component. The BMSSR model is indeed composed of BSSR components, each of them is described by the parameters $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \mathbf{B}_k^T, \sigma_k^2, \xi_k^2)^T$ and a mixing proportion π_k. Therefore, conditional on the mixture component k, the parameter priors are defined similarly as in the BSSR model (4.30) presented in the previous section. For the BMSSR model, we therefore just need to specify the distribution on the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ which follow the Multinomial distribution in the generative model of the non-Bayesian mixture. I use a conjugate prior as for the other parameters, thats is, a Dirichlet prior with hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. The hierarchical prior from for the BMSSR model parameters is therefore given by:

$$\begin{array}{lcl} \pi & \sim & \mathcal{D}(\alpha_1, \ldots, \alpha_K) \\ \boldsymbol{\beta}_k & \sim & \mathcal{N}(\boldsymbol{\beta}_k|\boldsymbol{\mu_0}, \Sigma_0) \\ \mathbf{b}_{ik}|\xi_k^2 & \sim & \mathcal{N}(\mathbf{b}_{ik}|\mathbf{0}_d, \xi_k^2\mathbf{I}_d) \\ \xi_k^2 & \sim & \mathcal{IG}(\xi_k^2|a_0, b_0) \\ \sigma_k^2 & \sim & \mathcal{IG}(\sigma_k^2|g_0, h_0). \end{array} \tag{4.37}$$

Bayesian inference using Gibbs sampling

Once the model prior is defined, here I derive the full conditional posterior distributions needed for the Gibbs sampler to infer the model parameters. Further mathematical calculation details for these posterior distributions are given in [J-11]. Consider the vector of augmented parameters, which is the vector of parameters $(\boldsymbol{\pi}^T, \boldsymbol{\beta}^T, \mathbf{B}^T, \boldsymbol{\sigma}^{2^T}, \boldsymbol{\xi}^{2^T})^T$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)^T$, and $\boldsymbol{\xi}^2 = (\xi_1^2, \ldots, \xi_K^2)^T$, augmented by the unknown components labels $\mathbf{z} = (z_1, \ldots, z_n)$ and the observed data $\{\mathbf{S}_i, \boldsymbol{y}_i\}$. Let us also introduce the binary latent component-indicators Z_{ik} such that $Z_{ik} = 1$ iff $Z_i = k$, Z_i being the hidden label of the mixture component from which the ith individual is generated. Then, the full conditional distributions are given as follows:

$$Z_i|... \sim \mathcal{M}(1; \tau_{i1}, \ldots, \tau_{iK}) \text{ with } \tau_{ik}(1 \leq k \leq K) = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \mathbf{S}_i; \boldsymbol{\Psi}) = \frac{\pi_k \, \mathcal{N}\left(\boldsymbol{y}_i|\mathbf{S}_i(\boldsymbol{\beta}_k + \mathbf{b}_{ik}), \sigma_k^2 \mathbf{I}_{m_i}\right)}{\sum_{l=1}^K \pi_l \, \mathcal{N}\left(\boldsymbol{y}_i|\mathbf{S}_i(\boldsymbol{\beta}_l + \mathbf{b}_{il}), \sigma_l^2 \mathbf{I}_{m_i}\right)} \tag{4.38}$$

$$\boldsymbol{\pi}|... \sim \mathcal{D}\left(\alpha_1 + n_1, \ldots, \alpha_K + n_K\right) \text{ with } n_k = \sum_{i=1}^n Z_{ik} \tag{4.39}$$

$$\boldsymbol{\beta}_k|... \sim \mathcal{N}(\boldsymbol{\nu}_0, \mathbf{V}_0) \text{ with } \mathbf{V}_0^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_k^2} \sum_{i=1}^n Z_{ik} \, \mathbf{S}_i^T \mathbf{S}_i, \; \boldsymbol{\nu}_0 = \mathbf{V}_0 \left(\frac{1}{\sigma_k^2} \sum_{i=1}^n Z_{ik} \, \mathbf{S}_i^T \left(\boldsymbol{y}_i - \mathbf{S}_i \mathbf{b}_{ik}\right) - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu_0} \right) \tag{4.40}$$

$$\mathbf{b}_{ik}|... \sim \mathcal{N}(\boldsymbol{\nu}_1, \mathbf{V}_1) \text{ with } \mathbf{V}_1^{-1} = \frac{1}{\sigma_k^2} \mathbf{S}_i^T \mathbf{S}_i + \frac{1}{\xi_k^2} \mathbf{I}, \; \boldsymbol{\nu}_1 = \mathbf{V}_1 \left(\frac{1}{\sigma_k^2} \mathbf{S}_i^T (\boldsymbol{y}_i - \mathbf{S}_i \boldsymbol{\beta}_k) \right), \tag{4.41}$$

$$\sigma_k^2|... \sim \mathcal{IG}(g_1, h_1) \text{ with } g_1 = g_0 + \frac{1}{2} \sum_{i=1}^n Z_{ik}, \; h_1 = h_0 + \frac{\sum_{i=1}^n Z_{ik} \left(\boldsymbol{y}_i - \mathbf{S}_i \boldsymbol{\beta}_k - \mathbf{S}_i \mathbf{b}_{ik}\right)^T \left(\boldsymbol{y}_i - \mathbf{S}_i \boldsymbol{\beta}_k - \mathbf{S}_i \mathbf{b}_{ik}\right)}{2} \tag{4.42}$$

$$\xi_k^2|... \sim \mathcal{IG}\left(a_1, b_1\right) \text{ with } a_1 = a_0 + \frac{n}{2}, \; b_1 = b_0 + \frac{\sum_{i=1}^n \mathbf{b}_{ik}^T \mathbf{b}_{ik}}{2}. \tag{4.43}$$

The Gibbs sampler for the BMSSR then cycles by sampling from each of the above posterior distributions until a sufficiently large number of samples is reached.

The label switching problem

Here I open a parenthesis to discuss a well-known problem encountered in Bayesian inference of mixtures, that is, the one of label switching. The statistical inference is meaningful if the notion of identifiability is established. The estimation of $\boldsymbol{\Psi}$ is therefore meaningful if the model $f(.|\boldsymbol{\Psi})$ is identifiable, that is, when $f(\boldsymbol{y}_i|\boldsymbol{\Psi}) = f(\boldsymbol{y}_i|\boldsymbol{\Psi}^\star)$ if and only if $\boldsymbol{\Psi} = \boldsymbol{\Psi}^\star$. It is well known that mixture models are not identifiable in the strict sense, but a weak identifiability can be established for them, that is, identifiability up to a permutation. As discussed for example in (McLachlan and Peel., 2000, Section 1.14), this problem is not of concern in maximum likelihood fitting of mixtures via the EM algorithm. However, identifiably in mixtures is of concern in the Bayesian framework where in the posterior simulation the mixture component labels can be interchanged from one sample to another. This problem is known as the label-switching problem. Different strategies were proposed in the literature to deal with this problem. One simple way to deal with label switching is to impose constraints on the model parameters to force an unique labeling in the MCMC sampling, and hence ensure identifiability. For example one may use ordering constrains on the parameters as in Richardson and Green (1997) for the case of univariate Gaussian mixtures, e.g., constraints on the means, the variances, or the mixing proportions. This was also discussed in Marin et al. (2005). However, Celeux (1999); Celeux et al. (2000) showed that this strategy of forcing constrains on the model parameters is not efficient and, if it works, does not scale to higher dimensions. Another approach is to post-process the posterior parameter samples by searching for the labels permutation that minimizes some loss function as in Stephens (2000b). As discussed in Celeux (1999) and Celeux et al. (2000), while this procedure works well, it can be numerically demanding as it is an offline algorithm needing storing significant amount of data samples, and it is also restricted to the limited framework of Bayesian analysis of latent structure models with conjugate prior distributions. Celeux (1999); Celeux et al. (2000) proposed a better solution in the same spirit of the one of Stephens which consists of a sequential k-means like algorithm to cluster the posterior samples and which has several advantages. It is quite simple, not specific to Bayesian analysis with conjugate prior distributions or to the mixture context, and it is not numerically demanding. So what is suggested here is to relabel the obtained posterior parameter samples when the label switching happens by the K-means-like algorithm of Celeux (1999); Celeux et al. (2000).

Model-based surface clustering using the BMSSR

In addition to Bayesian density estimation, The previously presented BMSSR model can also be used for Bayesian model-based surface clustering to provide a partition of the data into K clusters. Model-based clustering using the BMSSR model consists in assuming that the observed data $\{\mathbf{S}_i, \boldsymbol{y}_i\}_{i=1}^n$ are generated from a K component mixture of spatial spline regressions with mixed-effects with parameter vector $\boldsymbol{\Psi}$. The mixture components can be interpreted as clusters and hence each cluster can be associated with a mixture component. The problem of clustering therefore becomes the one of estimating the BMSSR parameters $\boldsymbol{\Psi}$. This is performed here by Gibbs sampling which provides a MAP estimator $\hat{\boldsymbol{\Psi}}_{\text{MAP}}$, which can be obtained by averaging the Gibbs posterior sample after removing some initial samples corresponding to a burn-in period. A partition of the data can then be obtained from the posterior memberships by applying the Bayes' optimal allocation rule, that is, by maximizing the posterior component probabilities to assign each surface to a component (cluster): $\hat{z}_i = \arg\max_{k=1}^{K} \tau_{ik}(\hat{\boldsymbol{\Psi}}_{\text{MAP}})$ where $\hat{z}_i$ represents the estimated cluster label for the ith surface.

4.3.5 Experiments

Two proposed Bayesian models were experimented in [J-11] on simulated surfaces and real surfaces issued from a handwritten character recognition problem by considering real images from the MNIST data set (LeCun et al., 1998) to test it in terms of surface approximation and clustering. I first considered bi-dimensional arbitrary non-linear functions and I attempted to approximate it from a sample of simulated noisy surfaces generated on a square domain in order to test the model in terms of surface approximation. The simulated data include mixed effects. The fitted mean surfaces using a reasonable number of basis functions, is very close to the true one. This is confirmed by the obtained small values (i.e 0.0865) of the empirical sum of squared error (SSE) between the true surface and the fitted one.

Figure 4.3 shows an example of actual arbitrary mean function before the noise and the random effects are added, an example of simulated surface the fitted mean surface $\hat{\mu}(\boldsymbol{x}) = \mathbf{S}_i\hat{\boldsymbol{\beta}}$ from a set of 100 surfaces with $d = 15 \times 15$ NBFs.

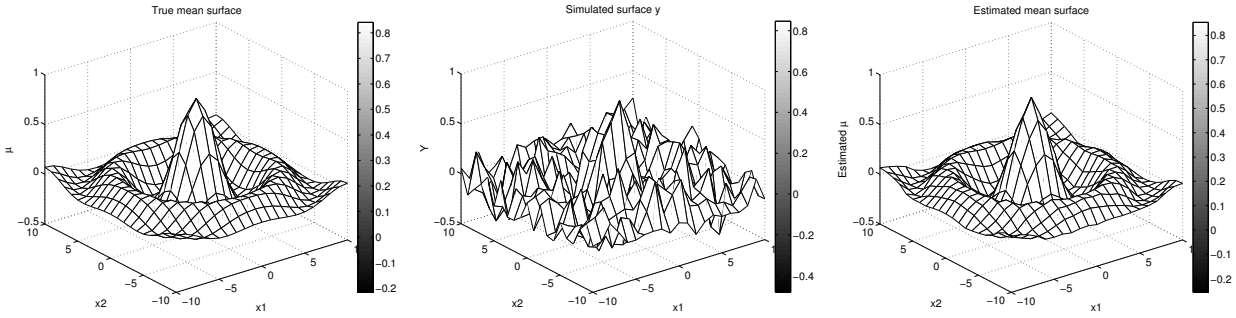

Figure 4.3: True mean surface (left), an example of noisy surface (middle), A BSSR fit from 100 surfaces using 15×15 NBFs (right).

Then, the second model, that is the BMSSR, was applied on a subset of the ZIPcode data set Hastie et al. (2010), which is issued from the MNIST data set. The data set contains 9298 16 by 16 pixel gray scale images of Hindu-Arabic handwritten numerals. Each individual $\boldsymbol{y}_i$ contains $m_i = 256$ observations and the Gibbs sampler is run with different numbers of clusters on a subset of 1000 digits randomly chosen from the Zipcode testing set. The best solution is selected in terms of the Adjusted Rand Index (ARI) values, which promotes a partition with $K = 12$ clusters. The cluster means for the partition obtained by the proposed Bayesian model (BMSSR) clearly shows that the model is able to recover the ten digits, and not surprisingly has revealed subgroups of some digits (0 and 5).

Figure 4.4 shows the cluster means for the obtained clusters by the proposed Baysian model (BMSSR).

Figure 4.4: Cluster mean images obtained by the proposed BMSSR model on an MNIST set with 12 mixture components.

4.3.6 Conclusion

In this section I first presented a probabilistic Bayesian model for homogeneous spatial data based on spatial spline regression with mixed-effects (BSSR). The model is able to accommodate individuals with both fixed and random effect variability. Then, motivated by a model-based surface clustering perspective, I introduced the Bayesian mixture of spatial spline regressions with mixed-effects (BMSSR) for spatial functional data dispersed into groups. I derived Gibbs samplers to infer the models. Application on simulated surfaces illustrates the surface approximation using the BSSR model. Then, application on real data in a handwritten digit recognition framework shows the potential benefit of the proposed BMSSR model for Bayesian surface clustering. The BMSSR can be extended to be used for supervised surface classification. This can be performed without difficulty by modeling each class by a BMSSR model and then applying the Bayes rule to assign a new unlabeled surface to the class corresponding to the highest posterior probability. One future work might also concern the assessment of the performance of the Bayesian mixture model in the case where the data (e.g. the handwritten character images) are sparsely sampled by introducing missing data as in Nguyen et al. (2014). Since the BMSSR is a latent (missing) data model, it can be applied directly without data imputation unlike other competitors. Then, another interesting perspective is to derive a Bayesian non-parametric model by relying on Dirichlet Process mixture models where the number of mixture components can be directly inferred from the data.

Chapter 5

Bayesian non-parametric parsimonious mixtures for multivariate data

Contents

5.1 Introduction 65
5.1.1 Personal contribution 67
5.2 Finite mixture model model-based clustering 67
5.2.1 Bayesian model-based clustering 68
5.2.2 Parsimonious Gaussian mixture models 68
5.3 Dirichlet Process Parsimonious Mixtures 69
5.3.1 Dirichlet Process Parsimonious Mixtures 69
5.3.2 Chinese Restaurant Process parsimonious mixtures 71
5.3.3 Bayesian inference via Gibbs sampling 72
5.3.4 Bayesian model comparison via Bayes factors 73
5.3.5 Experiments 74
5.4 Conclusion 77

Related PhD thesis:

[1] M. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015

Related journal papers:

[1] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet Process Parsimonious Gaussian Mixture for clustering. *arXiv:1501.03347*, 2015. URL `http://arxiv.org/pdf/1501.03347.pdf`. Submitted

[2] F. Chamroukhi et al. Bayesian non-parametric models for unsupervised decomposition of whale songs. *Journal of Acoustical Society of America*, 2015. In preparation

Related conference papers:

[1] M. Bartcus, F. Chamroukhi, and H. Glotin. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In *The International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015

[2] F. Chamroukhi, M. Bartcus, and H. Glotin. Bayesian non-parametric parsimonious Gaussian mixture for clustering. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, August 2014a

[3] F. Chamroukhi, Marius Bartcus, and Herve Glotin. Bayesian non-parametric parsimonious clustering. In *Proceedings of 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, April 2014b

[4] M. Bartcus and F. Chamroukhi. Hierarchical Dirichlet Process Hidden Markov Model for unsupervised learning from bioacoustic data. In *Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML)*, Beijing, June 2014

[5] Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture. In *Proceedings of the NIPS4B workshop, Neural Information Processing Systems (NIPS)*, pages 205–211, Nevada, USA, 2013

[6] M. Bartcus, F. Chamroukhi, and H. Glotin. Clustering Bayésien Parcimonieux Non-Paramétrique. In *Extraction et Gestion des Connaissances (EGC), Atelier CluCo : Clustering et Co-clustering*, pages 3–13, Rennes, France, Jan 2014

I initiated this research direction in 2012, with the beginning of the PhD thesis of Marius Bartcus, for whom I was the principal supervisor. In this research I investigate the mixture models for multivariate data in a fully Bayesian framework. It is structured into two parts. The first one corresponds to the investigation of what can be called parametric Bayesian mixtures and their inference using mainly Bayesian sampling, with a particular focus on the finite parsimonious mixtures which offer a great modeling flexibility. The second one however addresses the problem from a non-parametric perspective by investigating the Dirichlet process Mixture derivation for Bayesian mixtures which can be interpreted as an infinite mixture model, with particularly the derivation of new Dirichlet process parsimonious mixtures. This research has lead, until this day, to the following publications: [J-10] [C-1] [C-2] [C-4] [C-3] [C-6] [C-5] and an application paper [J-17] is in preparation for submission to a specialized journal.

5.1 Introduction

In this axis, I consider the problem of Bayesian inference for fitting multivariate Gaussian mixtures. The framework of Bayesian inference was already introduced in the second part of the previous chapter dedicated to the Bayesian models for spatial functional data. In this Chapter, I revisit the classical problem of fitting Gaussian mixtures from multivariate data and I'll focus on the parsimonious mixtures which are promoted to fit flexible structures to high dimensional data and can be considered as a dimensionality reduction method. The angle of approach compared to the previous Chapter is different though, since here I will be placed mainly in the Bayesian non-parametric framework where the number of mixture components is unbounded, that is, by considering the infinite mixture modeling using Dirichlet Process mixture models or by equivalence the Chinese Restaurant Process mixtures. The considered application is clustering which is one of the essential tasks in statistics and machine learning. Model-based clustering, that is the clustering approach based on the parametric finite mixture model (McLachlan and Peel., 2000), is one of the most popular and successful approaches in cluster analysis (McLachlan and Basford, 1988; Banfield and Raftery, 1993; Fraley and Raftery, 2002). The finite mixture model decomposes the density of the observed data as a weighted sum of a finite number of K component densities. Most often, the used model for multivariate real data is the finite Gaussian mixture model (GMM) in which each mixture component is Gaussian. This chapter will be focusing on Gaussian mixture modeling for multivariate real data. In Banfield and Raftery (1993) and Celeux and Govaert (1995), the authors developed a parsimonious GMM clustering approach by exploiting an eigenvalue decomposition of the group covariance matrices of the GMM components, which provides a wide range of very flexible models with different clustering criteria. It was also demonstrated in Fraley and Raftery (2002) that the parsimonious mixture model-based clustering framework provides very good results in density estimation as well as in cluster and discriminant analyses. In model-based clustering using GMMs, the parameters of the Gaussian mixture are usually estimated in a maximum likelihood estimation (MLE) framework by maximizing the observed data likelihood. This is usually performed by the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) or EM extensions (McLachlan and Krishnan, 2008). The parameters of the parsimonious Gaussian mixture models can also be estimated in a MLE framework by using the EM algorithm (Celeux and Govaert, 1995). However, a possible issue in the MLE approach using the EM algorithm for normal mixtures is that it may fail due to singularities or degeneracies, as highlighted namely in Stephens (1997); Snoussi and Mohammad-Djafari (2001, 2005); Fraley and Raftery (2005) and Fraley and Raftery (2007). The Bayesian estimation methods for mixture models have lead to intensive research in the field for dealing with the problems encountered in MLE for mixtures (Diebolt and Robert, 1994; Escobar and West, 1994; Robert, 2007; Richardson and Green, 1997; Stephens, 1997; Bensmail et al., 1997; Bensmail and Meulman, 2003; Marin et al., 2005; Gelman et al., 2003) which rely on a Bayesian formulation of the the mixture model. They allow to avoid these problems by replacing the MLE by the maximum a posterior (MAP) estimator. This is namely achieved by introducing a regularization over the model parameters via prior parameter distributions, which are assumed to be uniform in the case of MLE. The MAP estimation for the Bayesian Gaussian mixture is performed by maximizing the posterior parameter distribution. This can be performed, in some situations by an EM-MAP scheme as in Fraley and Raftery (2005) and Fraley and Raftery (2007) where the authors proposed an EM algorihtm for estimating Bayesian parsimonious Gaussian mixtures. However, the common estimation approach in the case of Bayesian mixtures is still the one based on Bayesian sampling such as Markov Chain Monte Carlo (MCMC), namely Gibbs sampling (Diebolt and Robert, 1994; Stephens, 1997; Bensmail et al., 1997)

when the number of mixture components K is known, or by reversible jump MCMC introduced by Green (1995) as in Richardson and Green (1997) and Stephens (1997), when this one is unknown. The principle of Bayesian inference using MCMC was described in Section 4.3.1.
The flexible eigenvalue decomposition of the group covariance matrix described previously was also exploited in Bayesian parsimonious model-based clustering by Bensmail et al. (1997); Bensmail and Meulman (2003) where the authors used a Gibbs sampler for the model inference. For these model-based clustering approaches, the number of mixture components is usually assumed to be known. Another issue in the finite mixture model-based clustering approach, including the MLE approach as well as the MAP approach, is therefore the one of selecting the optimal number of mixture components, that is the problem of model selection. The model selection is in general performed through a two-fold strategy by selecting the best model from pre-established inferred model candidates. For the MLE approach, the choice of the optimal number of mixture components can be performed via penalized log-likelihood criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Akaike Information Criterion (AIC) (Akaike, 1974), the Approximate Weight of Evidence (AWE) criterion (Banfield and Raftery, 1993), or the Integrated Classification Likelihood criterion (ICL) (Biernacki et al., 2000), etc. For the MAP approach, this can still be performed via modified penalized log-likelihood criteria such as a modified version of BIC as in (Fraley and Raftery, 2007) computed for the posterior mode, and more generally the Bayes factors (Kass and Raftery, 1995) as in Bensmail et al. (1997) for parsimonious mixtures. Bayes factors are indeed the natural Bayesian criterion for model selection and comparison in the Bayesian framework and for which the criteria such as BIC, AWE, etc represent indeed approximations. There is also Bayesian extensions for mixture models that analyze mixtures with unknown number of components, for example as mentioned before the one of Richardson and Green (1997) using RJMCMC and the one of Stephens (2000a, 1997) using the birth and death process. They are referred to as fully Bayesian mixture models (Richardson and Green, 1997) as they consider the number of mixture components as a parameter to be inferred from the data, jointly with the mixture model parameters, based on the posterior distributions.

However, these standard finite mixture models, including the non-Bayesian and the Bayesian ones, are parametric and may not be well adapted in the case of unknown and complex data structure. Recently, the Bayesian-non parametric (BNP) formulation of mixture models, that goes back to Ferguson (1973) and Antoniak (1974), have took much attention as a nonparametric alternative for formulating mixtures. The BNP methods (Robert, 2007; Hjort et al., 2010) have indeed recently become popular due to their flexible modeling capabilities and advances in inference techniques, in particular for mixture models, by using namely MCMC sampling techniques (Neal, 2000; Rasmussen, 2000) or variational inference ones (Blei and Jordan, 2006). BNP methods for clustering, including Dirichlet Process Mixtures (DPM) and Chinese Restaurant Process (CRP) mixtures (Ferguson, 1973; Antoniak, 1974; Pitman, 1995; Wood and Black, 2008; Samuel and Blei, 2012) which can be represented as infinite Gaussian mixture models as in Rasmussen (2000), provide a principled way to overcome the issues in standard model-based clustering and classical Bayesian mixtures for clustering. They are fully Bayesian approaches that offer a principled alternative to jointly infer the number of mixture components (i.e clusters) and the mixture parameters, from the data. By using general processes as priors, they allow to avoid the problem of singularities and degeneracies of the MLE, and to simultaneously infer the optimal number of clusters from the data, in a one-fold scheme, rather than in a two-fold approach as in standard model-based clustering. They also avoid assuming restricted functional forms and thus allow the complexity and accuracy of the inferred models to grow as more data is observed. They also represent a good alternative to the difficult problem of model selection in parametric mixture models. Note that the term non-parametric does not mean that there are no parameters, it rather means that one would have more and more parameters, as more data are observed.

In this chapter, I present a new BNP formulation of the Gaussian mixture with the eigenvalue decomposition of the group covariance matrix of each Gaussian component which has proven its flexibility in cluster analysis for the parametric case (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 2002; Bensmail et al., 1997). We develop new Dirichlet Process mixture models with parsimonious covariance structure, which results in Dirichlet Process Parsimonious Mixtures (DPPM). They represent a Bayesian nonparametric formulation of these parsimonious Gaussian mixture models. The proposed DPPM models are Bayesian parsimonious mixture models with a Dirichlet Process prior and thus provide a principled way to overcome the issues encountered in the parametric Bayesian and non-Bayesian case and allow to automatically and simultaneously infer the model parameters and the optimal model structure from the data, from different models, going from simplest spherical ones to the

more complex standard general one. We develop a Gibbs sampling technique for maximum a posteriori (MAP) estimation of the various models and provide an unifying framework for model selection and models comparison by using namely Bayes factors, to simultaneously select the optimal number of mixture components and the best parsimonious mixture structure. The proposed DPPM are more flexible in terms of modeling and their use in clustering, and automatically infer the number of clusters from the data.

5.1.1 Personal contribution

My contribution in this direction is two-fold. The first one consists in investigating Bayesian mixtures and their inference using mainly MCMC sampling, with a particular focus on the finite parsimonious mixtures which offer great modeling flexibilities. The second one, however, addresses the problem from a non-parametric perspective by investigating the Dirichlet process mixtures. I developed a Bayesian nonparametric formulation for the parsimonious mixture models. By relying on Dirichlet Process mixtures, or by equivalence the Chinese Restaurant Process mixtures, I introduced Dirichlet Process Parsimonious mixture models (DPPMs), which provide a flexible framework for modeling different data structures as well as a good alternative to tackle the problem of model selection. I derive a Gibbs sampler to infer the models and use Bayes Factors for Bayesian model comparison. Applications and comparisons on several data sets highlight the effectiveness of the proposed nonparametric parsimonious mixture models as a good nonparametric alternative for the parametric parsimonious models. The models have also shown very encouraging performance in a challenging problem of unsupervised bioacoustic signals decomposition application.

This chapter is organized as follows. Section 5.2 describes and discusses previous work on model-based clustering. Then, section 5.3 presents the proposed models and the learning technique. In section 5.3.5, we give experimental results to evaluate the proposed models on simulated data and real data. Finally, Section 5.4 is devoted to a discussion and concluding remarks.

5.2 Finite mixture model model-based clustering

Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a sample of n i.i.d observations in $\mathbb{R}^d$, and let $\mathbf{z} = (z_1, \ldots, z_n)$ be the corresponding unknown cluster labels where $z_i \in \{1, \ldots, K\}$ represents the cluster label of the ith data point $\boldsymbol{x}_i$, K being the possibly unknown number of clusters.
Parametric Gaussian clustering, also called model-based clustering (McLachlan and Basford, 1988; Fraley and Raftery, 2002), is based on the finite GMM (McLachlan and Peel., 2000) in which the probability density function of the data is given by:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{x}_i|\theta_k) \tag{5.1}$$

where the π_k's are the mixing proportions, $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are respectively the mean vector and the covariance matrix for the kth Gaussian component density and $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T, \text{vech}(\boldsymbol{\Sigma}_1)^T, \ldots, \text{vech}(\boldsymbol{\Sigma}_K)^T)^T$ is the GMM parameter vector. From a generative point of view, the generative process of the data for the finite mixture model can be stated as follows. First, a mixture component z_i is sampled independently from a Multinomial distribution given the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Then, given the mixture component $z_i = k$, and the corresponding parameters $\boldsymbol{\theta}_k$, the individual $\boldsymbol{x}_i$ is generated independently from a Gaussian with parameters $\boldsymbol{\theta}_k$, that is:

$$z_i \sim \mathcal{M}(\boldsymbol{\pi}) \tag{5.2}$$

$$\boldsymbol{x}_i|\theta_{z_i} \sim \mathcal{N}(\boldsymbol{x}_i|\theta_{z_i}). \tag{5.3}$$

The mixture model parameters $\boldsymbol{\theta}$ is usually estimated in a Maximum Likelihood Estimation (MLE) framework by maximizing the observed data likelihood (5.4):

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{x}_i|\theta_k). \tag{5.4}$$

via the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) or EM extensions (McLachlan and Krishnan, 2008).

5.2.1 Bayesian model-based clustering

As mentioned in the introduction, the MLE approach using the EM algorithm for normal mixtures may fail in some situations due to singularities or degeneracies (Stephens, 1997; Fraley and Raftery, 2005, 2007). The Bayesian approach of mixture models avoids the problems associated with the MLE via a MAP estimation framework by maximizing the posterior parameter distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X}), \tag{5.5}$$

$p(\boldsymbol{\theta})$ being a chosen prior distribution over the model parameters $\boldsymbol{\theta}$. The prior distribution in general takes the following form for the GMM:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0)p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \boldsymbol{\Lambda}_0, \nu) = p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)p(\boldsymbol{\Sigma}_k).$$

where $(\boldsymbol{\alpha}, \boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Lambda}_0, \nu_0)$ are hyperparameters. A common choice for the GMM is to assume conjugate priors, that is Dirichlet distribution for the mixing proportions as in Richardson and Green (1997) and Ormoneit and Tresp (1998), and a multivariate normal Inverse-Wishart prior distribution for the Gaussian parameters, that is a multivariate normal for the means and an Inverse-Wishart for the covariance matrices, for example as in Bensmail et al. (1997), Fraley and Raftery (2005) and Fraley and Raftery (2007).

From a generative point of view, to generate data from the Bayesian GMM, a first step is to sample the model parameters from the prior, that is to sample the mixing proportions from their conjugate Dirichlet prior distribution, and the mean vectors and the covariance matrices of the Gaussian components from the corresponding conjugate multivariate normal Inverse-Wishart prior. Then, the generative procedure remains the same as in the previously described generative process for the non-Bayesian finite mixture, and is summarized by the following steps:

$$\begin{aligned} \boldsymbol{\pi}|\boldsymbol{\alpha} &\sim \mathcal{D}(\boldsymbol{\alpha}) \\ z_i|\boldsymbol{\pi} &\sim \mathcal{M}(\boldsymbol{\pi}) \\ \boldsymbol{\theta}_{z_i}|G_0 &\sim G_0 \\ \boldsymbol{x}_i|\boldsymbol{\theta}_{z_i} &\sim \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}_{z_i}) \end{aligned} \tag{5.6}$$

where $\boldsymbol{\alpha}$ are hyperparameters of the Dirichlet prior distribution, and G_0 is a prior distribution for the parameters of the Gaussian component, that is a multivariate Normal Inverse-Wishart:

$$\boldsymbol{\Sigma}_k \sim \mathcal{IW}(\nu_0, \Lambda_0) \tag{5.7}$$

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}}{\kappa_0}) \tag{5.8}$$

where the $\mathcal{IW}$ stands for the Inverse-Wishart distribution.

The parameters $\boldsymbol{\theta}$ of the Bayesian Gaussian mixture are estimated by MAP estimation by maximizing the posterior parameter distribution (5.5). The MAP estimation can still be performed by EM, namely in the case of conjugate priors where the prior distribution is only considered for the parameters of the Gaussian components, as in Fraley and Raftery (2005) and Fraley and Raftery (2007). However, in general, the common estimation approach in the case the Bayesian GMM described above, is the one using Bayesian sampling such as MCMC sampling techniques, namely the Gibbs sampler (Geyer, 1991; Neal, 1993; Diebolt and Robert, 1994; Bensmail et al., 1997; Ormoneit and Tresp, 1998; Stephens, 1997).

5.2.2 Parsimonious Gaussian mixture models

The GMM clustering has been extended to parsimonious GMM clustering (Banfield and Raftery, 1993; Celeux and Govaert, 1995) by exploiting an eigenvalue decomposition of the group covariance matrices,

which provides a wide range of very flexible models with different clustering criteria. In these parsimonious models, the group covariance matrix $\mathbf{\Sigma}_k$ for each cluster k is decomposed as

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{5.9}$$

where $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$, $\mathbf{D}_k$ is an orthogonal matrix of eigenvectors of $\mathbf{\Sigma}_k$ and $\mathbf{A}_k$ is a diagonal matrix with determinant 1 whose diagonal elements are the normalized eigenvalues of $\mathbf{\Sigma}_k$ in a decreasing order. As described in Celeux and Govaert (1995), the scalar λ_k determines the volume of cluster k, $\mathbf{D}_k$ its orientation and $\mathbf{A}_k$ its shape. Thus, this decomposition leads to several flexible models going from simplest spherical models to the complex general one and hence is adapted to various clustering situations.

The parameters $\boldsymbol{\theta}$ of the parsimonious Gaussian mixture models are estimated in a MLE framework by using the EM algorithm. The details of the EM algorithm for the different parsimonious finite GMMs are given in Celeux and Govaert (1995). The parsimonious GMMs have also took much attention under the Bayesian prospective. For example, in Bensmail et al. (1997), the authors proposed a fully Bayesian formulation for inferring the previously described parsimonious finite Gaussian mixture models. This Bayesian formulation was applied in model-based cluster analysis (Bensmail et al., 1997; Bensmail and Meulman, 2003). The model inference in this Bayesian formulation is performed in a MAP estimation framework by using MCMC sampling techniques, see for example (Bensmail et al., 1997; Bensmail and Meulman, 2003). Another Bayesian regularization for the parsimonious GMM was proposed by Fraley and Raftery (2005, 2007) in which the maximization of the posterior can still be performed by the EM algorithm in the MAP framework (EM-MAP). Here we consider the parsimonious GMMs (PGMMs) mainly in a Bayesian non-parametric framework as well see in what follows, instead of into a finite (Bayesian) mixture. This as will see helps namely to tackle the problem of model selection from non-parametric prospective.

Model selection in finite mixture models Finite mixture model-based clustering requires to specify the number of mixture components (i.e., clusters) and, in the case of parsimonious models, the type of the model. The main issues in this parametric model are therefore the one of selecting the number of mixture components (clusters), and possibly the type of the model, that fit at best the data. This problem can be tackled by penalized log-likelihood criteria such as BIC (Schwarz, 1978) or penalized classification log-likelihood criteria such as AWE (Banfield and Raftery, 1993) or ICL (Biernacki et al., 2000), etc, or more generally by using Bayes factors (Kass and Raftery, 1995) which provide a general way to select and compare models in (Bayesian) statistical modeling, namely in Bayesian mixture models.

5.3 Dirichlet Process Parsimonious Mixtures

The Bayesian and non-Bayesian finite mixture models described previously are however in general parametric and may not be well adapted to represent complex and realistic data sets. Recently, the Bayesian-non parametric (BNP) mixtures, in particular the Dirichlet Process Mixture (DPM) (Ferguson, 1973; Antoniak, 1974; Wood and Black, 2008; Samuel and Blei, 2012) or by equivalence the Chinese Restaurant Process (CRP) mixture (Aldous, 1985; Pitman, 2002; Samuel and Blei, 2012), which can be seen as an infinite mixture model (Rasmussen, 2000), provide a principled way to overcome the issues in standard model-based clustering and classical Bayesian mixtures for clustering. They are fully Bayesian approaches and offer a principled alternative to jointly infer the number of mixture components (i.e clusters) and the mixture parameters, from the data. BNP mixture approaches for clustering assume general process as prior on the infinite possible partitions, which is not restrictive as in classical Bayesian inference. Such a prior can be a Dirichlet Process (Ferguson, 1973; Antoniak, 1974; Samuel and Blei, 2012) or, by equivalence, a Chinese Restaurant Process (Pitman, 2002; Samuel and Blei, 2012). In the next section, we rely on the Dirichlet Process Mixture (DPM) formulation to derive the proposed Bayesian non-parametric formulation of the parsimonious models.

5.3.1 Dirichlet Process Parsimonious Mixtures

A Dirichlet Process (DP) (Ferguson, 1973) is a distribution over distributions and has two parameters, the concentration parameter $\alpha_0 > 0$ and the base measure G_0. We denote it by $\text{DP}(\alpha, G_0)$. Assume there is a parameter $\tilde{\boldsymbol{\theta}}_i$ following a distribution G, that is $\tilde{\boldsymbol{\theta}}_i | G \sim G$. Modeling with DP means that we

assume that the prior over G is a DP, that is, G is itself generated from a DP: $G \sim \text{DP}(\alpha, G_0)$. This can be summarized by the following generative process:

$$\tilde{\boldsymbol{\theta}}_i|G \sim G,\ \forall i \in 1,\ldots,n \tag{5.10}$$
$$G|\alpha, G_0 \sim \text{DP}(\alpha, G_0)\cdot \tag{5.11}$$

The DP has two properties (Ferguson, 1973). First, random distributions drawn from DP, that is $G \sim \text{DP}(\alpha, G_0)$, are discrete. Thus, there is a strictly positive probability of multiple observations taking identical values within the set $(\tilde{\boldsymbol{\theta}}_1, \cdots, \tilde{\boldsymbol{\theta}}_n)$. Suppose we have a random distribution G drawn from a DP followed by repeated draws $(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n)$ from that random distribution, Blackwell and MacQueen (1973) introduced a Pólya urn representation of the joint distribution of the random variables $(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n)$, that is

$$p(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n) = p(\tilde{\boldsymbol{\theta}}_1)p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1)p(\tilde{\boldsymbol{\theta}}_3|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)\ldots p(\tilde{\boldsymbol{\theta}}_n|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{n-1}), \tag{5.12}$$

which is obtained by marginalizing out the underlying random measure G:

$$p(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n|\alpha, G_0) = \int \left(\prod_{i=1}^{n} p(\tilde{\boldsymbol{\theta}}_i|G)\right) \mathrm{d}p(G|\alpha, G_0) \tag{5.13}$$

and results in the following Pólya urn representation for the calculation of the predictive terms of the joint distribution (5.12):

$$\tilde{\boldsymbol{\theta}}_i|\tilde{\boldsymbol{\theta}}_1, \ldots\tilde{\boldsymbol{\theta}}_{i-1} \sim \frac{\alpha_0}{\alpha_0 + i - 1}G_0 + \sum_{j=1}^{i-1}\frac{1}{\alpha_0 + i - 1}\delta_{\tilde{\boldsymbol{\theta}}_j} \tag{5.14}$$
$$\sim \frac{\alpha_0}{\alpha_0 + i - 1}G_0 + \sum_{k=1}^{K_{i-1}}\frac{n_k}{\alpha_0 + i - 1}\delta_{\boldsymbol{\theta}_k} \tag{5.15}$$

where K_{i-1} is the number of clusters after $i-1$ samples, n_k denotes the number of times each of the parameters $\{\boldsymbol{\theta}_k\}_{k=1}^{\infty}$ occurred in the set $\{\tilde{\boldsymbol{\theta}}_i\}_{i=1}^{n}$. The DP therefore places its probability mass on a countability infinite collection of points, also called atoms, that is an infinite mixture of Dirac deltas (Ferguson, 1973; Sethuraman, 1994; Samuel and Blei, 2012):

$$G = \sum_{k=1}^{\infty}\pi_k\delta_{\boldsymbol{\theta}_k} \quad \boldsymbol{\theta}_k|G_0 \sim G_0,\ k = 1, 2, ..., \tag{5.16}$$

where π_k represents the probability assigned to the kth atom, and the set satisfy $\sum_{k=1}^{\infty}\pi_k = 1$, and $\boldsymbol{\theta}_k$ is the location or value of that component (atom). These atoms are drawn independently from the base measure G_0. Hence, according to the DP process, the generated parameters $\tilde{\boldsymbol{\theta}}_i$ exhibit a clustering property, that is, they share repeated values with positive probability where the unique values of $\tilde{\boldsymbol{\theta}}_i$ shared among the variables are independent draws for the base distribution G_0 (Ferguson, 1973; Samuel and Blei, 2012). The Dirichlet process therefore provides a very interesting approach for a clustering perspective, when we do not have a fixed number of clusters, in other words having an infinite mixture, say K tends to infinity. Consider a set of observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ to be clustered. Clustering with DP adds a third step to the DP (5.11), that is we assume that the random variables $\boldsymbol{x}_i$, given the distribution parameters $\tilde{\boldsymbol{\theta}}_i$ which are generated from a DP, are generated from a conditional distribution $f(.|\tilde{\boldsymbol{\theta}}_i)$. This is the DP mixture (DPM) model (Antoniak, 1974; Escobar, 1994; Wood and Black, 2008; Samuel and Blei, 2012). The DPM adds therefore a third step to the DP, that is the of generating random variables $\boldsymbol{x}_i$ given the distribution parameters $\tilde{\boldsymbol{\theta}}_i$. The generative process of the DP Mixture (DPM) is therefore as follows:

$$G|\alpha, G_0 \sim \text{DP}(\alpha, G_0) \tag{5.17}$$
$$\tilde{\boldsymbol{\theta}}_i|G \sim G \tag{5.18}$$
$$\boldsymbol{x}_i|\tilde{\boldsymbol{\theta}}_i \sim f(\boldsymbol{x}_i|\tilde{\boldsymbol{\theta}}_i) \tag{5.19}$$

where $f(\boldsymbol{x}_i|\tilde{\boldsymbol{\theta}}_i)$ is a cluster-specific density, for example a multivariate Gaussian density in the case of DP multivariate Gaussian mixture, in which $\tilde{\boldsymbol{\theta}}_i$ is composed of a mean vector and a covariance matrix. In that

case, the base measure G_0 corresponds to the prior parameters distribution which may be a multivariate normal Inverse-Wishart conjugate prior. When K tends to infinity, it can be shown that the finite mixture model (5.1) - (5.6) converges to a Dirichlet process mixture model (Ishwaren and Zarepour, 2002; Neal, 2000; Rasmussen, 2000). The Dirichlet process has a number of properties which make inference based on this nonparametric prior computationally tractable. It also has a interpretation in term of the CRP mixture (Pitman, 2002; Samuel and Blei, 2012) which explicitly shows its suitability to clustering thanks to the integration of the hidden component labels z_i in the generative process. Indeed, the second property of the DP, that is the fact that random parameters drawn from a DP share identical values and thus exhibit a clustering property, connects the DP to the CRP. Consider a random distribution drawn from a DP $G \sim DP(\alpha, G_0)$ followed by repeated draws from that random distribution $\tilde{\boldsymbol{\theta}}_i \sim G$, $\forall i \in 1, \ldots, n$. The structure of the shared values defines a partition of the integers from 1 to n, and the distribution of this partition is a CRP (Ferguson, 1973; Samuel and Blei, 2012). This is defined in the following section

5.3.2 Chinese Restaurant Process parsimonious mixtures

Consider the unknown cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$ where each value z_i is an indicator random variable that represents the label of the unique value $\boldsymbol{\theta}_{z_i}$ of $\tilde{\boldsymbol{\theta}}_i$ such that $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_{z_i}$ for all $i \in \{1, \ldots, n\}$. The CRP provides a distribution on the infinite partitions of the data, that is a distribution over the positive integers $1, \ldots, n$. Consider the following joint distribution of the unknown cluster assignments $(z_1, \ldots, z_n)$:

$$p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1) \ldots p(z_n|z_1, z_2, \ldots, z_{n-1}) \cdot \tag{5.20}$$

From the Pólya urn distribution (5.15), each predictive term of the joint distribution (5.20) can be computed as:

$$p(z_i = k|z_1, ..., z_{i-1}; \alpha_0) = \frac{\alpha_0}{\alpha_0 + i - 1}\delta(z_i, K_{i-1} + 1) + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha_0 + i - 1}\delta(z_i, k) \cdot \tag{5.21}$$

where $n_k = \sum_{j=1}^{i-1} \delta(z_j, k)$ is the number of indicator random variables taking the value k after $i-1$ observations, and $K_{i-1}+1$ is the previously unseen value. From this distribution, one can therefore allow assigning new data to possibly previously unseen (new) clusters as the data are observed, after starting with one cluster. The distribution on partitions induced by the sequence of conditional distributions in Eq. (5.21) is commonly referred to as the Chinese Restaurant Process (CRP). It can be interpreted as follows. Suppose there is a restaurant with an infinite number of tables and in which customers are entering and sitting at these tables. We assume that customers are social, so that the ith customer sits at table k with probability proportional to the number of already seated customers n_k ($k \leq K_{i-1}$ being a previously occupied table), and may choose a new table ($k > K_{i-1}$, k being a new table to be occupied) with a probability proportional to a small positive real number α, which represents the CRP concentration parameter.

In clustering with the CRP, customers correspond to data points and tables correspond to clusters. In CRP mixture, the prior $\text{CRP}(z_1, \ldots, z_{i-1}; \alpha)$ (5.21) is completed with a likelihood with parameters $\boldsymbol{\theta}_k$ for each table (cluster) k (i.e., a multivariate Gaussian likelihood with mean vector and covariance matrix in the GMM case), and a prior distribution (G_0) for the parameters. For example, in the GMM case, one can use a conjugate multivariate normal Inverse-Wishart prior distribution for the mean vectors and the covariance matrices. This process therefore corresponds to the fact that the ith customer sits at table $Z_i = k$, chooses a dish (the parameter $\boldsymbol{\theta}_{z_i}$) from the prior of that table (cluster). The CRP mixture can be summarized according to the following generative process.

$$\begin{aligned} z_i &\sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) && (5.22)\\ \boldsymbol{\theta}_{z_i}|G_0 &\sim G_0 && (5.23)\\ \boldsymbol{x}_i|\boldsymbol{\theta}_{z_i} &\sim f(.|\boldsymbol{\theta}_{z_i}) \cdot && (5.24) \end{aligned}$$

where the CRP distribution is given by Eq. (5.20), G_0 is a base measure (the prior distribution) and $f(\boldsymbol{x}_i|\boldsymbol{\theta}_{z_i})$ is a cluster-specific density. In the DPM and CRP mixtures with multivariate Gaussian components, the parameters $\boldsymbol{\theta}$ of each cluster density are composed of a mean vector and a covariance matrix. In that case, a common base measure G_0 is a multivariate normal Inverse-Wishart conjugate prior.

We note that in the proposed DP parsimonious mixture, or by equivalence, CRP parsimonious mixture, the cluster covariance matrices are parametrized in terms of an eigenvalue decomposition to provide more flexible clusters with possibly different volumes, shapes and orientations. In terms of a CRP interpretation, this can be seen as a variability of dishes for each table (cluster). We indeed use the eigenvalue value decomposition described in section 5.2.2 which until now has been considered only in the case of parametric finite mixture model-based clustering (eg. see Celeux and Govaert (1995) and Banfield and Raftery (1993)), and Bayesian parametric finite mixture model-based clustering (eg. see Bensmail et al. (1997), Bensmail and Meulman (2003), Fraley and Raftery (2005), and Fraley and Raftery (2007).) We investigate twelve parsimonious models and implemented and experimented the following nine models, covering the three families of the mixture models: the general, the diagonal and the spherical family. The parsimonious models therefore go from the simplest spherical one to the more general full model. Table 5.1 summarizes the considered parsimonious Gaussian mixture models, the corresponding prior distribution for each model and the corresponding number of free parameters for a mixture model with K components for data of dimension d. We used conjugate priors, that is Dirichlet distribution for the

Model	Type	Prior	Applied to	# free parameters
$\lambda\mathbf{I}$	Spherical	$\mathcal{IG}$	λ	$\upsilon + 1$
$\lambda_k\mathbf{I}$	Spherical	$\mathcal{IG}$	λ_k	$\upsilon + d$
$\lambda\mathbf{A}$	Diagonal	$\mathcal{IG}$	diag. elements of $\lambda\mathbf{A}$	$\upsilon + d$
$\lambda_k\mathbf{A}$	Diagonal	$\mathcal{IG}$	diag. elements of $\lambda_k\mathbf{A}$	$\upsilon + d + K - 1$
$\lambda\mathbf{DAD}^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$	$\upsilon + \omega$
$\lambda_k\mathbf{DAD}^T$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$	$\upsilon + \omega + K - 1$
$\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	General	$\mathcal{IG}$	diag. elements of $\lambda\mathbf{A}$	$\upsilon + K\omega - (K-1)d$
$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	General	$\mathcal{IG}$	diag. elements $\lambda_k\mathbf{A}$	$\upsilon + K\omega - (K-1)(d-1)$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	$\upsilon + K\omega$

Table 5.1: Considered Parsimonious models, the associated prior for the covariance structure and number of free parameters where $\upsilon = (K-1) + Kd$ and $\omega = d(d+1)/2$.

mixing proportions (Richardson and Green, 1997; Ormoneit and Tresp, 1998), and a multivariate Normal for the mean vectors and and an Inverse-Wishart or an Inverse-Gamma prior for the covariance matrix depending on the parsimonious model as in (Fraley and Raftery, 2007) and Bensmail et al. (1997).

5.3.3 Bayesian inference via Gibbs sampling

Given a sample of n i.i.d observations $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ modeled by one of the proposed Dirichlet process parsimonious mixture models (DPPMs), the aim is to infer the number K of latent clusters underlying the observed data, their parameters $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and the latent cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$. We developed an MCMC Gibbs sampling technique, as in Neal (2000), Rasmussen (2000), and Wood and Black (2008) for the Bayesian inference of the nonparametric parsimonious mixture models.

Recall that, as presented in Section 4.3.1 the Gibbs sampler for mixtures performs in an iterative way as follows. Given an initial mixture parameters $\boldsymbol{\theta}^{(0)}$, and the prior over the missing labels $\mathbf{z}$ (here the CRP), the Gibbs sampler draws the missing labels $\mathbf{z}^{(t)}$ from their posterior distribution $p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$ at each iteration t, which is in this case a Multinomial distribution whose parameters are the posterior component probabilities. Then, given the completed data and the prior distribution $p(\boldsymbol{\theta})$ over the mixture parameters, the Gibbs sampler generates the mixture parameters $\boldsymbol{\theta}^{(t+1)}$ from the corresponding posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{z}^{(t+1)})$, which is in this conjugate prior case a multivariate Normal Inverse-Wishart, or a Normal-Inverse-Gamma distribution, depending on the parsimonious model. This Bayesian sampling procedure produces namely an ergodic Markov chain of samples $(\boldsymbol{\theta}^{(t)})$ with a stationary distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$. Therefore, after initial M burn-in samples in N Gibbs samples, the variables $(\boldsymbol{\theta}^{(M+1)}, ..., \boldsymbol{\theta}^{(N)})$, can be considered to be approximately distributed according to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. The Gibbs sampler consists in sampling the couple $(\mathbf{\Theta}, \mathbf{z})$ from their corresponding posterior distribution. The posterior distribution for $\boldsymbol{\theta}_k$ given all the other variables is given by

$$p(\boldsymbol{\theta}_k|\mathbf{z}, \mathbf{X}, \mathbf{\Theta}_{-k}, \alpha; H) \propto \prod_{i|Z_i=k} f(\boldsymbol{x}_i|Z_i = k; \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k; H) \tag{5.25}$$

where $\mathbf{\Theta}_{-k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_{K_{i-1}})$ and $p(\boldsymbol{\theta}_k; H)$ is the prior distribution for $\boldsymbol{\theta}_k$, that is G_0,

with H being the hyperparameters of the model. The cluster labels z_i are similarly sampled from the posterior distribution which is given, up to a constant, by:

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \boldsymbol{\Theta}, \alpha) \propto f(\boldsymbol{x}_i|z_i; \boldsymbol{\Theta})p(z_i|\mathbf{z}_{-i}; \alpha) \tag{5.26}$$

where $\mathbf{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, and $p(z_i|\mathbf{z}_{-i}; \alpha)$ is the prior predictive distribution corresponds which to the CRP distribution computed as in Equation (5.21). The prior distribution, and the resulting posterior distribution, for each of the considered models, are close to those in Bensmail et al. (1997) and are provided in detail in [J-10].

Sampling the hyperparameter α of the DPPM

The number of mixture components in the models depends on the concentration hyperparameter α of the Dirichlet Process (Antoniak, 1974). We therefore choose to sample it to avoid fixing an arbitrary value for it. We follow the method introduced by Escobar and West (1994) which consists in sampling it by assuming a prior Gamma distribution $\alpha \sim \mathcal{G}(a, b)$ with a shape hyperparameter $a > 0$ and scale hyperparameter $b > 0$. Then, a variable η is introduced and sampled conditionally on α and the number of clusters K_{i-1}, according to a Beta distribution, that is, $\eta|\alpha, K_{i-1} \sim \mathcal{B}(\alpha+1, n)$. The resulting posterior distribution for the hyperparameter α is given by:

$$p(\alpha|\eta, K) \sim \vartheta_\eta \mathcal{G}\left(a + K_{i-1}, b - \log\left(\eta\right)\right) + \left(1 - \vartheta_\eta\right) \mathcal{G}\left(a + K_{i-1} - 1, b - \log\left(\eta\right)\right)$$

where the weights $\vartheta_\eta = \frac{a+K_{i-1}-1}{a+K_{i-1}-1+n(b-\log(\eta))}$. Finally, after a sufficiently large number of samples, the retained solution is the one corresponding to the posterior mode of the number of mixture components, that is the one that appears the most frequently during the sampling.

Complexity The cost of the method is mainly related to the sampling of the labels z_i and hence to the sample size and the number of components, and model parameters $\boldsymbol{\theta}_i$. More specifically, the complexity related to each Gibbs sample is proportional to the current value of the number of mixture components K and hence varies randomly from one iteration to another. Since asymptotically K tends to $\alpha \log(n)$ when n tends to infinity (Antoniak, 1974), therefore, each sample requires $O(\alpha n \log(n))$ operations for sampling the class labels z_i. The parameter simulation (the mean vector and the covariance matrix) requires in the worst case (when the covariance matrix is full, that is a non-parsimonious model) approximatively $O\left(\alpha \log(n)\left(d + d^3\right)\right)$. This gives us a complexity in $O\left(\alpha n \log(n) d^3\right)$.

5.3.4 Bayesian model comparison via Bayes factors

This section provides the used strategy for model comparison, that is, the selection of the best model from the different parsimonious models. We use Bayes factors (Kass and Raftery, 1995; Basu and Chib, 2003) which provide a general way to compare models in (Bayesian) statistical modeling, and has been widely studied in the case of mixture models (Kass and Raftery, 1995; Bensmail et al., 1997; Gelfand and Dey, 1994; Carlin and Chib, 1995; Basu and Chib, 2003). Suppose that we have two model candidates M_1 and M_2, if we assume that the two models have the same prior probability $p(M_1) = p(M_2)$, the Bayes factor is given by

$$BF_{12} = \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)} \tag{5.27}$$

which corresponds to the ratio between the marginal likelihoods of the two models M_1 and M_2. It is a summary of the evidence for model M_1 against model M_2 given the data $\mathbf{X}$. The marginal likelihood $p(\mathbf{X}|M_m)$ for model $M_m,\ m \in \{1, 2\}$, also called the integrated likelihood, is given by

$$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)d\boldsymbol{\theta}_m \tag{5.28}$$

where $p(\mathbf{X}|\boldsymbol{\theta}_m, M_m)$ is the likelihood of model M_m with parameters $\boldsymbol{\theta}_m$ and $p(\boldsymbol{\theta}_m|M_m)$ is the prior density of the mixture parameters $\boldsymbol{\theta}_m$ for model M_m. As it is difficult to compute analytically the marginal likelihood (5.28), several approximations have been proposed to approximate it. One of the most used approximations is the Laplace-Metropolis approximation (Lewis and Raftery, 1994) given by

$$\hat{p}_{\text{Laplace}}(\mathbf{X}|M_m) = (2\pi)^{\frac{\nu_m}{2}} |\hat{\mathbf{H}}|^{\frac{1}{2}} p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m) \tag{5.29}$$

where $\hat{\boldsymbol{\theta}}_m$ is the posterior estimation of $\boldsymbol{\theta}_m$ (posterior mode) for model M_m, ν_m is the number of free parameters of the mixture model M_m as given in Table 5.1, and $\hat{\mathbf{H}}$ is minus the inverse Hessian of the function $\log(p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m))$ evaluated at the posterior mode of $\boldsymbol{\theta}_m$, that is $\hat{\boldsymbol{\theta}}_m$. The matrix $\hat{\mathbf{H}}$ is asymptotically equal to the posterior covariance matrix (Lewis and Raftery, 1994), and is computed as the sample covariance matrix of the posterior simulated sample. We note that, in the proposed DPPM models, as the number of components K is itself a parameter in the model and is changing during the sampling, which leads to parameters with different dimension, we compute the Hessian matrix $\hat{\mathbf{H}}$ in Eq. (5.29) by taking the posterior samples corresponding to the posterior mode of K. Once the estimation of Bayes factors is obtained, it can be interpreted as described in Table 5.2 as suggested by Jeffreys (1961), see also Kass and Raftery (1995).

BF_{12}	$2\log \mathrm{BF}_{12}$	Evidence for model M_1
< 1	< 0	Negative (M_2 is selected)
$1 - 3$	$0 - 2$	Not bad
$3 - 12$	$2 - 5$	Substantial
$12 - 150$	$5 - 10$	Strong
> 150	> 10	Decisive

Table 5.2: Model comparison using Bayes factors.

5.3.5 Experiments

In [J-10][C-2](Bartcus, 2015), the proposed DPPMs was applied and evaluated by performing experiments on both simulated and real data, including complex data from a challenging bioacoustic application. We assess their flexibility in terms of modeling, their use for clustering and inferring the number of clusters from the data. We show how the proposed DPPM approach is able to automatically and simultaneously select the best model with the optimal number of clusters by using the Bayes factors. We also perform comparisons with the finite model-based clustering approach of Bensmail et al. (1997), , which will be abbreviated as PGMM approach. Note that also the one in Fraley and Raftery (2007) was considered. We also use the Rand index to evaluate and compare the provided partitions, and the misclassification error rate when the number of estimated components equals the actual one.

For the simulations, we consider several situations of simulated data, from different models, and with different levels of cluster separations, in order to assess the capability of the proposed approach to retrieve the actual partition with the actual number of clusters. We also assess the stability of our proposed DPPMs models regarding the choice of the hyperparameters values, by considering several situations and varying them.

Simulation results: Varying the clusters shapes, orientations, volumes and separation In this experiment, we apply the proposed models on simulated data simulated according to different models, and with different level of mixture separation, going from poorly separated mixtures to very-well separated mixtures. To simulate the data, we first consider an experimental protocol close to the one used by Celeux and Govaert (1995). We performed extensive experiments involving all the models and many Monte Carlo simulations for several data structure situations. Furthermore, for each type of model structure, we consider three different levels of mixture separation, that is: poorly separated, well separated, and very-well separated mixture. We compare the proposed DPPMs to the parametric PGMM approach in model-based clustering Bensmail et al. (1997) based on the finite parsimonious mixtures. In summary, the simulation results are very satisfactory for all the considered situations. The proposed DPPMs, in almost all the situations (except for one situation) retrieve the actual model, with the actual number of clusters. We also observed that the selected DPPM model, has the highest log-marginal likelihood value, compared to the finite PGMM alternative. Furthermore, we also observe that the solutions provided by the proposed DPPM are, in some cases more parsimonious than those provided by the PGMM, and, in the other cases, the same as those provided by the PGMM.

Also in terms of misclassification error, the proposed DPPM models, compared to the PGMM ones, provide partitions with the lower miscclassification error, for situations with poorly, well or very-well separated clusters, and for clusters with equal and different volumes (except for one situation). On the other hand, for the DPMM models, the evidence of the selected model, compared to the majority of

the other alternatives is, according to Table 5.2, in general decisive. Indeed, the value $2\log \mathrm{BF}_{12}$ of the Bayes Factor between the selected model, and the other models, is more than 10, which corresponds to a decisive evidence for the selected model. Also, in terms of evidence of the selected model against the more competitive one, for the situations with very bad mixture separation, with clusters having the same volume, the evidence is not bad (0.3). However, for all the other situations, the optimal model is selected with an evidence going from an almost substantial evidence (a value of 1.7), to a strong and decisive evidence, especially for the models with different cluster volumes. We can also conclude that the models with different cluster volumes may work better in practice. This was also highlighted by Celeux and Govaert (1995) for the finite parsimonious models in the MLE framework.

Stability with respect to the variation of the hyperparameters values In order to examine the effect of the choice of the hyperparameters values of the mixture on the estimations, we considered two-class situations identical to those used in the parametric parsimonious mixture approach proposed in Bensmail et al. (1997). In order to assess the stability of the models with respect to the values of the hyperparameters, we consider four situations with different hyperparameter values. The obtained log marginal likelihood values for the four models for each of the situations of the hyperparameters show that, for all the situations, the selected model is the actual model with the correct number of clusters. Also, the Bayes factor values ($2\log \mathrm{BF}$), between the selected model, and the more competitive one, for each of the four situations, according to Table 5.2, correspond to a decisive evidence of the selected model. These results confirm that the DPPM are quite stable with respect to the variation of the hyparameters values.

Then, we performed experiments on several real data sets and provide numerical results in terms of comparisons of the Bayes factors (via the log marginal likelihood values) and as well the Rand index and the misclassification error rate for data sets with known actual partition.

Experiments on benchmarks To confirm the results previously obtained on simulated data, we have conducted several experiments on freely available well-known real data sets: Iris, Old Faithful Geyser, Crabs and Diabetes (also Trees, Wine etc). We compare the proposed DPPM models to the PGMM models. For the four data sets, the proposed DPPMs they outperform the alternative finite mixture approach in terms of the Bayes factor value (marginal likelihoods) as well in terms of classification error or Rand index values. The best model in always selected with the actual number of clusters (For Iris, the DPPM approach selects two components as well as the PGMM alternative) and the majority of the parsimonious models (even those which are not selected), retrieve in general the correct number of clusters. Also, the evidence of the selected DPPM models, compared to the other ones, for the four real data sets, is significant. The evidence of the selected model, according to Table 5.2 is indeed strong for Old Faithful geyser data, and very decisive for Crabs, Diabetes and Iris data. Also, the model selection by the proposed DPMM for these latter three data sets, is made with a greater evidence, compared to the PGMM approach. For illustration, Figure (5.1) shows the Diabetes data set, the optimal model partition provided by the DPPM model ($\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$) and the distribution of the number of clusters K. We can observe that the partition is quite well defined (the misclassification rate in this case is 17.24 ± 2.47) and the posterior mode of the number of clusters equals the actual number of clusters ($K = 3$).

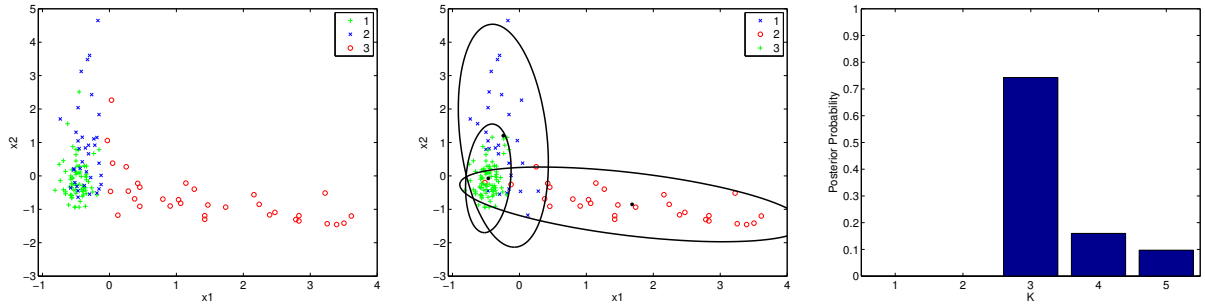

Figure 5.1: Diabetes data set in the space of the components 1 (glucose area) and 3 (SSPG) and the actual partition (left), the optimal partition obtained by the DPPM model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ (middle) and the empirical posterior distribution for the number of mixture components (right).

Challenging application to real-world bioacoustic data We also applied the proposed DPPMs to a real dataset in the framework of a challenging problem of humpback whale song decomposition. The analysis of such data are the core of CNRS MASTODONS SABIOD project[1]. Humpback whales produce songs with a specific structure and the study of that songs is very useful for bio-acousticians and scientists to namely understand how do whales communicate (possibly according to which vocabulary) and to have an idea about their geographical origin. The analysis of such complex signals that aims at discovering the call units (which can be considered as a kind of whale alphabet), can be seen as a problem of unsupervised classification as in Pace et al. (2010) to automatically retrieve possible call units. We therefore reformulate the problem of whale song decomposition as a unsupervised non-parametric classification problem. Contrary to the approach used in Pace et al. (2010), in which the number of states (call units in this case) has been fixed manually, here, we apply the DPPM to automatically find possible call units in the whale song, and automatically infer the number of such song units. The used data are available in the framework of our SABIOD project publicly. They consist of pre-extracted features (mainly MFCC parameters) of 8.6 minutes of a Humpback whale song recordings (51336 observations) produced at few meters distance from the whale in La Reunion - Indian Ocean by the "Darewin" group in 2013. The results obtained by the BNP parsimonious models on these difficult data, are, according to experts very satisfactory. The models are indeed able to find quite satisfactory decomposition compared to the literature in the application field, as well as compared to the finite parsimonious mixture, which select decomposition with large number of components (sometime more than 50) which is not plausible. For the proposed DPPMs however, the decomposition consists in a plausible number of few clusters (not much more that 10 in general) which may correspond to likely call units. The obtained results clearly highlighted the interest of using parsimonious Bayesian non-parametric modeling. For illustration, for example Figure 5.2 shows the spectrograms of two signal portions of 15 seconds each (the algorithm was applied on the whole data set) and the corresponding obtained partition with the parsimonious diagonal model $\lambda_k \mathbf{A}$. We can see that the partition for the two portions is quite satisfactory, and among the obtained classes, there is at least four clearly informative call units for the first example and at least three for the second example. The states 1, 2, 8 and 11 would correspond to up and down sweeps. State 9 corresponds to silence. The seventh state is also the silence that generally ends the ninth state. Also upon visual inspection, it can be seen that for the second example, the states 4 and 8 are clearly different and may correspond to two different call units. This is also the case for the states 7 and 8 for the first example. So the results are very encouraging to continue exploring this data from a BNP prospective.

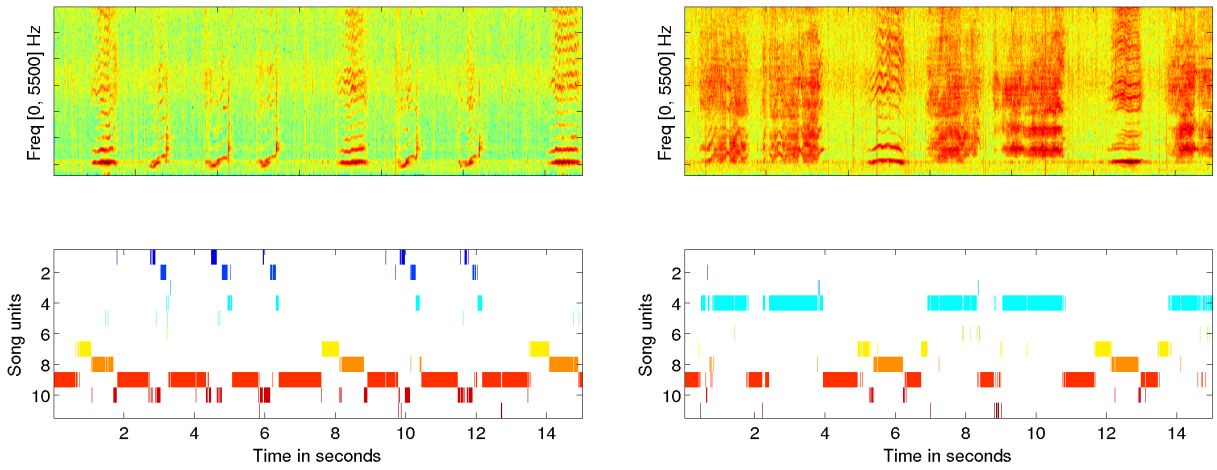

Figure 5.2: Obtained song units by applying or DPPM model with the parametrization $\lambda_k \mathbf{A}$ (diagonal) to two different signals with top: the spectrogram of the part of the signal starting at 280 seconds and it's corresponding partition, and bottom those for the part of signal starting at 295 seconds.

[1]Scaled Acoustic BIODiversity: `http://sabiod.univ-tln.fr/data_samples.html`

5.4 Conclusion

In this chapter I presented Bayesian nonparametric parsimonious mixture models for clustering. They are based on an infinite Gaussian mixture with an eigenvalue decomposition of the cluster covariance matrix and Dirichlet Process prior, or by equivalence a Chinese Restaurant Process prior. This allows deriving several flexible models and avoids the problem of model selection encountered in the standard maximum likelihood-based and Bayesian parametric Gaussian mixture. We also described a Bayesian model comparison framework to automatically select the best model with the best structure by using Bayes factors. Experiments on simulated data highlighted that the proposed DPPMs represent a good nonparametric alternative to the standard parametric Bayesian and non-Bayesian finite mixtures. They simultaneously and accurately estimate partitions with the optimal number of clusters and the best structure from the data. We also applied the proposed approach on real data sets. The obtained results show the interest of using the Bayesian parsimonious clustering models and the potential benefit of using them in practical applications. A future work related to this proposal may concern other parsimonious models such us those recently proposed by Biernacki and Lourme (2014) based on a variance-correlation decomposition of the group covariance matrices, which are stable and visualizable and have desirable properties. Also, until now we have only considered the problem of clustering. A perspective of this work is to extend it to the case of model-based co-clustering (Govaert and Nadif, 2013) with block mixture models, which consists in simultaneously cluster individuals and variables, rather that only individuals. The nonparametric formulation of these models may represent a good alternative to select the number of latent blocks or co-clusters. Note that, while the DPPMs model assume that data are i.i.d (exchangeable), they provided quite satisfactory results in the analysis of the bioacoustc sequential data. Thus, this application opens a perspective on the extension of the DPPM models to the sequential case, say further integrating them into a Markovian framework.
In [C-1][C-3], we investigated the BNP formulation for the standard HMM, that is the HDP-HMM in this application of unsupervised learning from bioacoustic data and the results have been reveled improved, which is promising for the Markovian perspective of the DPPMs.

Chapter 6

Non-normal mixtures of experts

Contents

6.1 Introduction 81
6.1.1 Personal contribution 82
6.1.2 Mixture of experts for continuous data 83
6.1.3 The normal mixture of experts model and its MLE 83
6.2 The skew-normal mixture of experts model 84
6.2.1 The model 84
6.2.2 Maximum likelihood estimation via the ECM algorithm 85
6.3 The t mixture of experts model 88
6.3.1 The model 88
6.3.2 Maximum likelihood estimation 89
6.3.3 MLE via the EM algorithm 89
6.3.4 MLE via the ECM algorithm 91
6.4 The skew t mixture of experts model 91
6.4.1 The model 91
6.4.2 Identifiability of the STMoE model 92
6.4.3 Maximum likelihood estimation via the ECM algorithm 92
6.5 Prediction, clustering and model selection with the non-normal MoE . . . 94
6.6 Experiments 96
6.7 Conclusion 98

Related papers:

[1] F. Chamroukhi. Non-Normal Mixtures of Experts. *arXiv:1506.06707*, 2015c. URL `http://arxiv.org/pdf/1506.06707.pdf`. 61 pages

[2] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015g. URL `http://chamroukhi.univ-tln.fr/papers/TMoE.pdf`. submitted

[3] F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015f. URL `http://chamroukhi.univ-tln.fr/papers/STMoE.pdf`. submitted

Related conference presentation:

[1] F. Chamroukhi. Robust non-normal mixtures of experts. ERCIM 2015 : The 8th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2015h. London, UK

I initiated this research direction very recently (in May 2015) to investigate mixture of experts (MoE) for continuous data, in the case where the expert components are non-normal, that is, do not follow the Normal distribution. MoE being a popular framework for modeling heterogeneity in data in the computer science field particularly machine learning, as well as in statistics. Indeed, the previously developed models in my research, as well as those classically used in learning for the analysis of continuous data, the models are very often based on the normal hypothesis regarding the distribution of the data or a group of the data. However, 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 of the MoE model. In this research I attempt to overcome these (well-known) limitations of modeling with the normal distribution. I proposed three non-normal derivations including two robust mixture of experts (MoE) models. The proposed models are suitable to accomodate data which exhibit additional features such as skewness, heavy-tails and which may be affected by atypical data. I derived dedicated EM and ECM algorithms for model fitting. This research has lead to the following pre-publications [J-12][J-13][J-14] ([J-12] also includes all the developed MoE models in this framework).

6.1 Introduction

Mixture of experts (MoE) introduced by (Jacobs et al., 1991) are widely studied in statistics and machine learning. They consist in a fully conditional mixture model where both the mixing proportions, known as the gating functions, and the component densities, known as the experts, are conditional on some predictors. MoE have been investigated, in their simple form, as well as in their hierarchical form (Jordan and Jacobs, 1994) (e.g., Section 5.12 of McLachlan and Peel. (2000)) for regression and model-based cluster and discriminant analyses and in different application domains. A complete review of the MoE models can be found in Yuksel et al. (2012). For continuous data, which I consider here in the context of non-linear regression and model-based cluster analysis, MoE usually use normal experts, that is, expert components following the Gaussian distribution. Along this chapter, I will call it the normal mixture of experts, abbreviated as NMoE. However, it is well-known that the normal distribution is sensitive to outliers. Moreover, for a set of data containing a group or groups of observations with heavy tails or asymmetric behavior, the use of normal experts may be unsuitable and can unduly affect the fit of the MoE model. In this proposal, I attempt to overcome these limitations in MoE by proposing more adapted and robust mixture of experts models which can deal with possibly skewed, heavy-tailed and atypical data.

Recently, the problem of sensitivity of NMoE to outliers have been considered by Nguyen and McLachlan (2016) where the authors proposed a Laplace mixture of linear experts (LMoLE) for a robust modeling of non-linear regression data. The model parameters are estimated by maximizing the observed-data likelihood via a minorization-maximization (MM) algorithm. Here, I propose alternative MoE models, by relaying on other non-normal distributions that generalize the normal distribution, that is, the skew-normal, t-, and the skew-t distributions. I call these proposed NNMoE models, respectively, the skew-normal MoE (SNMoE), the t MoE (TMoE), and the skew-t MoE (STMoE). Indeed, in these last years, the use of the skew-normal distribution, firstly proposed by Azzalini (1985, 1986), has been shown beneficial in dealing with asymmetric data in various theoretic and applied problems. This has been studied in the finite mixture literature by namely Lin et al. (2007b) for modeling asymmetric univariate data with the univariate skew-normal mixture. On the other hand, the t distribution provides a natural robust extension of the normal distribution to model data with possible outliers. This has been integrated to develop the t mixture model proposed by Mclachlan and Peel (1998) for robust cluster analysis of multivariate data. Recently, Bai et al. (2012) proposed a robust mixture modeling in the regression context on univariate data, by using a univariate t-mixture model. Moreover, in many practical problems, the robustness of t mixtures may however be not sufficient in the presence of asymmetric observations. To deal with this issue, Lin et al. (2007a) proposed the univariate skew-t mixture model which allows for accommodation of both skewness and thick tails in the data, by relying on the skew-t distribution recently introduced by Azzalini and Capitanio (2003). For the general multivariate case using t, skew-normal and skew-t mixtures, one can refer to Mclachlan and Peel (1998); Peel and Mclachlan (2000), Pyne et al. (2009), (Lin, 2010), Lee and McLachlan (2013b), Lee and McLachlan (2013a), Lee and McLachlan (2014), and recently, the unifying framework for previous restricted and unrestricted skew-t mixtures, using the CFUST distribution (Lee and McLachlan, 2015). The inference in the previously described approaches is

performed by maximum likelihood estimation via expectation-maximization (EM) or extensions (Dempster et al., 1977; McLachlan and Krishnan, 2008), in particular the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993). For the Bayesian framework, Frühwirth-Schnatter and Pyne (2010) have considered the Bayesian inference for both the univariate and the multivariate skew-normal and skew-t mixtures. For the regression context, the robust modeling of regression data has been studied namely by Wei (2012) who considered a t-mixture model for regression analysis of univariate data, as well as by Bai et al. (2012) who relied on the M-estimate in mixture of linear regressions using the t-distribution. In the same context of regression, recently Song et al. (2014) proposed the mixture of Laplace regressions, which has been then extended by Nguyen and McLachlan (2016) to the case of mixture of experts, by introducing the Laplace mixture of linear experts (LMoLE). More recently, Zeller et al. (2015) introduced the scale mixtures of skew-normal distributions for robust mixture regressions. However, unlike our proposed NNMoE models, the regression mixture models of Wei (2012), Bai et al. (2012), Song et al. (2014), Zeller et al. (2015) do not consider conditional mixing proportions, that is, mixing proportions depending on some input variables, as in the case of mixture of experts, which I investigate here. In addition, the models of Wei (2012), Bai et al. (2012) and Song et al. (2014) do not consider both the problem of robustness to outliers and the one to deal with possibly asymmetric data. Indeed, here I consider the mixture of experts framework for non-linear regression problems and model-based clustering of regression data, and I attempt to overcome the limitations of the NMoE model in dealing with asymmetric, heavy-tailed data and which may contain outliers. I investigate the use of the skew-normal, t and skew t distributions for the experts, rather than the commonly used normal distribution. First, the skew-normal mixture of experts (SNMoE) is proposed to accommodate data with possible asymmetric behavior. For heavy tailed or possibly noisy data, that is, data with atypical observations, I first propose the t-mixture of experts model (TMoE) to handle the issues regarding namely the sensitivity of the NMoE to outliers. Finally, I propose the skew-t mixture of experts model (STMoE) which allows for accommodation of both skewness and heavy tails in the data and which is also robust to outliers. These models correspond to extensions of the unconditional mixture of skew-normal (Lin et al., 2007b), t (Mclachlan and Peel, 1998; Wei, 2012), and skew t (Lin et al., 2007a) mixture models, to the mixture of experts (MoE) framework, where the mixture means are regression functions and the mixing proportions are covariate-varying. For the models inference, I develop dedicated expectation-maximization (EM) and expectation conditional maximization (ECM) algorithms to estimate the parameters of the proposed models by monotonically maximizing the observed data log-likelihood. The EM algorithms are indeed very popular and successful estimation algorithms for mixture models in general and for mixture of experts in particular. Moreover, the EM algorithm for MoE has been shown by Ng and McLachlan (2004) to be monotonically maximizing the MoE likelihood. The authors have showed that the EM (with IRLS in this case) algorithm has stable convergence and the log-likelihood is monotonically increasing when a learning rate smaller than one is adopted for the IRLS procedure within the M-step of the EM algorithm. They have further proposed an expectation conditional maximization (ECM) algorithm to train MoE, which also has desirable numerical properties. The MoE has also been considered in the Bayesian framework, for example one can cite the Bayesian MoE (Waterhouse et al., 1996; Waterhouse, 1997) and the Bayesian hierarchical MoE (Bishop and Svensén, 2003). Beyond the Bayesian parametric framework, the MoE models have also been investigated within the Bayesian non-parametric framework. We cite for example the Bayesian non-parametric MoE model (Rasmussen and Ghahramani, 2001) and the Bayesian non-parametric hierarchical MoE approach of Shi et al. (2005) using Gaussian Processes experts for regression. For further account on Bayesian MoE for regression, the reader can be referred to for example the book of Shi and Choi (2011). In this chapter, I investigate semi-parametric models under the maximum likelihood estimation framework.

6.1.1 Personal contribution

To overcome the limitations of modeling with the normal mixture of experts (NMoE), I introduced three new non-normal mixture of experts (NNMoE) that can better accommodate data exhibiting non-normal features, including asymmetry, heavy-tails, and the presence of outliers. The proposed models are the skew-normal MoE [J-12] and the robust t MoE [J-12] [J-13] and skew t MoE [J-12] [J-14], respectively named SNMoE, TMoE and STMoE. I developed dedicated E(C)M algorithms to estimate the parameters of the proposed models by monotonically maximizing the observed data log-likelihood. I describe how the presented models can be used in prediction in regression as well as in model-based clustering of regression

data. Numerical experiments carried out on simulated data show the effectiveness and the robustness of the proposed models in terms of modeling non-linear regression functions as well as in model-based clustering. Then, to show their usefulness for practical applications, the proposed models have been applied to the real-world data of tone perception for musical data analysis, and the one of temperature anomalies for the analysis of climate change data. The obtained results are very satisfactory compared to standard NMoE and the alternative mixture models.

The remainder of this chapter is organized as follows. In Section 6.1.2 I briefly recall the MoE framework, the NMoE model and its maximum-likelihood estimation via EM. In Section 6.2, I present the SNMoE model and in Section 6.2.2 I present its inference technique using the ECM algorithm. Then, in Section 6.3 I present the TMoE model and derive its parameter estimation technique using the EM algorithm in Section 6.3.2. Then, in Section 6.4, I present the STMoE model and in Section 6.4.3 the parameter estimation technique using the ECM algorithm. In Section 6.5, I also show how the model selection can be performed for these NNMoE models. I then investigate in Section 6.5 the use of the proposed models for fitting non-linear regression functions as well for prediction on future data. I also show in Section 6.5 how the models can be used in a model-based clustering prospective. In Section 6.6, I perform experiments to assess the proposed models. Finally, in Section 6.7, conclusions are drawn and a future work

6.1.2 Mixture of experts for continuous data

Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994) are used in a variety of contexts including regression, classification and clustering. Here I consider the MoE framework for fitting (non-linear) regression functions and clustering of univariate continuous data . The univariate MoE model assumes that each of the observed pairs of data $(\boldsymbol{x}, y)$ where $y \in \mathbb{R}$ is the response for some covariate $\boldsymbol{x} \in \mathbb{R}^p$, is generated from one of K parametric regression functions with conditional density $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)$ where $(k = 1, \ldots, K)$ governed by a hidden categorical random variable Z indicating from which component each observation is generated. Furthermore, MoE for regression analysis (Jacobs et al., 1991; Jordan and Jacobs, 1994) explore the relationship between the component membership variable Z as function of some predictors $\boldsymbol{r} \in \mathbb{R}^q$. More specifically, the model of the responses Z, known as the gating network in the context of MoE, is a multinomial logistic model and is defined by:

$$\mathbb{P}(Z = k|\boldsymbol{r};\boldsymbol{\alpha}) = \pi_k(\boldsymbol{r};\boldsymbol{\alpha}) = \frac{\exp\left(\boldsymbol{\alpha}_k^T\boldsymbol{r}\right)}{\sum_{\ell=1}^{K}\exp\left(\boldsymbol{\alpha}_\ell^T\boldsymbol{r}\right)} \tag{6.1}$$

where $\boldsymbol{\alpha}_k \in \mathbb{R}^q$ is the coefficient vector associated with $\boldsymbol{r}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T)^T$ is the parameter vector of the logistic model, with $\boldsymbol{\alpha}_K$ being the null vector. Thus, the MoE model decomposes the nonlinear regression model density $f(y|\boldsymbol{x})$ into a convex weighted sum of K regression component models $f_k(y|\boldsymbol{x})$ and can be defined by:

$$f(y|\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K}\pi_k(\boldsymbol{r};\boldsymbol{\alpha})f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k) \tag{6.2}$$

where the π_k's are covariate-varying mixing proportions. The model parameter vector is given by $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$, $\boldsymbol{\Psi}_k$ being the parameter vector of the kth component density. Thus, the MoE model consists in a fully conditional mixture model where both the mixing proportions (the gating functions) and the component densities (the experts) are conditional on some covariate variables (respectively $\boldsymbol{r}$ and $\boldsymbol{x}$).

6.1.3 The normal mixture of experts model and its MLE

In the case of mixture of experts for regression, it is usually assumed that the experts $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)$ are normal. A K-component normal mixture of experts (NMoE) $(K > 1)$ has the following formulation:

$$f(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K}\pi_k(\boldsymbol{r};\boldsymbol{\alpha})\mathrm{N}\big(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big) \tag{6.3}$$

which involves, in the semi-parametric case, component means defined as parametric (non-)linear regression functions $\mu(\boldsymbol{x};\boldsymbol{\beta}_k)$. The NMoE model parameters are estimated by maximizing the observed data log-likelihood given an i.i.d sample of n observations $(y_1,\ldots,y_n)$ with their respective associated covariates $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and $(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)$:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}) \mathrm{N}\big(y_i;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big) \tag{6.4}$$

by using the EM algorithm (Dempster et al., 1977; Jacobs et al., 1991; Jordan and Jacobs, 1994; Jordan and Xu, 1995; Ng and McLachlan, 2004; McLachlan and Krishnan, 2008). The E-Step at the mth iteration of the EM algorithm for the NMoE model requires the calculation of the following posterior probability that the observation $(y_i,\boldsymbol{x}_i,\boldsymbol{r}_i)$ belongs to expert k, given a parameter estimation $\boldsymbol{\Psi}^{(m)}$:

$$\tau_{ik}^{(m)} = \mathbb{P}(Z_i = k|y_i,\boldsymbol{x}_i,\boldsymbol{r}_i;\boldsymbol{\Psi}^{(m)}) = \frac{\pi_k(\boldsymbol{r};\boldsymbol{\alpha}^{(m)})\mathrm{N}\Big(y_i;\mu_k(\boldsymbol{x}_i;\beta_k^{(m)}),\sigma_k^{2\,(m)}\Big)}{f(y_i|\boldsymbol{r}_i,\boldsymbol{x}_i;\boldsymbol{\Psi}^{(m)})}. \tag{6.5}$$

Then, the M-step calculates the parameter vector update $\boldsymbol{\Psi}^{(m+1)}$ by maximizing the well-known Q-function, that is the expected complete-data log-likelihood:

$$\boldsymbol{\Psi}^{(m+1)} = \arg\max_{\boldsymbol{\Psi}\in\boldsymbol{\Omega}} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) \tag{6.6}$$

where $\boldsymbol{\Omega}$ is the parameter space. For example, in the case of normal mixture of linear experts (NMoLE) where each expert's mean has the flowing linear form:

$$\mu(\boldsymbol{x};\boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^T\boldsymbol{x}, \tag{6.7}$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^p$ is the vector of regression coefficients of component k, the updates for each of the expert component parameters consist in analytically solving a weighted Gaussian linear regression problem and are given by:

$$\boldsymbol{\beta}_k^{(m+1)} = \Big[\sum_{i=1}^{n}\tau_{ik}^{(m)}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}\sum_{i=1}^{n}\tau_{ik}^{(q)}y_i\boldsymbol{x}_i, \tag{6.8}$$

$$\sigma_k^{2\,(m+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(m)}\Big(y_i-\boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i\Big)^2}{\sum_{i=1}^{n}\tau_{ik}^{(m)}}. \tag{6.9}$$

For the mixing proportions, the parameter vector update $\boldsymbol{\alpha}^{(m+1)}$ cannot however be obtained in a closed form. It is calculated by Iteratively Reweighted Least Squares (IRLS) (Jacobs et al., 1991; Jordan and Jacobs, 1994; Chen et al., 1999a; Green, 1984)[C-14][J-1].

However, the normal distribution is not adapted to deal with asymmetric and heavy tailed data. It is also known that the normal distribution is sensitive to outliers. In the proposal, I first propose to address the issue regarding the skewness, by proposing the skew-normal mixture of experts (SNMoE). Then, I propose a robust fitting of the MoE, which is adapted to heavy-tailed data, by using the t distribution, that is, the t mixture of experts (TMoE). Finally, the proposed skew-t mixture of experts (STMoE) allows for simultaneously accommodating asymmetry and heavy tails in the data and is also robust to outliers.

6.2 The skew-normal mixture of experts model

6.2.1 The model

The skew-normal mixture of experts (SNMoE) model uses the skew-normal distribution as density for the expert components. As introduced by Azzalini (1985, 1986), a random variable Y follows a univariate skew-normal distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 \in (0,\infty)$ and skewness parameter $\lambda \in \mathbb{R}$ if it has the density

$$f(y;\mu,\sigma^2,\lambda) = \frac{2}{\sigma}\phi(\frac{y-\mu}{\sigma})\Phi\left(\lambda(\frac{y-\mu}{\sigma})\right) \tag{6.10}$$

where $\phi(.)$ and $\Phi(.)$ denote, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution. It can be seen from (6.10) that when the skewness parameter $\lambda = 0$, the skew-normal reduces to the normal distribution.

The presented skew-normal mixture of experts (SNMoE) extends the skew-normal mixture model (Lin et al., 2007b) to the case of mixture of experts framework, by considering conditional distributions for both the mixing proportions and the means of the mixture components. The SNMoE is therefore a MoE model with skew-normal experts and is defined as follows. Let $\text{SN}(\mu, \sigma^2, \lambda)$ denotes a skew-normal distribution with location parameter μ, scale parameter σ and skewness parameter λ. A K-component SNMoE is then defined by:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \text{SN}\big(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\big) \tag{6.11}$$

where each expert component k has indeed a skew-normal distribution, whose density is defined by (6.10). The parameter vector of the model 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 (6.11) reduces to the NMoE model (6.3).

Before going on the model inference, I first present its stochastic and hierarchical representations, which will serve to derive the ECM algorithm for maximum likelihood parameter estimation. The SNMoE model is characterized as follows.

Let U and E be independent univariate random variables following the standard normal distribution $\mathcal{N}(0, 1)$ with pdf $\phi(.)$. Given some covariates $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$, a random variable Y_i is said to follow the SNMoE model (6.11) if it has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \delta_{z_i}\sigma_{z_i}|U_i| + \sqrt{1 - \delta_{z_i}^2}\,\sigma_{z_i}E_i. \tag{6.12}$$

In (6.12), $|U|$ denotes the magnitude of U and $\delta_{z_i} = \frac{\lambda_{z_i}}{\sqrt{1+\lambda_{z_i}^2}}$ where $Z_i \in \{1, \ldots, K\}$ is a categorical variable Z_i which follows the multinomial distribution, that is:

$$Z_i|\boldsymbol{r}_i \sim \text{Mult}(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})) \tag{6.13}$$

where each of the probabilities $\pi_{z_i}(\boldsymbol{r}_i; \boldsymbol{\alpha}) = \mathbb{P}(Z_i = z_i|\boldsymbol{r}_i)$ is given by the logistic function (6.1). In this incomplete data framework, Z_i represents the hidden label of the component generating the ith observation. The stochastic representation (6.12) of the SNMoE leads to the following hierarchical representation, which, as it will be presented in Section 6.2.2, greatly facilitates the model inference.

By introducing the binary latent component-indicators Z_{ik} such that $Z_{ik} = 1$ iff $Z_i = k$, a hierarchical model for the SNMoE can be derived from its stochastic representation (6.12) and is as follows

$$\begin{aligned} Y_i|u_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \text{N}\Big(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k) + \delta_k|u_i|, (1 - \delta_k^2)\sigma_k^2\Big), \\ U_i|Z_{ik} = 1 &\sim \mathcal{N}(0, \sigma_k^2), \\ \boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \text{Mult}\left(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\right) \end{aligned} \tag{6.14}$$

where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ and $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$. In this hierarchical representation, in addition to the hidden component labels Z_i, the variables U_i are also hidden. This hierarchical incomplete data representation facilitates the inference scheme by using the ECM algorithm.

6.2.2 Maximum likelihood estimation via the ECM algorithm

The unknown parameter vector $\boldsymbol{\Psi}$ of the SNMoE model can be estimated by maximizing the observed-data log-likelihood. Given an observed i.i.d sample of n observations $(y_1, \ldots, y_n)$ with their respective associated covariates $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$, under the SNMoE model (6.11), the observed data log-likelihood for the parameter vector $\boldsymbol{\Psi}$ is given by:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \text{SN}\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\right). \tag{6.15}$$

The maximization of this log-likelihood in this incomplete data framework can not be performed in a closed form. It can be performed via EM-type algorithms (McLachlan and Krishnan, 2008). More specifically, I propose a dedicated Expectation Conditional Maximization (ECM) algorithm to monotonically maximize (6.15). 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.

Deriving the ECM algorithm requires the definition of the complete-data log-likelihood. From the hierarchical representation (6.14) of the SNMoE, 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:

$$\log L_c(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \log L_c(\boldsymbol{\Psi}_k), \tag{6.16}$$

with

$$\begin{aligned}
\log L_c(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),\\
\log L_c(\boldsymbol{\Psi}_k) &= \sum_{i=1}^{n} Z_{ik}\Big[-\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ u_i}{(1-\delta_k^2)\sigma_k} - \frac{u_i^2}{2(1-\delta_k^2)\sigma_k^2}\Big],
\end{aligned}$$

where $d_{ik} = \frac{y_i - \mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)}{\sigma_k}$. Then, the proposed ECM algorithm for the SNMoE model performs as follows. It starts with an initial parameter vector $\boldsymbol{\Psi}^{(0)}$ and alternates between the E- and CM- steps until a convergence criterion is satisfied.

E-Step The E-Step of the ECM algorithm for the SNMoE calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood (6.16), given the observed data $\{(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)\}_{i=1}^n$ and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$, m being the current iteration:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = \mathbb{E}\big[\log L_c(\boldsymbol{\Psi}) | \{y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n; \boldsymbol{\Psi}^{(m)}\big] = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K} Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}), \tag{6.17}$$

with

$$Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}), \tag{6.18}$$

$$Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) + \frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k} - \frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2} - \frac{d_{ik}^2}{2(1-\delta_k^2)} \right] \tag{6.19}$$

for $k = 1, \ldots, K$, where the required conditional expectations are given by:

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

The $\tau_{ik}^{(m)}$'s represent the posterior distribution of the hidden component labels Z_i and correspond to the posterior memberships of the observed data. The conditional expectations $e_{1,ik}^{(m)}$ and $e_{2,ik}^{(m)}$ correspond to the posterior distribution of the hidden variables U_i and U_i^2, respectively. From (6.17), (6.18), and (6.19), it follows that the Q-function is calculated by analytically calculating these conditional expectations as shown in [J-12].

M-Step Then, the M-step calculates the parameter vector $\boldsymbol{\Psi}^{(m+1)}$ as in (6.6), that is by maximizing the Q-function (6.17) with respect to $\boldsymbol{\Psi}$. This can be performed by separately maximizing $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$ with respect to logistic parameters $\boldsymbol{\alpha}$ and, for each component k ($k = 1,\ldots,K$), the function $Q(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ with respect to the skew-normal expert parameters $\boldsymbol{\Psi}_k$ where $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$. I adopt the ECM extension of the EM algorithm. The M-step in this case consists of four conditional-maximization (CM)-steps, corresponding to the decomposition of the parameter vector $\boldsymbol{\Psi}$ into four sub-vectors $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda})^T$. Thus, this leads to the following CM steps.

CM-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$: $\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}} Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$. Unlike in standard skew-normal mixture model and skew-normal regression mixture model, this maximization in the case of the proposed SNMoE does not exist in closed form. It is performed iteratively by Iteratively Reweighted Least Squares (IRLS).

The Iteratively Reweighted Least Squares (IRLS) algorithm: The IRLS algorithm is used to maximize $Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})$ given by (6.18) with respect to the parameter $\boldsymbol{\alpha}$ in the M step at each iteration m of the ECM algorithm. The IRLS consists in starting with a vector $\boldsymbol{\alpha}^{(0)}$, and, at the $l+1$ iteration, updating the estimation of $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l)} - \Big[\frac{\partial^2 Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\Big]^{-1}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \frac{\partial Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \tag{6.20}$$

where $\frac{\partial^2 Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}$ and $\frac{\partial Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}}$ are respectively the Hessian matrix and the gradient vector of $Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})$. At each IRLS iteration the Hessian and the gradient are evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(l)}$ and are computed similarly as in [J-1][J-2]. The parameter update $\boldsymbol{\alpha}^{(m+1)}$ is taken at convergence of the IRLS algorithm (6.20). Then, for $k = 1\ldots, K$,

CM-Step 2 Calculate $\boldsymbol{\beta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (6.19) w.r.t $\boldsymbol{\beta}_k$. Here I focus on the common linear case for the experts where each expert-component mean function is the one of a linear regression model and has the form (6.7). It can be easily shown that the maximization problem for this resulting skew-normal mixture of linear of experts (SNMoLE) can be solved analytically and has the following solution:

$$\boldsymbol{\beta}_k^{(m+1)} = \Big[\sum_{i=1}^n \tau_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^n \tau_{ik}^{(q)} \left(y_i - \delta_k^{(m)} e_{1,ik}^{(m)}\right) \boldsymbol{x}_i. \tag{6.21}$$

CM-Step 3: Calculate $\sigma_k^{2(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (6.19) w.r.t σ_k^2. Similarly to the update of $\boldsymbol{\beta}_k$, the analytic solution of this problem is given by:

$$\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(m)} \left[\left(y_i - \boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)} e_{1,ik}^{(m)}(y_i - \boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i) + e_{2,ik}^{(m)}\right]}{2\left(1-\delta_k^{2(m)}\right)\sum_{i=1}^n \tau_{ik}^{(m)}}. \tag{6.22}$$

CM-Step 4 Calculate $\lambda_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (6.19) w.r.t λ_k, with $\boldsymbol{\beta}_k$ and σ_k^2 fixed at $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2(m+1)}$, respectively. This consists in solving the following equation in λ_k (recall that $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$) to obtain $\lambda_k^{(m+1)}$ ($k = 1,\ldots,K$) as the solution of:

$$\sigma_k^{2(m+1)}\delta_k(1-\delta_k^2)\sum_{i=1}^n \tau_{ik}^{(m)} + (1+\delta_k^2)\sum_{i=1}^n \tau_{ik}^{(m)}(y_i - \boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i)\, e_{1,ik}^{(m)} - \delta_k \sum_{i=1}^n \tau_{ik}^{(m)}\left[e_{2,ik}^{(m)} + \left(y_i - \boldsymbol{\beta}_k^{T(m+1)}\boldsymbol{x}_i\right)^2\right] = 0\cdot \tag{6.23}$$

This scalar equation can be solved with a root finding algorithm, such as Brent's method (Brent, 1973). Then, given the update of the skewness parameter $\lambda_k^{(m+1)}$, the update of δ_k is calculated as $\delta_k^{(m+1)} = \frac{\lambda_k^{(m+1)}}{\sqrt{1+\lambda_k^{2};(m+1)}}$.

It is obvious to see that when the skewness parameter $\lambda_k = \delta_k = 0$ for all k, the parameter updates for the SNMoE corresponds to those of the standard NMoE. Hence, compared to the standard NMoE, the SNMoE model is characterized by an additional flexibility feature, that is the one to be handle possibly skewed data. However, while the 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 NMoE and the SNMoE may thus be affected by outliers. In the next section, I address the problem of sensitivity of normal mixture of experts to outliers and heavy tails. I first propose a robust mixture of experts modeling by using the t distribution.

6.3 The t mixture of experts model

The proposed t mixture of experts (TMoE) model is based on the t distribution, which is known as a robust generalization of the normal distribution. The use of the t distribution for mixture components has been indeed shown to be more robust than the normal distribution to handle outliers in the data and accommodate data with heavy tailed distribution. This has been shown in terms of density modeling and cluster analysis for multivariate data (Mclachlan and Peel, 1998; Peel and Mclachlan, 2000) as well as for univariate data (Lin et al., 2007a) and regression mixtures (Bai et al., 2012; Wei, 2012; Ingrassia et al., 2012). The t-distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 \in (0, \infty)$ and degrees of freedom $\nu \in (0, \infty)$ has the probability density function

$$f(y; \mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{d_y^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{6.24}$$

where $d_y^2 = \left(\frac{y-\mu}{\sigma}\right)^2$ denotes the squared Mahalanobis distance between y and μ (σ being the scale parameter), and Γ is the Gamma function given by $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}\,dx$.

6.3.1 The 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), and Ingrassia et al. (2012) to the MoE framework. Wei (2012); Bai et al. (2012); Ingrassia et al. (2012) considered the t-mixture model for the regression context on univariate data where the component means are (linear) regression functions of the form $\mu(\boldsymbol{x}; \boldsymbol{\beta}_k)$. However, this model do not explicitly model the mixing proportions as function the predictors; They are assumed to be constant.

The proposed TMoE is a MoE model with t-distributed experts and is defined as follows. Let $t_\nu(\mu, \sigma^2, \nu)$ denotes a t distribution with location parameter μ, scale parameter σ and degrees of freedom ν, whose density is given by (6.24). A K-component TMoE model is then defined by:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha})\; t_{\nu_k}\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right). \tag{6.25}$$

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 $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \nu_k)^T$ is the parameter vector for the kth t expert component which has a t distribution. One can see that when the robustness parameter $\nu_k \to \infty$ for each expert k, the TMoE model (6.25) approaches the NMoE model (6.3).

The stochastic representation for the t mixture of experts (TMoE) is as follows. Let E be a univariate random variable following the standard normal distribution $E \sim \phi(.)$. Suppose that, conditional on the hidden variable $Z_i = z_i$, a random variable W_i is distributed as 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 (6.25) if it has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_{z_i}}}, \tag{6.26}$$

where the categorical variable Z_i conditional on the covariate $\boldsymbol{r}_i$ follows the multinomial distribution (6.13).

Similarly to the case of the previously presented SNMoE model, the stochastic representation (6.26) leads to the following hierarchical representation of the TMoE, which facilitates the model inference as it will be presented in Section 6.3.2.

The hierarchical representation of the TMoE model is written as:

$$\begin{aligned} Y_i|w_i, Z_{ik}=1, \boldsymbol{x}_i &\sim \mathrm{N}\left(\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k), \frac{\sigma_k^2}{w_i}\right), \\ W_i|Z_{ik}=1 &\sim \mathrm{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\ \boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \mathrm{Mult}\left(1; \pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha})\right). \end{aligned} \tag{6.27}$$

This hierarchical representation involving the hidden variables Z_i and W_i facilitates the ML inference of model parameters $\boldsymbol{\Psi}$ via the EM or the ECM algorithm.

6.3.2 Maximum likelihood estimation

Given an i.i.d sample of n observations, the unknown parameter vector $\boldsymbol{\Psi}$ can be estimated by maximizing the observed-data log-likelihood, which, under the TMoE model, is given by:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}) t\nu_k\left(y; \mu(\boldsymbol{x};\boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right). \tag{6.28}$$

To perform this maximization, I first use the EM algorithm and then describe an ECM extension (Meng and Rubin, 1993) as in Liu and Rubin (1995) for a single t distribution and as in Mclachlan and Peel (1998) and Peel and Mclachlan (2000) for the mixture of t-distributions.

6.3.3 MLE via the EM algorithm

To maximize the log-likelihood function (6.28), the EM algorithm for the TMoE model starts with an initial parameter vector $\boldsymbol{\Psi}^{(0)}$ and alternates between the E- and M- steps until convergence. The E-step computes the expected completed data log-likelihood (the Q-function) and the M-Step maximizes it. From the hierarchical representation of the TMoE (6.27), 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 (6.27) and the latent labels $(z_1, \ldots, z_n)$. Thus, the complete-data log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \left[\log L_{2c}(\boldsymbol{\Psi}_k) + \log L_{3c}(\nu_k)\right], \tag{6.29}$$

where

$$\log L_{1c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}), \tag{6.30}$$

$$\log L_{2c}(\boldsymbol{\Psi}_k) = \sum_{i=1}^{n} Z_{ik}\Big[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{1}{2} w_i d_{ik}^2\Big], \tag{6.31}$$

$$\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik}\Big[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}-1\right)\log(w_i) - \left(\frac{\nu_k}{2}\right) w_i\Big]. \tag{6.32}$$

E-Step The E-Step of the EM algorithm for the TMoE calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood (6.44), given the observed data and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$. It can be seen from (6.30), (6.31) and (6.32) that computing the Q-function, given by:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K}\left[Q_2(\boldsymbol{\theta}_k, \boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k, \boldsymbol{\Psi}^{(m)})\right], \tag{6.33}$$

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

$$\begin{aligned}
Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}), \\
Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n} \tau_{ik}^{(m)} \Big[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{1}{2}\, w_{ik}^{(m)} d_{ik}^2 \Big]. \\
Q_3(\nu_k; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right)\, w_{ik}^{(m)} + \left(\frac{\nu_k}{2} - 1\right) e_{1,ik}^{(m)} \right]
\end{aligned}$$

requires the following conditional expectations:

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

These required conditional expectations are calculated analytically as given in [J-12][J-13].

M-Step In the M-step, as it can be seen from (6.33), the Q-function can be maximized by independently maximizing $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$, and, for each k, $Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)})$, $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$, with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\theta}_k$ and ν_k, respectively. Thus, on the $(m+1)$th iteration of the M-step, the model parameters are updated as follows.

M-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\alpha}$. This can be performed iteratively via IRLS (6.20) 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 $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2)^T$. This is achieved by first maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$ with respect to $\boldsymbol{\beta}_k$ and then with respect to σ_k^2. For the t mixture of linear experts (TMoLE) case where the expert means are of the form (6.7), this maximization can be performed analytically and provides the following updates:

$$\boldsymbol{\beta}_k^{(m+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} y_i \boldsymbol{x}_i, \tag{6.34}$$

$$\sigma_k^{2(m+1)} = \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \left(y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2. \tag{6.35}$$

Here, I note that, following Kent et al. (1994) in the case of ML estimation for single component t distribution and Mclachlan and Peel (1998); Peel and Mclachlan (2000) for mixture of multivariate t distributions, the EM algorithm can be modified slightly by replacing the divisor $\sum_{i=1}^{n} \tau_{ik}^{(m)}$ in (6.35) by $\sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)}$. The modified algorithm may converge faster than the conventional EM algorithm. This is was also observed in practice for the proposed TMoE.

M-Step 3 Calculate $\nu_k^{(m+1)}$ by maximizing $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$ w.r.t ν_k. The degrees of freedom update $\nu_k^{(m+1)}$ is therefore obtained by iteratively solving the following equation for ν_k:

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} \Big(\log(w_{ik}^{(m)}) - w_{ik}^{(m)}\Big) + \psi\left(\frac{\nu_k^{(m)} + 1}{2}\right) - \log\left(\frac{\nu_k^{(m)} + 1}{2}\right) = 0. \tag{6.36}$$

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

It can be seen that, as mentioned previously, if the number of degrees of freedom ν_k approaches ∞ for all k, then the parameter updates for the TMoE model are exactly those of the NMoE model (since w_{ik} tends to 1 in this case). The TMoE model constitutes therefore a robust generalization of the NMoE model that is able to model data with density heaving longer tails than those of the NMoE model.

After deriving the EM algorithm for the TMoE parameter estimation, now I described an ECM extension.

6.3.4 MLE via the ECM algorithm

Following the ECM extension of the EM algorithm for a single t distribution proposed by Liu and Rubin (1995) and the one of the EM algorithm for the t-mixture model (Mclachlan and Peel, 1998; Peel and Mclachlan, 2000), the EM algorithm for the TMoE model can also be modified to give an ECM version by adding an additional E-Step between the two M-steps 2 and 3. This additional E-step consists in taking the parameter vector $\boldsymbol{\Psi}$ with $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{(m+1)}$ instead of $\boldsymbol{\Psi}_k^{(m)}$, that is

$$Q_2(\nu_k; \boldsymbol{\Psi}^{(m)}) = Q_2(\nu_k; \boldsymbol{\alpha}^{(m)}, \boldsymbol{\theta}_k^{(m+1)}, \nu_k^{(m)}).$$

Thus, the M-Step 3 in the above is replaced by a Conditional-Maximization (CM)-Step in which the degrees of freedom update (6.36) is calculated with the conditional expectation $w_{ik}^{(m)}$ and $e_{1,ik}^{(m)}$ computed with the updated parameters $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2(m+1)}$ respectively given by (6.34) and (6.35).

The SNMoE presented before allows to deal with asymmetric data. The TMoE handles the problem of heavy tailed data possibly affected by outliers. Now, I propose the skew t mixture of experts (STMoE) model which attempts to simultaneously accommodate heavy tailed data with possible outliers and with asymmetric distribution.

6.4 The skew t mixture of experts model

The proposed skew t mixture of experts (STMoE) model is a MoE model in which the expert components have a skew-t density, rather than the standard normal one as in the NMoE model, or the previously presented skew-normal and t ones as in the SNMoE and the TMoE, respectively.

The skew t distribution

The skew t distribution, introduced by Azzalini and Capitanio (2003) in 2003, can be characterized as follows. Let U be an univariate random variable with a standard skew-normal distribution $U \sim \text{SN}(0, 1, \lambda)$ (which can be shortened as $U \sim \text{SN}(\lambda)$) with pdf given by (6.10). Then, let W be an univariate random variable independent of U and following the Gamma distribution: $W \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$. A random variable Y having the following representation:

$$Y = \mu + \sigma \frac{U}{\sqrt{W}} \tag{6.37}$$

follows the skew t distribution $\text{ST}(\mu, \sigma^2, \lambda, \nu)$ with location parameter μ, scale parameter σ, skewness parameter λ and degrees of freedom ν, whose density is defined by:

$$f(y; \mu, \sigma^2, \lambda, \nu) = \frac{2}{\sigma}\, t_\nu(d_y)\, T_{\nu+1}\left(\lambda\, d_y \sqrt{\frac{\nu+1}{\nu + d_y^2}}\right) \tag{6.38}$$

where $d_y = \frac{y-\mu}{\sigma}$ and $t_\nu(.)$ and $T_\nu(.)$ respectively denote the pdf and the cdf of the standard t distribution with degrees of freedom ν.

6.4.1 The model

The proposed skew t mixture of experts (STMoE) model extends the univariate skew t mixture model, which was first introduced by Lin et al. (2007a), to the MoE framework. In the skew-t mixture model, the mixing proportions and the expert means are constant, that is, they are not function of predictors. In the proposed STMoE, I consider skew-t expert components with regression mean functions, and covariate varying mixing proportions. A K-component mixture of skew t experts (STMoE) is therefore defined by:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) \quad = \quad \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha})\, \text{ST}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k, \nu_k)\cdot \tag{6.39}$$

The parameter vector of the STMoE model is $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ where $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k, \nu_k)^T$ is the parameter vector for the kth skew t expert component whose density is defined by

$$f\big(y|\boldsymbol{x}; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma^2, \lambda, \nu\big) = \frac{2}{\sigma}\, t_\nu(d_y(\boldsymbol{x}))\, T_{\nu+1}\left(\lambda\, d_y(\boldsymbol{x}) \sqrt{\frac{\nu+1}{\nu + d_y^2(\boldsymbol{x})}}\right) \tag{6.40}$$

where $d_y(\boldsymbol{x}) = \frac{y-\mu(\boldsymbol{x};\boldsymbol{\beta}_k)}{\sigma}$. It can be seen that, when the robustness parameter $\nu_k \to \infty$ for each k, the STMoE model (6.39) reduces to the SNMoE model (6.11). On the other hand, if the skewness parameter $\lambda_k = 0$ for each k, the STMoE model reduces to the TMoE model (6.25). Moreover, when $\nu_k \to \infty$ and $\lambda_k = 0$ for each k, it approaches the standard NMoE model (6.3). This therefore makes the STMoE flexible as it generalizes the previously described models to accommodate situations with asymmetry, heavy tails, and outliers.

The STMoE model is characterized as follows. Suppose that conditional on a categorical variable $z_i \in \{1, \ldots, K\}$ representing the hidden label of the component generating the ith observation and following the multinomial distribution (6.13), a random variable has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_i}} \tag{6.41}$$

where E_i and W_i are independent univariate random variables with, respectively, a standard skew-normal distribution $E_i \sim \text{SN}(\lambda_{z_i})$, and a Gamma distribution $W_i \sim \text{Gamma}(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2})$, and $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$ are some given covariate variables. Then, the variable Y_i is said to follow the STMoE defined by (6.39).

From the hierarchical representation of the skew t distribution, a hierarchical model for the proposed STMoE model (6.39) can be derived from its stochastic representation (6.41) and is as follows:

$$\begin{aligned}
Y_i|u_i, w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \text{N}\left(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k) + \delta_k|u_i|, \frac{1-\delta_k^2}{w_i}\sigma_k^2\right), \\
U_i|w_i, Z_{ik} = 1 &\sim \text{N}\left(0, \frac{\sigma_k^2}{w_i}\right), \\
W_i|Z_{ik} = 1 &\sim \text{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\
\boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \text{Mult}\big(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\big).
\end{aligned} \tag{6.42}$$

The variables U_i and W_i are hidden in this hierarchical representation, which facilitates the inference scheme and will be used to derive the maximum likelihood estimation of the STMoE model parameters $\boldsymbol{\Psi}$ by using the ECM algorithm.

6.4.2 Identifiability of the STMoE model

Jiang and Tanner (1999) have established that ordered, initialized, and irreducible MoEs are identifiable. Ordered implies that there exist a certain ordering relationship on the experts parameters $\boldsymbol{\Psi}_k$ such that $(\boldsymbol{\alpha}_1^T, \boldsymbol{\Psi}_1^T)^T \prec \ldots \prec (\boldsymbol{\alpha}_K^T, \boldsymbol{\Psi}_K^T)^T$; initialized implies that $\boldsymbol{\alpha}_K$, the parameter vector of the Kth logistic proportion, is the null vector, and irreducible implies that $\boldsymbol{\Psi}_k \neq \boldsymbol{\Psi}_{k\prime}$ for any $k \neq k\prime$. For the proposed STMoE, which generalizes the previously seen MoE models, ordered implies that there exist a certain ordering relationship such that $(\boldsymbol{\beta}_1^T, \sigma_1^2, \lambda_1, \nu_1)^T \prec \ldots \prec (\boldsymbol{\beta}_K^T, \sigma_K^2, \lambda_K, \nu_K)^T$; initialized implies that $\mathbf{w}_K$ is the null vector, as assumed in the model, and finally, irreducible implies that if $k \neq k\prime$, then one of the following conditions holds: $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k\prime}$, $\sigma_k \neq \sigma_{k\prime}$, $\lambda_k \neq \lambda_{k\prime}$ or $\nu_k \neq \nu_{k\prime}$. Then, we can establish the identifiability of ordered and initialized irreducible STMoE models by applying Lemma 2 of Jiang and Tanner (1999), which requires the validation of the following nondegeneracy condition. The set $\{\text{ST}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_1), \sigma_1^2, \lambda_1, \nu_1), \ldots, \text{ST}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_{4K}), \sigma_{4K}^2, \lambda_{4K}, \nu_{4K})\}$ contains $4K$ linearly independent functions of y, for any $4K$ distinct quadruplet $(\mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k, \nu_k)$ for $k = 1, \ldots, 4K$. Thus, via Lemma 2 of Jiang and Tanner (1999) we have any ordered and initialized irreducible STMoE is identifiable.

6.4.3 Maximum likelihood estimation via the ECM algorithm

The unknown parameter vector $\boldsymbol{\Psi}$ of the STMoE model is estimated by maximizing the following observed-data log-likelihood given an observed i.i.d sample of n observations, that is, the responses

$(y_1, \ldots, y_n)$ and the corresponding predictors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \text{ST}(y; \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k, \nu_k). \tag{6.43}$$

This is performed iteratively by a dedicated ECM algorithm. The complete data consist of the observations, as well as the latent variables $(u_1, \ldots, u_n)$ and $(w_1, \ldots, w_n)$ and the latent component labels $(z_1, \ldots, z_n)$. Then, from the hierarchical representation of the STMoE (6.42), the complete-data log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\log L_c(\boldsymbol{\Psi}) \quad = \quad \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \left[\log L_{2c}(\boldsymbol{\theta}_k) + \log L_{3c}(\nu_k) \right] \tag{6.44}$$

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

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

The ECM algorithm for the STMoE model starts with an initial parameter vector $\boldsymbol{\Psi}^{(0)}$ and alternates between the E- and CM- steps until convergence.

E-Step The E-Step of the CEM algorithm for the STMoE calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood (6.44), 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:

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K} \left[Q_2(\boldsymbol{\theta}_k, \boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k, \boldsymbol{\Psi}^{(m)}) \right], \tag{6.45}$$

where

$$\begin{aligned}
Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}), \\
Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{w_{ik}^{(m)}\ d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k} - \frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2} \right], \\
Q_3(\nu_k; \boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right)\ w_{ik}^{(m)} + \left(\frac{\nu_k}{2}\right) e_{3,ik}^{(m)} \right].
\end{aligned}$$

From (6.44), it can be seen that computing the Q-function only requires the following conditional expectations:

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

These conditional expectations are calculated analytically as shown in [J-12][J-14], except for $e_{3,ik}^{(m)}$ for which I adopted a one-step-late (OSL) approach as described in Lee and McLachlan (2014), rather than using a Monte Carlo approximation as in Lin et al. (2007a). 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.

M-Step The M-step maximizes the Q-function (6.45) with respect to $\boldsymbol{\Psi}$ and provides the parameter vector update $\boldsymbol{\Psi}^{(m+1)}$. From (6.45), it can be seen that the maximization of Q can be performed by separately maximizing Q_1 with respect to the parameters $\boldsymbol{\alpha}$ of the mixing proportions, and for each expert k ($k = 1, \ldots, K$), Q_2 with respect to $(\boldsymbol{\beta}_k^T, \sigma_k^2)^T$ and λ_k, and Q_3 with respect to ν_k. The maximization of Q_2 and Q_3 is carried out by conditional maximization (CM) steps by updating $(\boldsymbol{\beta}_k, \sigma_k^2)$ and then updating (λ, ν_k) with the given updated parameters. This leads to the following CM steps. On the $(m+1)$th iteration of the M-step, the STMoE model parameters are updated as follows.

CM-Step 1 Calculate the parameter $\boldsymbol{\alpha}^{(m+1)}$ maximizing the function $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ given by (6.34) by using IRLS (6.20). Then, for $k = 1 \ldots, K$,

CM-Step 2 Calculate $(\boldsymbol{\beta}_k^{T(m+1)}, \sigma_k^{2(m+1)})^T$ by maximizing $Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)})$ w.r.t $(\boldsymbol{\beta}_k^T, \sigma_k^2)^T$. For the skew-$t$ mixture of linear experts (STMoLE) case, where the expert means are linear regressors, that is, of the form (6.7), this maximization can be performed in a closed form and provides the following updates:

$$\boldsymbol{\beta}_k^{(m+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(w_{ik}^{(m)} y_i - \boldsymbol{e}_{1,ik}^{(m)} \delta_k^{(m+1)}\right) \boldsymbol{x}_i, \tag{6.46}$$

$$\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \Big[w_{ik}^{(m)} \left(\boldsymbol{y}_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)} \boldsymbol{e}_{1,ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i) + \boldsymbol{e}_{2,ik}^{(m)}\Big]}{2\left(1 - \delta_k^{2(m)}\right) \sum_{i=1}^{n} \tau_{ik}^{(m)}}. \tag{6.47}$$

CM-Step 3 The skewness parameters λ_k are updated by maximizing $Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)})$ w.r.t λ_k, with $\boldsymbol{\beta}_k$ and σ_k^2 fixed at the update $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2(m+1)}$, respectively. It can be easily shown that the maximization to obtain $\lambda_k^{(m+1)}$ $(k = 1, \ldots, K)$ consists in solving the following equation in λ_k (recall we have $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$):

$$\delta_k(1 - \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} + (1 + \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} \frac{d_{ik}^{(m+1)} e_{1,ik}^{(m)}}{\sigma_k^{(m+1)}} - \delta_k \sum_{i=1}^{n} \tau_{ik}^{(m)} \Big[w_{ik}^{(m)} d_{ik}^{2\ (m+1)} + \frac{e_{2,ik}^{(m)}}{\sigma_k^{2(m+1)}}\Big] = 0\cdot \tag{6.48}$$

CM-Step 4 Similarly, the degrees of freedom ν_k are updated by maximizing $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$ w.r.t ν_k with $\boldsymbol{\beta}_k$ and σ_k^2 fixed at $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2(m+1)}$, respectively. An update $\nu_k^{(m+1)}$ is calculated as solution of the following equation in ν_k:

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left(\boldsymbol{e}_{3,ik}^{(m)} - w_{ik}^{(m)}\right)}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} = 0. \tag{6.49}$$

The two scalar non-linear equations (6.48) and (6.49) can be solved similarly as in the TMoE model, that is, with a root finding algorithm, such as Brent's method (Brent, 1973).

As mentioned before, one can see that, when the robustness parameter $\nu_k \to \infty$ for all the components, the parameter updates for the STMoE model correspond to those of the SNMoE model. On the other hand, when the skewness parameters $\lambda_k = 0$, the STMoE parameter updates correspond to those of the TMoE model. Finally, when both the degrees of freedom $\nu_k \to \infty$ and the skewness parameters $\lambda_k = 0$, we obtain the parameter updates of the standard NMoE model. The STMoE therefore provides a more general framework for inferring flexible MoE models.

6.5 Prediction, clustering and model selection with the non-normal MoE

Prediction The goal in regression is to be able to make predictions for the response variable(s) given some new value of the predictor variable(s) on the basis of a model trained on a set of training data. In regression analysis using mixture of experts, the aim is therefore to predict the response y given new values of the predictors $(\boldsymbol{x}, \boldsymbol{r})$, on the basis of a MoE model characterized by a parameter vector $\hat{\boldsymbol{\Psi}}$

inferred from a set of training data, here, by maximum likelihood via E(C)M. These predictions can be expressed in terms of the predictive distribution of y, which is obtained by substituting the maximum likelihood parameter $\hat{\boldsymbol{\Psi}}$ into (6.2) to give:

$$f(y|\boldsymbol{x},\boldsymbol{r};\hat{\boldsymbol{\Psi}}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}}) f_k(y|\boldsymbol{x};\hat{\boldsymbol{\Psi}}_k).$$

Using f, we might then predict y for a given set of $\boldsymbol{x}$'s and $\boldsymbol{r}$'s as the expected value under f, that is by calculating the prediction $\hat{y} = \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r},\boldsymbol{x})$. I thus need to compute the expectation of the mixture of experts model. It is easy to show (see for example Section 1.2.4 in Frühwirth-Schnatter (2006)) that the mean and the variance of a mixture of experts distribution are respectively given by

$$\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r},\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}}_n)\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x}), \tag{6.50}$$

$$\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r},\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}}_n)\big[\left(\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x})\right)^2 + \mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x})\big] - \big[\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r},\boldsymbol{x})\big]^2, \tag{6.51}$$

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. The calculations of the mean and the variance, for each of the developed MoE models, are derived respectively in [J-12] and [J-13][J-14].

Model-based clustering The MoE models can also be used for a model-based clustering perspective to provide a partition of the regression data into K clusters. Model-based clustering using the NNMoE consists in assuming that the observed data $\{\boldsymbol{x}_i,\boldsymbol{r}_i,y_i\}_{i=1}^n$ are generated from a K component mixture of, respectively, skew-normal, t or skew t experts, with parameter vector $\boldsymbol{\Psi}$. The mixture components can be interpreted as clusters and hence each cluster can be associated with a mixture component. The problem of clustering therefore becomes the one of estimating the MoE parameters $\boldsymbol{\Psi}$, which is performed here by using dedicated EM algorithms. Once the parameters are estimated and we get $\hat{\boldsymbol{\Psi}}$, the provided posterior component memberships τ_{ik} given by

$$\tau_{ik}(\hat{\boldsymbol{\Psi}}) = \frac{\pi_k(\boldsymbol{r};\hat{\boldsymbol{\Psi}}) f_k\Big(y_i|\boldsymbol{r}_i,\boldsymbol{x}_i;\hat{\boldsymbol{\Psi}}\Big)}{\sum_{k'=1}^{K}\pi_{k'}(\boldsymbol{r};\hat{\boldsymbol{\alpha}}) f_{k'}\Big(y_i|\boldsymbol{r}_i,\boldsymbol{x}_i;\hat{\boldsymbol{\Psi}}_{k'}\Big)} \tag{6.52}$$

represent a fuzzy partition of the data. A hard partition of the data can then be obtained from the posterior memberships by applying the Bayes' optimal allocation rule, that is, by maximizing the posterior component memberships to assign each observation to a cluster: $\hat{z}_i = \arg\max_{k=1}^{K} \hat{\tau}_{ik}(\hat{\boldsymbol{\Psi}})$ where $\hat{z}_i$ represents the estimated cluster label for the ith observation.

Model selection One of the issues in mixture model-based clustering is model selection. The problem of model selection for the NNMoE models presented here in their general forms, is equivalent to the one of choosing the optimal number of experts K, the degree p of the regression and the degree q for the logistic regression. The optimal value of (K,p,q) can be computed by using some model selection criteria such as AIC, BIC, ICL, which are used here. The AIC and BIC are are penalized observed data log-likelihood criteria which can be defined as functions to be maximized and are respectively given by: $\text{AIC}(K,p,q) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}}\log(n)}{2}$, $\text{BIC}(K,p,q) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}}\log(n)}{2}$. The ICL criterion consists in a penalized complete-data log-likelihood and can be expressed as follows: $\text{ICL}(K,p,q) = \log L_c(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}}\log(n)}{2}$. In the above, $\log L(\hat{\boldsymbol{\Psi}})$ and $\log L_c(\hat{\boldsymbol{\Psi}})$ are respectively the incomplete (observed) data log-likelihood and the complete data log-likelihood, obtained at convergence of the E(C)M algorithm for the corresponding mixture of experts model and $\eta_{\boldsymbol{\Psi}}$ is the number of free model parameters. The number of free parameters $\eta_{\boldsymbol{\Psi}}$ is given by $\eta_{\boldsymbol{\Psi}} = K(p+q+3)-q-1$ for the NMoE model, $\eta_{\boldsymbol{\Psi}} = K(p+q+4)-q-1$ for both the SNMoE and the TMoE models, and $\eta_{\boldsymbol{\Psi}} = K(p+q+5)-q-1$ for the STMoE model.

However, note that in MoE it is common to use mixing proportions modeled as logistic transformation of linear functions of the covariates, that is the covariate vector in (6.1) is given by $\boldsymbol{r}_i = (1,r_i)^T$ (corresponding to $q=2$), r_i being an univariate covariate variable. This is also adopted in this work. Moreover, for the case of linear experts, that is when the experts are linear regressors with parameter

vector $\boldsymbol{\beta}_k$ for which the corresponding covariate vector $\boldsymbol{x}_i$ in (6.7) is given by $\boldsymbol{x}_i = (1, x_i)^T$ (corresponding to $p = 2$), x_i being an univariate covariate variable, the model selection reduces to choosing the number of experts K. Here I mainly consider this linear case with linear components to consider comparisons with approaches that considered the linear case, but for the non-linear (polynomial) case (the equations are given) and the code is also implemented.

6.6 Experiments

In [J-12] [J-13][J-14] I evaluated the performance of proposed EM algorithms for the three NNMoE models in terms of modeling, robustness to outliers and clustering, on both simulated and real data sets.
The simulations showed that the three models provide estimates which converge to the true parameters. In addition, the estimated fitted mean curves, mixing proportions, and obtained partitions are very close to the true counterparts for all the situations, including when the data are generated according to the model in question, or according to the standard Normal one. This supports the fact that the proposed algorithms perform well and the corresponding proposed models are good generalizations of the normal mixture of experts (NMoE).

Robustness of the NNMoE I examined the robustness of the proposed models to outliers versus the standard NMoE one. For that, I considered each of the four models for data generation and inference, where the data include , with a probability c a class of outliers for $c = 0\%, 1\%, 2\%, 3\%, 4\%, 5\%$. I considered the same class of outliers as in Nguyen and McLachlan (2016), that is the predictor x is generated uniformly over the interval $(-1, 1)$ and the response y is set the value -2. As a criterion of evaluation of the impact of the outliers on the quality of the results, I considered the MSE between the true regression mean function and the estimated one.
When there is no outliers ($c = 0\%$), the error of the TMoE is less than those of the other models, for the four situations, that is including the case where the data are not generated according to it, which is somewhat surprising. This includes the case where the data are generated according to the NMoE model, for which the TMoE error is slightly less than the one of the NMoE model. Then, it can be seen that when there is outliers, the TMoE model outperforms the other models for almost all the situations, except the one in which the data are generated according to the STMoE model. When the data do not contain outliers and are generated from the STMoE, this one indeed outperforms the NMoE and SNMoE models. For the situation when there is no outliers and the data are generated according to the TMoE or the STMoE, these two models may provide quasi-identical results. In the case of presence of outliers in data generated from the STMoE, this one outperforms the NMoE and SNMoE models for all the situations, and outperforms the TMoE for the majority of situations, namely when the number of the outliers is more than 2%. Also, for all the situations with outliers, as expected, the TMoE and STMoE models always provide the best results. These two models are indeed much more robust to outliers compared to the normal and skew-normal ones because the expert components in these two models follow a robust distribution, that is the t distribution for the TMoE, and the skew t distribution for the STMoE. The NMoE and SNMoE are sensitive to outliers. When there is outliers, the SNMoE behavior is comparable to the one of the NMoE. But the SNMoE is more adapted to skewed data compared to the standard NMoE model. However, when the number of outliers is increasing, the increase in the error of the NMoE and SNMoE model is more pronounced compared to the one of the TMoE and STMoE models. The error for both the TMoE and STMoE may indeed slightly increase, remain stable or even decrease in some situations. This supports the expected robustness of the TMoE and STMoE models.

Figure 6.1 shows an example of results obtained on a data set simulated according to the NMoE model and contain $c = 5\%$ of outliers. In this example, we clearly see that the NMoE is severely affected by the outliers and provides a rough fit especially for the red component. However, both the TMoE and the STMoE model are clearly robust and provide a precise fit.

Application to two real-world data sets I considered an application to two real-world data sets: the tone perception data set and the temperature anomalies data set (see for for example [J-12][J-13][J-14] for more detailed description of the data).

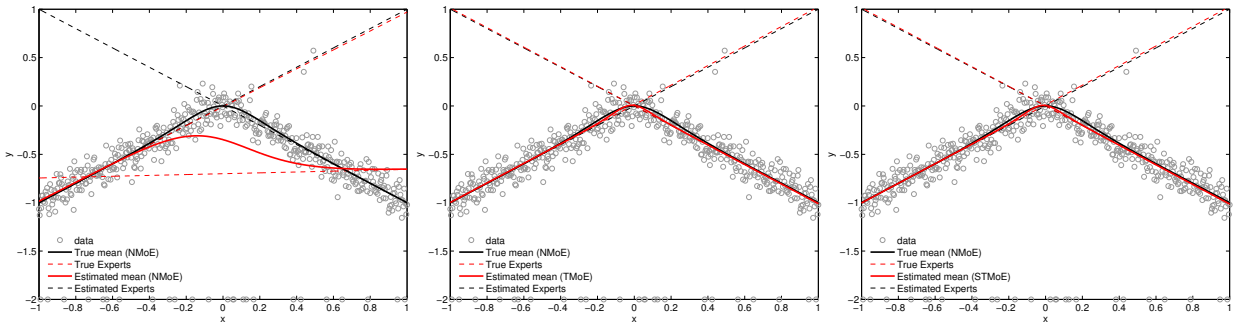

Figure 6.1: Fitted MoE to a data set of $n = 500$ observations generated according to the NMoE model and including 5% of outliers $(x; y = -2)$, with NMoE fit (left), TMoE fit (midde) and STMoE fit (right).

Tone perception data set The first analyzed data set is the real tone perception data set[1] which goes back to Cohen (1984). It was recently 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. To apply the proposed MoE models, we set the response $y_i (i = 1, \ldots, 150)$ as the "strech ratio" variables and the covariates $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ where x_i is the "tuned" variable of the ith observation. For the original data I obtain a good fit with all the models; The NMoE and SNMoE are quasi-identical, and differ very slightly from those of the TMoE and STMoE, which are very similar. The two regression lines may correspond to correct tuning and tuning to the first overtone, respectively, as analyzed in Bai et al. (2012). I also performed a model selection procedure on this data set to choose the best number of MoE components for a number of components between 1 and 5. I used BIC, AIC, and ICL. The NMoE model overestimates the number of components. AIC performs poorly for all the models. BIC provides the correct number of components for the three proposed models. ICL too estimated the correct number of components for both the SNMoE and STMoE models, but hesitates between 2 (the correct number) and 3 components for the TMoE model.
I also examined the sensitivity of the MoE models to outliers based on this real data set. For this, I adopt the same scenario used in Bai et al. (2012) and Song et al. (2014) (the last and more difficult scenario) by adding 10 identical pairs $(0, 4)$ to the original data set as outliers in the y-direction, considered as high leverage outliers. In this situation, the normal and the skew-normal mixture of experts provide almost identical fits and are sensitive to outliers. However, in both cases, compared to the normal regression mixture result in Bai et al. (2012), and the Laplace regression mixture and the t regression mixture results in Song et al. (2014), the fitted NMoE and SNMoE model are affected less severely by the outliers. This may be attributed to the fact that the mixing proportions here are depending on the predictors, which is not the case in these regression mixture models, namely the ones of Bai et al. (2012) and Song et al. (2014). However, as it can be seen on Figure 6.2 the TMoE and the STMoE provide robust fits, which are quasi-identical to the fit obtained on the original data without outliers. Moreover, I notice that, as showed in Song et al. (2014), for this situation with outliers, the t mixture of regressions fails; The fit is affected severely by the outliers. However, the proposed TMoE and STMoE, the ten high leverage outliers have no impact on the fitted experts.

Temperature anomalies data set

This real-world data set[2] relates climate change analysis. The data consist of $n = 135$ yearly measurements of the global annual temperature anomalies (in degrees C) computed using data from land meteorological stations for the period of $1882 - 2012$. The response $y_i (i = 1, \ldots, 135)$ is set as the temperature anomalies and the covariates $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ where x_i is the year of the ith observation. These data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2016) by using the Laplace mixture of linear experts (LMoLE). The four models are successfully applied on the data set and provide very similar results. These results are also similar to those found by Nguyen and McLachlan (2016) who used a Laplace mixture of linear experts. Both the TMoE and STMoE fits

[1]Source: `http://artax.karlin.mff.cuni.cz/r-help/library/fpc/html/tonedata.html`
[2]source: from Ruedy et al., `http://cdiac.ornl.gov/ftp/trends/temp/hansen/gl_land.txt`

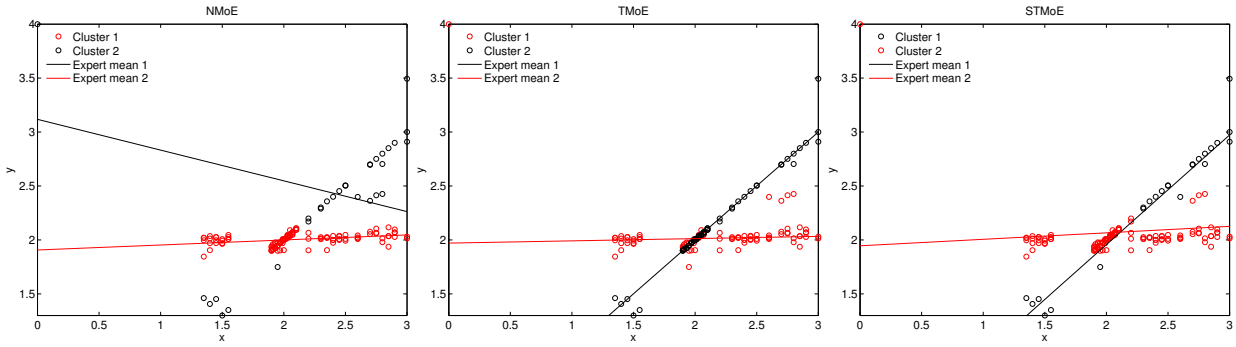

Figure 6.2: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$. Left: NMoE fit, Middle: TMoE fit and Right: STMoE fit. The predictor x is the actual tone ratio and the response y is the perceived tone ratio.

provide a degrees of freedom more than 17, which tends to approach a normal distribution. On the other hand, the regression coefficients are also similar to those found by Nguyen and McLachlan (2016) who used a Laplace mixture of linear experts. I performed a model selection procedure on the temperature anomalies data set to choose the best number of MoE components from values between 1 and 5. Except the result provided by AIC for the NMoE model which overestimates the number of components, all the others results provide evidence for two components in the data.

6.7 Conclusion

In this chapter, I proposed new non-normal MoE models, which generalize the normal MoE. They are based on the skew-normal, t and skew t distribution and are respectively the SNMoE, TMoE, and STMoE. The SNMoE model is suggested for non-symmetric data, the TMoE for data with possibly outliers and heavy tail, and the STMoE is suggested for both possibly non-symmetric, heavy tailed and noisy data. I developed EM-type algorithms to infer each of the proposed models and described the use of the models in non-linear regression and prediction as well as in model-based clustering. The developed models are successfully applied on simulated and real data sets. The results obtained on simulated data confirm the good performance of the models in terms of density estimation, non-linear regression function approximation and clustering. In addition, the simulation results provide evidence of the robustness of the TMoE and STMoE models to outliers, compared to the normal alternative models. The proposed models were also successfully applied to two different real data sets, including a situation with outliers. The model selection using information criteria tends to promote using BIC and ICL against AIC which may perform poorly in the analyzed data. The obtained results support the potential benefit of the proposed approaches for practical applications.

In this chapter, I only considered the MoE in their standard (non-hierarchical) version. One interesting future direction is therefore to extend the proposed models to the hierarchical mixture of experts framework (Jordan and Jacobs, 1994). Furthermore, a natural future extension of this work is to consider the case of MoE for multiple regression on multivariate data rather than simple regression on univariate data.

Chapter 7

Conclusion and perspectives

7.1 Conclusion

The previous chapters presented my research during the last five years as well as my ongoing research on the problems of statistical learning of flexible models for complex data analysis. This involved research in statistics in the related fields of classification, high dimensional and functional data analysis, statistical signal processing, machine learning and pattern recognition, and in the field of statistical inference. The focus in the latter field has been on the methodology and applications of latent data models, particularly mixture models, and on maximum likelihood estimation via EM algorithms as well as maximum a posteriori estimation via Bayesian sampling, including in the Bayesian non-parametric paradigm. A particular attention was given to the statistical methodology and its computational aspects, which constitute a common theme of my research.

7.2 Perspectives

Each part of the manuscript ends with a part where the perspectives and extensions related to the described work are presented. So the perspectives I open here are not taking back (I hope anyway) these developments but instead, propose perspectives less correlated with my previous work.
So beyond the previously discussed perspectives related to each Chapter, here I provide some future avenues I will pursue in the future. They relate some already effectively started ones or others I intend to start in the near future.

7.2.1 Advanced mixtures for complex data (My ongoing CNRS research leave project)

My research on model-based cluster and discriminant analyses extend my interests to further investigating the subject by including (Bayesian) model-based co-cluster and discriminant analyses as well as feature and model selection in unsupervised classification of high dimensional data including functional data. The developed BNP approach for parsimonious models might also be investigated for the recently developed parsimonious models based on a variance-correlation decomposition of the group covariance matrices (Biernacki and Lourme, 2014) which have more desirable properties compared to the flexible parsimonious GMMs baed rather on an eigenvalue decomposition.

These perspectives mainly relate my CNRS research leave I was awarded this year starting from the first of September and which I will spend in the Probability and Statistics team of the lab of mathematics Paul-Painlevé UMR CNRS 8524 in Lille where I will work with mainly Pr. Christope Biernacki. The description of the project is available here `http://chamroukhi.univ-tln.fr/FChamroukhi-projet-delegation-CNRS.pdf`. Possible applications relate namely computational biology (gene expression data), which also involves feature selection and classification challenges, as well as text classification in the discrete case for example.

7.2.2 LEarning from biG cOmplex FunctIonal daTa - LegoFit (2015 - an ANR proposal)

This perspective is actually ongoing and consists in the ANR (french research agency) proposal LEarning from biG cOmplex FunctIonal daTa - "LegoFit" I initiated and submitted this year. I'm the Principal Investigator of the project. LegoFit is an academic research project that aims at developing models and algorithms for transforming big data into knowledge. The considered data are massive functional data with complex hidden structure. The key tenet of the proposal is to develop an original probabilistic methodology that links between statistical learning and functional data analysis (FDA) at large scale. The proposal will be focused on the field of Functional Data Analysis (FDA). Two pilot applications are considered. The first one is in collaboration with the leader AIRBUS and concerns large scale time series data of aircraft condition monitoring. The second one concerns massive transportation data derived from vehicle sharing systems (bikes and car) and is in collaboration with IFSTTAR, the national leader institute on transportation research. The proposal will be focused on the field of Functional Data Analysis (FDA). The overall scientific objective of this project is therefore to significantly improve the automatic analysis and decision making from massive data available as (discretized) values of smooth functions and to apply them in the framework of two main pilot applications. This mainly involves tackling the problems of functional regression, classification and clustering. Particularly, we will focus on the unsupervised context in which some information/data may be missing or hidden and therefore the data completeness is of great interest. A special focus will be given to non-linear dynamical functions that may be subject to multiple changes in regime. We propose to address these issues from a probabilistic prospective through specific latent models with sound statistical framework. In some circumstances, an a priori available knowledge on the data, including on its structure, has to be taken into account as it is likely to improve the models accuracy. In LegoFit this will be formulated within a Bayesian learning framework. A particular focus will be attributed to the non-parametric Bayesian approaches for functional data to provide more flexible and generic probabilistic models adapted to big data. The scalability of the developed algorithms will also be central to the project. The main scientific originality of our proposal covers therefore the development of new statistical learning and data analysis techniques for big data with functional complex structure, and scaling them up. The main questions involved in the proposal form a scientific breakthrough for the partners of the project. We have structured our project around the following tasks for the analysis of big functional data with complex hidden structure:

- Task 1: Model-based clustering and discrimination
- Task 2: Model-based co-clustering
- Task 3: Model-based (co)-clustering under topographic considerations
- Task 4: Bayesian non-parametric clustering
- Task 5: Computational scalability of the proposed algorithms
- Task 6: Two pilot applications for the validation of the proposed methods: AIRBUS data and IFSTTAR data

The consortium is composed of four research units: LSIS (project coordinator, PI Faicel Chamroukhi), LIPN, IFSTTAR-GRETTIA, LIPADE and AIRBUS Research and Technology. It gathers experts in the area of statistical learning, data analysis, computer science and signal and image processing, one industrial leader in aircraft condition monitoring applications, and the national leader institute on transportation research.

7.2.3 Non-normal mixture modeling

The framework of non-normal mixture modeling is receiving increasing attention in these recent years in particular, the skew normal and skew t-mixture models, are emerging as promising extensions to the traditional normal and t-mixture models (Lin et al., 2007b,a; Lin, 2010; Pyne et al., 2009; Lee and McLachlan, 2013b,a, 2014, 2015) This what led me to develop the mixture of experts with more flexible parametric expert components that can better accommodate data exhibiting non-normal features, including asymmetry, heavy-tails, and the presence of outliers. In the future, I intend to pursue this direction by further investigating the non-normal mixture modeling framework for potentially functional data and in a more flexible hierarchical setting, such as hierarchical mixture of experts, as well us possibly into a BNP framework to provide a non-parametric framework for model inference and selection.

7.2.4 Feature selection in model-based clustering

Until now in the problem of model-based clustering, I was mainly focusing on classifying individuals. To deal with high dimensional problems, I either tackled the problem by parsimonious models through the re-parametrization the cluster-specific covariance matrix, or directly performed the classification in the input data space via functional data models. There is another known and quite new framework to deal with the issue of unsupervised classification of high dimensional data in model-based clustering, that is the one of feature selection in clustering which in general consists in classifying individuals while figuring out and keeping only variables which describe at best the individuals. These last years this problem took an important interest in the community (Law et al., 2004; Raftery and Dean, 2006; Zhou et al., 2009; Maugis et al., 2009b,a; Witten and Tibshirani, 2010; Celeux et al., 2011) I intend to investigate it as it opens interesting challenges in modeling as well the computational aspects related to inference.

7.2.5 Bayesian latent variable models for sparse representations

My research in Bayesian inference for unsupervised data classification with the focus on dealing with high dimensional data, extended my interests to Bayesian inference for unsupervised data representation, particularly Bayesian leaning of sparse representations. Some work concerning this perspective is already ongoing and relates a not filled Master internship position I proposed few months ago. The problem of finding sparse representations of a "signal" given a dictionary of possibly overcomplete basis vectors is an important task in several scientific domains including signal processing, computer vision and for many application area such as signal and image compression/restoration, object recognition, etc (Olshausen and Field, 1996, 1997, 2004)(Dobigeon and Tourneret, 2010). Several methods have been proposed for sparse coding, for example the ones based on $l1$-norm regularized regression known as the LASSO (Tibshirani, 1996), also often referred to as Basis Pursuit (BP) (Chen et al., 1999b), FOCUS (Gorodnitsky and Bhaskar, 1997), as well as explicitly formulated Bayesian methods for finding sparse representations (Wipf, 2006) namely l_1-norm Bayesian sparse representations (Lin, 2008) and (Dobigeon and Tourneret, 2010) where, roughly, the sparse codes are modeled by as Bernoulli-Gaussian processes, or related Bayesian pursuit algorithms (Herzet and Drémeau, 2014; Drémeau and Herzet, 2011). The Bayesian inference framework for finding sparse representations offers a principled general framework for sparse coding as in many cases, cost error functions related to deterministic sparse coding approaches are particular cases for maximum a posteriori (MAP) criteria of corresponding Bayesian models. The Bayesian algorithms for sparse coding allow therefore for taking explicitly and in a principled way a prior knowledge on a formulated probabilistic model to encourage sparsity and they include namely latent data models (e.g. see (Wipf, 2006)). The statistical inference for the resulting Bayesian regularized models is tackled by using dedicated statistical inference tools, that is, MCMC as for example in Dobigeon and Tourneret (2010) as well as from a usual frequentist-like point of view by using Variational Bayes EM as in Drémeau and Herzet (2011).

One of the fast and efficient developed approaches for sparse representations if the Predictive Sparse Decomposition (PSD) (Kavukcuoglu et al., 2008; Kavukcuoglu, 2011) which jointly learns a dictionary and approximates the sparse representations by a predictive function (rather than computing exact sparse representations). A first avenue to derive efficient Bayesian sparse representations might be to formulate the PSD in a Bayesian framework which leads to a Bayesian Predictive Sparse Decomposition (BPSD). Then, the BPSD might be reformulated as a latent variable model by integrating a mixture prior over the codes in order to control the sparsity into a probabilistic way. This may lead to a MAP criterion which may be solved by an EM-type algorithm. I also would be interested in applying them to images and/or sounds representation for recognition.

7.2.6 Unsupervised learning of feature hierarchies: Deep learning

My research in the field of classification of multivariate data is on deriving flexible models that provide accurate partitions of unlabeled data or make prediction for future data based on labeled data. In both cases, the input data are assumed to be (hopefully) well-constructed features so that the main focus is on the classification. However, the focus on providing representations that represent at best the data is without doubt beneficial to (easily achieve) the classification task. Good representations indeed eliminate irrelevant variabilities of the input data, while preserving the information that is useful for the final task (e.g. recognition or prediction). This generates challenges regarding learning representations and one

family of methods is for example the one stated before, that is, sparse coding techniques where the purpose is to produce sparse representations in an unsupervised way. One recent successful and very popular technique is the one based on hierarchical neural networks, known as deep learning which aim to produce deep feature hierarchies as proposed by Hinton and Salakhutdinov (2006); Hinton et al. (2006); Bengio and LeCun (2007); Ranzato et al. (2008); Bengio (2009). Deep architectures for producing high level representations consist, roughly, in stacking unsupervised neural network modules on top of each other so that the input of each module is the output of the one at the level just below. This gives them the ability to capture high-level dependencies in the data and the learned representations provide very accurate classification results when using standard classification techniques such as a Support Vector Classifier or even a logistic regression. However, one main question is, at least from a probabilistic point of view is why do they work. Deep networks have been developed as biologically plausible models for approximating "humans" in recognizing objects and hence have already at least biological foundations to explain how do they work. Very recently, Patel et al. (2015), introduced a probabilistic theory of deep learning that seems to answer the question from a probabilistic point of view. The power of such models in representations with both biological and probabilistic foundations greatly motivates me to investigate the deep learning in the future. I namely want to develop a platform to KOGITATE - KnOwledGe, learnIng and arTificiAl InTElligence `http://cogiter.univ-tln.fr/` which is namely dedicated to learning latent data models and deep feature hierarchies for artificial intelligence problems.

Chapter 8

Personal bibliography

My citations can be found in my Google Scholar profile: `https://scholar.google.com/citations?user=AOxrOsOAAAAJ&hl=fr`

8.1 Monograph and editorials

[B-1] F. Chamroukhi. *Probabilistic Learning From Longitudinal Data: Background, Novel theoretical models, Classifiers and Algorithms*. Lap Lambert Academic Publishing, 2011a. ISBN 978-3844311372

[B-2] H. Glotin, F. Chamroukhi, and T. Maillot. *Representations and Decisions in Cognitive Vision*. Proceedings of the International Summer School ERMITES 2012, 2012. ISBN 979-10-90821-00-2. pdf

[B-3] F. Chamroukhi and H. Glotin. *Unsupervised learning from big bioacoustic data (uLearnBio)*. Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML), 2014. ISBN 979-10-90821-06-4. home page

8.2 Journal papers

8.2.1 Publications (9)

[J-1] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009d. URL `http://chamroukhi.univ-tln.fr/papers/Chamroukhi_Neural_Networks_2009.pdf`

[J-2] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_neucomp_2010.pdf`

[J-3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1: 15–32, Jan 2011c. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_same_govaert_aknin_rnti.pdf`

[J-4] A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011. URL `http://chamroukhi.univ-tln.fr/papers/adac-2011.pdf`

[J-5] F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013a.pdf`

[J-6] F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013b.pdf`

[J-7] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on Hidden Markov Model Regression. *IEEE Transactions on Automation Science and Engineering*, 3(10):829–335, 2013. URL `http://arxiv.org/pdf/1312.6965.pdf`

[J-8] F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015e. doi: 10.1080/00949655.2015.1109096. URL `http://chamroukhi.univ-tln.fr/papers/Chamroukhi-JSCS-2015.pdf`. Published online: 05 Nov 2015

[J-9] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015d. URL `http://arxiv.org/pdf/1312.6974v2.pdf`. Accepted

8.2.2 Submitted papers (6)

[J-10] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet Process Parsimonious Gaussian Mixture for clustering. *arXiv:1501.03347*, 2015. URL `http://arxiv.org/pdf/1501.03347.pdf`. Submitted

[J-11] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, 2015a. URL `http://arxiv.org/pdf/1508.00635.pdf`

[J-12] F. Chamroukhi. Non-Normal Mixtures of Experts. *arXiv:1506.06707*, 2015c. URL `http://arxiv.org/pdf/1506.06707.pdf`. 61 pages

[J-13] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015g. URL `http://chamroukhi.univ-tln.fr/papers/TMoE.pdf`. submitted

[J-14] F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015f. URL `http://chamroukhi.univ-tln.fr/papers/STMoE.pdf`. submitted

[J-15] F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 2015. URL `http://chamroukhi.univ-tln.fr/papers/Sensors-2015.pdf`. submitted

8.2.3 Papers in preparation (2)

[J-16] F. Chamroukhi. Mixture of hidden Markov model regressions for functional data clustering and segmentation. *Neural Networks*, 2015b. In preparation

[J-17] F. Chamroukhi et al. Bayesian non-parametric models for unsupervised decomposition of whale songs. *Journal of Acoustical Society of America*, 2015. In preparation

8.3 International conference papers

[C-1] M. Bartcus, F. Chamroukhi, and H. Glotin. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In *The International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015

[C-2] F. Chamroukhi, M. Bartcus, and H. Glotin. Bayesian non-parametric parsimonious Gaussian mixture for clustering. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, August 2014a

[C-3] M. Bartcus and F. Chamroukhi. Hierarchical Dirichlet Process Hidden Markov Model for unsupervised learning from bioacoustic data. In *Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML)*, Beijing, June 2014

[C-4] F. Chamroukhi, Marius Bartcus, and Herve Glotin. Bayesian non-parametric parsimonious clustering. In *Proceedings of 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, April 2014b

[C-5] F. Chamroukhi. Robust EM algorithm for model-based curve clustering. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Dallas, Texas, August 2013

[C-6] Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture. In *Proceedings of the NIPS4B workshop, Neural Information Processing Systems (NIPS)*, pages 205–211, Nevada, USA, 2013

[C-7] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Classification automatique de données temporelles en classes ordonnées. In *Actes des 44 ème Journées de Statistique*, Bruxelles, Belgique, Mai 2012c

[C-8] F. Chamroukhi and H. Glotin. Mixture model-based functional discriminant analysis for curve classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Brisbane, Australia, June 2012a

[C-9] F. Chamroukhi, H. Glotin, and C. Rabouy. Functional Mixture Discriminant Analysis with hidden process regression for curve classification. In *Proceedings of XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 281–286, Bruges, Belgium, April 2012a

[C-10] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Supervised and unsupervised classification approaches for human activity recognition using body-mounted sensors. In *Proceedings of the XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 417–422, Bruges, Belgium, April 2012

[C-11] F. Chamroukhi, A. Samé, P. Aknin, and G. Govaert. Model-based clustering with Hidden Markov Model regression for time series with regime changes. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 2814–2821, Jul-Aug 2011a

[C-12] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Activity Recognition Using Hidden Markov Models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, september 2011

[C-13] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A dynamic probabilistic modeling of railway switches operating states. In *Proccedings of the 9th World Congress on Railway Research (WCRR)*, Lille, May 2011b

[C-14] 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)*, pages 489–496, Atlanta, GA, June 2009c

[C-15] R. Onanena, F. Chamroukhi, L. Oukhellou, D. Candusso, P. Aknin, and D. Hissel. Supervised learning of a regression model based on latent process. Application to the estimation of fuel cell lifetime. In *Proceeding of the Eighth IEEE International Conference on Machine Learning and Applications (IEEE ICMLA)*, pages 632–637, Miami, Flordia, USA, 2009. IEEE Computer Society

[C-16] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A regression model with a hidden logistic process for signal parameterization. *Proceedings of XVIIth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 503–508, 2009b

[C-17] F. Chamroukhi, A. Samé, and P. Aknin. A probabilistic approach for the classification of railway switch operating states. In *The VIth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Dublin, UK, 2009a. IEEE

[C-18] F. Chamroukhi, A. Samé, and P. Aknin. Switch mechanism diagnosis using a pattern recognition approach. In *The 4th IET International Conference on Railway Condition Monitoring RCM (IEEE)*, pages 1–4, Derby, UK, June 2008b. IEEE

8.4 Invited talks in international conferences

[C-1] F. Chamroukhi. Robust non-normal mixtures of experts. ERCIM 2015 : The 8th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2015h. London, UK

[C-2] F. Chamroukhi. Model-based cluster and discriminant analysis for functional data. ERCIM 2014 : The 7th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2014a. Pisa, Italy

[C-3] F. Chamroukhi. Mixture models for cluster analysis: from model-based inference to Bayesian nonparametrics. uLearnBio workshop of the International Conference on Machine Learning (ICML), June 2014b

[C-4] F. Chamroukhi. Learning probabilistic latent process models from temporal data. VIIth International Summer School ERMITES 2012 on Representations and Decisions in Cognitive Vision, september 2012d

8.5 Francophone conferences

[C-1] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Classification automatique de données temporelles en classes ordonnées. In *Actes des 44 ème Journées de Statistique*, Bruxelles, Belgique, Mai 2012c

[C-2] F. Chamroukhi, A. Samé, and P. Aknin. Régression à variable latente pour la modélisation de signaux de manœuvres d'aiguillage. In In J. Marais and editors M. Berbineau, editors, *Actes INRETS : Communiquer, naviguer, surveiller. Innovations pour des transports plus sûrs, plus efficaces et plus attractifs*, volume 112, pages 57–64, 2008a

[C-3] A. Samé, F. Chamroukhi, and P. Aknin. Détection séquentielle de défauts sur des signaux de manœuvres d'aiguillage. In *Workshop Surveillance, Sûreté et Sécurité des Grands Systèmes 3SGS'08*, Troyes, France, June 2008

[C-4] A. Samé, F. Chamroukhi, and G. Govaert. Algorithme EM et modèle à processus latent pour la régression non linéaire. In *Actes des 41èmes Journées de Statistique de la SFDS*, Bordeaux, 2009

[C-5] M. Bartcus, F. Chamroukhi, and H. Glotin. Clustering Bayésien Parcimonieux Non-Paramétrique. In *Extraction et Gestion des Connaissances (EGC), Atelier CluCo : Clustering et Co-clustering*, pages 3–13, Rennes, France, Jan 2014

8.6 Theses

[Th-1] F. Chamroukhi. *Hidden process regression for curve modeling, classification and tracking.* Ph.D. thesis, Université de Technologie de Compiègne, Compiègne, France, 2010a

[Th-2] F. Chamroukhi. Reconnaissance de formes pour le diagnostic et le suivi de point de fonctionnement. Mster of engenieering thesis, Paris 6 university, Paris, France, 2007

8.7 Award and distinctions

[A-1] F. Chamroukhi. Titulaire de la PEDR Prime de Recherche et d'Encadrement Doctoral (ex. PES Prime d'Excellence Scientifique) 2014-2017

[A-2] F. Chamroukhi. Modèle probabiliste à processus latent pour la description, la segmentation et la classification de signaux unidimensionnels. Best Poster Award of the Winter School "Statistical Learning and Data Mining", Feb 2010b

8.8 Invited and contributed seminars

[S-1] F. Chamroukhi. Statistical learning of latent data models for complex data analysis. Séminaire du Laboratoire de Mathématiques Paul Painlevé, UMR CNRS 8524, Université Lille 1, 04 Nov 2015i

[S-2] F. Chamroukhi. Apprentissage de modèles probabilistes à processus latent à partir de données temporelles. Seminaire GFD-LIPADE Université Paris 5, july 2012e

[S-3] F. Chamroukhi. Apprentissage de modèles génératifs à partir de données temporelles. Séminaire du Laboratoire LSIS, Jan 2012a. Toulon

[S-4] F. Chamroukhi. Apprentissage de modèles génératifs à partir de données temporelles. Séminaire du Laboratoire LSIS, Feb 2012c. Toulon

[S-5] F. Chamroukhi. Apprentissage de modèles probabilistes génératifs : Approches Bayésiennes Parcimonieuses pour des données fonctionnelles et des données multidimensionnelles. Séminaire du Laboratoire LSIS, Jan 2012b. Toulon

[S-6] F. Chamroukhi. Diagnostic par analyse de données longitudinales. TISIC, Séminaire n 13, "Analyse de Données Longitudinales", March 2011b

[S-7] F. Chamroukhi. Modèle probabiliste à base de processus latent pour la description et la classification de signaux. Application au diagnostic d'un système ferroviaire. The Computer Science lab of Paris 13 University (LIPN), April 2010c

[S-8] F. Chamroukhi. Modélisation probabiliste à base de processus latent de signaux monodimensionnels. Utilisation dans le diagnostic d'un composant de l'infrastructure ferroviaire: l'aiguillage. Business Intelligence lab, Télécom-ParisTech, October 2009a

[S-9] F. Chamroukhi. Diagnostic par suivi de point de fonctionnement. Journée des doctorants Heudiasyc, July 2009b. Compiègne

[S-10] F. Chamroukhi. Modèle à processus latent pour la paramétrisation et l'apprentissage de signaux évolutifs. Application au diagnostic d'aiguillages. Séminaire DIAG, Inrets-LTN, April 2009c. Marne-La-Vallée

[S-11] F. Chamroukhi. Diagnostic par suivi de point de fonctionnement. Journée des doctorants Heudiasyc, July 2008a. Compiègne

[S-12] F. Chamroukhi. Régression à variable latente pour la modélisation des signaux de manœvre d'aiguillages. Journée des doctorants SPI, INRETS, July 2008b. Lille

Bibliography

C. Abraham, P. A. Cornillon, E. Matzner-Lober, and N. Molinari. Unsupervised Curve Clustering using B-Splines. *Scandinavian Journal of Statistics*, 30(3):581–595, 2003.

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

D. J. Aldous. Exchangeability and Related Topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Math. 1117.

J. Alon, S. Sclaroff, G. Kollios, and V. Pavlovic. Discovering Clusters in Motion Time-Series Data. In *Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition (CVPR)*, pages 375–381, Los Alamitos, CA, USA, 2003.

K. Altun, B. Barshan, and O. Tuncel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43:3605–3620, October 2010.

A. Antoniadis, J. Berruyer, and R. Carmona. *Rgression non linaire et applications*. Economica, 1992.

A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi. Functional Clustering using Wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(1), 2013.

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 2015. URL `http://chamroukhi.univ-tln.fr/papers/Sensors-2015.pdf`. submitted.

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.

M. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015.

M. Bartcus and F. Chamroukhi. Hierarchical Dirichlet Process Hidden Markov Model for unsupervised learning from bioacoustic data. In *Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML)*, Beijing, June 2014.

M. Bartcus, F. Chamroukhi, and H. Glotin. Clustering Bayésien Parcimonieux Non-Paramétrique. In *Extraction et Gestion des Connaissances (EGC), Atelier CluCo : Clustering et Co-clustering*, pages 3–13, Rennes, France, Jan 2014.

M. Bartcus, F. Chamroukhi, and H. Glotin. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In *The International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015.

Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture. In *Proceedings of the NIPS4B workshop, Neural Information Processing Systems (NIPS)*, pages 205–211, Nevada, USA, 2013.

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

S. Basu and S. Chib. Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of the American Statistical Association*, 98:224–235, 2003.

L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the Association for Computing Machinery (CACM)*, 4(6):284, 1961.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1): 1–127, 2009. doi: 10.1561/2200000006. Also published as a book. Now Publishers, 2009.

Yoshua Bengio and Yann LeCun. Scaling Learning Algorithms towards AI. In Léon Bottou, Olivier Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press, 2007.

H. Bensmail and Jacqueline J. Meulman. Model-based Clustering with Noise: Bayesian Inference and Estimation. *Journal of Classification*, 20(1):049–076, 2003.

H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

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. Biernacki, G. Celeux, and G. Govaert. Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.

Christophe Biernacki and Alexandre Lourme. Stable and visualizable Gaussian parsimonious clustering models. *Statistics and Computing*, 24(6):953–969, 2014.

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

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

David M. Blei and Michael I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

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

V. L. Brailovsky and Y. Kempner. Application of piecewise regression to detecting internal structure of signal. *Pattern recognition*, 25(11):1361–1370, 1992.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Wadsworth, New York, 1984.

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.

Bradley P. Carlin and Siddhartha Chib. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B*, 57(3):473–484, 1995.

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

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

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

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

G. Celeux, M. Hurn, and C. P. Robert. Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

G. Celeux, O. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5:1–25, 2005.

Gilles Celeux, Marie-Laure Martin-Magniette, Cathy Maugis, and Adrian E. Raftery. Letter to the editor: "A framework for feature selection in clustering". *Journal of the American Statistical Association*, 106: 383, 2011. doi: 10.1198/jasa.2011.tm10681.

F. Chamroukhi. Titulaire de la PEDR Prime de Recherche et d'Encadrement Doctoral (ex. PES Prime d'Excellence Scientifique) 2014-2017.

F. Chamroukhi. Reconnaissance de formes pour le diagnostic et le suivi de point de fonctionnement. Mster of engenieering thesis, Paris 6 university, Paris, France, 2007.

F. Chamroukhi. Diagnostic par suivi de point de fonctionnement. Journée des doctorants Heudiasyc, July 2008a. Compiègne.

F. Chamroukhi. Régression à variable latente pour la modélisation des signaux de manœvre d'aiguillages. Journée des doctorants SPI, INRETS, July 2008b. Lille.

F. Chamroukhi. Modélisation probabiliste à base de processus latent de signaux monodimensionnels. Utilisation dans le diagnostic d'un composant de l'infrastructure ferroviaire: l'aiguillage. Business Intelligence lab, Télécom-ParisTech, October 2009a.

F. Chamroukhi. Diagnostic par suivi de point de fonctionnement. Journée des doctorants Heudiasyc, July 2009b. Compiègne.

F. Chamroukhi. Modèle à processus latent pour la paramétrisation et l'apprentissage de signaux évolutifs. Application au diagnostic d'aiguillages. Séminaire DIAG, Inrets-LTN, April 2009c. Marne-La-Vallée.

F. Chamroukhi. *Hidden process regression for curve modeling, classification and tracking*. Ph.D. thesis, Université de Technologie de Compiègne, Compiègne, France, 2010a.

F. Chamroukhi. Modèle probabiliste à processus latent pour la description, la segmentation et la classification de signaux unidimensionnels. Best Poster Award of the Winter School "Statistical Learning and Data Mining", Feb 2010b.

F. Chamroukhi. Modèle probabiliste à base de processus latent pour la description et la classification de signaux. Application au diagnostic d'un système ferroviaire. The Computer Science lab of Paris 13 University (LIPN), April 2010c.

F. Chamroukhi. *Probabilistic Learning From Longitudinal Data: Background, Novel theoretical models, Classifiers and Algorithms*. Lap Lambert Academic Publishing, 2011a. ISBN 978-3844311372.

F. Chamroukhi. Diagnostic par analyse de données longitudinales. TISIC, Séminaire n 13, "Analyse de Données Longitudinales", March 2011b.

F. Chamroukhi. Apprentissage de modèles génératifs à partir de données temporelles. Séminaire du Laboratoire LSIS, Jan 2012a. Toulon.

F. Chamroukhi. Apprentissage de modèles probabilistes génératifs : Approches Bayésiennes Parcimonieuses pour des données fonctionnelles et des données multidimensionnelles. Séminaire du Laboratoire LSIS, Jan 2012b. Toulon.

F. Chamroukhi. Apprentissage de modèles génératifs à partir de données temporelles. Séminaire du Laboratoire LSIS, Feb 2012c. Toulon.

F. Chamroukhi. Learning probabilistic latent process models from temporal data. VIIth International Summer School ERMITES 2012 on Representations and Decisions in Cognitive Vision, september 2012d.

F. Chamroukhi. Apprentissage de modèles probabilistes à processus latent à partir de données temporelles. Seminaire GFD-LIPADE Université Paris 5, july 2012e.

F. Chamroukhi. Robust EM algorithm for model-based curve clustering. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Dallas, Texas, August 2013.

F. Chamroukhi. Model-based cluster and discriminant analysis for functional data. ERCIM 2014 : The 7th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2014a. Pisa, Italy.

F. Chamroukhi. Mixture models for cluster analysis: from model-based inference to Bayesian nonparametrics. uLearnBio workshop of the International Conference on Machine Learning (ICML), June 2014b.

F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, 2015a. URL `http://arxiv.org/pdf/1508.00635.pdf`.

F. Chamroukhi. Mixture of hidden Markov model regressions for functional data clustering and segmentation. *Neural Networks*, 2015b. In preparation.

F. Chamroukhi. Non-Normal Mixtures of Experts. *arXiv:1506.06707*, 2015c. URL `http://arxiv.org/pdf/1506.06707.pdf`. 61 pages.

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015d. URL `http://arxiv.org/pdf/1312.6974v2.pdf`. Accepted.

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

URL http://chamroukhi.univ-tln.fr/papers/Chamroukhi-JSCS-2015.pdf. Published online: 05 Nov 2015.

F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015f. URL http://chamroukhi.univ-tln.fr/papers/STMoE.pdf. submitted.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015g. URL http://chamroukhi.univ-tln.fr/papers/TMoE.pdf. submitted.

F. Chamroukhi. Robust non-normal mixtures of experts. ERCIM 2015 : The 8th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2015h. London, UK.

F. Chamroukhi. Statistical learning of latent data models for complex data analysis. Séminaire du Laboratoire de Mathématiques Paul Painlevé, UMR CNRS 8524, Université Lille 1, 04 Nov 2015i.

F. Chamroukhi and H. Glotin. Mixture model-based functional discriminant analysis for curve classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 1–8, Brisbane, Australia, June 2012a.

F. Chamroukhi and H. Glotin. Mixture model-based functional discriminant analysis for curve classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, June 2012b.

F. Chamroukhi and H. Glotin. *Unsupervised learning from big bioacoustic data (uLearnBio)*. Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML), 2014. ISBN 979-10-90821-06-4. home page.

F. Chamroukhi, A. Samé, and P. Aknin. Régression à variable latente pour la modélisation de signaux de manœuvres d'aiguillage. In In J. Marais and editors M. Berbineau, editors, *Actes INRETS : Communiquer, naviguer, surveiller. Innovations pour des transports plus sûrs, plus efficaces et plus attractifs*, volume 112, pages 57–64, 2008a.

F. Chamroukhi, A. Samé, and P. Aknin. Switch mechanism diagnosis using a pattern recognition approach. In *The 4th IET International Conference on Railway Condition Monitoring RCM (IEEE)*, pages 1–4, Derby, UK, June 2008b. IEEE.

F. Chamroukhi, A. Samé, and P. Aknin. A probabilistic approach for the classification of railway switch operating states. In *The VIth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Dublin, UK, 2009a. IEEE.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A regression model with a hidden logistic process for signal parameterization. *Proceedings of XVIIth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 503–508, 2009b.

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)*, pages 489–496, Atlanta, GA, June 2009c.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009d. URL http://chamroukhi.univ-tln.fr/papers/Chamroukhi_Neural_Networks_2009.pdf.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL http://chamroukhi.univ-tln.fr/papers/chamroukhi_neucomp_2010.pdf.

F. Chamroukhi, A. Samé, P. Aknin, and G. Govaert. Model-based clustering with Hidden Markov Model regression for time series with regime changes. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, pages 2814–2821, Jul-Aug 2011a.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A dynamic probabilistic modeling of railway switches operating states. In *Proccedings of the 9th World Congress on Railway Research (WCRR)*, Lille, May 2011b.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011c. URL http://chamroukhi.univ-tln.fr/papers/chamroukhi_same_govaert_aknin_rnti.pdf.

F. Chamroukhi, H. Glotin, and C. Rabouy. Functional Mixture Discriminant Analysis with hidden process regression for curve classification. In *Proceedings of XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 281–286, Bruges, Belgium, April 2012a.

F. Chamroukhi, H. Glotin, and C. Rabouy. Functional Mixture Discriminant Analysis with hidden process regression for curve classification. In *Proceedings of XXth European Symposium on Artificial Neural Networks ESANN*, pages 281–286, Bruges, Belgium, April 2012b.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Classification automatique de données temporelles en classes ordonnées. In *Actes des 44 ème Journées de Statistique*, Bruxelles, Belgique, Mai 2012c.

F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013a.pdf`.

F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b. URL `http://chamroukhi.univ-tln.fr/papers/chamroukhi_et_al_neucomp2013b.pdf`.

F. Chamroukhi, M. Bartcus, and H. Glotin. Bayesian non-parametric parsimonious Gaussian mixture for clustering. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, August 2014a.

F. Chamroukhi, Marius Bartcus, and Herve Glotin. Bayesian non-parametric parsimonious clustering. In *Proceedings of 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, April 2014b.

F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet Process Parsimonious Gaussian Mixture for clustering. *arXiv:1501.03347*, 2015. URL `http://arxiv.org/pdf/1501.03347.pdf`. Submitted.

K. Chen, L. Xu, and H. Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252, 1999a.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit,. *SIAM Journal on Scientific Computing*, 1999b.

Raymond J. Cho, Michael J. Campbell, Elizabeth A. Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G. Wolfsberg, Andrei E. Gabrielian, David Landsman, David J. Lockhart, and Ronald W. Davis. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, 2(1):65–73, 1998.

Elizabeth A. Cohen. Some effects of inharmonic partials on interval perception. *Music Perception*, 1, 1984.

Sophie Dabo-Niang, Frédéric Ferraty, and Philippe Vieu. On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis*, 51(10):4878 – 4890, 2007.

C. Deboor. *A Practical Guide to Splines*. Springer-Verlag, 1978.

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.

Wayne DeSarbo and William Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, 1988.

E. Devijver. Model-based clustering for high-dimensional data. Application to functional data. Technical report, Département de Mathématiques, Université Paris-Sud, 2014.

F.X. Diebold, J.-H. Lee, and G. Weinbach. Regime Switching with Time-Varying Transition Probabilities. *Nonstationary Time Series Analysis and Cointegration. (Advanced Texts in Econometrics)*, pages 283–302, 1994.

Jean Diebolt and C. P. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society, Series B*, 56(2):363–375, 1994.

Nicolas Dobigeon and Jean-Yves Tourneret. Bayesian orthogonal component analysis for sparse representation. *IEEE Transactions on Signal Processing*, 58(5):2675–2685, 2010.

Nicolas Dobigeon, Jean-Yves Tourneret, and Jeffrey D. Scargle. Joint segmentation of multivariate astronomical time series : Bayesian sampling with a hierarchical model. *IEEE Transactions on Signal Processing*, 55(2):414–423, 2007.

A. Drémeau and C. Herzet. Soft bayesian pursuit algorithm for sparse representations. In *in IEEE International Workshop on Statistical Signal Processing SSP'11)*, 2011.

David Hart Eamonn Keogh, Selina Chu and Michael Pazzani. *Segmenting Time Series: A Survey and Novel Approach*, chapter In an Edited Volume, Data mining in Time Series Databases, pages 1–22. Published by World Scientific, 1993.

Michael D. Escobar. Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1994.

F. Chamroukhi et al. Bayesian non-parametric models for unsupervised decomposition of whale songs. *Journal of Acoustical Society of America*, 2015. In preparation.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, February 2010.

P. Fearnhead and Z. Liu. Online Inference for Multiple Changepoint Problems. *Journal of the Royal Statistical Society, Series B*, 69:589–605, 2007.

Paul Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213, 2006.

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

F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4):545–564, 2002.

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

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

Mário A. T. Figueiredo and Anil K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24:381–396, 2000.

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

C. Fraley and A. E. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 24(2):155–181, 2007.

Chris Fraley and Adrian E. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Technical Report 486, Departament of Statistics, Box 354322, University of Washington Seattle, WA 98195-4322 USA, August 2005.

M. Fridman. Hidden Markov Model Regression. Technical report, Institute of mathematics, University of Minnesota, 1993.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models (Springer Series in Statistics)*. Springer Verlag, New York, 2006.

S. Frühwirth-Schnatter and S. Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336, 2010.

S. J. Gaffney. *Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models*. PhD thesis, Department of Computer Science, University of California, Irvine, 2004.

S. J. Gaffney and P. Smyth. Joint probabilistic curve clustering and alignment. In *In Advances in Neural Information Processing Systems (NIPS)*, 2004.

Scott Gaffney and Padhraic Smyth. Trajectory Clustering with Mixtures of Regression Models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72. ACM Press, 1999.

A. E. Gelfand and D. K. Dey. Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B*, 56(3):501–514, 1994.

Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, June 1990.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984.

C. Geyer. Markov Chain Monte Carlo maximum likelihood. In *Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension. *Biometrics*, 69(1):31–40, 2013.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

H. Glotin, F. Chamroukhi, and T. Maillot. *Representations and Decisions in Cognitive Vision*. Proceedings of the International Summer School ERMITES 2012, 2012. ISBN 979-10-90821-00-2. pdf.

I. Gorodnitsky and D. R. Bhaskar. Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

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

P. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some robust and resistant alternatives. *Journal of The Royal Statistical Society, B*, 46(2):149–192, 1984.

Peter J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82:711–732, 1995.

J. Gui and H. Li. Mixture functional discriminant analysis for gene function classification based on time course gene expression data. In *Proc. Joint Stat. Meeting (Biometric Section)*, 2003.

J. Hansen, R. Ruedy, J. Glascoe, and M. Sato. GISS analysis of surface temperature change. *Journal of Geophysical Research*, 104:30997–31022, 1999.

J. Hansen, R. Ruedy, Sato M., M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl. A closer look at United States and global surface temperature change. *Journal of Geophysical Research*, 106:23947–23963, 2001.

T. Hastie and R. Tibshirani. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society, B*, 58:155–176, 1996.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition edition, January 2010.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized Discriminant Analysis. *Annals of Statistics*, 23:73–102, 1995.

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.

C. Herzet and A. Drémeau. Bayesian pursuit algorithms. In *CoRR abs/1401.7538*, 2014.

G.E. Hinton and R.R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006.

N. Hjort, C. Holmes, P. Muller, and S. G. Waller. *Bayesian Non Parametrics: Principles and practice*. Cambrige University Press, 2010.

J. P. Hughes, P. Guttorp, and S. P. Charles. A non-homogeneous hidden Markov model for precipitation occurrence. *Applied Statistics*, 48:15–30, 1999.

B. Hugueney, G. Hébrail, Y. Lechevallier, and F. Rossi. Simultaneous Clustering and Segmentation for Functional Data. In *Proceedings of XVIIth European Symposium on Artificial Neural Networks (ESANN)*, pages 281–286, Bruges, Belgium, April 2009.

DR Hunter and DS Young. Semiparametric Mixtures of Regressions. *Journal of Nonparametric Statistics*, 24(1):19–38, 2012.

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.

H. Ishwaren and M. Zarepour. Exact and approximate representations for the sum Dirichlet process. *Canadian Journal of Statistics*, 30:269–283, 2002.

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.

Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.

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.

H. Jeffreys. *Theory of Probability*. Oxford, third edition, 1961.

Wenxin Jiang and Martin A. Tanner. On the Identifiability of Mixtures-of-Experts. *Neural Networks*, 12:197–220, 1999.

P. N. Jones and G. J. McLachlan. FITTING FINITE MIXTURE MODELS IN A REGRESSION CON-

TEXT. *Australian Journal of Statistics*, 34(2):233–240, June 1992.

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

M. I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995.

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Koray Kavukcuoglu. *Learning Feature Hierarchies for Object Recognition*. PhD thesis, Department of Computer Science, New York University, 2011.

Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. Fast Inference in Sparse Coding Algorithms with Applications to Object Recognition. Technical Report CBLL-TR-2008-12-01, Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 2008.

J.T. Kent, D.E. Tyler, and Y Vardi. A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics - Simulation and Computation*, 23:441–453, 1994.

Charles Kooperberg and Charles J. Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, December 1982.

M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, September 2004.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Sharon X. Lee and Geoffrey J. McLachlan. On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, 7(3):241–266, 2013a.

Sharon X. Lee and Geoffrey J. McLachlan. Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods and Applications*, 22(4):427–454, 2013b.

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.

P. Lenk and W. DeSarbo. Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1):93–119, 2000.

Steven M. Lewis and Adrian E. Raftery. Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association*, 92:648–655, 1994.

J. F-S. Lin and D. Kulić. Automatic Human Motion Segmentation and Identification using Feature Guided HMM for Physical Rehabilitation Exercises. In *In Robotics for Neurology and Rehabilitation, Workshop at the IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, California, 2011.

Tsung I. Lin. Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, 20(3):343–356, 2010.

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, 2007a.

Tsung I. Lin, Jack C. Lee, and Shu Y Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17:909–927, 2007b.

Yuanquing Lin. *l_1-Norm sparse Bayesian learning: Theory and applications*. PhD thesis, University of Pennsylvania, 2008.

Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

X. Liu and M.C.K. Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53(4):1361–1376, 2009.

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

Jean-Michel Marin, Kerrie L. Mengersen, and Christian Robert. Bayesian modelling and inference on mixtures of distributions. In D. Dey and C.R. Rao, editors, *Handbook of Statistics: Volume 25*. Elsevier, 2005.

C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009a.

Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65(3):701–709, 2009b.

V. E. McGee and W. T. Carleton. Piecewise regression. *Journal of the American Statistical Association*, 65:1109–1124, 1970.

G. J. McLachlan. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3): 318–324, 1978.

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.

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.

R. M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

S. K. Ng, G. J. McLachlan, K. Wang adn L. Ben-Tovim Jones, and S.-W. Ng. A Mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22(14): 1745–1752, 2006.

Shu-Kay Ng and Geoffrey J. McLachlan. Using the EM algorithm to train neural networks: misconceptions and a new algorithm for multiclass classification. *IEEE Transactions on Neural Networks*, 15(3): 738–749, 2004.

S.K. Ng and G.J. McLachlan. Mixture models for clustering multilevel growth trajectories. *Computational Statistics & Data Analysis*, 71(0):43– 51, 2014.

Hien D. Nguyen and Geoffrey J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93:177–191, 2016. doi: http://dx.doi.org/10.1016/j.csda.2014.10.016.

Hien D. Nguyen, Geoffrey J. McLachlan, and Ian A. Wood. Mixtures of spatial spline regressions for clustering and classification. *Computational Statistics and Data Analysis*, 2014. doi: http://dx.doi.org/10.1016/j.csda.2014.01.011.

B. A. Olshausen and D. J. Field. Emergence of simple celle receptive field properties by learning a sparse code for nature images. *Nature*, 381:607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 1997.

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 2004.

R. Onanena, F. Chamroukhi, L. Oukhellou, D. Candusso, P. Aknin, and D. Hissel. Supervised learning of a regression model based on latent process. Application to the estimation of fuel cell lifetime. In *Proceeding of the Eighth IEEE International Conference on Machine Learning and Applications (IEEE ICMLA)*, pages 632–637, Miami, Flordia, USA, 2009. IEEE Computer Society.

D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9 (4):639–650, 1998.

Federica Pace, Frederic Benard, Herve Glotin, Olivier Adam, and Paul White. Subunit definition and analysis for humpback whale call classification. *Applied Acoustics*, 71(11):1107 – 1112, 2010.

Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A Probabilistic Theory of Deep Learning. Technical Report Technical Report No 2015-1, Rice University Electrical and Computer Engineering Dept., April 2015. URL `http://arxiv.org/abs/1504.00641v1`.

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

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995. ISSN 0178-8051.

J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. of Statistics. UC, Berkeley, 2002.

S Pyne, X Hu, K Wang, E Rossin, TI Lin, LM Maier, C Baecher-Allan, GJ McLachlan, P Tamayo, DA Hafler, PL De Jager, and JP Mesirow. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences USA*, 106(21):8519–8524, 2009.

R. E. Quandt. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972.

R. E. Quandt and J. B. Ramsey. Esimating Mixtures of Normal Distributions and Switching Regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 1986.

Adrian E. Raftery and Nema Dean. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101:168–178, 2006.

Adrian E. Raftery and Steven Lewis. How Many Iterations in the Gibbs Sampler? In *In Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer, 2002.

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.

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

M. A. Ranzato, Y. Boureau, and Y. L. Cun. Sparse Feature Learning for Deep Belief Networks. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 20*, pages 1185–1192, 2008.

C. Rasmussen. The Infinite Gaussian Mixture Model. *Advances in neuronal Information Processing Systems*, 10:554–560, 2000.

Carl Edward Rasmussen and Zoubin Ghahramani. Infinite Mixtures of Gaussian Process Experts. In *In Advances in Neural Information Processing Systems 14*, pages 881–888. MIT Press, 2001.

N. Ravi, N. Dandekar, P. Mysore, and M. Littman. Activity Recognition from Accelerometer Data. *Proceedings of the National Conference on Artificial Intelligence, MIT Press*, pages 1541–1546, 2005.

Chandan K. Reddy, Hsiao-Dong Chiang, and Bala Rajaratnam. TRUST-TECH-Based Expectation Maximization for Learning Finite Mixture Models. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(7):1146–1157, 2008. ISSN 0162-8828.

Sylvia Richardson and Peter J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society*, 59(4):731–792, 1997.

C. Robert and G. Casella. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102–115, Feb 2011.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, second edition, 2007.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999. ISBN 1441919392.

R. Ruedy, M. Sato, and K. Lo. NASA GISS Surface Temperature (GISTEMP) Analysis. DOI: 10.3334/CDIAC/cli.001. Center for Climate Systems Research, NASA Goddard Institute for Space Studies 2880 Broadway, New York, NY 10025 USA.

M.P. Ruppert, D. Wand and R.J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

A. Samé, F. Chamroukhi, and P. Aknin. Détection séquentielle de défauts sur des signaux de manœuvres d'aiguillage. In *Workshop Surveillance, Sûreté et Sécurité des Grands Systèmes 3SGS'08*, Troyes, France, June 2008.

A. Samé, F. Chamroukhi, and G. Govaert. Algorithme EM et modèle à processus latent pour la régression

non linéaire. In *Actes des 41èmes Journées de Statistique de la SFDS*, Bordeaux, 2009.

A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011. URL `http://chamroukhi.univ-tln.fr/papers/adac-2011.pdf`.

J. Gershman Samuel and David M. Blei. A tutorial on Bayesian non-parametric model. *Journal of Mathematical Psychology*, 56:1–12, 2012.

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

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

A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27: 387–397, 1971.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

J. Q. Shi and T. Choi. *Gaussian Process Regression Analysis for Functional Data*. Chapman & Hall/CRC Press, 2011.

J. Q. Shi and B. Wang. Curve prediction and clustering with mixtures of Gaussian process functional regression models. *Statistics and Computing*, 18(3):267–283, 2008.

J. Q. Shi, R. Murray-Smith, and D. M. Titterington. Hierarchical Gaussian process mixtures for regression. *Statistics and Computing*, 15(1):31–41, 2005.

P. Smyth. Clustering Sequences with Hidden Markov Models. In *Advances in Neural Information Processing Systems 9, NIPS*, pages 648–654, 1996.

Hichem Snoussi and Ali Mohammad-Djafari. Penalized maximum likelihood for multivariate Gaussian mixture. pages 36–46, august 2001.

Hichem Snoussi and Ali Mohammad-Djafari. Degeneracy and Likelihood Penalization in Multivariate Gaussian Mixture Models. Technical report, University of Technology of Troyes ISTIT/M2S, 2005.

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

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

M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford, 1997.

M. Stephens. Bayesian Analysis of Mixture Models with an Unknown Number of Components – an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74, 2000a.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809, 2000b.

H. Stone. Approximation of curves by line segments. *Mathematics of Computation*, 15(73):40–47, 1961.

Martin A. Tanner and Wing Hung Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–550, 1987.

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, 1985.

D. Trabelsi. *Contribution à la reconnaissance non-intrusive d'activités humaines*. Ph.D. thesis, Université Paris-Est Créteil, Laboratoire Images, Signaux et Systèmes Intelligents (LiSSi), June 2013.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Activity Recognition Using Hidden Markov Models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, september 2011.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Supervised and unsupervised classification approaches for human activity recognition using body-mounted sensors. In *Proceedings of the XXth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 417–422, Bruges, Belgium, April 2012.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on Hidden Markov Model Regression. *IEEE Transactions on Automation Science and Engineering*, 3(10):829–335, 2013. URL `http://arxiv.org/pdf/1312.6965.pdf`.

Richard D. De Veaux. Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8(3): 227–245, 1989.

Geert Verbeke and Emmanuel Lesaffre. A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

K. Viele and B. Tong. Modeling with Mixtures of Linear Regressions. *Statistics and Computing*, 12: 315–330, 2002.

A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

S. R. Waterhouse. *Classification and regression using Mixtures of Experts*. PhD thesis, Department of Engineering, Cambridge University, 1997.

Steve Waterhouse, David Mackay, and Tony Robinson. Bayesian Methods for Mixtures of Experts. In *NIPS*, pages 351–357. MIT Press, 1996.

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

David Paul Wipf. *Bayesian Methods for Finding Sparse Representations*. PhD thesis, University of California, San Diego, 2006.

Daniela M. Witten and Robert Tibshirani. A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

F. Wood and M. J. Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, 173(1):1–12, 2008.

Yimin Xiong and Dit-Yan Yeung. Time series clustering with ARMA mixtures. *Pattern Recognition*, 37 (8):1675–1689, 2004.

W Xu and D Hedeker. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics*, 11(4):253–73, 2001.

C. C. Yang and Y. L. Hsu. A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring. *Sensors*, 10:7772–7788, 2010.

Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, November 2012. ISSN 0031-3203.

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

DS Young and DR Hunter. Mixtures of Regressions with Predictor-Dependent Mixing Proportions. *Computational Statistics and Data Analysis*, 55(10):2253–2266, 2010.

Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty Years of Mixture of Experts. *IEEE Trans. Neural Netw. Learning Syst.*, 23(8):1177–1193, 2012.

C. B. Zeller, V. H. Lachos, and C.R. Cabral. Robust Mixture Regression Modelling based on Scale Mixtures of Skew-Normal Distributions. *Test (revision invited)*, 2015.

Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics [electronic only]*, 3:1473–1496, electronic only, 2009. ISSN 1935-7524.