Université de Lille
École doctorale des Sciences Pour l'Ingénieur

Habilitation à Diriger des Recherches (HDR)

présentée par **Vincent Vandewalle**
Maître de Conférences, Université de Lille
le 7 janvier 2021 à Villeneuve d'Ascq

Discipline : **Mathématiques Appliquées**

Contribution to model-based clustering of heterogeneous data

Jury :

Examinateur	Christophe Biernacki	Professeur	Université de Lille (Détaché DR Inria)
Rapporteur	Sophie Lambert-Lacroix	Professeure	Université Grenoble Alpes
Examinateur	Mohamed Lemdani	Professeur	Université de Lille
Rapporteur	Catherine Matias	Dir. de recherches	CNRS
Rapporteur	Paul D. McNicholas	Professeur	Université McMaster, Canada
Garant	Cristian Preda	Professeur	Université de Lille
Président	Jérôme Saracco	Professeur	Institut Polytechnique de Bordeaux

Acknowledgments

First of all, I would like to thank Sophie Lambert-Lacroix, Catherine Matias et Paul McNicholas who have accepted to take on their time to review my research work. I would also like to thank Christophe Biernacki, Mohamed Lemdani, and Jérôme Saracco for accepting to participate in my defense as examiners. Finally, I would like to thank Cristian Preda for accepting to be my guarantor. I am honored to have all of them as members of my jury.

Je remercie toutes les personnes avec qui j'ai eu l'occasion d'échanger au cours de toutes ces années et qui ont contribué de près ou de loin aux travaux présentés dans ce manuscrit. Le travail présenté est le fruit d'un ensemble d'échanges et de collaborations au cours de ces dix dernières années, ainsi que d'un contexte professionnel et familial.

Je remercie particulièrement Christophe Biernacki avec qui j'ai le plaisir de travailler depuis maintenant de nombreuses années. Je le remercie pour sa bienveillance, pour les échanges toujours riches, pour toute l'énergie qu'il a déployé dans la création et le fonctionnement de l'équipe MODAL. Il est pour moi un véritable exemple dans tous les aspects inhérents à la carrière d'un enseignant chercheur et au-delà.

Je remercie Cristian Preda pour m'avoir transmis le goût de l'analyse des données fonctionnelles, pour m'avoir fait découvrir la société Roumaine de Probabilités et Statistique, et pour sa bonne humeur en toutes circonstances. C'est un plaisir d'échanger avec lui au quotidien.

Je remercie Matthieu Marbac, que j'ai eu plaisir de co-encadrer en thèse, pour l'ensemble des travaux que nous avons mené ensemble pendant sa thèse, et que nous continuons à mener. Une bonne partie des travaux présentés dans ce manuscrit sont issus de travaux communs.

Je remercie Adrien Ehrhardt, que j'ai eu plaisir à co-encadrer en thèse, ce fut très instructif pour moi d'échanger avec lui sur la pratique du *credit-scoring* en entreprise et d'avancer ensemble sur les problématiques de recherche inhérentes. Je remercie Guillemette Marot pour tous les projets communs, et les discussions autour des données omiques. Je me réjouis que nous travaillions plus activement ensemble depuis quelques années, et en particulier à travers nos travaux avec Florence Pinet et Christophe Bauteurs que je tiens eux aussi à remercier. Je remercie Sophie Dabo pour les travaux que nous avons menés avec Cristian Preda sur les données fonctionnelles spatiales. Je remercie Serge Iovleff d'avoir monté le projet CloHé, et ce fut un plaisir de l'avoir comme collègue de bureau. Je remercie Florent Dewez, Benjmain Guedj et Arthur Talpaert pour les travaux que nous avons menés ensemble dans le cadre du projet PERF-AI. J'en profite aussi pour remercier l'entreprise SafetyLine avec qui nous avons travaillé sur ce projet (Pierre Jouniaux, Baptiste Gregorutti, ...). Je remercie Mohammed Sekdi pour nos travaux en cours sur le *clustering*-prédictif avec Christophe Biernacki et Matthieu Marbac. Je remercie

Vlad Barbu pour l'ensemble de nos réflexions en cours sur les modèles de semi-Markov. Je remercie Alexandre Lourme pour toutes les discussions que nous avons eu sur l'interprétation probabiliste du *clustering* spectral. Je remercie Wilfried Heyse et Axel Potier de m'accorder leur confiance pour la co-supervision de leurs thèses.

Je remercie l'ensemble des membres de l'équipes MODAL avec qui j'ai pris plaisir à échanger au cours des dix dernières années : les membres actuels Christophe Biernacki, Cristian Preda, Guillemette Marot, Sophie Dabo, Benjamin Guedj, Serge Iovleff, Hemant Tyagi, Alain Celisse, Wilfried Heyse, Florent Dewez, Yaroslav Averianov, Axel Potier, Issam Moindji, Guillaume Braun, Julien Vandromme, ... Mais aussi les anciens membres ou membres extérieurs de l'équipe : Pascal Germain, Julien Jacques, Arthur Tarlpaert, Adrien Ehrhardt, Quentin Grimonprez, Vincent Kubicki, Matthieu Marbac, Anne-Lise Bedenel, Maxime Baelde, Etienne Gofinnet, Maxime Brunin, Faicel Chamrouki, Philippe Heinrich, Bhargav Srinivasa, Parmeet Bhatia, Florence Loingeville, Clément Théry, ... et surement de nombreux autres membres que j'oublie. Je voudrais aussi remercier l'ensemble des assistantes d'équipes de MODAL pour leur disponibilité et leur réactivité : Sandrine Meilen, Corinne Jamroz et Anne Rejl.

Je remercie l'ensemble des membres de l'équipe METRICS, et en particulier Alexandre Caron et Benoît Dervaux qui m'ont initié à la question des problèmes d'utilisabilité, Alain Duhamel pour son accueil au moment de mon recrutement et son soutien au cours des années. Je remercie aussi Génia Babikina et Cyrielle Dumont pour les nombreux échanges que nous avons eus, et pour les projets de collaboration en cours. Je remercie aussi Grégoire Ficheur, Emmanuel Chazard et Jean-Basptiste Beuscart pour leur dynamisme et les échanges riches autour des projets d'analyse du parcours patient à l'hôpital qui offrent de nombreuses perspectives de recherche. Je remercie aussi Mohamed Lemdani, Michaël Génin, Julien Hamonier, Camille Ternynck, Benjamin Guinhouya, Antoine Lamer ... pour les échanges que nous avons pu avoir. Je remercie Renaud Perichon et Mélanie Steffe pour leur gentillesse et leur support technique et administratif.

Je remercie bien sûr l'ensemble des membres de l'IUT C, et en particulier l'ensemble des membres du département STID. C'est un vrai plaisir pour moi d'appartenir à ce département dans lequel je parviens à concilier mes activités d'enseignement et de recherche. Je remercie en particulier Fatma Bouali et Larbi Aït Hennani pour la confiance qu'ils m'ont accordé depuis mon recrutement. Je remercie Valérie Duhamel pour sa gestion admirable du département STID et son engagement total auprès des étudiants. Je tiens aussi à remercier l'ensemble de mes collègues actuels au sein du département : Saïd Serbouti, Fatima Belkouch, Joséphine Combes, Caroline Herrmann, Marie-Christine Desmaret, Carole Sieradski, Jenna Boller, Khalid Gaber, Emmanuelle Hugonnard, mais aussi mes anciens collègues : Marie-Christine Mourier, Sylvia Canonne, Emmanuel N'Goe, ... Je remercie aussi l'ensemble des secrétaires du département STID/ LP SID pour leur professionalisme et leur dynamisme : Marie-Christine Demeester, Rahma Benslimane et Elodie Dillies. Je

remercie aussi Eric Denneulin, Karim El Bachiri et Pascal Bauch pour leur support informatique.

Je remercie l'ensemble de mes collègues d'autres départements STID avec qui j'ai eu l'occasion d'échanger en réunions pédagogiques STID, ACD STID, ou session STID. Je remercie en particulier : Chloé Friguet, Frédérique Letuet, Antoine Roland, Elisabeth Le Saux, Delphine Blanke, Fançois-Xavier Jollois, Florence Muri, Jean-Michel Poggi, ...

Je remercie Gilles Celeux et Gérard Govaert pour les riches discussions que nous avons eues pendant ma thèse et qui habitent encore mes réflexions actuelles. Je remercie aussi l'ensemble des enseignants et maîtres de stage avec qui j'ai fait mes premiers pas en statistique : Stéphane Robin, Olivier David, Sylvain Billiard, Emilie Lebarbier, Tristan Mary-Huard, Jean-Jacques Daudin, Avner Bar-Hen, Gilbert Saporta, Pascal Massart, Elisabeth Gassiat, ... Je remercie aussi l'ensemble de mes co-auteurs, et j'espère aussi que les personnes que j'ai omis de remercier ne m'en tiendront pas rigueur.

Je remercie aussi le service des affaires doctorales de l'Université et en particulier Céline Delohen, Thi Nguyen secrétaire de l'école doctorale des sciences pour l'ingénieur, les membres de la commission de thèse en mathématiques, Alexi Virelizier et Serge Dachian, Patrick Popescu-Pampu responsable de l'ED SPI pour les mathématiques, ainsi que Ludovic Macaire directeur de l'ED SPI, pour leur aide dans ce processus compliqué qu'est l'HDR.

Enfin, le dernier mais non le moindre, je remercie mon épouse Aurélie et nos trois enfants Timéo, Charlie et Léonie pour leur soutien, c'est un plaisir au quotidien de partager chaque instant avec eux (même si ce n'est pas toujours de tout repos ...). Je remercie bien sûr mes parents pour l'éducation qu'ils m'ont transmise, ma sœur, et l'ensemble de ma famille. J'ai aussi une pensée particulière pour mes grands-parents.

À Aurélie, Timéo, Charlie et Léonie.

Contents

1 Introduction **1**
- 1.1 Statistical education . . . 1
- 1.2 Synthesis of my contributions . . . 2
  - 1.2.1 Summary of my contributions . . . 2
  - 1.2.2 Main collaborations . . . 3
  - 1.2.3 Scientific production fields . . . 3
  - 1.2.4 Packages related to my work . . . 3
  - 1.2.5 Publications . . . 5
- 1.3 Outline of the manuscript . . . 6
- 1.4 Overview of Part I . . . 7
- 1.5 Overview of Part II . . . 10

I Contribution to model-based clustering **15**

2 Introduction of model-based clustering **17**
- 2.1 Introduction to mixture models . . . 18
  - 2.1.1 Mixture density . . . 18
  - 2.1.2 Latent partition . . . 19
- 2.2 Parameters estimation . . . 20
  - 2.2.1 Maximum likelihood through the EM algorithm . . . 20
  - 2.2.2 Bayesian estimation through Gibbs sampling . . . 21
- 2.3 Model selection . . . 23

3 Model for clustering of categorical and mixed data **29**
- 3.1 Introduction . . . 29
- 3.2 Conditional dependency per block . . . 30
  - 3.2.1 State of the art . . . 30
  - 3.2.2 Mixture of intermediate dependency (CCM) . . . 31
  - 3.2.3 Dependency per mode (CMM) . . . 38
- 3.3 Gaussian copulas for mixed data . . . 41
  - 3.3.1 State of the art for mixed data . . . 41
  - 3.3.2 Conclusion . . . 47
- 3.4 Conclusion . . . 48

4 Clustering of functional data **53**
- 4.1 Introduction . . . 54
- 4.2 Clustering categorical functional data . . . 54
  - 4.2.1 Introduction to categorical functional data . . . 54
  - 4.2.2 Extension of multiple correspondence analysis . . . 55

4.2.3 Mixture of Markov processes . . . 57
4.2.4 Conclusion and perspectives . . . 59
4.3 Clustering of spatial functional data . . . 60
4.3.1 Introduction . . . 60
4.3.2 Model-based clustering for spatial functional data . . . 61
4.3.3 Application . . . 64
4.4 Conclusion and perspectives . . . 67

5 Multiple partition clustering 73
5.1 Introduction . . . 73
5.2 Multiple partition by blocks of variables . . . 74
5.2.1 Introduction . . . 74
5.2.2 Multiple partitions mixture model . . . 76
5.2.3 Comments . . . 77
5.2.4 Inference of the model and model selection . . . 78
5.2.5 Illustration on real data . . . 81
5.2.6 Conclusion . . . 86
5.3 Multiple partition by linear projections . . . 87
5.3.1 Introduction . . . 87
5.3.2 Multi-Partition Subspace Mixture Model . . . 89
5.3.3 Estimation of the Parameters of the Model . . . 91
5.3.4 Experiments on real data . . . 92
5.3.5 Conclusions . . . 93
5.4 Conclusion and perspectives . . . 94

6 Contribution to general issues in model-based clustering 103
6.1 Introduction . . . 104
6.2 Dealing with the label switching . . . 104
6.2.1 State of the art . . . 105
6.2.2 Reminding of the label switching problem . . . 105
6.2.3 Our proposal: posterior distribution restricted by the partition 106
6.2.4 Sampling according to a Gibbs algorithm . . . 107
6.2.5 Conclusion . . . 108
6.3 Missing data . . . 108
6.3.1 Gaussian mixture with missing data . . . 108
6.3.2 Distance estimation with missing data . . . 109
6.3.3 Degeneracy in mixture . . . 109
6.4 Visualization in mixture . . . 113
6.4.1 Introduction . . . 113
6.4.2 Possible mapping strategies . . . 114
6.4.3 New proposed method: controlling the distribution family . . . 116
6.4.4 Example . . . 117
6.4.5 Conclusion . . . 119
6.5 Conclusion . . . 119

II Contribution to some applications 127

7 Contribution to usability study 129
7.1 Introduction . . . 129
7.2 Bayesian modeling of the discovery matrix . . . 130
7.2.1 Motivation from a practical point of view . . . 130
7.2.2 Proposed solution . . . 131
7.2.3 Performance of the method . . . 134
7.2.4 Discussion . . . 134
7.3 Number of additional subjects needed . . . 134
7.3.1 Validation study in usability testing . . . 135
7.3.2 Modeling the economic consequences of undetected problems 135
7.4 Conclusion and perspectives . . . 137

8 Artificial intelligence for aviation 141
8.1 Introduction . . . 141
8.2 Flight data and aviation practices . . . 142
8.2.1 Flight data . . . 142
8.2.2 Aviation practices . . . 142
8.3 Inferring performance tables from flight data . . . 144
8.4 Optimizing flight trajectory through a functional approach . . . 146
8.4.1 General trajectory optimization problem . . . 146
8.4.2 A model for instantaneous consumption . . . 147
8.4.3 Decomposition of the trajectory in a basis and resulting quadratic optimization problem . . . 148
8.4.4 Satisfying constraints . . . 148
8.5 Conclusion . . . 149

9 Contributions to Credit scoring 153
9.1 Introduction . . . 154
9.2 Rational review of reject inference methods . . . 154
9.2.1 Problem presentation . . . 154
9.2.2 General parametric model . . . 155
9.2.3 General EM principle . . . 156
9.2.4 Several reintegration algorithm . . . 156
9.2.5 Concluding remarks . . . 158
9.3 Model-embedded feature quantization . . . 158
9.3.1 Motivation . . . 158
9.3.2 Quantization as a combinatorial challenge . . . 159
9.3.3 State of the art . . . 160
9.3.4 Proposed criterion . . . 161
9.3.5 A relaxation of the optimization problem . . . 162
9.3.6 A neural network-based estimation strategy . . . 165
9.3.7 Conclusion . . . 167

9.4 Conclusion and perspectives 167

III Research Project **173**

10 Research project **175**
10.1 Projects in the scope of model-based clustering 175
10.1.1 Multiple partition clustering dissemination 175
10.1.2 Predictive clustering 176
10.1.3 Clustering of recurrent event data and integration of covariates 176
10.2 Perspectives motivated by applications 177
10.2.1 Application to medical data 177
10.2.2 Application to the industry/retail 178
10.3 Reducing the gap between data and end-use 179

Appendices **183**

A Publications and scientific activities **185**
A.1 Publications 185
A.1.1 Articles published in peer-reviewed journals 185
A.1.2 Book chapter 187
A.1.3 Talks in conferences 187
A.2 Synthesis of scientific activities 190
A.2.1 Participation in research projects 190
A.2.2 Doctoral and scientific supervision 191
A.2.3 Participation in thesis juries 192
A.2.4 Participation in selection committees 192
A.2.5 Contracts with companies 193
A.3 Scientific responsibilities 193
A.3.1 Participation in scientific societies 193
A.3.2 Organization of national and international conferences 193
A.3.3 Reviewing activities 194

B Teaching activities and responsibilities **195**
B.1 Teaching activities 195
B.1.1 Teachings 195
B.1.2 Internship and project monitoring 196
B.2 Teaching responsibilities 196
B.2.1 Head of STID Department (2012-2015) 196
B.2.2 Promotional of STID department (2013-2014) 197
B.2.3 Head of Tutored Projects (2014-2015) 198
B.2.4 Director of Studies (2011-2012) 198

Chapter 1

Introduction

I am very glad to present my HDR dissertation. The work presented in this manuscript is the fruit of many formal and informal discussions on various occasions, such as daily work, conferences, and other collaborative projects. In this particular period, I wish that such a way of working will become fluent again, even if any way of working together will succeed. Thus beyond the diploma, I hope that it will give to me the opportunity to continue and initiate new collaborative projects on various hot statistical topics with potential impacts for the field of statistics and beyond statistics.

1.1 Statistical education

I have always been interested in the sciences. More precisely mathematics and biology. After two years of preparatory classes I had the opportunity to enter into the AgroParisTech (2003-2006), an engineering school focused on agronomy, the food industry, and ecology. This was for me a great experience, while focusing on biological aspects it also has a strong statistical content from the agronomical/ecology perspective.
My first true experience of statistical modeling came in my first year of AgroParisTech when I followed the module "randomness modeling in biology", this was for me a very exciting experience. It was already mixing some important tools that are part of my daily work such as probability, computer programming, simulations. In my second year, I enjoyed particularly the "linear model" teaching, the module "micro-arrays and bioinformatic" and the module "neurons and models". This convinced me to make three small internships during this year, one with Olivier David at INRA of Jouy-en-Josas on the temporal variations of allelic frequencies, one with Stéphane Robin at INRA of Paris on pattern detection in RNA sequences, and the third one with Sylvain Billiard at the ecology Lab of the University of Lille on detection of isolation by distance. All these internships convinced me to follow the Master 2 in Probability in Orsay as the third year of AgroParisTech. This helped me to deepen my statistical knowledge from a theoretical point of view. I had the chance to follow Gilles Celeux teaching on model-based clustering which I found very interesting, such tools as mixture models and EM algorithm were fascinating for me, where they enable to "model the heterogeneity in the data". Then, I had the pleasure to make my M2 internship with Christophe Biernacki at the University of Lille to investigate the question of semi-supervised learning using mixture models.

Then I had the chance to follow with a Ph.D. thesis on this subject (2006-2009) under to supervision of Christophe Biernacki, Gilles Celeux, and Gérard Govaert. I then obtained my actual position of Assistant Professor at the University of Lille, at the IUT C to teach statistics to undergraduate students specialized in statistics and business intelligence (2010-today). Since this date, I am affiliated with the now ULR 2694 METRIC teams of the University of Lille focused on evaluating health technologies and medical practices. I am also affiliated with the MODAL (MOdels for Data Analysis and Learning) research team of Inria Lille where I work among others on designing mixture models to deal with complex multivariate and heterogeneous data. My actual position allows me to encounter applied problems from several points of view (follow students performing their internships in companies, research contract between Inria and companies, and medical with the METRICS team). Belonging to both MODAL and METRICS teams allows me to still develop model-based clustering approaches from a general way and also to investigate problems applied to medicine.

1.2 Synthesis of my contributions

1.2.1 Summary of my contributions

My research work is in the field of applied statistics and more particularly in the field of classification (supervised, unsupervised, and semi-supervised), considered from the perspective of probabilistic models. In this context, I have been interested in several issues. Firstly, in the continuity of my thesis work on semi-supervised learning, I have been interested in the proposal of model choice criteria (Vandewalle, Biernacki, et al., 2013), the use of probabilistic models for distance estimation (Eirola et al., 2014) and the proposal of a solution to the label switching problem in the mixtures setting (Biernacki and Vandewalle, 2011). Then, I have been interested in taking into account the dependency between categorical variables to propose multivariate parametric distributions necessary for the clustering of categorical data (Marbac, Biernacki, and Vandewalle, 2015; Marbac, Biernacki, and Vandewalle, 2016). In this context, we have also proposed a copula-based model to take into account the dependence between continuous, binary, and ordinal variables (Marbac, Biernacki, and Vandewalle, 2017). I have been interested in the issue of multiple partition classification (*i.e.* when several latent class variables are considered) for which we have proposed two models and their associated estimation and model choice procedures (Marbac and Vandewalle, 2019; Vandewalle, 2020). I have also been interested in proposing a generic method to visualize the output of a mixture (Biernacki, Marbac, and Vandewalle, 2020).
My methodological research has also been fed by applied problems. On the one hand in the medical field, such as the implementation of mixture models for the classification of the patients' path at the hospital (Dhaenens et al., 2018), the use of high-dimensional regression models in proteomics (Cuvelliez et al., 2019), or the prediction of the number of problems in the field of usability (Vandewalle, Caron, et

al., 2020). On the other hand, I have been interested in the industrial and retail field through the proposal of automatic quantization models in credit scoring (Ehrhardt et al., 2018) and the use of machine learning in the field of aviation (Dewez, Guedj, and Vandewalle, 2020).

1.2.2 Main collaborations

Most of my collaborators are members of the MODAL team of Inria and of the METRICS team of the University of Lille. During my Ph.D. thesis, I have worked with Christophe Biernacki, Gilles Celeux, and Gérard Goveart. Since my Ph.D. defense, I still work actively with Christophe Biernacki (MODAL) on various mixture topics. We have supervised together the Ph.D. thesis of Matthieu Marbac, and I still work with Matthieu Marbac (now ENSAI) on multiple partition and predictive clustering. I have also supervised Adrien Ehrhardt's Ph.D. thesis with Christophe Biernacki (MODAL) and Phillipe Heinrich (Laboratoire Paul Painlevé, University of Lille) on credit scoring. For several years I also work with Cristian Preda (MODAL) and Sophie Dabo (MODAL) on functional data. I also work with Guillemette Marot (MODAL/METRICS), Christophe Bauters (INSERM Lille), and Florence Pinet (INSERM Lille) on omics data and I am co-supervising with Christophe Bauters and Guillemette Marot the Ph.D. thesis of Wilfried Heyse on taking into account time-varying high dimensional proteomic data. For a few years, I work with Alexandre Caron (METRICS) and Benoit Dervaux (METRICS) on usability issues on medical devices and their economical evaluation. Since the starting of the PERF-AI project (2018), I work with Benjamin Guedj (MODAL) and with Florent Dewez (Post-doc MODAL) on machine learning applied to aviation. I have also work in progress with Genia Babikina (METRICS) and Cyrielle Dumont (METRICS) on joint modeling and clustering recurrent event data

1.2.3 Scientific production fields

Most of my scientific production is on model-based clustering, with a particular focus on the model proposal and related model choice perspective. This can be a central point to select the relevant structure of the model and the number of clusters. I have considered the question of model choice from both using asymptotic criteria, but also by considering a Bayesian setting. Some of my scientific production is also motivated by applications in various fields. Table 1.1 gives some summary of these various topics in my publications.

1.2.4 Packages related to my work

There are some package that I have contributed in

- MGMM : `http://r-forge.r-project.org/projects/mgmm`, for multiple partition clustering

Article	Clustering	Model choice	Bayesian setting	Motivated by application
Dewez, Guedj, and Vandewalle (2020) *DCE*				✓
Vandewalle, Caron, et al. (2020) *BMC Medical Research Methodology*			✓	✓
Biernacki, Marbac, and Vandewalle (2020) *JoC*	✓			
Vandewalle (2020) *Mathematics*	✓			
Vandewalle, Preda, and Dabo (2020) *Book chapter*	✓			
Cuvelliez et al. (2019) *Scientific reports*				✓
Marbac and Vandewalle (2019) *CSDA*	✓	✓		
Dhaenens et al. (2018) *IRBM*	✓			✓
Marbac, Biernacki, and Vandewalle (2017) *Comm. Statist. Theory Methods*	✓		✓	
Marbac, Biernacki, and Vandewalle (2016) *ADAC*	✓	✓	✓	
Marbac, Biernacki, and Vandewalle (2015) *JoC*	✓			
Eirola et al. (2014) *Neurocomputing*	✓			
Vandewalle, Biernacki, et al. (2013) *CSDA*	✓	✓		
Biernacki and Vandewalle (2011) *AIP Conference Proceedings*	✓			
Vandewalle (2009) *MODULAD*	✓	✓		

Table 1.1: Cross-table between my articles and some main topics

- ClusVis : `https://cran.r-project.org/web/packages/ClusVis`, generic visualization of the output of any mixture
- GLMDISC : `https://cran.r-project.org/web/packages/glmdisc` (R) + `https://pypi.org/project/glmdisc` (Python), automatic discretization of variable for logisitc regression
- Clustericat : `https://rdrr.io/rforge/Clustericat`, clustering of categorical variables based on intermediate dependency models
- CoMode : `https://rdrr.io/rforge/CoModes`, clustering of categorical variables based on dependence per modes
- cfda : `https://rdrr.io/github/modal-inria/cfda`, analysis of categorical functional data through functional MCA

- pyrotor : https://github.com/bguedj/pyrotor (available soon), optimization of trajectories on some basis
- useval : https://github.com/alexandre-caron/useval (available soon), risk assesement based on the discovery matrix

1.2.5 Publications

Post-thesis articles

1. V. Vandewalle (2020). "Multi-Partitions Subspace Clustering". In: *Mathematics* 8.4, p. 597
2. C. Biernacki, M. Marbac, and V. Vandewalle (2020). "Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering". In: *Journal of Classification*. DOI: 10.1007/s00357-020-09369-y. URL: https://doi.org/10.1007/s00357-020-09369-y
3. F. Dewez, B. Guedj, and V. Vandewalle (2020). "From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning". In: *Data-Centric Engineering* 1, e11. DOI: 10.1017/dce.2020.12
4. V. Vandewalle, A. Caron, C. Delettrez, R. Périchon, S. Pelayo, A. Duhamel, and B. Dervaux (2020). "Estimating the number of usability problems affecting medical devices: modelling the discovery matrix". In: *BMC Medical Research Methodology* 20.234
5. V. Vandewalle, C. Preda, and S. Dabo (2020). "Clustering spatial functional data". In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons
6. M. Cuvelliez, V. Vandewalle, M. Brunin, O. Beseme, A. Hulot, P. de Groote, P. Amouyel, C. Bauters, G. Marot, and F. Pinet (2019). "Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction". In: *Scientific reports* 9.1, pp. 1–12
7. M. Marbac and V. Vandewalle (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179
8. C. Dhaenens, J. Jacques, V. Vandewalle, M. Vandromme, E. Chazard, C. Preda, A. Amarioarei, P. Chaiwuttisak, C. Cozma, G. Ficheur, et al. (2018). "ClinMine: Optimizing the management of patients in hospital". In: *IRBM* 39.2, pp. 83–92
9. M. Marbac, C. Biernacki, and V. Vandewalle (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656

10. M. Marbac, C. Biernacki, and V. Vandewalle (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207

11. M. Marbac, C. Biernacki, and V. Vandewalle (2015). "Model-based clustering for conditionally correlated categorical data". In: *Journal of Classification* 32.2, pp. 145–175

12. E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki (2014). "Mixture of Gaussians for distance estimation with missing data". In: *Neurocomputing* 131, pp. 32–42

Conference with proceedings

1. C. Biernacki and V. Vandewalle (2011). "Label switching in mixtures". In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401

Other articles and conference talks are detailed in Appendix A.

1.3 Outline of the manuscript

My thesis work led to the article Vandewalle, Biernacki, et al. (2013) proposing a new criterion for model selection, the AIC_{cond} criterion in semi-supervised classification. It measures the predictive capacity of the generative model, based on an approximation of predictive deviance, and is an interesting alternative to the criteria usually used such as AIC or BIC. When using BIC or AIC in semi-supervised learning the selected model tends to focus on the unlabeled data thus leading to potentially poor performance from the predictive point of view.
After my thesis, I have continued to work on mixture models, on the one hand by proposing mixture models able to take into account various kinds of variables, or to consider several latent partitions, on the other hand by studying mixture models from a general point of view. I have also worked on several issues coming from applications.
Thus the manuscript is divided into three main parts. Part I presents my contributions to model-based clustering, while Part II presents my contributions motivated by several applications. Part III gives the perspectives of my research work. In appendices, I give additional information about my research and teaching activities. The bibliography of each chapter is given separately at the end of the chapter.
Part I of the manuscript presents my contribution to model-based clustering. It is composed of five chapters. Chapter 2 presents a short introduction to model model-based clustering, it aims at introducing the main notations that will be used all along Part I. Chapter 3 presents my contribution to the proposal of models for the clustering of categorical and mixed data. Chapter 4 presents my contribution to the proposal of models for clustering functional data. Chapter 5 presents my

contribution to the proposal of models considering several latent partitions. Chapter 6 presents my contribution to general issues in model-based clustering such as dealing with the label switching problem, with missing data, and visualization in mixture.
Part II of the manuscript presents methodological contribution motivated by three applications. Chapter 7 presents a contribution to the statistical analysis of usability studies through the modeling of the discovery matrix and the assessment of economic consequences. Chapter 8 presents a contribution to the optimization of flight trajectory based on flight data, through the use of regression models and optimization in a functional space. Chapter 9 presents a contribution to credit scoring principally by proposing a method for embedding feature quantization inside the predictive model fitting.

1.4 Overview of Part I

Chapter 2: Introduction to model-based clustering

I tried as possible to make notations homogeneous along Part I, even if it is hard to keep the same notations for the result of several articles. Thus, Chapter 2 presents a short introduction to model model-based clustering, it aims at introducing the main notations that will be used all along Part I.

Chapter 3: Model for clustering categorical and mixed data

From 2011 to 2014, I co-supervised Matthieu Marbac's thesis (Marbac-Lourdelle, 2014) with Christophe Biernacki. In this thesis we were interested in taking into account correlations between categorical data on the one hand and between mixed data, on the other hand, this conditionally to the class in the clustering framework. Indeed, for these data, the hypothesis of conditional independence is often made, which can lead to a bad estimation of the clusters in particular through an overestimation of the number of clusters. The articles Marbac, Biernacki, and Vandewalle (2015) and Marbac, Biernacki, and Vandewalle (2016) propose to take into account the dependencies between categorical variables by making blocks of variables and by proposing a parsimonious model on the distribution of these block of variables. The article Marbac, Biernacki, and Vandewalle (2017) proposes to take into account the dependencies between variables of different natures using copulas.
In the article Marbac, Biernacki, and Vandewalle (2015) we have proposed an intermediate dependency model (a mixture between a total dependency distribution and an independence one) to model the distribution of a group of categorical variables given the class. We proposed an algorithm for estimating parameters with a fixed block structure, as well as an algorithm for stochastic search of the best block structure of the variables, the model selection is performed using a BIC type criterion.
In the article Marbac, Biernacki, and Vandewalle (2016) we propose a significantly different model for modeling the distribution of a block of class-conditional cate-

gorical variables. This model assumes that the distribution of a crossing of the categorical variables can be modeled by a set of modal crossings, with identical probabilities for all the other crossings. In this framework, the block search question arises again, which is again solved using a stochastic algorithm. A contribution of this paper is also the proposal of an integrated likelihood criterion for the choice of the number of modes, whose performance exceeds that of its commonly used BIC approximation.
In the article Marbac, Biernacki, and Vandewalle (2017) we have proposed a model in the mixed data setting (quantitative, discrete, ordinal). In this framework, the assumption of conditional independence is usual since it is difficult to take into account dependencies between variables of different kinds. In this framework, we propose to take into account the dependencies between variables conditionally to the class thanks to a copula. The margins are modeled by standard distributions, while the copula allows taking into account the dependencies between variables of different kinds. Here we place ourselves in a Bayesian framework for the estimation of the parameters. We proposed a Gibbs algorithm to find the mode of the posterior distribution.
Chapter 3 presents an overview of these articles.

Chapter 4: Clustering functional data

From 2013 to 2017, I took part in the ANR ClinMine: "Optimisation of Patient Care at the Hospital" project. In this ANR project, I worked with Cristian Preda on the implementation of mixture models for the clustering of categorical functional data. Indeed, we had a lot of hospital data on categorical variables over time (the type of pathology, state of mail processing, ...), but few models exist for their classification, especially when the period over which the observation is carried out varies from one individual to another. The originality of our approach is to propose a mixture of Markov processes with continuous time and discrete state space. This model facilitates the management of categorical functional data and gives an interpretation of groups in terms of the distribution of time spent in each state and the probabilities of transition from one state to another. This work gave rise to conference presentations (Vandewalle, Cozma, and Preda, 2015; Vandewalle and Preda, 2016; Vandewalle and Preda, 2017), as well as to the article Dhaenens et al. (2018). I continue to work with Cristian Preda on this issue, where we are currently working on the identifiability of the proposed mixture of Markov processes.
In the same scope, I have worked with Cristian Preda and Quentin Grimonpret on using the extension of the functional PCA to the categorical functional data setting. I have contributed to the R package `cfda` (`https://github.com/modal-inria/cfda/`), and we have submitted an article related to this work at the Journal of Statistical Software (Preda, Grimonprez, and Vandewalle, 2020).
In the book chapter Vandewalle, Preda, and Dabo (2020) we were interested in the unsupervised classification of spatial functional data. In this framework, we proposed a model taking into account the spatial information in the clustering.

This spatial information is taken into account as logistic weights in the prior group membership probabilities. This model then allows the clustering of curves while taking into account the spatial point of view.
Chapter 4 details these works.

Chapter 5: Multiple partition clustering

I am also interested in the issue of unsupervised classification when several latent class variables are considered (multiple partition clustering). Indeed, assuming that all heterogeneity in the data can be explained by a single variable is a very strong assumption, and it may be useful to consider that several blocks (or linear combinations) of variables can provide different partitions of individuals. This may reveal new lines of analysis in the data set. In this framework, we have proposed two approaches.
The first one assumes the existence of several groups of variables, each one leading to a different partition of the individuals. This work has been published in Marbac and Vandewalle (2019). The approach has the interest to propose an efficient algorithm allowing the search for blocks of variables as well as the estimation of the different partitions of the individuals. The key assumption is the independence of variables given the cluster in each block. This assumption allows at each step to reassign each variable to the most relevant block of variables at a low computation cost. This model makes it possible to classify the variables into blocks, each producing a specific grouping of individuals.
A second model assumes the existence of several classifying projections in the data and has been recently published (Vandewalle, 2020). For this approach, I have proposed a model and an estimation algorithm. The main idea is to assume that there are different linear combinations of variables in the data, each one explained by a different latent class variable. Thus the method allows obtaining different classifying projections and the associated partitions. The proposed approach remains limited to cases where the number of variables is less than the number of individuals but has the advantage of being invariant by any linear bijective transformation of the variables.
Chapter 5 details these two articles.

Chapter 6: Contribution to general issues in model-based clustering

In Chapter 6, I detail some contributions to general issues in model-based clustering. In this scope, I have been interested in the issue of label switching in mixtures, a problem that is important in the context of inferring the parameters of a mixture model in the Bayesian framework. In this framework, we proposed an approach based on latent partitioning which partially removes this problem. This work has been published in conference proceedings (Biernacki and Vandewalle, 2011).
I have also been interested in the problem of estimating distances when some variables are missing. In Eirola et al. (2014) a multivariate Gaussian mixture model,

allows us to easily take into account incomplete data for distance estimation. These estimated distances can then be used as input data for distance-based predictive models (RBF, SVM, ...). Study the EM behavior when considering missing data has raised the question of the study of the degeneracy (convergence to a degenerate solution) in this particular case, where we have observed that it is particularly slow. To avoid this phenomenon we have proposed a modified version of the EM algorithm (but untractable) and also some relevant approximations presented in conferences (Vandewalle and Biernacki, 2015; Biernacki, Castellan, et al., 2016) but are still a work in progress.
Finally, the multiple partition approach proposed in Vandewalle (2020) already makes it possible to consider simultaneously the problems of classification and visualization due to the study of classifying linear projections. However, most classifications based on mixed models do not easily allow data visualization. This is for instance the case when the variables considered are of heterogeneous natures. In the article Biernacki, Marbac, and Vandewalle (2020) we propose an approach that produces the closest Gaussian visualization of any estimated mixture distribution. Contrary to the usual paradigm which imposes the mapping family for visualization (typically linear transformations), we propose here to constrain the arrival distribution to be Gaussian. This proposal can be seen as a paradigm shift in the field of visualization.
More details about this work can be found in Chapter 6.

1.5 Overview of Part II

Chapter 7: Contribution to the analysis of usability study in medicine

Since 2018, I have been working with Alexandre Caron and Benoît Dervaux, both members of the METRICS team, on issues of estimating the number of problems and the value of information in the field of usability. Based on usability study of a medical device the objective is to determine the number of possible problems linked to the use of a medical device (*e.g.* insulin pump) as well as their respective occurrence probabilities. Estimating this number and the different probabilities is essential to determine whether or not an additional usability study should be conducted, and to determine the number of users to be included in this study to maximize the expected benefits.
The discovery process can be modeled by a binary matrix, a matrix whose number of columns depends on the number of defects discovered by users. In this framework, we have proposed probabilistic modeling of this matrix. We have included this modeling in a Bayesian framework where the number of problems and the probabilities of discovery are considered as random variables. In this framework, the article Vandewalle, Caron, et al. (2020) shows the interest of the approach compared to the approaches proposed in the state of the art in usability. The approach beyond point estimation also makes it possible to obtain the distribution of the number of problems and their respective probabilities given the discovery matrix.

The proposed model allows us to implement an approach aiming at measuring the value of additional information in relation to the discovery process. In this framework, we are currently writing a second paper and developing an R package that should help practitioners in the field of usability to better dimension their studies. More details about this work can be found in Chapter 7.

Chapter 8: Application of machine learning in aviation

Since November 2018, I have been participating in the European PERF-AI project (European PERF-AI project: Enhance Aircraft Performance and Optimization through the utilization of Artificial Intelligence) in partnership with the company Safety Line. In particular, using data collected during flights involves developing Machine Learning models to optimize the aircraft's trajectory concerning fuel consumption, for example. In this context, the article Dewez, Guedj, and Vandewalle (2020) explains how, using flight recording data, it is possible to implement learning models on variables that have not been directly observed, and in particular to predict the drag and lift coefficients as a function of the angle and speed of the aircraft.
A second article is being written about the optimization of the aircraft's trajectory based on a consumption model learned from the data. The originality of the approach consists in decomposing the trajectory on a functional basis, and thus carrying out the optimization on the coefficients of the decomposition on this basis, rather than approaching the problem from the angle of optimal control. Furthermore, to guarantee compliance with aeronautical constraints, we have proposed an approach penalized by a deviation term from reference flights. A generic Python module to solve such optimization problems is being developed in conjunction with the proposed approach.
More details about this work can be found Chapter 8.

Chapter 9: Application in credit scoring

From April 2016 to September 2019, I have co-supervised with Christophe Biernacki and Philippe Heinrich, Adrien Ehrhardt's CIFRE thesis at CACF in the field of supervised classification applied to credit scoring. In this framework, we have proposed a reinterpretation of the different methods for reintegrating rejected clients. This work has been presented at the SFdS days and an article has been submitted and another one is being finalized. Within the framework of this thesis, we also worked on a generative model of automatic discretization of variables. This model makes it possible to reduce considerably the manual pre-processing necessary for the design of a score, and an article is being finalized on this subject, and the associated R package (glmdisc) is already available on the CRAN.
More details about this work can be found Chapter 9.

Bibliography

Biernacki, C., Castellan, G., Chretien, S., Guedj, B., and Vandewalle, V. (2016). "Pitfalls in Mixtures from the Clustering Angle". In: *Working Group on Model-Based Clustering Summer Session*. Paris, France. URL: https://hal.archives-ouvertes.fr/hal-01419755.

Biernacki, C., Marbac, M., and Vandewalle, V. (2020). "Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering". In: *Journal of Classification*. DOI: 10.1007/s00357-020-09369-y. URL: https://doi.org/10.1007/s00357-020-09369-y.

Biernacki, C. and Vandewalle, V. (2011). "Label switching in mixtures". In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401.

Cuvelliez, M., Vandewalle, V., Brunin, M., Beseme, O., Hulot, A., Groote, P. de, Amouyel, P., Bauters, C., Marot, G., and Pinet, F. (2019). "Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction". In: *Scientific reports* 9.1, pp. 1–12.

Dewez, F., Guedj, B., and Vandewalle, V. (2020). "From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning". In: *Data-Centric Engineering* 1, e11. DOI: 10.1017/dce.2020.12.

Dhaenens, C., Jacques, J., Vandewalle, V., Vandromme, M., Chazard, E., Preda, C., Amarioarei, A., Chaiwuttisak, P., Cozma, C., Ficheur, G., et al. (2018). "ClinMine: Optimizing the management of patients in hospital". In: *IRBM* 39.2, pp. 83–92.

Ehrhardt, A., Vandewalle, V., Biernacki, C., and Heinrich, P. (2018). "Supervised multivariate discretization and levels merging for logistic regression". In: *23rd International Conference on Computational Statistics*. Iasi, Romania. URL: https://hal.archives-ouvertes.fr/hal-01949128.

Eirola, E., Lendasse, A., Vandewalle, V., and Biernacki, C. (2014). "Mixture of Gaussians for distance estimation with missing data". In: *Neurocomputing* 131, pp. 32–42.

Marbac, M., Biernacki, C., and Vandewalle, V. (2015). "Model-based clustering for conditionally correlated categorical data". In: *Journal of Classification* 32.2, pp. 145–175.

— (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207.

— (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656.

Marbac, M. and Vandewalle, V. (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179.

Marbac-Lourdelle, M. (2014). "Modèles de mélange pour la classification non supervisée de données qualitatives et mixtes". 2014LIL10068. PhD thesis. URL: http://www.theses.fr/2014LIL10068/document.

Preda, C., Grimonprez, Q., and Vandewalle, V. (2020). "cfda: an R Package for Categorical Functional Data Analysis". working paper or preprint. URL: https://hal.inria.fr/hal-02973094.

Vandewalle, V. (2009). "Les modèles de mélange, un outil utile pour la classification semi-supervisée." In: *Monde des Util. Anal. Données* 40, pp. 121–145.

— (2020). "Multi-Partitions Subspace Clustering". In: *Mathematics* 8.4, p. 597.

Vandewalle, V. and Biernacki, C. (2015). "An efficient SEM algorithm for Gaussian Mixtures with missing data". In: *8th International Conference of the ERCIM WG on Computational and Methodological Statistics*. Londres, United Kingdom. URL: https://hal.inria.fr/hal-01242588.

Vandewalle, V., Biernacki, C., Celeux, G., and Govaert, G. (2013). "A predictive deviance criterion for selecting a generative model in semi-supervised classification". In: *Computational Statistics & Data Analysis* 64, pp. 220–236.

Vandewalle, V., Caron, A., Delettrez, C., Périchon, R., Pelayo, S., Duhamel, A., and Dervaux, B. (2020). "Estimating the number of usability problems affecting medical devices: modelling the discovery matrix". In: *BMC Medical Research Methodology* 20.234.

Vandewalle, V., Cozma, C., and Preda, C. (2015). "Clustering categorical functional data Application to medical discharge letters". 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, Dec 2015, Londres, United Kingdom.

Vandewalle, V. and Preda, C. (2016). "Clustering categorical functional data Application to medical discharge letters Medical discharge letters". Working Group on Model-Based Clustering Summer Session: Paris, July 17-23, 2016 Paris, France.

— (2017). "Clustering categorical functional data: Application to medical discharge letters". 20th conference of the society of probability and statistics of Roumania, Brasov (Roumania), April 28 (invité).

Vandewalle, V., Preda, C., and Dabo, S. (2020). "Clustering spatial functional data". In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons.

Part I

Contribution to model-based clustering

Chapter 2

Introduction of model-based clustering

Contents

2.1 Introduction to mixture models 18

2.1.1 Mixture density 18

2.1.2 Latent partition 19

2.2 Parameters estimation 20

2.2.1 Maximum likelihood through the EM algorithm 20

2.2.2 Bayesian estimation through Gibbs sampling 21

2.3 Model selection 23

Clustering (Jajuga, Sokołowski, and Bock, 2002) serves to summarize (typically large) data sets by assessing a partition among observations, the latter being thus summarized by (typically few) characteristic classes. There exists a large number of clustering methods, one can distinguish between geometric methods, based on distances, and model-based clustering (MBC) methods, based on modeling of the data distribution as a finite mixture of distributions. The advantage of using MBC is to answer classical challenges by relying on theoretical statistics tools, *e.g.*, estimating the partition using an EM algorithm (Dempster, Laird, and Rubin, 1977), selecting the number of groups using information criteria such as BIC or ICL (Schwarz, 1978; Biernacki, Celeux, and Govaert, 2000), dealing with missing values among observations (Larose, 2015). Such an issue being tedious for geometric approaches which generally need to define ad-hoc criteria to perform such choices. Some geometric modeling can be re-interpreted as MBC such as the k-means algorithm which can be re-interpreted as an isotropic Gaussian mixture, estimated using the classification EM algorithm (CEM) (Celeux and Govaert, 1995). In this manuscript, I will not discuss further geometric approaches.

This chapter introduces the main notations and elements which will be used in the next chapters. It does not pretend exhaustiveness on MBC. In Section 2.1, I present the general notations used in the manuscript. In Section 2.2, I present the general issue of parameters estimation. In Section 2.3, I discuss the possible strategies for model selection to choose the model family or the number of clusters.

2.1 Introduction to mixture models

2.1.1 Mixture density

Data to cluster $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are composed of n observations $\boldsymbol{x}_i \in \mathcal{X}$, where $\mathcal{X}$ depends on the type of variable considered (for instance $\mathcal{X} = \mathbb{R}^d$ if d continuous variables are considered). Assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independent and identically distributed (i.i.d) realizations of a random variable $\boldsymbol{X}$. If $\boldsymbol{X}$ admits a density function denoted by p with respect to some reference measure on $\boldsymbol{X}$, then for all $\boldsymbol{x} \in \mathcal{X}$, it is assumed that $p(\boldsymbol{x})$ can be decomposed as a finite mixture of distributions

$$p(\boldsymbol{x}) = \sum_{k=1}^{g} \pi_k p_k(\boldsymbol{x}) \tag{2.1}$$

where g is the assumed number of cluster, π_k is the proportion of cluster k ($\pi_k > 0$ and $\sum_{k=1}^{g} \pi_k = 1$).
In parametric framework, we assume that $p_k(\cdot) = p(\cdot|\boldsymbol{\alpha}_k)$ where $p(\cdot|\boldsymbol{\alpha}_k)$ is a parametric density with parameter $\boldsymbol{\alpha}_k$. For instance in Gaussian setting $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\mu}_k$ the vector of means in cluster k and $\boldsymbol{\Sigma}_k$ the covariance matrix in cluster k. $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ groups the proportions, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ groups the class specific parameters, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ groups the model parameters. Equivalently we will also denote $\boldsymbol{\theta}_k = (\pi_k, \boldsymbol{\alpha}_k)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ depending of the context. We have

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}|\boldsymbol{\alpha}_k). \tag{2.2}$$

In the rest of manuscript the densities are interpreted according to its parameters. Moreover to keep homogeneity in the manuscript we keep $p(\cdot|\cdot)$ to separate data from parameters whatever frequentist or Bayesian setting is considered.
The mixture distribution assumes that the heterogeneity of the distribution of $\boldsymbol{X}$ can be explained by a finite number of homogeneous class-specific distributions. A cluster is defined as the set of individuals coming from the same component of the mixture. The choice of the parametric family $p(\cdot|\boldsymbol{\alpha}_k)$ defines what is considered as homogenous class-specific distribution since homogeneity is defined with respect to this family thus defining the shape of the clusters we are looking for.
Thus MBC (McLachlan and Peel, 2004; McNicholas, 2016; Biernacki, 2017) allows for the analysis of different types of data by "simply" adapting the cluster distribution $p(\cdot|\boldsymbol{\alpha}_k)$ see Banfield and Raftery (1993), Celeux and Govaert (1995), and McNicholas and Murphy (2008) for continuous data, Goodman (1974), Celeux and Govaert (1991), Gollini and Murphy (2014), and Marbac, Biernacki, and Vandewalle (2016) for categorical data, Kosmidis and Karlis (2015), McParland and Gormley (2016), Punzo and Ingrassia (2016), Marbac, Biernacki, and Vandewalle (2017), and Mazo (2017) for mixed data, Samé et al. (2011), Bouveyron and Jacques (2011), and Jacques and Preda (2014b) for functional data , (Daudin, Picard, and Robin, 2008; Zanghi, Ambroise, and Miele, 2008; Ambroise and Matias, 2012) for networks data.

Identifiability In order to identify each component as a cluster, it is needed that the parameter of the mixture distribution $p(\cdot|\boldsymbol{\theta})$ be identifiable. More precisely, the parameters of the model are identifiable if for two arbitrary parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, $p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}'), \quad \forall \boldsymbol{x} \in \mathcal{X} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}'$ up to a permutation of the classes (since clusters are defined up to a permutation). Among others Teicher (1963), Teicher (1967), and Yakowitz, Spragins, et al. (1968) have presented sufficient conditions of identifiability. A lot of parametric families are identifiable, however, some models such as mixtures of products of multinomial are known to be non-identifiable. This problem of identifiability can eventually only hold for a set of parameters of measure null, that is called generic identifiability (Allman, Matias, and Rhodes, 2009). Generic identifiability can be sufficient for the model to be useful in practice as for mixtures of products of multinomial distributions.

Remarks Non-parametric mixtures could also be considered (Benaglia, Chauveau, and Hunter, 2009) but are not presented here. Let also notice that when dealing with continuous functional data the density is not well defined, thus making it impossible to directly use standard mixtures. To solve this problem some solutions exist such as using the concept of surrogate density for functional data (Delaigle and Hall, 2010), or to decompose functional data on some basis of functions then applying standard clustering on the coefficients. For a review on functional data clustering see Jacques and Preda (2014a). This question is further discussed in Chapter 4.
The presentation here is limited to i.i.d data but could be completed by the case of networks data for instance (Daudin, Picard, and Robin, 2008; Zanghi, Ambroise, and Miele, 2008; Ambroise and Matias, 2012). Moreover one could consider block clustering models (Govaert and Nadif, 2013) for the simultaneous clustering of rows and columns.

2.1.2 Latent partition

From a clustering point of view it is important to introduce the partition $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ with $z_{ik} = 1$ if observation i belongs to cluster k and $z_{ik} = 0$ otherwise, $\mathbf{z} \in \mathcal{Z}$ with $\mathcal{Z}$ the partition space. Assume $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are n i.i.d realizations of $\boldsymbol{Z}$ the associated random variable with one individual, $\boldsymbol{Z}$ follows a multinomial distribution $\boldsymbol{Z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$.
The generative model assume that the joint distribution $(\boldsymbol{X}, \boldsymbol{Z})$ is generated as follows

- Sample $\boldsymbol{Z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$,
- Sample $\boldsymbol{X}|Z_k = 1 \sim p(\cdot|\alpha_k)$.

Thus obtaining mixture distribution for $\boldsymbol{X}$ as presented Equation (2.2).
In practice, the partition is not observed, and it is straightforward to compute the posterior of class membership denoted by $t_{ik}(\boldsymbol{\theta})$ based on the mixture model

$$t_{ik}(\boldsymbol{\theta}) = p(Z_{ik} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\pi_k p(\boldsymbol{x}; \boldsymbol{\alpha}_k)}{\sum_{k'=1}^{g} \pi_{k'} p(\boldsymbol{x}; \boldsymbol{\alpha}_{k'})}.$$

It is then possible to deduce the estimated partition by maximum a posteriori (MAP) applied separately for all i in $\{1, \dots, n\}$:

$$\hat{z}_{ik}(\boldsymbol{\theta}) = \begin{cases} 1 \text{ if } k = \arg\max_{k' \in \{1,\dots,g\}} t_{ik'}(\boldsymbol{\theta}), \\ 0 \text{ otherwise.} \end{cases}$$

2.2 Parameters estimation

The two most popular ways to perform parameters estimation in mixture models are either by maximum likelihood through the EM algorithm (Dempster, Laird, and Rubin, 1977) or in a Bayesian framework by considering a Gibbs sampling on the augmented data (Marin, Mengersen, and Robert, 2005). Many other estimation methods could be considered such as moment estimation but are not considered here since not used in the sequel of the manuscript. In this section, we consider that the number g of clusters is known.

2.2.1 Maximum likelihood through the EM algorithm

Likelihood Under the independence assumption, the log-likelihood can be written

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

Then the maximum likelihood is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{x}).$$

It is known to have good theoretical properties such as consistency (Wald, 1949; Redner and Walker, 1984). However, this maximum can be not defined, like in the heteroscedastic Gaussian where it is unbounded. Fortunately, in such a case, it is known that a root of the gradient of the likelihood is a consistent estimator.

EM algorithm The maximum likelihood cannot be obtained directly and require the use of an iterative algorithm. The most tractable solution is to use the EM algorithm (Expectation-Maximization) (Dempster, Laird, and Rubin, 1977). It is based on the missing data interpretation of the latent partition $\mathbf{z}$.
The completed likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \pi_k p(\boldsymbol{x}_i|\boldsymbol{\alpha}_k).$$

In this case, the maximum likelihood would be straightforward depending on the parametric model $p(\cdot|\boldsymbol{\alpha}_k)$.
The general principle of the algorithm is presented in Algorithm 1. For sake of simplicity lower case letters will be used either to designate realizations or random variables according to the context, as commonly used in the Bayesian setting. By linearity of the completed log-likelihood, the E step simply requires the

computation of $\mathbb{E}[Z_{ik}|\boldsymbol{x}_i;\boldsymbol{\theta}^{(r)}] = t_{ik}(\boldsymbol{\theta}^{(r)})$ which is straightforward. At the M step $\boldsymbol{\alpha}_k^{(r+1)} = \arg\max_{\boldsymbol{\alpha}_k} \sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(r)}) \ln p(\boldsymbol{x}_i|\boldsymbol{\alpha}_k)$, thus only requiring to solve a weighted maximum likelihood problem without mixture. Thus we obtain Algorithm 2. More refined strategy will be detailed in Chapter 3, where it is needed to adapt the algorithm in order perform the optimization over discrete parameters (block structure), while avoiding restarting the algorithm from scratch.

Algorithm 1 EM algorithm: formulation general formulation

start from $\boldsymbol{\theta}^{(0)}$
for $r = 0$ to $r_{\max} - 1$ **do**
  E step: $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}[\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})|\mathbf{x}; \boldsymbol{\theta}^{(r)}]$
  M step: $\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$
return $\boldsymbol{\theta}^{(r_{\max})}$

Algorithm 2 EM algorithm: formulation in mixture independent setting

start from $\boldsymbol{\theta}^{(0)}$
for $r = 0$ to $r_{\max} - 1$ **do**
  E step: Compute $t_{ik}(\boldsymbol{\theta}^{(r)})$
  M step: For all k in $\{1, \ldots, g\}$

$$\boldsymbol{\alpha}_k^{(r+1)} = \arg\max_{\boldsymbol{\alpha}_k} \sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(r)}) \ln p(\boldsymbol{x}_i|\boldsymbol{\alpha}_k) \quad \text{and} \quad \pi_k^{(r+1)} = \frac{\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(r)})}{n}.$$

return $\boldsymbol{\theta}^{(r_{\max})}$

The EM algorithm improves the likelihood at each step, in practice additional conditions can be added to stop the algorithm before $r_{\max}$. Let notice that the EM algorithm can converge to a local maximum, or may be trapped by a degenerated solution. Like any iterative algorithm in a non-convex setting it is sensitive to the starting value, and thus must be started with several different starting values. Moreover, its convergence is slow (linear convergence) compared with Newton-Raphson for instance. Many variants exist see for instance McLachlan and Krishnan (2008) in particular we can notice the Generalized EM algorithm which only improves the expectation of the completed likelihood at M step, the Stochastic EM (SEM), and the Classification EM (CEM) algorithms (Celeux and Govaert, 1992). Different strategies can be used to chain these algorithms to find the best solution.

2.2.2 Bayesian estimation through Gibbs sampling

In a Bayesian setting is possible to define some prior distribution on $\boldsymbol{\theta}$, this distribution is often factorized in the following way

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^g p(\boldsymbol{\alpha}_k),$$

assuming prior independence between parameters. A non-informative Jeffrey prior is often chosen for $p(\boldsymbol{\pi})$, and the prior distribution $p(\boldsymbol{\alpha}_k)$ is often chosen as conjugated distribution of $p(\boldsymbol{x}|\boldsymbol{\alpha}_k)$. However it possible to specify other prior distributions to take advantage of available prior information.

From a Bayesian setting one is interested in the posterior distribution of $\boldsymbol{\theta}|\mathbf{x}$. This posterior distribution has not closed form, but it is possible to sample from it using a Gibbs sample algorithm. The Gibbs sampler generally used is presented in Algorithm 3.

Algorithm 3 Gibbs sampling algorithm for mixture

start from $\boldsymbol{\theta}^{(0)}$
for $r = 0$ to $r_{\max}$ **do**
  for $i = 0$ to n **do**
    Sample $\boldsymbol{z}_i^{(r+1)}$ from $\boldsymbol{z}_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(r)}$
  Sample $\boldsymbol{\pi}^{(r+1)}$ from $\boldsymbol{\pi}|\mathbf{z}^{(r+1)}$
  for $k = 1$ to g **do**
    Sample $\boldsymbol{\alpha}_k^{(r+1)}$ from $\boldsymbol{\alpha}_k|\{\boldsymbol{x}_i/z_{ik}^{(r+1)} = 1\}$
return $(\mathbf{z}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\mathbf{z}^{(r_{\max})}, \boldsymbol{\theta}^{(r_{\max})})$

The Gibbs sampler has the advantage to get tractable full conditional distributions since the sampling of $\boldsymbol{z}_i^{(r)}$ can simply be performed through a multinomial distribution $\boldsymbol{z}_i^{(r+1)} \sim \mathcal{M}\left(1; t_{i1}\left(\boldsymbol{\theta}^{(r)}\right), \ldots, t_{ig}\left(\boldsymbol{\theta}^{(r)}\right)\right)$ based on the class posterior probabilities. Sampling of $\boldsymbol{\alpha}_k^{(r)}$ can be performed using standard conjugated prior just using data sampled in cluster k. Thus after some burn-in, the sampled values of $\boldsymbol{\theta}$ are expected to come from $\boldsymbol{\theta}|\mathbf{x}$. For more details about Bayesian inference for mixture see for instance Marin, Mengersen, and Robert (2005). If dealing non conjugated prior distributions, direct sampling from $\boldsymbol{\alpha}_k|\{\boldsymbol{x}_i/z_{ik}^{(r+1)} = 1\}$ can be replaced with a Metropolis-Hastings sampling step. In models developed in Chapter 3 we also use particular strategies to perform the discrete parameters estimation.

Label-switching problem Let notice that since the mixture model is defined up to a permutation of the clusters, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is invariant up to a renumbering of the components as soon as $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are invariant up to a renumbering. This exact symmetry of the posterior distribution, also called label switching problem, makes meaningless direct computation of many usual punctual estimators as the posterior mean. Solutions exist to deal with this problem (Jasra, Holmes, and D. A. Stephens, 2005; M. Stephens, 2000), however, in practice the Gibbs algorithm is often trapped by a mode of the posterior distribution such that label-switching may be not observed. This problem is further discussed in Chapter 6.

2.3 Model selection

Model selection is an important issue in MBC, it encompasses the choice of the number of cluster g, and the parametric model $p(\cdot|\boldsymbol{\alpha}_k)$, it may also include discrete parameters of the model such as block structure that can be viewed as a particular model. For a given parametric family, let denote by $\boldsymbol{m} = (g, \boldsymbol{\omega}_g)$ the model which includes g, the number of clusters, and $\boldsymbol{\omega}_g$ some discrete parameters related to the structure parameter related to the parametric family possibly depending on g (for instance dependence structure inside classes).

Integrated likelihood From a Bayesian perspective, one would like to select the model $\boldsymbol{m}$ with the highest probability $p(\boldsymbol{m}|\mathbf{x})$. One could consider sampling from $p(\boldsymbol{m}|\mathbf{x})$ however this would require an approach like the reversible jump (Green, 1995) since the parameter space depends on the model $\boldsymbol{m}$. Assuming equal prior probabilities for each model $\boldsymbol{m}$ the problem is to find the model $\boldsymbol{m}$ maximizing the integrated likelihood $p(\mathbf{x}|\boldsymbol{m})$:

$$p(\mathbf{x}|\boldsymbol{m}) = \int_{\Theta_{\boldsymbol{m}}} p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{m}) p(\boldsymbol{\theta}|\boldsymbol{m}) \, d\boldsymbol{\theta}. \tag{2.3}$$

with $\Theta_{\boldsymbol{m}}$ the parameter space related to model $\boldsymbol{m}$.
This can for instance be obtained from the output of the Gibbs sampler (Chib, 1995). For a given structure $\boldsymbol{\omega}_g$ the integrated likelihood can be obtained for each possible value of g. Computing the integrated likelihood for each model $\boldsymbol{m}$ can be intractable if the number of models considered is too large. For fixed $\boldsymbol{\omega}_g$ the integrated likelihood can be obtained for each possible value of g since g_{max} the maximal number of cluster is rarely bigger than 20. Strategies investigated can be to screen each possible value of g, and for each perform a sampling of $\boldsymbol{\omega}_g|\mathbf{x}, \mathbf{z}, g$, such strategy is presented in Chapter 3.

BIC criterion In practice the most popular criterion in MBC is to use the BIC criterion (Schwarz, 1978) which relies on an asymptotic approximation of the log of the integrated likelihood. It has consistency properties (Lebarbier and Mary-Huard, 2006), and even if it results from a Bayesian setting it does not need to specify any prior distribution.

$$\mathrm{BIC}(\boldsymbol{m}) = \ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n, \tag{2.4}$$

with $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$ the maximum likelihood estimator of the parameters of model $\boldsymbol{m}$ and $\nu_{\boldsymbol{m}}$ the number of parameters of model $\boldsymbol{m}$. From a clustering perspective BIC tends to overestimate the number of clusters when the model is ill specified.

ICL criterion A possible solution to overcome limitations of BIC is to use the integrated completed likelihood (ICL) in order to take into account the classification perspective (Biernacki, Celeux, and Govaert, 2000). It is a BIC-like criterion

penalized by the class overlap.

$$\mathrm{ICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n = \mathrm{BIC}(\boldsymbol{m}) + \ln p(\hat{\mathbf{z}}|\boldsymbol{x}; \hat{\boldsymbol{\theta}}_m) \tag{2.5}$$

where $\hat{\mathbf{z}}$ is the partition obtained by maximum a posteriori based on $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$. This criterion does not enjoys BIC consistency properties, however it often succeed in finding a more relevant number of clusters since the parametric model is often mispecified thus leading BIC to overestimate the number of clusters.

MICL criterion Another possible criterion proposed by Marbac and Sedki (2017) is to consider the model maximizing the (log of) maximum integrated likelihood (MICL)

$$\mathrm{MICL}(\boldsymbol{m}) = \max_{\mathbf{z} \in \mathcal{Z}} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}). \tag{2.6}$$

As ICL, this criterion takes into account the clustering focus by looking for well-separated clusters. For some conditional independence models under conjugated priors, the computation of $p(\mathbf{x}, \mathbf{z}|\boldsymbol{m})$ has closed form, then transferring the integration issue to the optimization over the partition $\mathcal{Z}$. This optimization can be performed by alternate optimizing the cluster for one individual, all the other class memberships being fixed. This search be can initialized from the partition coming from the maximum a posteriori. It has proved to have good behavior for variable selection in clustering (Marbac and Sedki, 2017). We have used it in the multiple partition framework in Chapter 5.

Bibliography

Allman, E., Matias, C., and Rhodes, J. (2009). "Identifiability of parameters in latent structure models with many observed variables". In: *The Annals of Statistics* 37.6A, pp. 3099–3132.

Ambroise, C. and Matias, C. (2012). "New consistent and asymptotically normal parameter estimates for random-graph mixture models". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 74.1, pp. 3–35. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2011.01009.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2011.01009.x.

Banfield, J. and Raftery, A. (1993). "Model-based Gaussian and non-Gaussian clustering". In: *Biometrics* 49.3, pp. 803–821. ISSN: 0006-341X. DOI: 10.2307/2532201. URL: http://dx.doi.org/10.2307/2532201.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009). "An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures". In: *Journal of Computational and Graphical Statistics* 18.2, pp. 505–526.

Biernacki, C. (2017). " Mixture models". In: *Choix de modèles et agrégation*. Ed. by J.-J. Droesbeke, G. Saporta, and C. Thomas-Agnan. Technip. URL: https://hal.inria.fr/hal-01252671.

Biernacki, C., Celeux, G., and Govaert, G. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.

Bouveyron, C. and Jacques, J. (2011). "Model-based clustering of time series in group-specific functional subspaces". In: *Advances in Data Analysis and Classification* 5.4, pp. 281–300.

Celeux, G. and Govaert, G. (1992). "A classification EM algorithm for clustering and two stochastic versions". In: *Computational Statistics & Data Analysis* 14.3, pp. 315–332.

Celeux, G. and Govaert, G. (1991). "Clustering criteria for discrete data and latent class models". In: *Journal of Classification* 8.2, pp. 157–176. ISSN: 1432-1343. DOI: 10.1007/BF02616237. URL: https://doi.org/10.1007/BF02616237.

— (1995). "Gaussian parsimonious clustering models". In: *Pattern recognition* 28.5, pp. 781–793.

Chib, S. (1995). "Marginal likelihood from the Gibbs output". In: *Journal of the american statistical association* 90.432, pp. 1313–1321.

Daudin, J.-J., Picard, F., and Robin, S. (2008). "A mixture model for random graphs". In: *Stat. Comput.* 18.2, pp. 173–183. ISSN: 0960-3174. DOI: 10.1007/s11222-007-9046-7. URL: http://dx.doi.org/10.1007/s11222-007-9046-7.

Delaigle, A. and Hall, P. (2010). "Defining probability density for a distribution of random functions". In: *The Annals of Statistics*, pp. 1171–1193.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

Gollini, I. and Murphy, T. (2014). "Mixture of latent trait analyzers for model-based clustering of categorical data". In: *Statistics and Computing* 24.4, pp. 569–588.

Goodman, L. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". In: *Biometrika* 61.2, pp. 215–231.

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, p. 711.

Jacques, J. and Preda, C. (2014a). "Functional data clustering: a survey". In: *Advances in Data Analysis and Classification* 8.3, pp. 231–255.

— (2014b). "Model-based clustering for multivariate functional data". In: *Computational Statistics and Data Analysis* 71, pp. 92–106.

Jajuga, K., Sokołowski, A., and Bock, H. (2002). *Classification, clustering and data analysis: recent advances and applications*. Springer Verlag.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling". In: *Statistical Science*, pp. 50–67.

Kosmidis, I. and Karlis, D. (2015). "Model-based clustering using copulas with applications". English. In: *Statistics and Computing*, pp. 1–21. ISSN: 0960-3174. DOI: 10.1007/s11222-015-9590-5. URL: http://dx.doi.org/10.1007/s11222-015-9590-5.

Larose, C. (2015). "Model-Based Clustering of Incomplete Data". PhD thesis. University of Connecticut.

Lebarbier, E. and Mary-Huard, T. (2006). "Une introduction au critère BIC : fondements théoriques et interprétation". In: *Journal de la SFdS* 147.1, pp. 39–57.

Marbac, M., Biernacki, C., and Vandewalle, V. (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207.

— (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656.

Marbac, M. and Sedki, M. (2017). "Variable selection for model-based clustering using the integrated complete-data likelihood". In: *Statistics and Computing* 27.4, pp. 1049–1063.

Marin, J., Mengersen, K., and Robert, C. (2005). "Bayesian modelling and inference on mixtures of distributions". In: *Handbook of statistics* 25, pp. 459–507.

Mazo, G. (2017). "A semiparametric and location-shift copula-based mixture model". In: *Journal of Classification* 34.3, pp. 444–464.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. 2. ed. Wiley series in probability and statistics. Wiley.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McNicholas, P. D. (2016). *Mixture model-based classification*. CRC press.

McNicholas, P. D. and Murphy, T. (2008). "Parsimonious Gaussian mixture models". In: *Stat. Comput.* 18.3, pp. 285–296. ISSN: 0960-3174. DOI: 10.1007/s11222-008-9056-0. URL: http://dx.doi.org/10.1007/s11222-008-9056-0.

McParland, D. and Gormley, I. C. (2016). "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2, pp. 155–169.

Punzo, A. and Ingrassia, S. (2016). "Clustering bivariate mixed-type data via the cluster-weighted model". In: *Computational Statistics* 31.3, pp. 989–1013.

Redner, R. A. and Walker, H. F. (1984). "Mixture densities, maximum likelihood and the EM algorithm". In: *SIAM review* 26.2, pp. 195–239.

Samé, A., Chamroukhi, F., Govert, G., and Aknin, P. (2011). "Model-based clustering and segmentation of time series with changes in regime". In: *Advances in Data Analysis Classification* 5, pp. 301–321.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *Annals of Statistics* 6, pp. 461–464.

Stephens, M. (2000). "Dealing with label switching in mixture models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, pp. 795–809.

Teicher, H. (1963). "Identifiability of Finite Mixtures". In: *The Annals of Mathematical Statistics*, pp. 1265–1269.

— (1967). "Identifiability of mixtures of product measures". In: *Annals of Mathematical Statistics* 38, pp. 1300–1302.

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate". In: *The Annals of Mathematical Statistics* 20.4, pp. 595–601.

Yakowitz, S., Spragins, J., et al. (1968). "On the identifiability of finite mixtures". In: *The Annals of Mathematical Statistics* 39.1, pp. 209–214.

Zanghi, H., Ambroise, C., and Miele, V. (2008). "Fast online graph clustering via Erdős–Rényi mixture". In: *Pattern Recognition* 41.12, pp. 3592–3599. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2008.06.019. URL: http://www.sciencedirect.com/science/article/pii/S0031320308002483.

Chapter 3

Model for clustering of categorical and mixed data

Contents

3.1 Introduction 29
3.2 Conditional dependency per block 30
3.2.1 State of the art 30
3.2.2 Mixture of intermediate dependency (CCM) 31
3.2.3 Dependency per mode (CMM) 38
3.3 Gaussian copulas for mixed data 41
3.3.1 State of the art for mixed data 41
3.3.2 Conclusion 47
3.4 Conclusion 48

Related articles

1. M. Marbac, C. Biernacki, and V. Vandewalle (2015). "Model-based clustering for conditionally correlated categorical data". In: *Journal of Classification* 32.2, pp. 145–175
2. M. Marbac, C. Biernacki, and V. Vandewalle (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207
3. M. Marbac, C. Biernacki, and V. Vandewalle (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656

3.1 Introduction

This chapter deals essentially with contributions related to Matthieu Marbac's thesis that I have co-supervised with Christophe Biernacki (Marbac, Biernacki, and Vandewalle, 2015; Marbac, Biernacki, and Vandewalle, 2016; Marbac, Biernacki, and

Vandewalle, 2017). The main issue is to go beyond the hypothesis of conditional independence by considering mainly categorical data, but also mixed data. Contrary to the continuous framework where there are many tractable multivariate distributions, such as the multivariate Gaussian, there are much fewer models to take into account intra-class dependence in the framework of categorical/mixed data. The solution we proposed in Matthieu Marbac's thesis is to extend some models to take into account the dependency given the cluster. When performing clustering, fewer clusters are needed than when considering the hypothesis of conditional independence. Moreover, the proposed refined dependency can help to interpret the dependency within a cluster. For categorical variables alone, the idea is to relax the conditional class independence assumption by assuming a block of dependency given the class and to propose a particular dependency model within the block. For mixed data, the idea is to use a copula to take into account the dependency between different types of variables.
The main technical difficulty related to these models is the search for the structure of the block which is solved by using a Gibbs sampler. It is also needed to use integrated likelihood to avoid the low precision of the BIC approximation in some cases. For the copula-based model, one of the main difficulties is to manage sampling from the posterior distribution while avoiding the need to compute multivariate integrals at each step of the Gibbs sampler, this requires more advanced strategies such as Metropolis-within-Gibbs.
In Section 3.2, I present our proposal for the categorical data framework, which can be decomposed into two main contributions (Marbac, Biernacki, and Vandewalle, 2015; Marbac, Biernacki, and Vandewalle, 2016) which consider two possible block dependency models. In Section 3.3, I present our contribution to the mixed data setting (continuous, binary, and ordinal) through a mixture of Gaussian copula.

3.2 Conditional dependency per block

3.2.1 State of the art

The latent class model (Goodman, 1974) which assumes the conditional independence between variables (further referred to as CIM for conditional independent model) is the most popular model-based approach to cluster categorical data. Its interpretation is easy since classes are explicitly described by the probability of each modality for each variable. Moreover, the sparsity involved by the conditional independence assumption is a great advantage since it circumvents the curse of dimensionality. In practice, this model obtains good results in lots of applications (Hand and Keming, 2001). However, it leads to severe biases when its main assumption is violated, like an overestimation of the number of components (Van Hattum and Hoijtink, 2009). Furthermore, the larger the number of variables, the higher the risk to observe conditionally correlated variables in a data set, and consequently the higher the risk to involve such biases by using CIM.
Different models relax the class conditional independence assumption. Among them,

the *multilevel latent class model* (Vermunt, 2003) assumes that conditional dependency between the observed variables can be explained by other unobserved variables. Another approach considers the intra-class dependencies by using a single latent continuous variable and a probit function (Qu, Tan, and Kutner, 1996). The mixture of latent trait analyzers (Gollini and Murphy, 2013; Bartholomew, Knott, and Moustaki, 2011) is a good challenger for CIM. It assumes that the distribution of the observed variables depends on many latent variables: one categorical variable (the class) and many continuous latent variables (modeling the intra-class dependencies between the observed categorical variables). Although this model is very flexible, the intra-class dependency is hardly interpretable since the intra-class correlations are interpreted throughout relationships with unobserved continuous variables.

The log-linear models' (Agresti, 2002) purpose is to model the individual log-probability by selecting interactions between variables. Thus, the most general mixture model is the *log-linear mixture model* where all the kinds of interactions can be considered. It has been used for a long time (Hagenaars, 1988) and it obtains good results in many applications (Espeland and Handelman, 1989; Van Hattum and Hoijtink, 2009). However, this model family is huge and the model selection stays a real challenge. In the literature, authors either fix the considered interactions in advance or they perform a deterministic search like the *forward* method which is sub-optimal. Furthermore, the number of parameters increases with the conditional modality crossings, so there is an over-fitting risk and interpretation becomes harder.

3.2.2 Mixture of intermediate dependency (CCM)

We propose in Marbac, Biernacki, and Vandewalle, 2015 to extend the classical latent class model (CIM) for categorical data, by a new latent class model which relaxes the conditional independence assumption. We refer to this new model as the *conditionally correlated model* (denoted by CCM). This model is a parsimonious version of the log-linear mixture model and thus benefits from its interpretative power. Furthermore, we propose a Bayesian approach to automatically perform model selection.

The CCM model groups the variables into conditionally independent blocks given the class. The main intra-class dependencies are thus shown by the repartition of the variables into blocks. This approach, allowing modeling of the main conditional interactions, was first proposed by Jorgensen and Hunt (1996) to cluster continuous and categorical data. For CCM, each block follows a particular dependency distribution which corresponds to our main contribution. This distribution consists in a bi-component mixture of an *independence* and a *maximal dependency* distribution. This specific distribution of the blocks allows summarizing the conditional dependencies of the variables with only one continuous parameter: the maximum dependency distribution proportion. Thus, the model underlines the main conditional dependencies and their strength.

The new model is a two-degree parsimonious version of a log-linear mixture model. The first degree of parsimony is introduced by grouping the variables which are conditionally dependent into the same block. This repartition of the variables per block defines the interactions considered by the model for each class. Moreover, the strength of the correlation is reflected by the proportion of maximum dependency distribution. The second degree of parsimony is induced by the specific distribution of the blocks. As for all log-linear mixture models, the selection of the pertinent interactions is a combinatorial problem. We perform this model selection via a Gibbs sampler to overcome the enumeration of all the models. Thus, this general approach could also select the interactions of any log-linear mixture model.

3.2.2.1 Model assumptions

Observations to be classified are described with d discrete variables $\boldsymbol{x} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d)$ defined on the probabilistic space $\mathcal{X}$. Each variable j has m_j response levels with $m_j \geq 2$ and is written $\boldsymbol{x}^j = (x^{j1}, \ldots, x^{jm_j})$ where $x^{jh} = 1$ if variable j takes modality h and $x^{jh} = 0$ otherwise.

It considers that *conditionally* on the class k, variables are grouped into B_k *independent blocks* and that each block follows a specific distribution. The repartition in blocks of the variables determines a partition $\boldsymbol{\sigma}_k = (\boldsymbol{\sigma}_{k1}, \ldots, \boldsymbol{\sigma}_{k\mathrm{B}_k})$ of $\{1, \ldots, d\}$ in B_k disjoint non-empty subsets where $\boldsymbol{\sigma}_{kb}$ represents the subset b of variables in the partition $\boldsymbol{\sigma}_k$. This partition defines $\boldsymbol{x}^{\{kb\}} = \boldsymbol{x}^{\boldsymbol{\sigma}_{kb}} = (\boldsymbol{x}^{\{kb\}j}; j = 1, \ldots, d^{\{kb\}})$ which is the subset of $\boldsymbol{x}$ associated to $\boldsymbol{\sigma}_{kb}$. The integer $d^{\{kb\}} = \mathrm{card}(\boldsymbol{\sigma}_{kb})$ is the number of variables in block b of component k and $\boldsymbol{x}^{\{kb\}j} = (x^{\{kb\}jh}; h = 1, \ldots, m_j^{\{kb\}})$ corresponds to variable j of block b for component k with $x^{\{kb\}jh} = 1$ if the individual takes modality h for variable $\boldsymbol{x}^{\{kb\}j}$ and $x^{\{kb\}jh} = 0$ otherwise and where $m_j^{\{kb\}}$ represents the number of modalities of $\boldsymbol{x}^{\{kb\}j}$. Note that different repartitions of the variables into blocks are allowed for each component and they are grouped into $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_g)$.

For each component k, each block b follows a specific parametric distribution denoted as $p(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\theta}_{kb})$ where $\boldsymbol{\theta}_{kb}$ groups the parameters of this distribution. The model pdf is written as

$$p(\boldsymbol{x}|\boldsymbol{\sigma}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) \qquad \text{with} \qquad p(\boldsymbol{x}|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) = \prod_{b=1}^{B_k} p(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\theta}_{kb}), \quad (3.1)$$

where $\boldsymbol{\theta}$ is redefined as $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ with $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{k\mathrm{B}_k})$.

A new block distribution: a mixture of two extreme distributions We propose to model the distribution of each block by a bi-components mixture between an *independence* distribution $\mathring{p}(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\alpha}_{kb})$ and a *maximum dependency* distribution

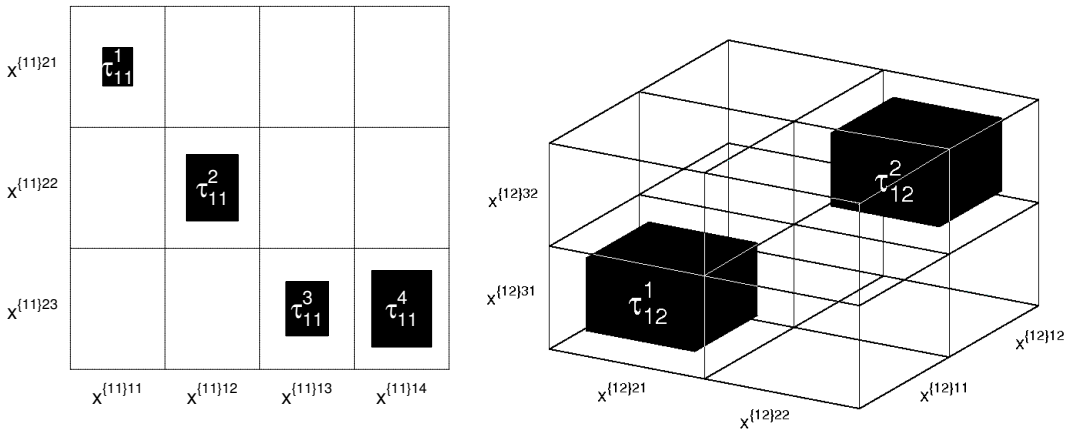
$\acute{p}(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$.

$$\mathring{p}(\boldsymbol{x}|\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}\prod_{h=1}^{m_j}(\alpha_k^{jh})^{x^{jh}} \qquad \text{and} \qquad \acute{p}(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}) = \sum_{h=1}^{m_1^{\{kb\}}} \tau_{kb}^h \mathbf{1}_{\left\{\boldsymbol{x}^{\{kb\}}=\boldsymbol{\delta}_{kb}^h\right\}}. \tag{3.2}$$

The maximum dependency distribution is illustrated Figure 3.1. It considers successive surjections of variables ordered in decreasing number of modalities. Denoting by $m_1^{\{kb\}}$ the number of modalities of the first variable of the block, it only needs to estimate the $m_1^{\{kb\}}$ locations of the non-null probabilities denoted by $\boldsymbol{\delta}_{kb}^h$ and their related probabilities $\boldsymbol{\tau}_{kb}^h$. This model is totally unrealistic alone but it can be useful to propose a distribution that moves slightly away from independence.
For block b of component k, the block distribution is modeled by:

$$p(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\theta}_{kb}) = (1-\rho_{kb})\mathring{p}(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\alpha}_{kb}) + \rho_{kb}\acute{p}(\boldsymbol{x}^{\{kb\}}|\boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}), \tag{3.3}$$

where $\boldsymbol{\theta}_{kb} = (\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$ and where ρ_{kb} is the proportion of the maximum dependency distribution in this mixture with $0 \leq \rho_{kb} \leq 1$. The proposed model requires few additional parameters compared with the conditional independence model. In addition, it is easily interpretable as explained in the next paragraph. Note that the limiting case where $\rho_{kb} = 0$ considers that the block follows an independence distribution. In this particular case, the parameters of the maximum dependency distribution are no longer defined.



(a) (b)

Figure 3.1: Two examples of the maximum dependency distributions. (a) A maximum dependency distribution for a block of two variables; (b) A maximum dependency distribution for a of three variables.

Under this distribution, the proportion of the maximum dependency distribution reflects the deviation from independence under the assumption that the other allowed

distribution is the maximum dependency distribution. The parameter ρ_{kb} gives an indicator of the *inter-variables correlation* of the block. It is not here a pairwise dependency among variables but a dependency between all variables of the block. Furthermore, it stays bounded when the number of variables is larger than two while the Cramer's V is not upper-bounded in this case. The *intra-variables dependencies* between the variables are defined by $\boldsymbol{\delta}_{kb}$. The strength of these dependencies is explained by $\boldsymbol{\tau}_{kb}$ since it gives the *weight of the over-represented modalities crossing* compared with the independence distribution.
Above, we interpreted the distribution by conditionally independent blocks as a parsimonious version of the log-linear mixture model because it determines the interactions to be modeled for each class. By choosing the proposed distribution for blocks, a second level of parsimony is added. Indeed, among the interactions allowed by this distribution with independent blocks, only those corresponding to the maximum dependency distribution will be modeled. Other interactions are considered null.

Properties:

- The CCM, stays parsimonious compared with CIM since, for each block with at least two variables, the number of the additional parameters depends only on the number of modalities of the first variable of the block and not on the number of variables in the block. By denoting by ν_{CIM} the number of parameters of the CIMmodel, the number of parameters of CCM is denoted ν_{CCM} by:
$$\nu_{\text{CCM}} = \nu_{\text{CIM}} + \sum_{\{(k,b)|d^{\{kb\}}>1\}} m_1^{\{kb\}}, \tag{3.4}$$
with $d^{\{kb\}}$ the number of variables of the block.

- The proposed distribution is identifiable under the condition that the block is composed by at least three variables ($d^{\{kb\}} > 2$) or that the modality number of the last variable of the block is more than two ($m_2^{\{kb\}} > 2$). This result is demonstrated in Appendix B of Marbac, Biernacki, and Vandewalle (2013). The parameter ρ_{kb} is a new indicator allowing measuring the correlation between variables, not limited to correlation between variable couples.

3.2.2.2 Estimation of the parameters

For a fixed model $(g, \boldsymbol{\sigma})$ (number of components and repartition of the variables in blocks for each class), two algorithms derived from the EM algorithm perform the estimation of the associated continuous parameters, since the proposed distribution CCM has two latent variables (the class membership and the intra-block distribution membership). The combinatorial problems arising from the consideration of the discrete parameters ($\boldsymbol{\delta}_{kb}$ the location of maximum dependency) are avoided by using a Metropolis-Hastings algorithm.

We have considered to use the GEM algorithm in order to find the maximum likelihood, since the maximization step in the EM algorithm requires estimating the continuous parameters for too many possible values of the discrete parameters in order to warrant the maximization of the complete-data log-likelihood expectation. Indeed, exhaustive enumeration for estimating the discrete parameters is generally impossible when a block contains variables with many modalities and/or many variables. Thus, a stochastic approach is proposed. Then, the estimation of the continuous parameters conditionally on the discrete parameters is performed via the classical EM algorithm since their estimation cannot be obtained in closed form. At iteration (r), the steps of the global GEM can be written as:

- **$\mathbf{E}_{\mathbf{global}}$ step:** $z_{ik}^{(r)} = \frac{\pi_k^{(r)} p(\boldsymbol{x}_i|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^{g} \pi_{k'}^{(r)} p(\boldsymbol{x}_i|\boldsymbol{\sigma}_{k'}, \boldsymbol{\theta}_{k'}^{(r)})}$,
- **$\mathbf{GM}_{\mathbf{global}}$ step:** $\pi_k^{(r+1)} = \frac{n_k^{(r)}}{n}$ where $n_k^{(r)} = \sum_{i=1}^{n} z_{ik}^{(r)}$ and $\forall (k,b)$ $\boldsymbol{\theta}_{kb}^{(r+1)}$ is updated under the constraint that the conditional expectation of complete-data log-likelihood increases.

$\mathrm{GM}_{\mathrm{global}}$ step can be performed by using a stochastic search using a Metropolis-Hastings algorithm, which is not detailed here. It considers some neighborhood of the current block structure.

3.2.2.3 Model selection

Since the number of components g determines the dimension of $\boldsymbol{\sigma}$, the model construction is done in two steps. Firstly, the selection of the number of components and, secondly, the determination of the variable repartition per blocks for each component. In a Bayesian context, the best model $(\hat{g}, \hat{\boldsymbol{\sigma}})$ is defined as (Robert, 2005):

$$(\hat{g}, \hat{\boldsymbol{\sigma}}) = \underset{g,\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(g, \boldsymbol{\sigma}|\mathbf{x}). \tag{3.5}$$

Thus, by considering that $p(g) = \frac{1}{g_{\max}}$ if $g \leq g_{\max}$ and 0 otherwise, where $g_{\max}$ is the maximum number of classes allowed by the user, and by assuming that $p(\boldsymbol{\sigma}|g)$ follows a uniform distribution, the best model is also defined as:

$$\hat{g} = \underset{g}{\operatorname{argmax}}\ p(\mathbf{x}|g, \hat{\boldsymbol{\sigma}}^g) \quad \text{with} \quad \hat{\boldsymbol{\sigma}}^g = \underset{\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(\mathbf{x}|g, \boldsymbol{\sigma}) \tag{3.6}$$

To obtain $(\hat{g}, \hat{\boldsymbol{\sigma}}^{\hat{g}})$, a Gibbs algorithm is used for estimating $\operatorname{argmax}_{\boldsymbol{\sigma}}\ p(\mathbf{x}|g, \boldsymbol{\sigma})$, for each value of $g \in \{1, \ldots, g_{max}\}$. Indeed, this method limits the combinatorial problem involved by the detection of the block structure of variables. We propose to use an easier Gibbs sampler-type having $p(\boldsymbol{\sigma}|\mathbf{x}, g)$ as stationary distribution. It alternates between two steps: the generation of a stochastic neighborhood $\Sigma^{[q]}$ conditionally on the current model $\boldsymbol{\sigma}^{[q]}$ by a proposal distribution and the generation of a new pattern $\boldsymbol{\sigma}^{[q+1]}$ included in $\Sigma^{[q]}$ with a probability proportional to its posterior probability. A full Bayesian approach could be considered, but here we have chosen to replace the integrated likelihood by its approximation based on BIC which

produces good results in practice. More details can be found in Marbac, Biernacki, and Vandewalle (2015).

3.2.2.4 Illustration on Calve data set

The results obtained by the CCM are compared to those obtained for the CIM by the RMixmod software (Lebret et al., 2012). The "Genes Diffusion" company has collected information from French breeders to cluster calves. The 4270 studied calves are described by nine variables related to behavior (aptitude for sucking *Apt*, behavior of the mother just before the calving *Iso*) and health (treatment against omphalitis *TOC*, respiratory disease *TRC* and diarrhea *TDC*, umbilicus disinfection *Dis*, umbilicus emptying *Emp*, mother preventive treatment against respiratory disease *TRM* and diarrhea *TDM*).
Table 3.1 displays the BIC criterion values and the number of parameters for the CIM and CCM models. Furthermore, the computing time in minutes (obtained with an Intel Core i5-3320M processor) to estimate CCM by starting 20 MCMC chains with a stopping criterion of $q_{\max} = 180$ while CIMneeds 3 seconds with the R package RMixmod (Lebret et al., 2012).

	g	1	2	3	4	5	6	7	8
CIM	BIC	-28589	-26859	-26526	-26333	-26238	-26235	-26226	**-26185**
	ν_{CIM}	17	35	53	71	89	107	125	**143**
CCM	BIC	-26653	-26289	-26173	-26038	**-26025**	-26059	-26045	-26058
	ν_{CCM}	24	48	80	89	**112**	131	148	163
	time (min)	0.97	3.32	6.16	6.56	**10.03**	11.76	12.31	14.92

Table 3.1: Results for the CIMand the CCM according to different class numbers. For both models, first row corresponds to the BIC criterion values and the second row indicates the continuous parameter number. For each model, the best results according to the BIC criterion are in bold. The computing time for the CCM estimation is given in minutes.

For the CIM, the BIC criterion selects a high number of classes, since it selected eight classes. The interpretation of the clusters is also difficult and we can assume that the estimator's quality is very poor. Figure 3.2 helps the interpretation for the CCM with five components (best model according to the BIC criterion). For example, this figure shows that the first class has a proportion of 0.29 and that it is composed of four blocks. The most correlated block of the first class has $\rho_{kb} \simeq 0.80$ and the strength of the biggest modalities link is also close to 0.85. This block consists of the variables *TDC* and *TRM*.
Here is now a possible interpretation of Class 1 (note that the others classes are also meaningful; see details in Marbac, Biernacki, and Vandewalle (2013)):

- **General:** This class has a proportion equal to 0.29 and consists of three blocks of dependency and one block of independence.

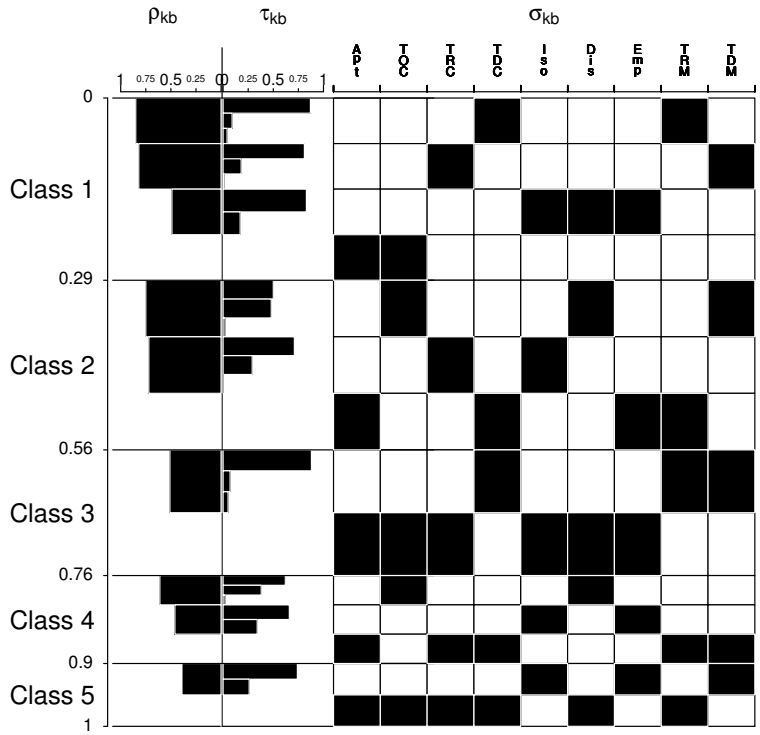


Figure 3.2: Summary of the best CCM according to BIC for the calves data set.

- **Block 1:** There is a strong correlation (ρ_{11}) between the variables diarrhea treatment of the calf and mother preventive treatment against respiratory disease, especially between the modality no treatment against the calf diarrhea and the absence of preventive treatment against respiratory diseases of its mother ($\boldsymbol{\tau}_{11}$ and $\boldsymbol{\delta}_{11}$).
- **Block 2:** There is a strong correlation (ρ_{12}) between the variables treatment against respiratory illness of the calf and mother preventive treatment against diarrhea, especially between the modality preventive treatment against respiratory illness of the calf and the presence of diarrhea preventive treatment of its mother ($\boldsymbol{\tau}_{12}$ and $\boldsymbol{\delta}_{12}$).
- **Block 3:** There exists another strong link between the behavior of the mother, the emptying of the umbilical and its disinfection ($\boldsymbol{\tau}_{13}$ and $\boldsymbol{\delta}_{13}$).
- **Block 4:** This block is characterized by the absence of preventive treatment against omphalitis and having 50% of the calves infected by this illness ($\boldsymbol{\alpha}_{14}$).

3.2.2.5 Conclusion

The block distribution is defined as a mixture between an independent distribution and a maximum dependency distribution. This specific distribution, which remains

parsimonious, is compared to the full latent class model and allows different levels of interpretation. The blocks of variables detect the conditional dependencies between variables while their strengths are reflected by the proportions of maximum dependency distribution. The parameters of the block distribution reflect the links and the strength between modalities.

The parameter estimation and the model selection are simultaneously performed via a Gibbs sample-type algorithm. It allows reducing the combinatorial problems of the block structure detection and the links between modalities search for the estimation of the maximum dependency distribution. The results are good when the number of modalities is small for each variable. For more than six modalities, the detection of other links encounters some persistent difficulties. So the algorithm can be slow in this case. The proposed approach to estimate the block structure is not adapted for data sets with many variables. The R package *Clustericat*[1] allows clustering categorical data sets by using CCM.

The proposed approach has the advantage to be very interpretable, however, it could lack flexibility since the maximal dependency distribution only covers a small part of the space limited to the number of modalities of the first variable of the block. Thus in the next section, we propose a solution that potentially covers a larger part of the space.

3.2.3 Dependency per mode (CMM)

In this section, contrary to CCM, we propose to model the distribution in a block by a multinomial distribution per modes which assumes that few levels, named *modes*, are characteristic whereas the other ones follow a uniform distribution. This distribution is considered on crossings of the variables of the block. The resulting multinomial distribution is parsimonious since its free parameters are limited to these few modes. In this model, the repartition of the variables into blocks is assumed to be the same in each cluster.

For a fixed number of components, the model selection (repartition of the variables into blocks and mode numbers) is the most challenging problem since the number of competing models is huge. Therefore, the model selection is carried out by an MCMC algorithm whose mode of the stationary distribution corresponds to the model having the highest posterior probability. This algorithm performs a random walk in the model space and requires the computation of the integrated complete-data likelihood. This quantity is not approximated by BIC-like methods since their results are poor. Indeed, the integrated complete-data likelihood is accessible and has closed-form through weakly informative conjugate prior. This approach provides an efficient model selection in a reasonable computational time since the parameters are estimated via an EM algorithm only for the single selected model.

[1] https://r-forge.r-project.org/R/?group%20id=1803

Conditional mode model (CMM) The repartition of the variables into the B blocks[2] is defined by the partition $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_B)$ of $\{1, \ldots, d\}$. The new variable resulting from the concatenation of the initial variables affiliated to block j (*i.e.* $\{\boldsymbol{x}_i^b; b \in \boldsymbol{\sigma}_j\}$) is itself a categorical variable whose levels are defined by the Cartesian product of the variables affiliated to block j. This new (block dependent) categorical variable defined for block j is denoted by $\tilde{\boldsymbol{x}}_i^j = (\tilde{x}_i^{jh}; h = 1, \ldots, \tilde{m}_j)$, such as $\tilde{x}_i^{jh} = 1$ if individual i has level h and $\tilde{x}_i^{jh} = 0$ otherwise, where $\tilde{m}_j$ is the number of levels for block j. Since this mapping of the variables is bijective, defining a probability on $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^d)$ is equivalent to defining a probability on $\tilde{\boldsymbol{x}}_i = (\tilde{\boldsymbol{x}}_i^1, \ldots, \tilde{\boldsymbol{x}}_i^B)$. The CMM model considers that each block follows a so-called *multinomial distribution per modes*. This distribution has only few free parameters corresponding to its *modes*, while the other parameters are equal. Thus, the free parameters are those of the levels having the greatest probabilities whereas uniformity holds for non-mode levels. The distribution of block j for component k has u_{kj} degrees of freedom (so u_{kj} modes) with $0 \leq u_{kj} < \tilde{m}_j$, and its mode locations are defined by the discrete parameter $\boldsymbol{\delta}_{kj} = \{\delta_{kjh}; h = 1, \ldots, u_{kj}\}$ with $\delta_{kjh} \in \{1, \ldots, \tilde{m}_j\}$ and $\delta_{kjh} \neq \delta_{kjh'}$ if $h \neq h'$. Its probabilities are given by $\boldsymbol{\alpha}_{kj} = (\alpha_{kjh}; h = 1, \ldots, u_{kj} + 1)$ where α_{kjh} denotes the probability of mode h for $h = 1, \ldots, u_{kj}$ and where $\alpha_{kju_{kj}+1}$ corresponds to the remaining probability mass. So, $\boldsymbol{\alpha}_{kj}$ is defined on a truncated simplex denoted by $S(u_{kj}, \tilde{m}_j)$ with

$$S(u_{kj}, \tilde{m}_j) = \left\{ \boldsymbol{\alpha}_{kj} : \sum_{h=1}^{u_{kj}+1} \alpha_{kjh} = 1 \text{ and for } 1 \leq h \leq u_{kj},\ \alpha_{kjh} \geq \frac{\alpha_{kju_{kj}+1}}{\tilde{m}_j - u_{kj}} > 0 \right\}. \tag{3.7}$$

Therefore, the pdf of block j for component k is

$$p(\tilde{\boldsymbol{x}}_i^j | u_{kj}, \boldsymbol{\delta}_{kj}, \boldsymbol{\alpha}_{kj}) = \left(\prod_{h=1}^{u_{kj}} (\alpha_{kjh})^{\tilde{x}_i^{j\delta_{kjh}}} \right) \left(\frac{\alpha_{kju_{kj}+1}}{\tilde{m}_j - u_{kj}} \right)^{1 - \sum_{h' \in \boldsymbol{\delta}_{kj}} \tilde{x}_i^{j\delta_{kjh'}}}. \tag{3.8}$$

An instance of distribution per modes is given Figure 3.3. Making blocks and considering multinomial distribution per block allows flexible modeling of the distribution in each cluster, it is particularly well adapted when in a block most of the probability mass relies on few characteristics crossings. These models generalize the parsimonious model proposed by Celeux and Govaert (1991) for clustering categorical data. For a more detailed interpretation see (Marbac, Biernacki, and Vandewalle, 2016).

Parameters estimation Given g, $\boldsymbol{\sigma}$ and $\boldsymbol{u}$ (the number of modes in each block) the parameters of the model are very easy to estimate by a standard EM algorithm, with closed-form formula at M step, just needing to order the modalities according to their frequency, then equalizing the lowest frequencies.
Some difficulties are first to find $\boldsymbol{u}$ (the number of modes) while g and $\boldsymbol{\sigma}$ (the block structure) supposed to be known. Our first intuition was to use BIC to choose such a parameter. However, this approximation is revealed to have poor behavior, even in the simple multinomial case without considering any mixture. Thus we have chosen

[2]Note that the repartition of the variables into blocks is identical between classes for identifiability reasons.

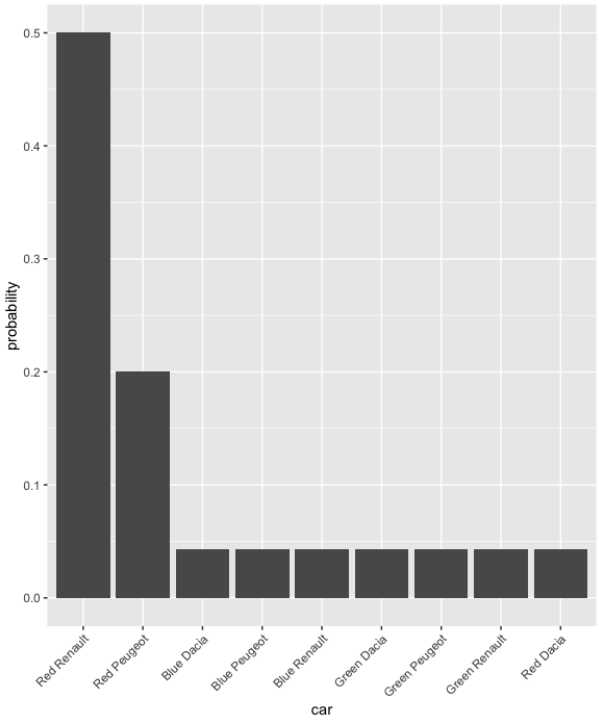


Figure 3.3: Example of distribution per modes. We consider the crossing of the color and the brand of cars. It has two modes and uniformity for other crossings.

to compute the integrated likelihood in the Bayesian framework. This approach was quite tractable since we are just considering parameters in the truncated simplex, which makes the exact computation of the integrated likelihood tractable by plugin the location of the modes.

Model choice The model defined by $g, \boldsymbol{\sigma}, \boldsymbol{u}$ is selected through a Gibbs sampler. The Gibbs sampler alternates between the conditional sampling of the partition $\mathbf{z}$ and the conditional sampling of the couple $(\boldsymbol{\sigma}, \boldsymbol{u})$. Thus obtaining a sample from $p(\boldsymbol{\sigma}, \boldsymbol{u}, \mathbf{z}|\mathbf{x}, g)$.Iteration $[q]$ of Gibbs sampler is written as

$$\mathbf{z}^{[q+1]} \sim p(\mathbf{z}|g, \boldsymbol{\sigma}^{[q]}, \boldsymbol{u}^{[q]}, \mathbf{x}) \tag{3.9}$$

$$(\boldsymbol{\sigma}^{[q+1]}, \boldsymbol{u}^{[q+1]}) \sim p(\boldsymbol{\sigma}, \boldsymbol{u}|g, \mathbf{x}, \mathbf{z}^{[q+1]}). \tag{3.10}$$

Particular details are omitted here but can be found in Marbac, Biernacki, and Vandewalle (2016). Moreover, for sake of brevity, we omit here numerical experiments.

Conclusion The distribution per modes on the variables crossing permits to take into account dependencies between variables, potentially in a sparse way when most of the probability mass relies on few characteristic crossings. Using this built-in model is particularly useful to take into account dependencies when considering the clustering of categorical variables. The combinatorial problems of the block

detection and the modes number selection is solved by a hybrid MCMC algorithm that uses the computation of the integrated complete-data likelihood. The R package `CoModes`[3] allows to perform the model selection and the parameter estimation.

3.3 Gaussian copulas for mixed data

This section considers a proposal to take into account dependencies between continuous, discrete, and ordinal variables in a model-based clustering framework, through the use of a Gaussian copula. The particular challenge in using Gaussian copulas with discrete and ordinal data is that likelihood computation requires numerical integration, thus we propose strategies below to avoid the computation of such integral.

3.3.1 State of the art for mixed data

The literature covering homogeneous data (composed of variables of the same type) is extensive and presents Gaussian mixture models (Banfield and Raftery, 1993), multinomial mixture models (Goodman, 1974) and Poisson mixture models (Karlis and Tsiamyrtzis, 2008) as the standard models used to cluster such data sets. Although many data sets contain mixed data (variables of different types), few mixture models can manage these data (Hunt and Jorgensen, 2011) due to the shortage of multivariate distributions.

The *locally independent mixture model* (Moustaki and Papageorgiou, 2005; Lewis, 1998; Hand and Keming, 2001) is a convenient approach for clustering mixed data since it assumes independence within-component between variables. Thus, each component is defined by a product of standard univariate distributions that facilitate their interpretation. However, this model can lead to severe bias when its main assumption is violated (Van Hattum and Hoijtink, 2009). Therefore, two models have been introduced to relax this assumption.

The *location mixture model* (Krzanowski, 1993; Willse and Boik, 1999) has been proposed for clustering a data set with continuous and categorical variables. It assumes that, for each component, the categorical variables follow a multinomial distribution and the continuous variables follow a multivariate Gaussian distribution conditionally on the categorical variables. Therefore, the intra-component dependencies are taken into account. However, the model requires too many parameters. Hunt and Jorgensen (1999) extended this approach by splitting the variables into within-component independent blocks. Each block contains no more than one categorical variable and follows a location model. The interpretation of this model can be complex since, for a given component, the univariate marginal of a continuous variable follows a Gaussian mixture model. Moreover, the estimation of the repartition of the variables into blocks is a difficult problem that the authors achieve with an ascending method that is sub-optimal.

[3] https://r-forge.r-project.org/R/?group_id=1809

The *underlying variables mixture model* (Everitt, 1988) has been proposed for clustering a data set with continuous and ordinal variables. It assumes that each ordinal variable arises from a latent continuous variable and that all continuous variables (observed and unobserved) follow a Gaussian mixture model. The distribution of observed variables is obtained by integrating each Gaussian component into the subset of latent variables. However, in practice, this computation is not feasible when there are more than two ordinal variables. To study data sets with numerous binary variables, Morlini (2012) expanded this model by estimating the scores of latent variables from those of binary variables. However, the interpretation of the mixture components can be complex since it is based on the score-related parameters (not those related to observed variables).

Previous models illustrate the difficulty of clustering mixed data with a model for which interpretation and inference are easy. Moreover, they do not take account of cases where some variables are integers. The main difficulty is due to a shortage of conventional distributions for mixed data. However, copulas are standard tools for systematically defining multivariate distributions, and they, therefore, have good potential for providing a sensible answer.

Copulas (Joe, 1997; Nelsen, 1999) can be used to build a multivariate model by defining, on the one hand, the *univariate marginal distributions*, and, on the other, the *dependency model*. Smith and Khaled (2012) and Murray et al. (2013) modeled the distribution of mixed variables using one Gaussian copula. As pointed out by Pitt, Chan, and Kohn (2006), the maximum likelihood inference is very difficult for a Gaussian copula with discrete margins. Therefore, it is often replaced by the *Inference Function for Margins* method performing the inference in two steps (Joe, 1997; Joe, 2005). When all the variables are continuous, the fixed-point-based algorithm proposed by Song, Fan, and Kalbfleisch (2005) achieves the maximum likelihood estimation, but this approach is not doable for mixed data. Therefore, as shown by Smith and Khaled (2012), it is more convenient to work in a Bayesian framework since this simplifies the inference by using the latent structure of the model.

3.3.1.1 Component modeled by a Gaussian copula

The Gaussian copula mixture model considers that each component follows a Gaussian copula. Component k is also parametrized by $\boldsymbol{\alpha}_k = (\boldsymbol{\Gamma}_k, \boldsymbol{\beta}_k)$ where $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{ke})$ groups the parameters of the univariate margin, $\boldsymbol{\beta}_{kj}$ being the parameters of the j-th univariate margin, and where $\boldsymbol{\Gamma}_k$ is the correlation matrix of size $e \times e$. The cumulative distribution function (cdf) of component k is written as

$$P(\boldsymbol{x}|\boldsymbol{\alpha}_k) = \Phi_e(\Phi_1^{-1}(u_k^1), \ldots, \Phi_1^{-1}(u_k^e)|\mathbf{0}, \boldsymbol{\Gamma}_k), \tag{3.11}$$

where $u_k^j = P(x^j|\boldsymbol{\beta}_{kj})$ is the value of the cdf of the univariate marginal distribution of variable j for component k evaluated at x^j, where $\Phi_e(.|\mathbf{0}, \boldsymbol{\Gamma}_k)$ is the cdf of the e-variate centred Gaussian distribution with correlation matrix $\boldsymbol{\Gamma}_k$ and where $\Phi_1^{-1}(.)$

is the inverse cumulative distribution function of the standard univariate Gaussian $\mathcal{N}_1(0,1)$.
The Gaussian copula mixture model involves the latent class variable $\boldsymbol{z}$ (in $\{0,1\}^g$) and a second latent variable $\boldsymbol{y} = (y^1, \ldots, y^e) \in \mathbb{R}^e$, such that $\boldsymbol{y}|z_k = 1$ follows an e-variate centered Gaussian distribution $\mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$. Thus, the Gaussian copula mixture can be interpreted as the marginal distribution of $\boldsymbol{x}$ based on the distribution of the variable triplet $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Conditionally on $(\boldsymbol{y}, z_k = 1)$, each element of $\boldsymbol{x}$ is defined by

$$x^j = P^{-1}(\Phi_1(y^j)|\boldsymbol{\beta}_{kj}),\ \forall j = 1, \ldots, e. \tag{3.12}$$

Thus, the generative model of the Gaussian copula mixture is written as

- Class membership *sampling*: $\boldsymbol{z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$,
- Gaussian copula *sampling*: $\boldsymbol{y}|z_k = 1 \sim \mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$,
- Observed data *deterministic computation*: $\boldsymbol{x}$ is obtained from (3.12).

Specific distributions for mixed-type variables The cdf of component k defined by (3.11) implies the cdf of the univariate marginal distributions. Hence, it requires the definition of the distributions of the univariate margins (*i.e.* distribution of $x^j|z_k = 1$). We use conventional parametric distributions to facilitate the component interpretation. The parameters of margin j for component k are denoted by β_{kj}. Hence,

- if x^j is continuous then $x^j|z_k = 1$ follows a *Gaussian* distribution with mean μ_{kj} and variance σ_{kj}^2 and $\beta_{kj} = (\mu_{kj}, \sigma_{kj})$,
- if x^j is integer then $x^j|z_k = 1$ follows a *Poisson* distribution with parameter $\beta_{kj} \in \mathbb{R}^{+*}$,
- if x^j is ordinal then $x^j|z_k = 1$ follows an *ordered multinomial* distribution with parameter β_{kj} defined on the simplex of size m_j. Note that the order between the levels is crucial since it permits the definition of the cdf.

Given that the first c variables of $\boldsymbol{x}$ ($\boldsymbol{x}^{\mathrm{C}}$) are continuous while the last d variables ($\boldsymbol{x}^{\mathrm{D}}$) are discrete, the pdf of component k can be decomposed as

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_k) = p(\boldsymbol{x}^{\mathrm{C}}|\boldsymbol{\alpha}_k) \times p(\boldsymbol{x}^{\mathrm{D}}|\boldsymbol{x}^{\mathrm{C}}, \boldsymbol{\alpha}_k). \tag{3.13}$$

We use the decomposition into sub-matrices $\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{k\mathrm{CC}} & \boldsymbol{\Gamma}_{k\mathrm{CD}} \\ \boldsymbol{\Gamma}_{k\mathrm{DC}} & \boldsymbol{\Gamma}_{k\mathrm{DD}} \end{bmatrix}$, for instance $\boldsymbol{\Gamma}_{k\mathrm{CC}}$ is the sub-matrix of $\boldsymbol{\Gamma}_k$ composed by the rows and the columns related to the observed continuous variables. Under component k, the knowledge of the continuous variable x^j implies that $y^j = \frac{x^j - \mu_{kj}}{\sigma_{kj}}$. Denoting $\boldsymbol{y}^{\mathrm{C}} = \left(\frac{x^j - \mu_{kj}}{\sigma_{kj}}; j = 1, \ldots, c\right)$, $p(\boldsymbol{x}^{\mathrm{C}}|\boldsymbol{\alpha}_k) = \frac{\phi_c(\boldsymbol{y}^{\mathrm{C}}|\mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^{c} \sigma_{kj}}$ where $\phi_c(.|\mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})$ denotes the pdf of c-variate Gaussian distribution with mean $\mathbf{0}$ correlation matrix $\boldsymbol{\Gamma}_{k\mathrm{CC}}$. If the variable j is discrete,

any value y^j in the interval $\mathcal{S}_k^j(x^j) =]b_k^{\ominus}(x^j), b_k^{\oplus}(x^j)]$ produces the same observation x^j under component k, where $b_k^{\ominus}(x^j) = \Phi_1^{-1}(P(x^j - 1|\boldsymbol{\beta}_{kj}))$ and $b_k^{\oplus}(x^j) = \Phi_1^{-1}(P(x^j|\boldsymbol{\beta}_{kj}))$. Under component k, the distribution of the continuous latent variable $\boldsymbol{y}^{\text{D}} = \left(y^j; j = c+1, \ldots, e\right)$ conditionally on $\boldsymbol{y}^{\text{C}}$ is a Gaussian distribution with mean $\boldsymbol{\mu}_k^{\text{D}} = \boldsymbol{\Gamma}_{k\text{DC}}\boldsymbol{\Gamma}_{k\text{CC}}^{-1}\Psi(\boldsymbol{x}^{\text{C}}; \boldsymbol{\alpha}_k)$ and covariance matrix $\boldsymbol{\sigma}_k^{\text{D}} = \boldsymbol{\Gamma}_{k\text{DD}} - \boldsymbol{\Gamma}_{k\text{DC}}\boldsymbol{\Gamma}_{k\text{CC}}^{-1}\boldsymbol{\Gamma}_{k\text{CD}}$. Thus, the pdf of component k is written as

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_k) = \frac{\phi_c(\boldsymbol{y}^{\text{C}}|\boldsymbol{0}, \boldsymbol{\Gamma}_{k\text{CC}})}{\prod_{j=1}^{c}\sigma_{kj}} \times \int_{\mathcal{S}_k(\boldsymbol{x}^{\text{D}})} \phi_d(\boldsymbol{u}|\boldsymbol{\mu}_k^{\text{D}}, \boldsymbol{\sigma}_k^{\text{D}})d\boldsymbol{u}, \tag{3.14}$$

where $\mathcal{S}_k(\boldsymbol{x}^{\text{D}}) = \mathcal{S}_k^{c+1}(x^{c+1}) \times \ldots \times \mathcal{S}_k^e(x^e)$. The Gaussian copula mixture model is identifiable if at least one variable is continuous or integer.

Related models The Gaussian copula mixture model generalizes many conventional mixture models, including the four cases mentioned below.

- If the correlation matrices are diagonal (*i.e.* $\boldsymbol{\Gamma}_k = \boldsymbol{I}$, $\forall k = 1, \ldots, g$), then the model is equivalent to the locally independent mixture model.
- If all the variables are continuous (*i.e.* $c = e$ and $d = 0$), then the model is equivalent to the Gaussian mixture model without constraints among parameters (Banfield and Raftery, 1993). In the spirit of the homoscedastic Gaussian mixture, we also propose a parsimonious version of the Gaussian copula mixture model by assuming equality between the correlation matrices over components. This model is named *homoscedastic* since the covariance matrices of the latent Gaussian variables are equal between components (*i.e.* $\boldsymbol{\Gamma}_1 = \ldots = \boldsymbol{\Gamma}_g$). The free correlation model will be now called the *heteroscedastic* model).
- The model is linked to the binned Gaussian mixture model. For example, when variables are ordinal, it is equivalent to the mixture model presented by Gouget (2006). In such cases, the model is stable through the fusion of modalities.
- If the variables are both continuous and ordinal, then the model is a new parametrization of the model proposed by Everitt (1988). It should be noted that Everitt directly estimates the space $\mathcal{S}_k(\boldsymbol{x}^{\text{D}})$ containing the antecedents of $\boldsymbol{x}^{\text{D}}$. Moreover, he uses a simplex algorithm to perform maximum likelihood inference, but this method dramatically limits the number of ordinal variables. The new parametrization of the proposed mixture allows the univariate marginal parameters $\boldsymbol{\beta}_{kj}$ of each component to be directly estimated, whereas Everitt's parametrization implies a difficult estimation of the bounds of integration. Thus, parameter inference is easier.

Standardized coefficient of correlation per class The Gaussian copula provides a user-friendly correlation coefficient for each pair of variables. Indeed, when both variables are continuous, it is equal to the upper boundary of the correlation

coefficients obtained by monotonic transformation of the variables (Klaassen and Wellner, 1997). Furthermore, when both variables are discrete, it is equal to the polychoric correlation coefficient (Olsson, 1979).

Data visualization per component: a by-product of Gaussian copulas By using the latent vectors of the Gaussian copulas $\boldsymbol{y}|\boldsymbol{z}$, a PCA-type method allows *visualization* of the individuals *per component* which permits the identification of main within-component dependencies. The visualization of component k is performed by computing the coordinates $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, z_k = 1; \boldsymbol{\alpha}_k]$ and then projecting them onto the PCA region associated with the Gaussian copula of component k. This space is obtained directly through spectral decomposition of $\boldsymbol{\Gamma}_k$. The individuals arising from component k follow a centered Gaussian distribution on this factorial map. Those arising from another component have an expectation not equal to zero. Therefore, an individual located far away from the origin arises from a distribution significantly different from the distribution of component k. Finally, the correlation circle summarizes the within-component correlations and avoids the direct interpretation of the correlation matrix $\boldsymbol{\Gamma}_k$, which can be tedious if e is large. The following example illustrates these properties.
Let three variables—one continuous, one integer and one binary—arise, in this order, from the bi-component Gaussian copula mixture model parametrized by

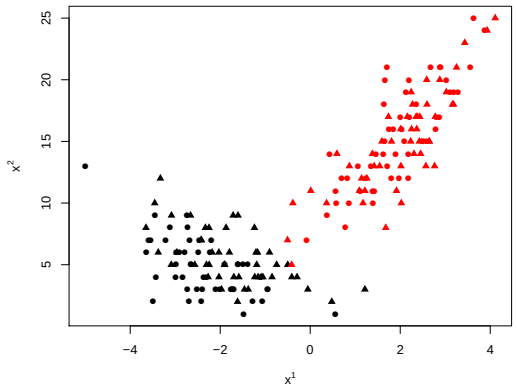
$$\boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\beta}_{11} = (-2, 1),\ \boldsymbol{\beta}_{12} = 5,\ \boldsymbol{\beta}_{13} = \boldsymbol{\beta}_{23} = (0.5, 0.5),\ \boldsymbol{\beta}_{21} = (2, 1),$$

$$\boldsymbol{\beta}_{22} = 15,\ \boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 & -0.4 & 0.4 \\ -0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Gamma}_2 = \begin{pmatrix} 1 & 0.8 & 0.1 \\ 0.8 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}.$$

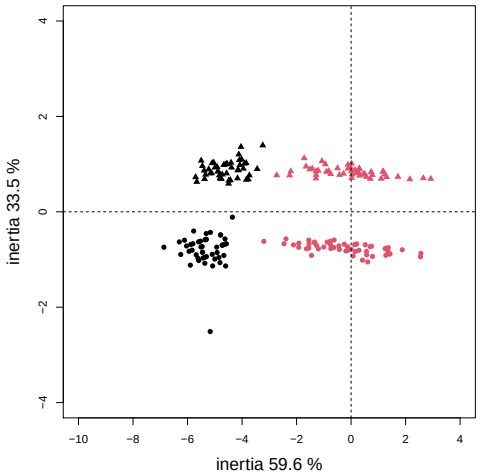
Figure 3.4 provides an example of data visualization. Figure 3.4a shows the scatterplot of the individuals in their native space. Figure 3.4b presents the scatterplot of the individuals in the first PCA-map of the second component (red). It allows two classes to be easily distinguished: a centered one (red) and a second one (black) located on the left side. More precisely, the first axis (explained by the continuous and the integer variables) is strongly discriminative while the second axis (explained exclusively by the binary variable) is not discriminative. Figure 3.4c shows the correlation circle of the first PCA-map of the red component. It allows a strong correlation to be identified, for the red component, between the continuous and the integer variables.

3.3.1.2 Bayesian parameters inference

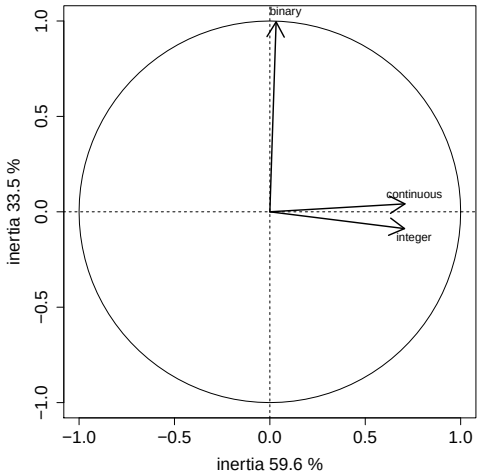
We observe the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ composed of n independent realizations $\boldsymbol{x}_i \in \mathbb{R}^c \times \mathcal{X}$ assumed to arise from a Gaussian copula mixture model. As pointed out by Smith and Khaled (2012), the Bayesian framework simplifies the inference considerably since it uses the latent structure of the model $(\boldsymbol{y}, \boldsymbol{z})$. Without prior information about the data, we assume independence between the prior distributions. The proportions and the parameters of the univariate marginal distributions



(a) Individuals described by three variables: one continuous (abscissa), one integer (ordinate) and one binary (symbol). Colors indicate the true class memberships



(b) Individuals in the first factorial map of component 2. Colors indicate the true class memberships, and symbols the value of the binary variable



(c) Variables in the first factorial map of component 2

Figure 3.4: Example of data visualization.

of each component $\boldsymbol{\beta}_{kj}$ follow the classical conjugate prior distributions (Robert, 2007). Finally, the conjugate prior of the covariance matrices is derived from an Inverse Wishart distribution as proposed by (Hoff, 2007).

Gibbs and Metropolis-within-Gibbs samplers The Bayesian estimation is managed by a Gibbs sampler (described in Algorithm 4). Its stationary distribution is $p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}|\mathbf{x})$ where $\mathbf{z}$ denotes the partition of $\mathbf{x}$ and where $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ denotes the Gaussian vector related to $\mathbf{x}$. Note that the Gaussian variable $\mathbf{y}$ is twice sampled

during one iteration of the algorithm to manage the strong dependencies between $\mathbf{y}$ and $\mathbf{z}$, and between $\boldsymbol{y}^j_{[rk]} = \{y^j_i : z^{(r)}_{ik} = 1\}$ and $\boldsymbol{\beta}_{kj}$. The stationary distribution stays unchanged. Thus, the sequence of parameters is sampled from the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$, and a consistent estimate of $\boldsymbol{\theta}$ can be obtained by taking the mean of the sampled parameters.

Algorithm 4 The Gibbs sampler for mixture au Gaussian copula.

Starting from an initial value $\boldsymbol{\theta}^{(0)}$, its iteration (r) consists in the following four steps ($k \in \{1, \ldots, g\}, j \in \{1, \ldots, e\}$)

$$\mathbf{z}^{(r)}, \mathbf{y}^{(r-1/2)} \sim \mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)} \tag{3.15}$$

$$\boldsymbol{\beta}^{(r)}_{kj}, \boldsymbol{y}^{j(r)}_{[rk]} \sim \boldsymbol{\beta}_{kj}, \boldsymbol{y}^j_{[rk]}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\Gamma}^{(r-1)}_k \tag{3.16}$$

$$\boldsymbol{\pi}^{(r)} \sim \boldsymbol{\pi}|\mathbf{z}^{(r)} \tag{3.17}$$

$$\boldsymbol{\Gamma}^{(r)}_k \sim \boldsymbol{\Gamma}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \tag{3.18}$$

where $\boldsymbol{y}_{[rk]} = \boldsymbol{y}_{\{i:z^{(r)}_i=k\}}$, $\boldsymbol{y}^{\bar{j}(r)}_i = (y^{1(r)}_i, \ldots, y^{j-1(r)}_i, y^{j+1(r-1/2)}_i, \ldots, y^{e(r-1/2)}_i)$ and $\boldsymbol{\beta}^{(r)}_{k\bar{j}} = (\boldsymbol{\beta}^{(r)}_{k1}, \ldots, \boldsymbol{\beta}^{(r)}_{kj-1}, \boldsymbol{\beta}^{(r-1)}_{kj+1}, \ldots, \boldsymbol{\beta}^{(r-1)}_{ke})$.

The samplings according to (3.17) and (3.18) are classical but the sampling from (3.15) and (3.16) are not easy. They are therefore replaced by one iteration of a Metropolis-Hastings algorithm that does not change the stationary distribution. The resulting algorithm is a Metropolis-within-Gibbs sampler (Robert and Casella, 2004).

Figure 3.4 gives a good illustration of the proposed model thus for sake of brevity experiments on real data are not detailed here, but can be found in Marbac, Biernacki, and Vandewalle, 2017.

3.3.2 Conclusion

A Gaussian copula mixture model has been introduced and used to cluster mixed data. Using Gaussian copulas, the univariate marginal distributions of each component follow conventional distributions, and within-class dependencies are effectively modeled. Thus, the model results can be easily interpreted. Using the continuous latent variables of Gaussian copulas, a PCA-type method allows for component-based visualization of individuals. Moreover, this approach provides a summary of within-component dependencies, which avoids the tedious interpretation of correlation matrices.

The number of parameters increases with the number of components and number of variables, particularly due to the correlation matrices of the Gaussian copulas. To overcome this drawback, we have proposed a homoscedastic version of the model which assumes equality between correlation matrices.

Since the distribution of all the variables is modeled, this model could be used to manage data sets with missing values. By assuming that values are missing at

random, the Gibbs sampler could also be adapted, but the underlying principle would remain roughly the same.
Finally, the proposed model cannot cluster non-ordinal categorical variables having more than two modalities. In such cases, the cumulative distribution function is not defined. An artificial order between modalities could be added to define a cumulative distribution function, but this method presents three potential difficulties that require attention: it assumes regular dependencies between the modalities of two variables, its estimation would slow down the estimation algorithm, and its stability would have to be verified.
The **MixCluster** R package (`https://r-forge.r-project.org/R/?group_id=1939`) performs the cluster analysis method described here.

3.4 Conclusion

Using the model dependency models presented in this chapter is interesting to propose a meaningful interpretation of the data at hand. The proposed models enable more flexibility, and makes it possible to interpret the obtained clusters on their own without requiring to interpret some meta-cluster resulting from merging of several clusters obtained using full conditional independence assumption.
However, these models, stay very difficult to estimate in practice and the estimation process is time-consuming. Thus I think that these models stay limited to a moderated number of variables. For a higher number of variables, the conditional independence assumption stays the best tractable solutions, coupled with variables selection (Marbac and Sedki, 2017; Marbac, Patin, and Sedki, 2018) it has the advantage to deal with a very large number of variables, while variable selection limiting the noise on the partition caused by non-classifying variables. This assumption of conditional independence given the cluster is the main assumption for one of the multi-partition clustering models presented in Chapter 5.

Bibliography

Agresti, A. (2002). *Categorical data analysis*. Vol. 359. John Wiley and Sons.

Banfield, J. and Raftery, A. (1993). "Model-based Gaussian and non-Gaussian clustering". In: *Biometrics*, pp. 803–821.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Vol. 899. Wiley. com.

Celeux, G. and Govaert, G. (1991). "Clustering criteria for discrete data and latent class models". In: *Journal of classification* 8.2, pp. 157–176.

Espeland, M. and Handelman, S. (1989). "Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements". In: *Biometrics* 45.2, pp. 587–599.

Everitt, B. (1988). "A finite mixture model for the clustering of mixed-mode data". In: *Statistics & Probability Letters* 6.5, pp. 305–309.

Gollini, I. and Murphy, T. (2013). "Mixture of latent trait analyzers for model-based clustering of categorical data". In: *Statistics and Computing* 24.4, pp. 1–20. ISSN: 0960-3174. DOI: 10.1007/s11222-013-9389-1.

Goodman, L. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". In: *Biometrika* 61.2, pp. 215–231.

Gouget, C. (2006). *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne.

Hagenaars, J. (1988). "Latent structure models with direct effects between indicators local dependence models". In: *Sociological Methods & Research* 16.3, pp. 379–405.

Hand, D. and Keming, Y. (2001). "Idiot's Bayes, not so stupid after all?" In: *International statistical review* 69.3, pp. 385–398.

Hoff, P. (2007). "Extending the rank likelihood for semiparametric copula estimation". In: *The Annals of Applied Statistics*, pp. 265–283.

Hunt, L. and Jorgensen, M. (1999). "Theory & Methods: Mixture model clustering using the MULTIMIX program". In: *Australian & New Zealand Journal of Statistics* 41.2, pp. 154–171.

— (2011). "Clustering mixed data". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.4, pp. 352–361.

Joe, H. (1997). *Multivariate models and dependence concepts*. Vol. 73. CRC Press.

Joe, H. (2005). "Asymptotic efficiency of the two-stage estimation method for copula-based models". In: *Journal of Multivariate Analysis* 94.2, pp. 401–419.

Jorgensen, M. and Hunt, L. (1996). "Mixture model clustering of data sets with categorical and continuous variables". In: *Proceedings of the Conference ISIS*. Vol. 96, pp. 375–384.

Karlis, D. and Tsiamyrtzis, P. (2008). "Exact Bayesian modeling for bivariate Poisson data and extensions". In: *Statistics and Computing* 18.1, pp. 27–40.

Klaassen, C. and Wellner, J. (1997). "Efficient estimation in the bivariate normal copula model: normal margins are least favourable". In: *Bernoulli* 3.1, pp. 55–77.

Krzanowski, W. (1993). "The location model for mixtures of categorical and continuous variables". In: *Journal of Classification* 10.1, pp. 25–49.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2012). "Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library". In: *preprint submitted* in press.

Lewis, D. (1998). "Naive (Bayes) at forty: The independence assumption in information retrieval". In: *Machine learning: ECML-98*. Springer, pp. 4–15.

Marbac, M., Biernacki, C., and Vandewalle, V. (2013). *Model-based clustering for conditionally correlated categorical data*. Rapport de recherche RR-8232. INRIA, p. 34.

Marbac, M., Biernacki, C., and Vandewalle, V. (2015). "Model-based clustering for conditionally correlated categorical data". In: *Journal of Classification* 32.2, pp. 145–175.

— (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207.

— (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656.

Marbac, M., Patin, E., and Sedki, M. (2018). "Variable selection for mixed data clustering: a model-based approach". In: *Journal of Classification* to appear.

Marbac, M. and Sedki, M. (2017). "Variable selection for model-based clustering using the integrated complete-data likelihood". In: *Statistics and Computing* 27.4, pp. 1049–1063.

Morlini, I. (2012). "A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model". English. In: *Advances in Data Analysis and Classification* 6.1, pp. 5–28. ISSN: 1862-5347.

Moustaki, I. and Papageorgiou, I. (2005). "Latent class models for mixed variables with applications in archaeometry". In: *Computational Statistics & Data Analysis* 48.3, pp. 659–675. ISSN: 0167-9473.

Murray, J., Dunson, D., Carin, L., and Lucas, J. (2013). "Bayesian Gaussian copula factor models for mixed data". In: *Journal of the American Statistical Association* 108.502, pp. 656–665.

Nelsen, R. (1999). *An introduction to copulas*. Springer.

Olsson, U. (1979). "Maximum likelihood estimation of the polychoric correlation coefficient". In: *Psychometrika* 44.4, pp. 443–460.

Pitt, M., Chan, D., and Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models". In: *Biometrika* 93.3, pp. 537–554.

Qu, Y., Tan, M., and Kutner, M. (1996). "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests". In: *Biometrics* 52.3, pp. 797–810. ISSN: 0006341X.

Robert, C. (2005). *Le choix bayesien: principes et pratique*. Springer France Editions.

— (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

Smith, M. and Khaled, M. (2012). "Estimation of copula models with discrete margins via Bayesian data augmentation". In: *Journal of the American Statistical Association* 107.497, pp. 290–303.

Song, P. X.-K., Fan, Y., and Kalbfleisch, J. D. (2005). "Maximization by parts in likelihood inference". In: *Journal of the American Statistical Association* 100.472, pp. 1145–1158.

Van Hattum, P. and Hoijtink, H. (2009). "Market Segmentation Using Brand Strategy Research: Bayesian Inference with Respect to Mixtures of Log-Linear Models". In: *Journal of Classification* 26.3, pp. 297–328. ISSN: 0176-4268.

Vermunt, J. (2003). "Multilevel latent class models". In: *Sociological methodology* 33.1, pp. 213–239.

Willse, A. and Boik, R. (1999). "Identifiable finite mixtures of location models for clustering mixed-mode data". In: *Statistics and Computing* 9.2, pp. 111–121.

Chapter 4

Clustering of functional data

Contents

4.1 Introduction 54

4.2 Clustering categorical functional data 54

4.2.1 Introduction to categorical functional data 54

4.2.2 Extension of multiple correspondence analysis 55

4.2.3 Mixture of Markov processes 57

4.2.4 Conclusion and perspectives 59

4.3 Clustering of spatial functional data 60

4.3.1 Introduction 60

4.3.2 Model-based clustering for spatial functional data 61

4.3.3 Application 64

4.4 Conclusion and perspectives 67

Related scientific production

1. C. Dhaenens, J. Jacques, V. Vandewalle, M. Vandromme, E. Chazard, C. Preda, A. Amarioarei, P. Chaiwuttisak, C. Cozma, G. Ficheur, et al. (2018). "ClinMine: Optimizing the management of patients in hospital". In: *IRBM* 39.2, pp. 83–92

2. V. Vandewalle, C. Preda, and S. Dabo (2020). "Clustering spatial functional data". In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons

3. C. Preda, Q. Grimonprez, and V. Vandewalle (2020). "cfda: an R Package for Categorical Functional Data Analysis". working paper or preprint. URL: `https://hal.inria.fr/hal-02973094`

4.1 Introduction

From 2013 to 2017, I took part in the ANR ClinMine: 'Optimisation of Patient Care at the Hospital" project. In this ANR project, I worked with Cristian Preda among others to on the implementation of mixture models for the clustering of categorical functional data. The originality of our approach is to propose a Markov process mixture model with continuous time and discrete state space. In the same scope, I have worked with Cristian Preda and Quentin Grimonpretz on the extension of the functional PCA to the categorical functional data setting. I have contributed to the R package `cfda` (`https://github.com/modal-inria/cfda/`), and we have submitted an article related to this work at the Journal of Statistical Software (submitted version available at Preda, Grimonprez, and Vandewalle (2020)).
In the scope of functional data, I have also worked with Sophie Dabo and Cristian Preda to propose a model for clustering spatial functional data. The spatial information is taken into account in the prior membership probability which is assumed to depend on the position based on a logistic model. This has been published as a book Chapter (Vandewalle, Preda, and Dabo, 2020).
In Section 4.2, I present the issue of categorical functional data, in particular the question of the multivariate canonical analysis (MCA) for functional categorical data and the clustering based on a mixture of Markov model. In Section 4.3, I present our proposal for clustering spatial functional data.

4.2 Clustering categorical functional data

4.2.1 Introduction to categorical functional data

Most literature devoted to functional data considers data as sample paths of a real-valued stochastic process, $\boldsymbol{X} = \{X_t, t \in \mathcal{T}\}$, $X_t \in \mathbb{R}^p$, $p \geq 1$ where $\mathcal{T}$ is some continuous set. Among a considerable record of papers on the subject, the monographs of Ramsay and Silverman (2005) and Ramsay and Silverman (2002) and Ferraty and Vieu (2006) still remain references presenting the main methodologies for visualization, denoising, classification and regression when dealing with functional data represented by real-valued functions.
We consider the case where the underlying stochastic model generating the data is a continuous-time stochastic process $\boldsymbol{X} = \{X_t, t \in \mathcal{T}\}$ such that for all $t \in \mathcal{T}$, X_t is a categorical random variable rather than a real-valued one.
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{S} = \{s_1, \ldots, s_m\}$, $m \geq 2$, be a set of m states and

$$\boldsymbol{X} = \{X_t \ ; \ X_t : \Omega \longrightarrow \mathcal{S}, \quad t \in \mathcal{T}\} \tag{4.1}$$

be a family of categorical random variables indexed by $\mathcal{T}$. Thus, a path of $\boldsymbol{X}$ is a sequence of states s_{i_j} and times points t_i of transitions from one state to another one : $\{(s_{i_1}, t_1), (s_{i_2}, t_2), \ldots\}$, with $s_{i_j} \in \mathcal{S}$ and $t_i \in \mathcal{T}$.

We call the sample paths of the process (4.1) *categorical functional data*. The Figure 4.1 presents graphically scalar and categorical functional data.

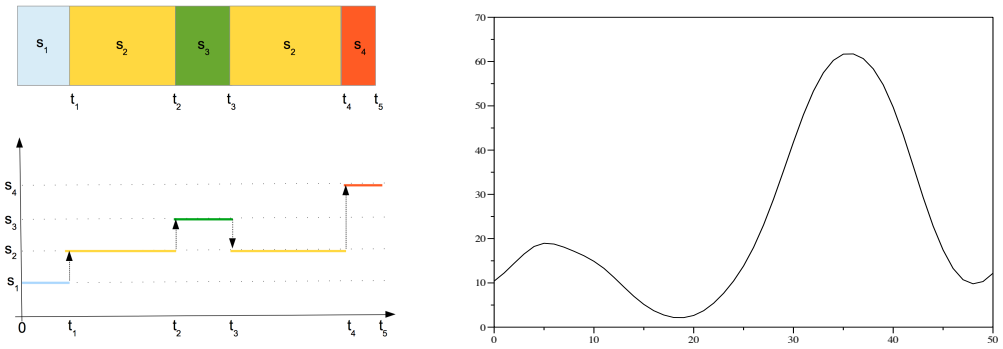


Figure 4.1: Examples of categorical (left) and scalar (right) functional data.

To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to this type of functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. As a start point in research on this topic we consider the works of Boumaza (1980), Deville (1982), Deville and Saporta (1983), Saporta (1981). These works are devoted to the extension of factorial techniques (canonical and multiple correspondences analysis) towards functional data. Applications of these techniques are presented in Heijden, Teunissen, and Orlé (1997) for analyzing career data and in Preda (1998) for studying spectral properties of the transition probability matrix of a the stationary Markovian jump process with continuous time.

4.2.2 Extension of multiple correspondence analysis

Optimal encoding Without loss of generality, let suppose that $\mathcal{T} = [0, T]$, with $T > 0$. For $x, y \in \mathcal{S}$ and $\forall t \in [0, T]$, let denote by:

- $\mathbf{1}_t^x = \begin{cases} 1 & \text{if } X_t = x, \\ 0 & \text{otherwise,} \end{cases}$
- $p^x(t) = p(X_t = x)$ and $p^{x,y}(t, s) = p(X_t = x, X_s = y)$.

The general hypotheses considered in that framework are:
H_1: the process $\boldsymbol{X}$ is continuous in probability,

$$\lim_{h \to 0} p(X_{t+h} \neq X_t) = 0$$

and

H_2: for each time $t \in [0,T]$ (except possibly a finite discrete set of time points), any state has a strictly positive probability to occur:

$$p^x(t) \neq 0, \forall x \in \mathcal{S}, \forall t \in [0,T].$$

In this framework, Saporta (1981) and Deville (1982) extend the multiple correspondence analysis to the process $\boldsymbol{X}$ (seen as infinite random variables). This is related the following egein-value problem

$$\int_0^T \sum_{y \in \mathcal{S}} p^{x,y}(t,s) a^y(s) ds = \lambda a^x(t) p^x(t), \quad \forall t \in [0,T], \forall x \in \mathcal{S}, \tag{4.2}$$

where $\{a^x\}_{x \in \mathcal{S}}$ are deterministic functions on $[0,T]$ that we call *optimal encoding* functions. Under the hypothesis H_1 and H_2 it admits the sequence of eigenvalues $\{\lambda_j\}_{j \geq 1}$ associated to the optimal encoding eigen-functions $\{a_j^x, x \in \mathcal{S}\}_{j \geq 1}$.
The j-th principal component y_j is derived from the j-th optimal encoding functions $\{a_j^x\}$ as

$$y_j = \int_0^T \sum_{x \in \mathcal{S}} a_j^x(t) \mathbf{1}_t^x dt, \quad \forall i \geq 1. \tag{4.3}$$

Dimension Reduction. Let $q \geq 1$, one obtains the best approximation of order q of $\boldsymbol{X}$ (viewed as a vector process $\boldsymbol{X} = \{\mathbf{1}^x, x \in \mathcal{S}\}$) under the L_2 norm, among all the linear expansions of type

$$\mathbf{1}_t^x \approx \sum_{j=1}^q z_j a_j^x(t) \frac{1}{p^x(t)}, \quad \forall x \in \mathcal{S}.$$

Thus, the q first principal components,

$$\{z_1, \ldots, z_q\}, \quad q \geq 1,$$

allow for

- graphical representation of sample paths of $\boldsymbol{X}$ in $\mathbb{R}^q$ (especially for $q = 2$, one obtains a 2-D representation of categorical functional data),
- fit of clustering and regression models with $\boldsymbol{X}$ as explanatory variables.

Discussion Technical details are not given here but can be found in Preda, Grimonprez, and Vandewalle (2020). The main idea is to consider and expansion of the $\{a^x\}_{x \in \mathcal{S}}$ on some basis, thus limiting the problem to some finite dimension problem thus solving a classical eigen-problem. One major interest for such dimension reduction is that is permits to consider data in $\mathbb{R}^q$ rather that the initial functional categorical data. Thus, it gives a first solution to perform clustering. The `cfda` R package (available on GitHub `https://github.com/modal-inria/cfda`) allows to perform this extension of MCA.

This approach has however some limitations, for instance, the paths need to have the same length. Moreover from a model-based clustering point of view, this needs some pre-processing before clustering, losing maybe some important information from the clustering point of view. Some interesting questions could be to investigate model-based clustering embedded in dimension reduction such as mixtures of factor analyzers, thus assuming some sparse model on the distribution of $\boldsymbol{X}$ given the cluster. In the next section, I present some work that we have performed with Cristian Preda by assuming some specific distribution on the process in each group, while it was supposed totally general in the previous section.

4.2.3 Mixture of Markov processes

4.2.3.1 State of the art

Clustering of categorical functional data is not so new, it has been considered in Blumen, Kogen, and McCarthy (1955) and Goodman (1961) for industrial mobility data, where it was assumed that data contained two groups; the movers and the stayers. To deal with such data, most approaches only consider discrete time Markov process such as Fougère and Kamionka (2003), Cadez et al. (2003), and Frühwirth-Schnatter and Pamminger (2009), however, this can lead to losing information about the time spent in each state. Continuous time Markov processes for clustering has been considered in Frydman (2005), by assuming constraints between the infinitesimal generator of the Markov processes such that only the speeds of transition change from a group to another. More recently Cardot et al. (2019) have proposed a mixture of semi-Markov processes with application to sensory data.

4.2.3.2 Motivation in the scope on the ClinMine project

The application which has motivated our work was in the scope of the ClinMine project where we have been interested in using data to improve patient care at the hospital. One possible way to improve patient care is to identify patients with similar paths. However, defining such paths from the application point of view is a difficult challenge. For instance, our first idea was to focus on some variable called homogeneous group of patients (some particular disease for instance), however, there are possibly 2000 different homogeneous groups of patients, and even reducing it to values lower than 100 was not relevant. An important point when working on categorical functional data is that the number of possible states must not be too high, otherwise, it may cause estimation problems as for the estimation of transition probabilities. Being able to deal with a high number of states is still a challenging task.
We have decided to focus on the state of medical discharge letters according to the time, these data were available thanks to the automatic letter dictation system. From the statistical point of view, the number of states stays limited (only eight) and thus tractable, while from the medical point of view it can provide important information since a discharge letter is mandatory for the patient to go home. Clus-

tering such types of data could help to improve patient management by detecting some atypical clusters. Thus identifying different groups of paths in these discharge letters may help the hospital to detect some problems such as a too long time in some state for instance, or some strange transition.
The data are provided from the GHICL hospital group in the scope of the ClinMine ANR project. The goal is to cluster patients' stays according to the evolution of the status of their discharge letters over time. The discharge letter can pass through eight different states:

1. the doctor is dictating the letter.
2. the letter is "waiting" to be type-writing by an assistant
3. the letter is type-writing by the assistant
4. the letter is "waiting" for doctor validation
5. the letter is in validation process by the doctor
6. the letter is "waiting" to be affected to an assistant
7. the letter is treated by the assistant
8. the letter is sent to the patient (end).

4.2.3.3 The model

Data Let take the n sample paths : $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, with the path i is $\boldsymbol{x}_i = (\mathbf{s}_{i0}, t_{i0}, \mathbf{s}_{i1}, t_{i1} \ldots, t_{i(d_i-1)}, \mathbf{s}_{id_i})$ where d_i is the number of jumps of the path i, t_{ij} the length of time spent in the j^{th} visited state of path i and $\mathbf{s}_{ij} = (s_{ijh}, \ldots, s_{ijm})$ the binary coding of the j^{th} state from the path i. It is supposed that the paths are uncensored, *i.e.* the paths are observed until they have reached the absorbing state. Assuming that the n paths come from g different processes characterized by parameters $\boldsymbol{\theta}_k$ ($k \in \{1, \ldots, g\}$). The likelihood function for the path i coming from cluster k is:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = p(\mathbf{s}_{i0}|\boldsymbol{\theta}_k)p(t_{i0}, \mathbf{s}_{i1}|\mathbf{s}_{i0}; \boldsymbol{\theta}_k) \prod_{j=1}^{d_i-1} p(t_{ij}, \mathbf{s}_{i(j+1)}|\mathbf{s}_{ij}, t_{i(j-1)}, \ldots, \mathbf{s}_{i1}, t_{i0}, \mathbf{s}_{i0}; \boldsymbol{\theta}_k)$$

Markovian assumptions The Markovian assumptions are the following:

H1: The distribution of $(t_{ij}, \mathbf{s}_{i(j+1)})$ is independent of the past given $\mathbf{s}_{ij}$

$$p(t_{ij}, \mathbf{s}_{i(j+1)}|\mathbf{s}_{ij}, t_{i(j-1)}, \ldots, \mathbf{s}_{i1}, t_{i0}, \mathbf{s}_{i0}; \boldsymbol{\theta}_k) = p(t_{ij}, \mathbf{s}_{i(j+1)}|\mathbf{s}_{ij}; \boldsymbol{\theta}_k)$$

H2: The distributions of t_{ij} and $\mathbf{s}_{i(j+1)}$ are independent given $\mathbf{s}_{ij}$

$$p(t_{ij}, \mathbf{s}_{i(j+1)}|\mathbf{s}_{ij}; \boldsymbol{\theta}_k) = p(t_{ij}|\mathbf{s}_{ij}; \boldsymbol{\theta}_k)p(\mathbf{s}_{i(j+1)}|\mathbf{s}_{ij}; \boldsymbol{\theta}_k)$$

H3: The distribution of t_{ij} given s_{ij} is an exponential distribution

H4: The distribution of the initial state does not depends on the cluster

$$p(\mathbf{s}_{i0}|\boldsymbol{\theta}_k) = p(\mathbf{s}_{i0})$$

Thus

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = p(\mathbf{s}_{i0}) \prod_{j=0}^{d_i-1} \underbrace{p(t_{ij}|\mathbf{s}_{ij};\boldsymbol{\theta}_k)}_{\text{time}} \underbrace{p(\mathbf{s}_{i(j+1)}|\mathbf{s}_{ij};\boldsymbol{\theta}_k)}_{\text{transition}}.$$

We see that the pdf of a path given the cluster can be decomposed into two parts, the first one about the time spent in each state, and the other one about the transitions between states. Thus this will allow performing a trade-off between these two aspects of the continuous time Markov process when performing the clustering. The parameters of cluster k: $\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k, \boldsymbol{\lambda}_k)$, can be decomposed in two parts, $\boldsymbol{\alpha}_k$ the transition probabilities matrix and $\boldsymbol{\lambda}_k$ the parameters of the distribution of the time in each state. Parameters estimation can simply be performed by a standard EM algorithm with closed-form at M step, and model choice can be performed by using the BIC criterion.

Illustration of clustering of medical discharge letters Let considers the paths of 443 325 discharge letters. Applying clustering in two clusters we have $\hat{\pi}_1 = 0.897$, $\hat{\pi}_2 = 0.103$, mean sojourn time is given in Table 4.1, transition probabilities given in Table 4.2.

state	1	2	3	4	5	6	7
cluster 1	288	290460	1136	373567	569	131702	712
cluster 2	863390	268556	215645	408716	380294	217268	48815

Table 4.1: Mean sojourn time (in seconds) in each state for each cluster

The main difference between the clusters is the sojourn time distribution. The second cluster is characterized by long sojourn times. The transition probabilities are roughly the same except for the transition from state 1 to state 8.

4.2.4 Conclusion and perspectives

The proposed mixture of Markov processes allows to take into account probabilities of transition and sojourn times related to categorical data, they also permit to take into account paths of different length. However, we were not yet able to prove the identifiability of such a model from a general point of view, especially we see on the medical discharge letters that in practice cases will null transition probabilities must be addressed.
Finding a model-based interpretation of the categorical functional data analysis presented in Section 4.2.2 is an interesting perspective of work, it would permit to perform clustering of categorical functional data from a more general point of view.

from \ to	2	3	4	5	6	7	8
1	0.20						0.80
2		0.96	0.03				
3			0.99		0.01		
4				0.87	0.06		0.07
5					0.83	0.01	0.16
6						0.96	0.04
7							1.00

(a) Transition probabilities in cluster 1

from \ to	2	3	4	5	6	7	8
1	0.38						0.62
2		0.98	0.02				
3			0.99				
4				0.88	0.08		0.04
5					0.80	0.03	0.17
6						0.97	0.03
7							1.00

(b) Transition probabilities in cluster 2

Table 4.2: Transition probabilities in each cluster

4.3 Clustering of spatial functional data

In this section, I present a model for clustering spatial functional data. This approach mainly relies on a surrogate density for functional random variables while taking into account the spatial features of the data: two observations that are spatially close share a common distribution of the associated random variables. More precisely we assume a spatial model for the prior weights of the mixture. This approach is illustrated by an application to air quality data. This work has been published in Vandewalle, Preda, and Dabo (2020), here we only focus on the model-based clustering part of the book chapter.

4.3.1 Introduction

Recent researches on the clustering of independent functional data are available in the literature devoted to functional data analysis (FDA). In particular, k-means techniques are adjusted to functional data, hierarchical algorithm and some of its variants are proposed as well, mainly for independent data (e.g Abraham, Biau, and Cadre (2006), Dabo-Niang, Ferraty, and Vieu (2007), Auder and Fischer (2012), Abraham, Cornillon, et al. (2003), Chiou and Li (2007), Cuevas, Febrero, and Fraiman (2001), Dabo-Niang, Ferraty, and Vieu (2007), García-Escudero and Gordaliza (2005), Romano, Mateu, and Giraldo (2015), Tarpey and Kinateder (2003), and Jacques and Preda (2014b). A revue of clustering methods for functional data under the independent model is provided in Jacques and Preda, 2014a. Other model-based approaches for clustering functional data are given in Floriello and Vitelli (2017), James and Sugar, 2003. In several domains, data are of spatio-functional nature,

observations may be dependent curves at some spatial locations, and clustering these data taking into account the spatial dependency can be more accurate. The independence hypothesis does not hold in this case. Few works exist on such dependent data: Dabo-Niang, Yao, et al. (2010) and Giraldo, Delicado, and Mateu (2012) extended some approaches on hierarchical clustering to the context of spatially correlated functional. Giraldo, Delicado, and Mateu (2012) measured the similarity between two curves by the trace-variogram (Giraldo, Delicado, and Mateu (2011)) while the spatial variation is taken into account by using kernel mode and density estimation in Dabo-Niang, Yao, et al. (2010). Other approaches for clustering spatial functional data are given recently in Romano, Mateu, and Giraldo (2015) and Romano, Balzanella, and Verde (2017).
We deal with a measurable spatial process $\boldsymbol{X} = \left(\boldsymbol{X}_s,\ s \in \mathbb{R}^N\right),\ N \geq 1$, defined on some probability space $(\Omega,\ \mathcal{A}, \mathbf{P})$. Assume that the process $\boldsymbol{X}$ is observed on some spatial region $\mathcal{I} \subseteq \mathbb{R}^N$ of cardinal n, $\mathcal{I} = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^N$, $i = 1 \ldots n$. We assume also that for each location $s \in \mathcal{I}$, the random variables X_s are valued in a metric space $(\mathcal{E},\ d)$ of eventually infinite dimension and are locally identically distributed (see for instance Klemelä, 2008). Here $d(.,.)$ is some measure of proximity, for instance, a metric or a semi-metric. This means that when a site u is close enough to site v, the variables $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ have the same distribution. This assumption is less restrictive than strict stationarity. It is motivated by the fact that one can imagine that, variables located at neighbors sites may be similar and have the same local distribution that may be different from the local distribution of another set of variables at other locations. In the classical framework of FDA, the space $\mathcal{E}$ is a space of functions, typically the space of squared integrable functions defined on some finite interval $\mathcal{T} = [0, T]$, $T > 0$.
Let denote with S the set of the n curves, $S = \{\boldsymbol{X}_s, s \in \mathcal{I}\}$ (renamed sometime in an arbitrary way, $S = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$).

4.3.2 Model-based clustering for spatial functional data

In a spatial dependency framework we model the distribution of $\boldsymbol{X}_s$ as a mixture of distribution. The probability density function of $\boldsymbol{X}_s$ evaluated in $\boldsymbol{x}$ is written as follows

$$p(\boldsymbol{x}|s;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_g(s;\boldsymbol{\beta}) p(\boldsymbol{x}|\boldsymbol{\theta}_k), \tag{4.4}$$

where $\pi_k(s;\boldsymbol{\beta}) = P(Z_k = 1|s;\boldsymbol{\beta})$ is the spatial prior and β its related parameters. Thus, conditionally to the cluster $Z_k = 1$, the distribution of observations within the cluster is independent of the location s, meaning that all spatial dependency is captured by the spatial prior $\pi_k(s;\boldsymbol{\beta})$. This idea is used in Cheam, Marbac, and McNicholas, 2017 for clustering spatio-temporal data. The authors propose the multinomial logistic regression as a model for the $\pi_k(s;\boldsymbol{\beta})$,

$$\ln \frac{\pi_k(s;\boldsymbol{\beta})}{\pi_g(s;\boldsymbol{\beta})} = \beta_{0k} + \langle \boldsymbol{\beta}_k, s \rangle_{\mathbb{R}^N}. \tag{4.5}$$

For functional random variables the notion of probability density in not well defined because of the infinite dimension of data. To overcome this difficulty, in James and Sugar (2003) and Bouveyron and Jacques (2011) use the expansion coefficients of $\boldsymbol{X}$ into some finite basis of functions. This approach allows them to get a well defined probability density function on the coefficients. In Delaigle and Hall (2010) the functional principal component analysis is used to define a surrogate of the probability density for functional data. This approach is used in the context of model based clustering in Jacques and Preda (2013) and Jacques and Preda (2014b). In a spatial setting Ruiz-Medina, Espejo, and Romano (2014) have proposed a mixed-effect model, in which the fix effect can take into account the spatial dependencies. Moreover, assuming a spatial autoregressive dynamic for the random effect, they propose a functional classification criterion to detect local spatially homogeneous regions. In what follows we assume that given $Z_k = 1$, $\boldsymbol{X}$ is a Gaussian process. Then, within the cluster k, we consider a modified version of the pseudo-density defined in Delaigle and Hall (2010):

$$p^{(q_k)}(\boldsymbol{x}|\boldsymbol{\theta}_k) = \prod_{j=1}^{q_k} p(c_{kj}(\boldsymbol{x})|\lambda_{kj}) \prod_{j'=q_k+1}^{d} p(c_{kj'}(\boldsymbol{x})|\bar{\lambda}_k), \tag{4.6}$$

where $p(\cdot|\lambda_{kj})$ is the probability density of the j-th principal component C_{kj} of $\boldsymbol{X}$ within the cluster k. The random variables C_{kj} $(j = 1, \ldots, q_k)$ are independent Gaussian zero-mean with variance equal to the eigen values λ_{kj} of the covariance operator of $\boldsymbol{X}$, and the random variables $C_{kj'}$ $(j' = q_k + 1, \ldots, d)$ are independent Gaussian zero-mean with variance equal to the mean $\bar{\lambda}_k$ of the eigen values $\lambda_{kj'}$ $(j' = q_k + 1, \ldots, d)$ of the covariance operator of $\boldsymbol{X}$. Thus the parameters $\boldsymbol{\theta}_k = (\lambda_{k1}, \ldots, \lambda_{kq_k}, \bar{\lambda}_k)$, q_k and d need to be defined. Let notice that compared to the definition of Delaigle and Hall (2010), we have added the term $\prod_{j'=q_k+1}^{d} p(c_{kj'}(\boldsymbol{x})|\bar{\lambda}_k)$.
The proposed surrogate density car be interpreted as a true density if the functional data belong to a finite dimensional space of functions spanned by some basis $\{\phi_1, \ldots, \phi_d\}$, $d \geq 1$, i.e.

$$X(t) = \sum_{j=1}^{d} \alpha_j \phi_j(t), \quad t \in [0, T], T > 0.$$

Thus we will take d as the dimension of the basis which has been used to perform the smoothing of the data. In this case, the principal components C_{kj} of the functional PCA can be obtained by performing PCA on the expansion coefficients of $\boldsymbol{X}$ in the metric M given by the inner product of the basis functions. Thus, if the learning data considered are now the expansion coefficients multiplied by $M^{1/2}$ then the proposed approach can simply be re-interpreted as learning a parsimonious high dimensional model (see Bouveyron, Girard, and Schmid (2007)) on these new data. Let notice that it is also possible to consider sparse versions of the mixture model such as for the homoscedastic setting (equal covariance process by cluster).

4.3.2.1 The EM algorithm

We are now ready to describe the EM algorithm for estimating $\boldsymbol{\theta}$ and therefore the clustering.
The EM algorithm is the same as the one presented in Chapter 2 expected that the spatial prior needs to be taken into account in the E step, and the parameters of spatial priors need to be updated at M step. At iteration (r)

E step. The E step consists in updating the posterior probabilities using the spatial priors:

$$t_{sk}(\boldsymbol{\theta}^{(r+1)}) = \frac{\pi_k(s;\boldsymbol{\beta}^{(r)})p^{(q_k)}(\boldsymbol{x}_s|\boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^{g}\pi_{k'}(s;\boldsymbol{\beta}^{(r)})p^{(q_{k'})}(\boldsymbol{x}_s|\boldsymbol{\theta}_{k'}^{(r)})}.$$

M step. The M step needs to also update the parameters of each cluster, for all k in $\{1,\ldots,g\}$

$$\boldsymbol{\theta}_k^{(r+1)} = \arg\max_{\boldsymbol{\theta}_k}\sum_{s\in\mathcal{I}} t_{ik}(\boldsymbol{\theta}^{(r)})\ln p^{(q_k)}(\boldsymbol{x}_s;\boldsymbol{\theta}_k) \quad \text{and} \quad \pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{(r)})}{n},$$

and the parameters related to the spatial priors

$$\boldsymbol{\beta}^{(r+1)} = \arg\max_{\boldsymbol{\beta}}\sum_{s\in\mathcal{I}}\sum_{k=1}^{g} t_{sk}(\boldsymbol{\theta}^{(r+1)})\log\pi_k(s;\boldsymbol{\beta}), \tag{4.7}$$

which can be obtained by a weighted multinomial logistic regression.

If parsimonious models such as homoscedastic models are considered, this leads to a modification of the update of $\boldsymbol{\theta}_k^{(r+1)}$, see Bouveyron, Girard, and Schmid (2007) for more details.

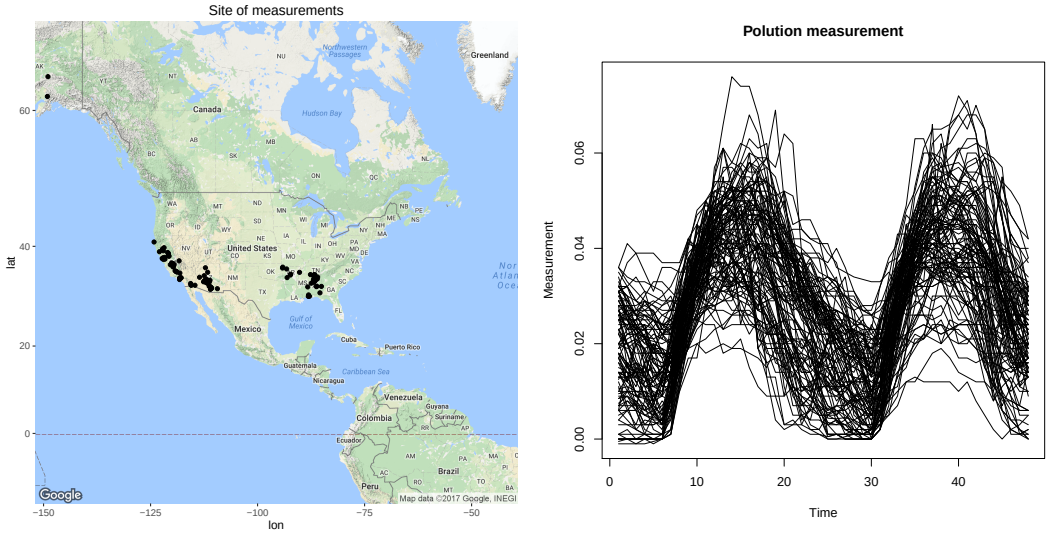
Model selection In order to select the number of cluster g when q_k $(k=1,\ldots,g)$ are known, we propose to maximize the BIC criterion.
When the values q_k $(k=1,\ldots,g)$ are unknown they can be selected in order to maximize the BIC criterion by considering the following modified M step which tries to maximize the conditional expectation of the BIC criterion:

$$(q_k^{(h+1)},\boldsymbol{\theta}_k^{(h+1)}) = \arg\max_{(q_k,\boldsymbol{\theta}_k)}\textstyle\sum_{s\in\mathcal{I}} t_{sk}(\boldsymbol{\theta}^{(h+1)})\ln p^{(q_k)}(\boldsymbol{x}_s;\boldsymbol{\theta}_k) - \frac{\nu_{q_k}}{2}\log n,$$

where $\nu_{q_k} = q_k(d-(q_k-1)/2)$ is the additional number of parameters required for the model with q_k principal components.
Let notice that if we consider the homoscedastic setting, the value of the BIC criterion can be easily computed at each step of the EM algorithm for each possible value of q which does not depends on g since in this case, this parameter is the same for each cluster.

4.3.3 Application

We illustrate our methodology using the ozone concentration data. The data is constituted of 48 pollution measurements (one per hour) on 108 US cities. The cities considered are mainly on the west and the east coast, and there are also two cities of Alaska which are considered (see Figure 4.2). The considered data have globally similar latitudes (except Alaska) but different longitude. From the spatial point of view, the city can be divided into three main clusters.



(a) Localization of the site of measurement

(b) Pollution curves

Figure 4.2: Example of data visualization.

We are given the ozone concentration for these 108 stations for every hour from 12am July 19 to 11pm July 20, 2015 (that is, 48 days). Since some of the stations had missing values, we use linear interpolation to estimate the missing values. We have implemented the above classification procedures. We denote the ozone concentration at time t by $X(t)$ where t refers to the day and the hour of observation. We suppose that $X(t)$ is observed for $t \in [1; 48)$ at each station (48 hours $\times$ 108 stations), see Figure 4.2. To apply the functional methodology, we organize the original space-time series into a set of daily functional data.

4.3.3.1 Results

To apply the EM algorithm for clustering spatial functional data presented in Section 4.3.2, we represented the curves using a Fourier basis of size $d = 25$. We have also removed the Alaska data since its extreme geographical position disturbs the results. A homoscedastic model has been applied since it gives more relevant results from the

classification point of view, and the value of q was selected during the EM algorithm by maximizing the BIC criterion computed at each step for each possible value of q. Under this setting, the BIC criterion (Figure 4.3 indicates that two ($g = 2$) or three ($g = 3$) clusters can be an appropriate choice of the number of clusters.

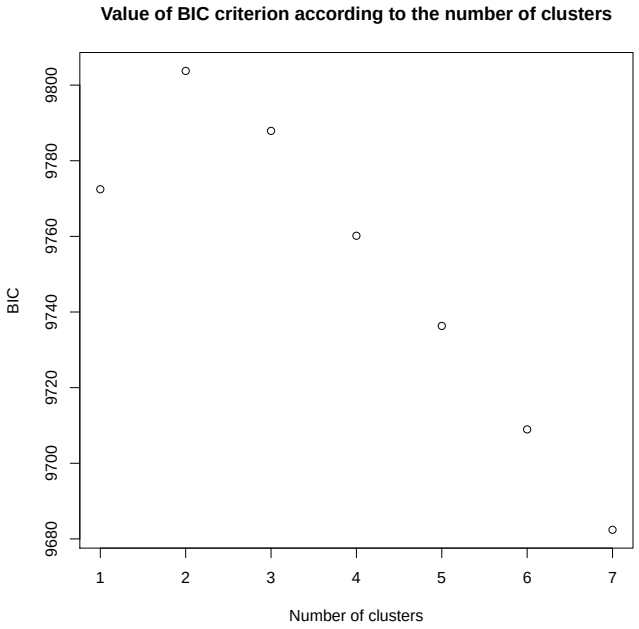


Figure 4.3: Choice of the number of cluster : BIC criterion

Figure 4.4 presents the clustering in two and respectively three clusters. For the clustering in two clusters, $q = 18$ principal components have been retained. We see on the map that the obtained clustering well separates the East cities from the West cities. Moreover, we see on the curves that the clusters are also well separated from the curves' point of view. On average we see in Figure 4.5 that West cities have higher pollution than Est cities. Mean curves per cluster are obtained based on the estimated means of the coefficients of the basis expansion. For the clustering in three clusters, $q = 17$ principal components have been retained. We see on the map that the obtained clustering still well separates the East curves from the West curves, but also the North from the South for the West side. When looking at these curves in Figure 4.5, we see that it is the six first hours that well separate cluster 1 (North) from cluster 3 (South).

As a conclusion of this application, for the clusterings (with two or three clusters) let observe that the method makes a trade-off between the geographical proximity and the common features of the curves, which allows taking into account spatial dependency while performing clustering. We see on the application that the obtained results are easily interpretable, and give a relevant spatial segmentation.

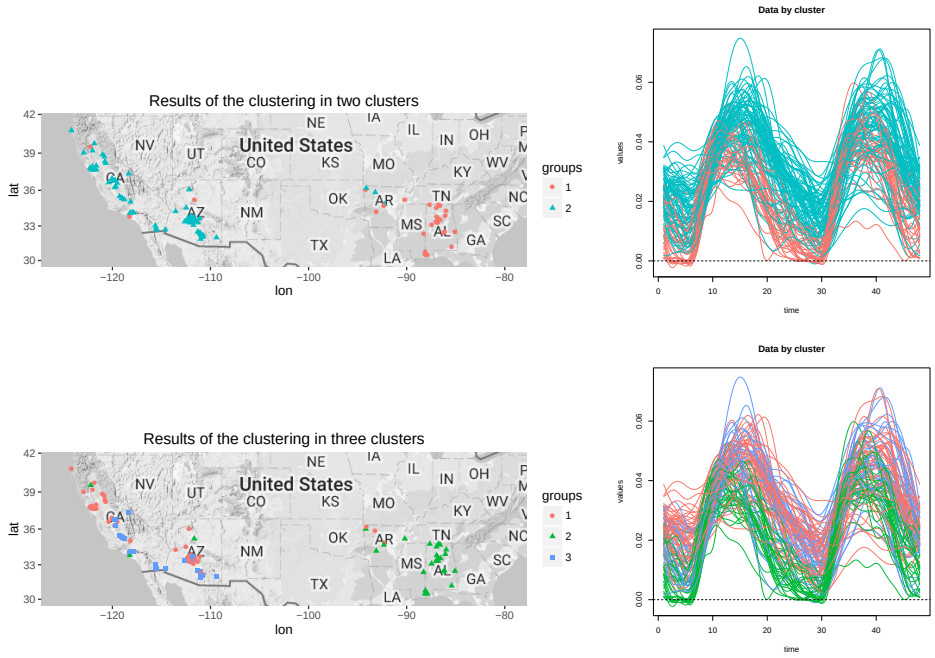


Figure 4.4: Two clusters (above) ; Three clusters (bellow)

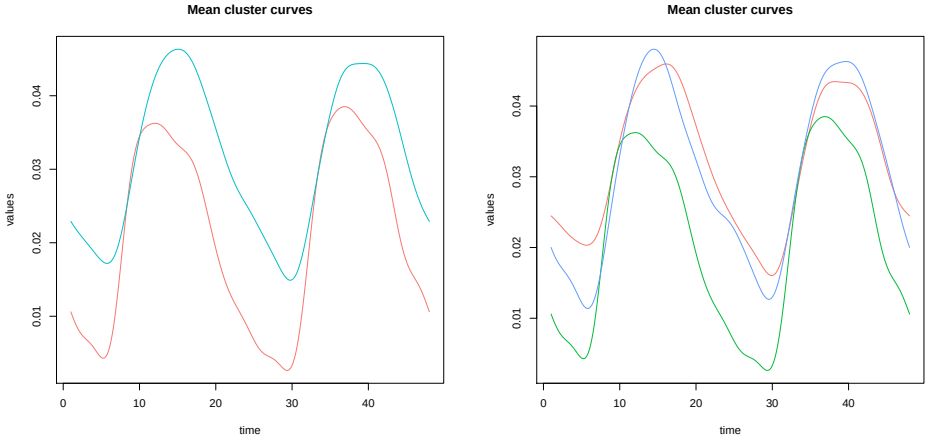


Figure 4.5: Two clusters (left) ; Three clusters (right)

4.4 Conclusion and perspectives

Two contributions to functional data have been presented. The part concerning the mixture of Markov processes is still a work in progress.
Data availability over time makes the issue of functional data analysis more and more up to date. Thus using and developing approaches in this context is a promising issue. From a statistical point of view, functional data are rarely directly observed since they are only observed in a finite set of points. This has the advantage of considering a finite dimensional setting making it possible to use standard mixture model tools such as in Samé et al. (2011). This may be not possible when the number of observed times being huge. In this case, the question of decomposing the data on some basis, thus going back to the finite dimensional setting is one of the most tractable solutions. However, the choice of the basis and its size stays an important issue.
Working on the ClinMine project has revealed to me the real challenge of studying the patient path in the hospital. It cannot be summarized by considering only one categorical variable over time, but also needs to consider a huge number of variables varying according to time. For being useful such an approach cannot be built based on the whole data available at the hospital, but rather should be focused on a particular case study, potentially usable in many other settings. It would be very exciting to work on such an issue while asking the question of global patient care, and to see how statistical models can bring some insights about interpreting those "complex data". From my work in usability (see Chapter 7), it would also be interesting to bring this in a cost-sensitive way to link the statistical model with the decision making process. More discussions are given in Chapter 10.

Bibliography

Abraham, C., Biau, G., and Cadre, B. (2006). "On the kernel rule for function classification". In: *Ann. Inst. Statist. Math.* 58.3, pp. 619–633. ISSN: 0020-3157.

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). "Unsupervised curve clustering using B-splines". In: *Scandinavian journal of statistics* 30.3, pp. 581–595.

Auder, B. and Fischer, A. (2012). "Projection-based curve clustering". In: *J. Stat. Comput. Simul.* 82.8, pp. 1145–1168. ISSN: 0094-9655.

Blumen, I., Kogen, M., and McCarthy, P. (1955). "The industrial mobility of workers as a probability process, vol. 6 of Cornell studies in industrial and and labor relations". In: *Cornell University, Ithaca, NY*.

Boumaza, R. (1980). "Contribution à l'étude descriptive d'une fonction aleatoire qualitative: these présentée a l'Université Paul Sabatier de Toulouse pour obtenir le grade de docteur de spécialité mathématiques appliquées". PhD thesis. Université Paul Sabatier.

Bouveyron, C., Girard, S., and Schmid, C. (2007). "High-Dimensional Data Clustering". In: *Computational Statistics and Data Analysis* 52.1, pp. 502–519.

Bouveyron, C. and Jacques, J. (2011). "Model-based clustering of time series in group-specific functional subspaces". In: *Advances in Data Analysis and Classification* 5.4, pp. 281–300.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2003). "Model-based clustering and visualization of navigation patterns on a web site". In: *Data mining and knowledge discovery* 7.4, pp. 399–424.

Cardot, H., Lecuelle, G., Schlich, P., and Visalli, M. (2019). "Estimating Finite Mixtures of Semi-Markov Chains: an Application to the Segmentation of Temporal Sensory Data". In: *Journal of the Royal Statistical Society C* 68.5, pp. 1281–1303. DOI: 10.1111/rssc.12356.

Cheam, A., Marbac, M., and McNicholas, P. D. (2017). "Model-based clustering for spatiotemporal data on air quality monitoring". In: *Environmetrics* 28.3.

Chiou, J.-M. and Li, P.-L. (2007). "Functional clustering and identifying substructures of longitudinal data". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69.4, pp. 679–699. ISSN: 1369-7412.

Cuevas, A., Febrero, M., and Fraiman, R. (2001). "Cluster analysis: a further approach based on density estimation". In: *Comput. Statist. Data Anal.* 36.4, pp. 441–459. ISSN: 0167-9473.

Dabo-Niang, S., Ferraty, F., and Vieu, P. (2007). "On the using of modal curves for radar waveforms classification". In: *Comput. Statist. Data Anal.* 51.10, pp. 4878–4890. ISSN: 0167-9473.

Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., and Gilbert, F. (2010). "Spatial mode estimation for functional random fields with application to bioturba-

tion problem". In: *Stochastic Environmental Research and Risk Assessment* 24.4, pp. 487–497.

Delaigle, A. and Hall, P. (2010). "Defining probability density for a distribution of random functions". In: *The Annals of Statistics*, pp. 1171–1193.

Deville, J.-C. and Saporta, G. (1983). "Correspondence analysis, with an extension towards nominal time series". In: *Journal of econometrics* 22.1-2, pp. 169–189.

Deville, J.-C. (1982). "Analyse de Données Chronologiques Qualitatives : Comment Analyser des Calendriers ?" In: *Annales de l'INSEE* 45, pp. 45–104. DOI: 10.2307/20076433.

Dhaenens, C., Jacques, J., Vandewalle, V., Vandromme, M., Chazard, E., Preda, C., Amarioarei, A., Chaiwuttisak, P., Cozma, C., Ficheur, G., et al. (2018). "ClinMine: Optimizing the management of patients in hospital". In: *IRBM* 39.2, pp. 83–92.

Ferraty, F. and Vieu, P. (2006). *Nonparameric Functional Data Analysis. Theory and Practice*. Springer-Verlag New York. ISBN: 978-0-387-36620-3. DOI: 10.1007/0-387-36620-2.

Floriello, D. and Vitelli, V. (2017). "Sparse clustering of functional data". In: *J. Multivariate Anal.* 154, pp. 1–18. ISSN: 0047-259X.

Fougère, D. and Kamionka, T. (2003). "Bayesian inference for the mover–stayer model in continuous time with an application to labour market transition data". In: *Journal of Applied Econometrics* 18.6, pp. 697–723.

Frühwirth-Schnatter, S. and Pamminger, C. (2009). *Bayesian clustering of categorical time series using finite mixtures of Markov chain models*. Tech. rep. NRN Working Paper, NRN: The Austrian Center for Labor Economics and the ...

Frydman, H. (2005). "Estimation in the mixture of Markov chains moving with different speeds". In: *Journal of the American Statistical Association* 100.471, pp. 1046–1053.

García-Escudero, L. A. and Gordaliza, A. (2005). "A proposal for robust curve clustering". In: *Journal of Classification* 22.2, pp. 185–201. ISSN: 0176-4268.

Giraldo, R., Delicado, P., and Mateu, J. (2011). "Ordinary kriging for function-valued spatial data". In: *Environ. Ecol. Stat.* 18.3, pp. 411–426. ISSN: 1352-8505.

Giraldo, R., Delicado, P., and Mateu, J. (2012). "Hierarchical clustering of spatially correlated functional data". In: *Statistica Neerlandica* 66.4, pp. 403–421.

Goodman, L. A. (1961). "Statistical methods for the mover-stayer model". In: *Journal of the American Statistical Association* 56.296, pp. 841–868.

Heijden, P., Teunissen, J., and Orlé, C. van (1997). "Multiple Correspondence Analysis as a Tool for Quantification or Classification of Career Data". In: *Journal of Educational and Behavioral Statistics* 22, pp. 447–477. DOI: 10.3102/10769986022004447.

Jacques, J. and Preda, C. (2013). "Funclust: A curves clustering method using functional random variables density approximation". In: *Neurocomputing* 112, pp. 164–171.

— (2014a). "Functional data clustering: a survey". In: *Advances in Data Analysis and Classification* 8.3, pp. 231–255.

— (2014b). "Model-based clustering for multivariate functional data". In: *Computational Statistics and Data Analysis* 71, pp. 92–106.

James, G. M. and Sugar, C. A. (2003). "Clustering for sparsely sampled functional data". In: *Journal of the American Statistical Association* 98.462, pp. 397–408.

Klemelä, J. (2008). "Density estimation with locally identically distributed data and with locally stationary data". In: *J. Time Ser. Anal.* 29.1, pp. 125–141. ISSN: 0143-9782.

Preda, C. (1998). "Analyse Harmonique Qualitative des Processus Markoviens des Sauts Stationnaires". In: *Scientific Annals of Computer Science* 7, pp. 5–18. URL: https://www.info.uaic.ro/en/sacs_articles/analyse-harmonique-qualitative-des-processus-markoviens-des-sauts-stationnaires/.

Preda, C., Grimonprez, Q., and Vandewalle, V. (2020). "cfda: an R Package for Categorical Functional Data Analysis". working paper or preprint. URL: https://hal.inria.fr/hal-02973094.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis*. Springer Series in Statistics. Methods and case studies. Springer-Verlag, New York, pp. x+190. ISBN: 0-387-95414-7.

— (2005). *Functional Data Analysis*. Springer-Verlag New York. ISBN: 978-0-387-22751-1. DOI: 10.1007/b98888.

Romano, E., Balzanella, A., and Verde, R. (2017). "Spatial variability clustering for spatially dependent functional data". In: *Stat. Comput.* 27.3, pp. 645–658. ISSN: 0960-3174.

Romano, E., Mateu, J., and Giraldo, R. (2015). "On the performance of two clustering methods for spatial functional data". In: *AStA Adv. Stat. Anal.* 99.4, pp. 467–492. ISSN: 1863-8171.

Ruiz-Medina, M. D., Espejo, R. M., and Romano, E. (2014). "Spatial functional normal mixed effect approach for curve classification". In: *Advances in Data Analysis and Classification* 8.3, pp. 257–285.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). "Model-based clustering and segmentation of time series with changes in regime". In: *Advances in Data Analysis and Classification* 5.4, pp. 301–321.

Saporta, G. (1981). *Méthodes Exploratoires d'Analyse de Données Temporelles*. Université Pierre et Marie Curie. Paris, France. URL: http://www.numdam.org/item/BURO_1981__37-38__7_0/.

Tarpey, T. and Kinateder, K. K. J. (2003). "Clustering functional data". In: *J. Classification* 20.1, pp. 93–114. ISSN: 0176-4268.

Vandewalle, V., Preda, C., and Dabo, S. (2020). "Clustering spatial functional data". In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons.

Chapter 5

Multiple partition clustering

Contents

5.1 Introduction 73
5.2 Multiple partition by blocks of variables 74
5.2.1 Introduction 74
5.2.2 Multiple partitions mixture model 76
5.2.3 Comments 77
5.2.4 Inference of the model and model selection 78
5.2.5 Illustration on real data 81
5.2.6 Conclusion 86
5.3 Multiple partition by linear projections 87
5.3.1 Introduction 87
5.3.2 Multi-Partition Subspace Mixture Model 89
5.3.3 Estimation of the Parameters of the Model 91
5.3.4 Experiments on real data 92
5.3.5 Conclusions 93
5.4 Conclusion and perspectives 94

Related articles

1. M. Marbac and V. Vandewalle (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179
2. V. Vandewalle (2020). "Multi-Partitions Subspace Clustering". In: *Mathematics* 8.4, p. 597

5.1 Introduction

Despite the increase of available models in clustering (McNicholas, 2016; McLachlan and Peel, 2004; Bouveyron, Celeux, et al., 2019), less attention has been paid to considering several latent class variables, latterly called multiple partitions framework (Galimberti and Soffritti, 2007; Galimberti, Manisi, and Soffritti, 2018). It can be mainly explained by the additional model interpretations and parameters estimation

complexities. However, in the area of massive data, with individuals described by possibly thousands of variables, it is difficult to believe that the whole heterogeneity in the data can be described by only one latent variable. Even if standard clustering gives some insights on the major source of variability, it may hide some other relevant information in the data. It can for instance be the case if variables related to some focus are more present than variables related to another one.
While in traditional clustering approaches the practitioner was selecting carefully variables to use in order to produce a summary according to some prior focus, nowadays data are collected on the fly, philosopher's stone being to put all variables into the model and let them speak. However, data rarely speak by themselves what renders difficult unsupervised framework contrary to the supervised one. This question is related to the way of evaluating clustering (Luxburg, Williamson, and Guyon, 2012; Hennig, 2015), and some challenging tasks such as benchmark for clustering [1]. An important difficulty in clustering being that it is claimed to be unsupervised, however at the end the clusters obtained are often used in a decision-making process. Thus one solution could be to connect these different frameworks, but from a practical point of view, it is often simpler to summarize data per cluster hopefully useful for several tasks.
The aim of the multiple partition framework is to propose possibly several clustering points of view with respect to the data. These points of view can raise some possible new knowledge of the data letting each possible part of the data give information to the user. We have considered two possible ways to perform this objective. The first one presented in Section 5.2 consists of grouping variables into blocks, the heterogeneity in each block being explained by some particular latent clustering variable, this approach has the advantage to take into account any kind of variables, it has been published in Marbac and Vandewalle (2019). The second one is presented in Section 5.3, it consists of looking for linear projections of the variables, each one being explained by some particular cluster variable, this approach allows to deal simultaneously with the multi-partition setting and to get a visualization of the data that we latter call clustering subspaces, it has been published in Vandewalle (2020).

5.2 Multiple partition by blocks of variables

5.2.1 Introduction

The problem of finding several partitions in the data, based on different groups of continuous variables, has been addressed by Galimberti and Soffritti (2007) in a model-based clustering framework (McLachlan and Peel, 2000). In this framework, the authors assume that the vector of variables can be partitioned into independent sub-vectors, each one following a particular Gaussian mixture model with a full covariance matrix. Then, they proposed a forward/backward search to perform

[1] see for instance cluster benchmark data repository of the IFCS https://ifcs.boku.ac.at/repository/

model selection based on the maximization of the BIC. More recently, Galimberti, Manisi, and Soffritti (2018) have proposed an extension of their previous works which relaxes the independence assumption between sub-vectors. This extension considers three types of variables, the classifying variables, the redundant variables with respect to the classifying variables, and the variables which are not classifying at all. This can be seen as an extension of the models proposed by Raftery and Dean (2006) and Maugis, Celeux, and Martin-Magniette (2009), in the framework of variable selection in clustering. In this framework, model selection is a difficult challenge because full Gaussian models are still considered, and many possible roles of the variables need to be considered, thus needing a lot of computation even for the re-affectation of only one variable. Therefore, they have to use forward/backward algorithms to maximize the BIC. However, these algorithms are suboptimal since they only converge to a local optimum of the BIC. Moreover, they are based on comparisons of the BIC between two models. Thus, they perform many calls of EM algorithm. Hence, these approaches only can deal with a limited number of variables (typically less than 100).

The problem of finding several partitions in the data has also been considered by Poon et al. (2013), in what they called facet determination. Their model is similar to Galimberti and Soffritti (2007) but it also allows tree dependency between latent variables, the resulting model is called pouch latent tree models (PLTMs). The best model is then selected using the BIC criterion by using a greedy search based on search operators such as node introduction or node deletion for instance. This model allows a rich interpretation of the data, however, the huge number of possibilities due to the tree structure search make it even more difficult to use than previous models when the number of variables is large.

In order to deal with large numbers of variables, the main idea is to use a more constrained model to be able to easily perform model selection. We assume that the distribution of the observed data can be factorized into several independent blocks of variables, each one following its own mixture distribution. The considered mixture distribution in a block is a latent class model (*i.e.,* each variable of a block is supposed to be independent of the others given the cluster variable associated with this block). This model is an extension of the approaches proposed by Marbac and Sedki (2017) and Marbac, Patin, and Sedki (2018) in the framework of variable selection in clustering, where only two blocks are considered, *i*.e. one block of classifying variables assuming conditional independence, and one block of non classifying variables assuming total independence. In the Gaussian setting, our model can also be seen as a simplified version of the model proposed by Galimberti, Manisi, and Soffritti (2018) where diagonal covariances matrices are assumed. However, our model also allows dealing with categorical data while it is not possible in Galimberti, Manisi, and Soffritti (2018). The simplicity of the model allows estimating the partition of the variables into blocks and the mixture parameters simultaneously like in Marbac and Sedki (2017) and Marbac, Patin, and Sedki (2018). We present a procedure for performing model selection (choice of the number of blocks, the number of clusters inside each block and the partition of variables into blocks) with the BIC (Schwarz,

1978) or the MICL (Marbac and Sedki, 2017). The BIC enjoys consistency properties and does not require to define prior distributions. However, in the clustering framework, it tends to over-estimate the number of clusters, and for small sample sizes, the asymptotic approximation on which it relies can be questionable. Thus, in the framework of variable selection, Marbac and Sedki (2017) have proposed the MICL criterion derived from the ICL criterion (Biernacki, Celeux, and Govaert, 2000). This criterion takes into account the classification purpose by computing the maximum integrated completed likelihood. Moreover, it is expected to well behave for small sample sizes, because it avoids the asymptotic approximations of the integrated completed likelihood by performing an exact integration over the parameter space thanks to conjugate priors. Depending on the context, either BIC or MICL can be preferred. In the context of multiple partitions clustering, it is possible to simultaneously perform parameter estimation (resp. partition estimation) and model selection with the BIC (resp. MICL) criterion like in Marbac and Sedki (2017) and Marbac, Patin, and Sedki (2018), thus avoiding to run EM algorithms for each partition of variables into blocks. Note that the proposed model allows to deal with mixed-data as in Marbac, Patin, and Sedki (2018), and it also includes the variable selection as a special case. Moreover, the proposed model can answer the problem of clustering mixed data in which continuous variables are often expected to dominate the clustering process. Allowing several partitions the categorical variables are now able, if necessary, to form their own clustering structure.

Let notice that the proposed framework has similarities with the bi-clustering framework, and in particular with the block clustering models proposed by Govaert and Nadif (2003). Block clustering consists of clustering the rows and the columns simultaneously while our approach makes blocks of variables, *i.e.* clustering of columns, and for each block of variables makes a clustering of the individuals, *i.e.* clustering of rows. However, instead of considering one partition in rows like the block clustering, our approach considers several partitions in rows. Moreover, block clustering is limited to deal with variables of the same kind assuming homogeneous distribution in each block while our approach allows dealing with heterogeneous data.

5.2.2 Multiple partitions mixture model

5.2.2.1 The model

As in previous chapters, data to cluster $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are composed of n observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ described by d variables potentially of different types (*i.e.,* each variable can be continuous, binary, count or categorical). Observations are assumed to independently arise from a multiple partitions model (MPM) which considers that variables are grouped into B independent blocks. The blocks of variables are defined by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$, where $\omega_j = b$ indicates that variable j belongs to block b. Moreover, MPM considers that variables of block b follow a g_b-component mixture assuming within-component independence. Thus, for a model $\boldsymbol{m} = (B, \boldsymbol{g}, \boldsymbol{\omega})$ with $\boldsymbol{g} = (g_1, \ldots, g_B)$, the probability distribution function (pdf) of

$\boldsymbol{x}_i$ is

$$p(\boldsymbol{x}_i|\boldsymbol{m},\boldsymbol{\theta}) = \prod_{b=1}^{B} p(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m},\boldsymbol{\theta}) \text{ with } p(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m},\boldsymbol{\theta}) = \sum_{k=1}^{g_b} \pi_{bk} \prod_{j\in\Omega_b} p(x_{ij}|\boldsymbol{\alpha}_{jk}), \quad (5.1)$$

where $\Omega_b = \{j : \omega_j = b\}$ denotes the indexes of variables of block b, $\boldsymbol{x}_{i\{b\}} = (x_{ij}; j \in \Omega_b)$ is the vector of observed variables of block b, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ groups the model parameters, $\boldsymbol{\pi} = (\boldsymbol{\pi}_b; b = 1, \ldots, B)$ groups the proportions with $\boldsymbol{\pi}_b = (\pi_{b1}, \ldots, \pi_{bg_b})$, $\pi_{bk} > 0$ and $\sum_{k=1}^{g_b} \pi_{bk} = 1$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d)$ and $\boldsymbol{\alpha}_j = (\boldsymbol{\alpha}_{j1}, \ldots, \boldsymbol{\alpha}_{jg_{\omega_j}})$. The univariate margin of a component for a continuous (respectively binary, count and categorical), denoted by $p(x_{ij}|\boldsymbol{\alpha}_{jk})$, is a Gaussian (Bernoulli, Poisson and multinomial) distribution with parameters $\boldsymbol{\alpha}_{jk}$ (Moustaki and Papageorgiou, 2005).
MPM provides B partitions among the observations (one partition per block of variables). The partition of block b is denoted by $\mathbf{z}_b = (\boldsymbol{z}_{1b}, \ldots, \boldsymbol{z}_{nb}) \in \mathcal{Z}_{g_b}$, where $\mathcal{Z}_{g_b}$ is the set of the partitions of n elements in g_b clusters, and $\boldsymbol{z}_{ib} = (z_{ib1}, \ldots, z_{ibg_b})$ with $z_{ibk} = 1$ if observation i belongs to cluster k for block b and $z_{ibk} = 0$ otherwise. The multiple partitions $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_B)$ for model $\boldsymbol{m}$ belongs to $\boldsymbol{\mathcal{Z}}_{\boldsymbol{m}} = \mathcal{Z}_{g_1} \times \ldots \times \mathcal{Z}_{g_B}$.

Example 1 *We consider $d = 4$ continuous variables arisen from* MPM *with $B = 2$ blocks of two variables with $\boldsymbol{\omega} = (1, 1, 2, 2)$ (*i.e., *the first two variables belong to block 1 and the last two variables belong to block 2). Moreover, each block follows a bi-component Gaussian mixture (*i.e., *$g_1 = g_2 = 2$) with equal proportions (*i.e., *$\pi_{bk} = 1/2$), mean $\mu_{j1} = 4$, $\mu_{j2} = -4$ and variance $\sigma^2_{jk} = 1$. Figure 5.1 gives the bivariate scatter-plot of the observations. Colors and symbols indicate the component memberships of block 1 and 2 respectively.*

5.2.3 Comments

Link with approaches of model-based clustering Standard methods of clustering consider that the observed variables explain a single partition among the observations. However, if this assumption is violated, MPM can circumvent this limit because it considers different partitions explained by different subsets of variables. Moreover, MPM generalizes approaches used for variable selection in model-based clustering. Indeed, if $B = 2$ and $g_1 = 1$ then variables belonging to block 1 are irrelevant for the clustering, while variables belonging to block 2 are relevant.

Model identifiability The model (5.1) is identifiable up to a switching of the component labels and a change in the order of the blocks. Identifiability of the distribution of each block leads to identifiability of (5.1). Identifiability holds for blocks containing at least one continuous or integer variable (Teicher, 1963; Teicher, 1967). Finally, identifiability holds for blocks only composed of categorical variables under mild conditions (Allman, Matias, and Rhodes, 2009).

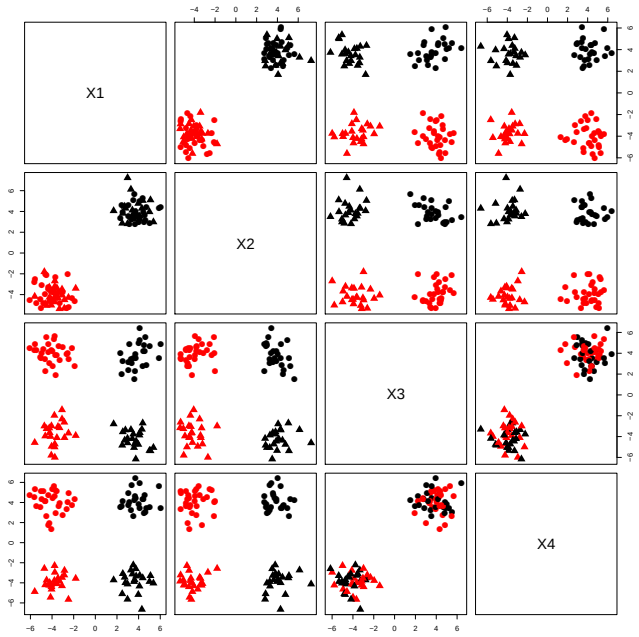


Figure 5.1: Sample generated from MPM where colors and symbols indicate the component memberships of block 1 and 2 respectively.

About the assumption of independence within components Contrary to Galimberti and Soffritti (2007) which assume full Gaussian covariance matrices, MPM assumes that variables are independent within components. This assumption is quite standard for clustering categorical or mixed-type data (Hand and Keming, 2001; Moustaki and Papageorgiou, 2005), and it limits the number of parameters. Hence, MPM has $\nu_{\boldsymbol{m}} = \sum_{b=1}^{B}(g_b - 1) + \sum_{j \in \Omega_b} \nu_j g_b$ parameters to be estimated, where $\nu_j = \dim(\Theta_j)$ and Θ_j is the space of the parameters of the univariate margin of one component of variable j (*e.g.*, $\nu_j = 2$ if the margin is a Gaussian distribution). Finally, it permits efficient approaches for model selection.

5.2.4 Inference of the model and model selection

Maximum likelihood inference For sample $\mathbf{x}$ and model $\boldsymbol{m}$, the observed-data log-likelihood is defined by

$$\ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}) = \sum_{b=1}^{B} \sum_{i=1}^{n} \ln p(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}). \tag{5.2}$$

The maximum likelihood estimates (MLE) can be obtained by an EM algorithm (Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 1997). Independence between the B blocks of variables permits to maximize the observed-data

log-likelihood on each block independently. A simple modification of this algorithms allows to estimate $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ simultaneously.

Model selection with the BIC Model has to be assessed from the data among a set of competing models $\mathcal{M}$ defined by

$$\mathcal{M} = \{\boldsymbol{m} = (B, \boldsymbol{g}, \boldsymbol{\omega}); 1 \leq B \leq B_{\max},\ \forall b,\ 1 \leq g_b \leq g_{\max}, \omega_j \in \{1, \ldots, B\}\}, \quad (5.3)$$

where $B_{\max}$ is the maximum number of blocks and $g_{\max}$ is the maximum number of components within block. The number of competing models is $\text{card}(\mathcal{M}) = \sum_{B=1}^{B_{\max}} S(d, B) g_{\max}^B$ where $S(d, B)$ denotes the Stirling number of the second kind. Model selection can be done by the BIC (Schwarz, 1978) where

$$\text{BIC}(\boldsymbol{m}) = \max_{\boldsymbol{\theta}_{\boldsymbol{m}}} \ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}) \text{ with } \ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}) = \ell(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n. \quad (5.4)$$

Model selection with the BIC consists in maximizing this criterion with respect to $\boldsymbol{m}$. Obviously, this is equivalent to maximizing the penalized likelihood on the couple $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{m}})$. Thus, model and parameter inference leads to search

$$(\boldsymbol{m}^\star, \hat{\boldsymbol{\theta}}_{\boldsymbol{m}^\star}) = \arg\max_{(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{m}})} \ell_{pen}(\boldsymbol{\theta}_m|\boldsymbol{m}, \mathbf{x}). \quad (5.5)$$

Due to the number of competing models, an exhaustive approach that consists of computing BIC for each competing model is not doable. However, holding $(B, \boldsymbol{g})$ fixed, model selection with BIC and maximum likelihood inference implies maximizing the penalized likelihood with respect to $(\boldsymbol{\omega}, \boldsymbol{\theta})$. This maximization can be carried out by a modified version of the EM algorithm (Green, 1990). Thus, the combinatorial problem of the estimation of the blocks of variables can be circumvented if the maximum number of blocks is small. Considering $B_{\max}$ small (*i.e.,* $B_{\max} < 5$) can seem restrictive. However, classical clustering methods consider $B_{\max} = 1$. Moreover, if $B_{\max}$ is wanted to be more than five, then the model stays well defined but the proposed methods of model selection suffer from combinatorial issues. Then, in this case, other algorithms (like forward/backward search) should be used for model estimation. Indeed, $(\boldsymbol{m}^\star, \hat{\boldsymbol{\theta}}_{\boldsymbol{m}^\star})$ can be found by running this algorithm for each value of $(B, \boldsymbol{g})$ allowed by $\mathcal{M}$. Therefore, less than $\sum_{B=1}^{B_{\max}} g_{\max}^B$ calls of the EM algorithm should be done.
Note that the number of calls of EM algorithm does not depend on the number of variables. To implement this modified EM algorithm, we introduce the penalized complete-data likelihood

$$\ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = \ell(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) - \frac{\boldsymbol{\nu}_{\boldsymbol{m}}}{2} \log n \quad (5.6)$$

$$= \sum_{b=1}^{B} \left(\ln p(\mathbf{z}_b|\boldsymbol{\pi}_b) - \frac{g_b - 1}{2} \ln n + \sum_{j \in \Omega_b} \ln p(\mathbf{x}_j|\mathbf{z}_b, \boldsymbol{\alpha}_j) - \frac{\nu_j g_b}{2} \ln n \right), \quad (5.7)$$

where $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})$ Holding $(B, \boldsymbol{g})$ fixed and starting from $(\boldsymbol{\omega}^{[0]}, \boldsymbol{\theta}^{[0]})$, its iteration $[r]$ is composed of two steps:
E-step Computation of the fuzzy partitions $t_{ibk}^{[r]} := \mathbb{E}[Z_{ibk}|\boldsymbol{x}_i, \boldsymbol{m}^{[r-1]}, \boldsymbol{\theta}^{[r-1]}]$, hence for $b = 1, \ldots, B$, for $k = 1, \ldots, g_b$, for $i = 1, \ldots, n$

$$t_{ibk}^{[r]} = \frac{\pi_{bk}^{[r-1]} \prod_{j \in \Omega_b^{[r-1]}} p(x_{ij}|\boldsymbol{\alpha}_{jk}^{[r-1]})}{\sum_{k'=1}^{g_b} \pi_{bk'}^{[r-1]} \prod_{j \in \Omega_b^{[r-1]}} p(x_{ij}|\boldsymbol{\alpha}_{jk'}^{[r-1]})},$$

M-step1 Updating the affectation of the variables to blocks

$$\omega_j^{[r]} = \underset{b \in \{1,\ldots,B\}}{\arg\max} \left(\sum_{k=1}^{g_b} \max_{\boldsymbol{\alpha}_{jk} \in \Theta_j} Q(\boldsymbol{\alpha}_{jk}|\mathbf{x}_j, \mathbf{t}_{bk}^{[r]}) - \frac{\nu_j g_b}{2} \ln n \right),$$

where $Q(\boldsymbol{\alpha}_{jk}|\mathbf{x}_j, \mathbf{t}_{bk}) = \sum_{i=1}^{n} t_{ibk} \ln p(x_{ij}|\boldsymbol{\alpha}_{jk})$ and $n_{bk}^{[r]} = \sum_{i=1}^{n} t_{ibk}^{[r]}$.
M-step2 Updating the model parameters

$$\pi_{bk}^{[r]} = \frac{n_{bk}^{[r]}}{n} \text{ and } \boldsymbol{\alpha}_{jk}^{[r]} = \underset{\boldsymbol{\alpha}_{jk} \in \Theta_j}{\arg\max}\, Q(\boldsymbol{\alpha}_{jk}|\mathbf{x}_j, \mathbf{t}_{\omega_j^{[r]} k}^{[r]}).$$

Like for the standard EM algorithm, the objective function (*i.e.,* $\ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}, \mathbf{z})$) increases at each iteration but the global optimum is not achieved in general. Hence, different random initializations must be done. Finally, note that the algorithm can return empty blocks. Indeed, M-step1 is done without constraining each block to contain at least one variable. Thus, each $\omega_j^{[r]}$ can be obtained independently.

Model selection by MICL Criteria based on the integrated complete-data likelihood are popular for model-based clustering. Indeed, they take into account the clustering purpose (modeling the data distribution and providing well-separated components). Moreover, integrated complete-data likelihood has closed-form when components belong to exponential family and conjugate priors are used. The integrated complete-data likelihood is defined by

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}. \tag{5.8}$$

We assume independence between the prior distributions, so

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = \prod_{b=1}^{B} p(\boldsymbol{\pi}_b|g_b) \prod_{j=1}^{d} p(\boldsymbol{\alpha}_j|\boldsymbol{g}, \omega_j) \text{ where } p(\boldsymbol{\alpha}_j|\boldsymbol{g}, \omega_j) = \prod_{k=1}^{g_{\omega_j}} p(\boldsymbol{\alpha}_{jk}).$$

Thus, the integrated complete-data likelihood has the form defined by

$$\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \sum_{b=1}^{B} \ln p(\mathbf{z}_b|g_b) + \sum_{j \in \Omega_b} \ln p(\mathbf{x}_j|\mathbf{z}_b, g_b), \tag{5.9}$$

where $p(\mathbf{z}_b|g_b) = \int_{\mathcal{S}(g_b)} p(\mathbf{z}_b|g_b, \boldsymbol{\pi}_b)p(\boldsymbol{\pi}_b|g_b)d\boldsymbol{\pi}_b$, $\mathcal{S}(g_b)$ denotes the simplex of dimension g_b and $p(\mathbf{x}_j|\mathbf{z}_b, g_b) = \int_{\Theta_j^{g_b}} p(\mathbf{x}_j|\mathbf{z}_b, \boldsymbol{\alpha}_j)p(\boldsymbol{\alpha}_j|\boldsymbol{g}, \omega_j)d\boldsymbol{\alpha}_j$. We use conjugate prior distributions, thus integrals $p(\mathbf{z}_b|g_b)$ and $p(\mathbf{x}_j|\mathbf{z}_{\omega_j}, g_{\omega_j})$ have closed forms.
The MICL (maximum integrated complete-data likelihood) criterion corresponds to the largest value of the integrated complete-data likelihood among all the possible partitions. Thus, the MICL is defined by

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ with } \mathbf{z}^\star_{\boldsymbol{m}} = \underset{\mathbf{z}\in\boldsymbol{\mathcal{Z}}_{\boldsymbol{m}}}{\arg\max} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}). \tag{5.10}$$

Model selection with MICL consists in finding $\boldsymbol{m}^\star = \arg\max_{\boldsymbol{m}\in\mathcal{M}} \text{MICL}(\boldsymbol{m})$.
Holding $(B, \boldsymbol{g})$ fixed, maximizing MICL corresponds to maximizing the integrated complete-data likelihood with respect to the affectation of the variables into blocks $\boldsymbol{\omega}$ and to the partition $\mathbf{z}$. Starting at the initial value $\boldsymbol{\omega}^{[0]}$ where each ω_j is uniformly sampled among $\{1, \ldots, B\}$, the algorithm alternates between two steps defined at iteration $[r]$ by
Partition step: find $\mathbf{z}_b^{[r]}$ such that for all $b = 1, \ldots, B$

$$\sum_{j\in\Omega_b^{[r-1]}} \ln p(\mathbf{x}_j, \mathbf{z}_b^{[r]}|g_b) \geq \sum_{j\in\Omega_b^{[r-1]}} \ln p(\mathbf{x}_j, \mathbf{z}_b^{[r-1]}|g_b),$$

where $\Omega_b^{[r-1]} = (j; \omega_j^{[r-1]} = b)$.
Model step: find $\boldsymbol{\omega}^{[r]}$ such that for $j = 1, \ldots, d$

$$\omega_j^{[r]} = \underset{b\in\{1,\ldots,B\}}{\arg\max}\ p(\mathbf{x}_j|\mathbf{z}_b^{[r]}, g_b).$$

Optimization at the Partition step is not obvious, despite that it is done on each block independently. So, the partition $\mathbf{z}_b^{[r]}$ is defined as a partition that increases the value of the integrated complete-data likelihood for the current model for block b. It is obtained by an iterative method where each iteration consists of optimizing the integrated complete-data likelihood for block b on the class membership of a single individual while the partition among the other observations stays hold. Optimization at the Model Step can be performed independently for each variable because of the intra-component independence assumption. The optimization algorithm converges to a local optimum of the integrated complete-data likelihood. Thus, many different initializations should be done.

5.2.5 Illustration on real data

Here we only present the numerical experiments on real data analyzed by considering $B_{\max} = g_{\max} = 5$, in order to give the interpretation of the results made possible by the multiple partition framework. Detailed simulations are given in Marbac and Vandewalle (2019).

5.2.5.1 NBA team data

Data description Data to cluster contain statistics of National Basketball Association (NBA) teams for the season 2016/2017[2]. Each team is described by 16 numerical variables including total minutes played (min), field goals made (fgm), field goals attempted (fga), three-pointers made (3pm), three-pointer attempted (3pa), free throws made (ftm), free throw attempted (fta), offensive rebounds (or), total rebounds (tr), assists (as), steals (st), turnovers (to), blocks (bk), personal fouls (pf), technical fouls (tc), and points (pts). Note that we substitute the variable fgm (respectively 3pm, ftm and or) by field goals made rate fgmr=fmg/fta (respectively 3pmr=3pm/3pa, ftmr=ftm/fta and orr=or/tr) because of its dependency with fta (respectively with 3pa, fta and tr).

Model selection with BIC The model selected by the BIC considers three blocks of variables. Moreover, if more than three blocks are allowed, then the model selected by the BIC only fills three blocks while the other blocks are empty. The models selected by the BIC for different numbers of blocks are presented in Table 5.1.

B	BIC	Time	Block	$\boldsymbol{g}$	variables
1	-1932	7	1	2	all the variables
2	-1915	47	1	3	fgmr, fga, 3pmr, 3pa, tr, as, to, pts
			2	1	min, ftmr, fta, orr, st, bk, pf, tc
3	-1909	170	1	2	fgmr, 3pmr, pf
			2	2	fga, 3pa, tr, as, st, to, pts
			3	1	min, ftmr, fta, orr, bk, tc

Table 5.1: Models selected by the BIC for different numbers of blocks for the NBA team data: BIC (BIC), computing time in seconds (Time), number of components ($\boldsymbol{g}$) and variables of each block (variables)

The model selected by the BIC defines two partitions among NBA teams. The first block is composed of three features (field goals made rate, three points made rate, and personal fouls) and permits to distinguish the style of playing of the teams (offensive or defensive). Based on the parameters of block 1, presented in Table 5.2, there are two types of teams: offensive teams characterized by high shooting ability and low personal fouls (Boston Celtics, Cleveland Cavaliers, Denver Nuggets, GS Warriors, Houston Rockets, Indiana Pacers, LA Clippers, Miami Heat, Milwaukee Bucks, Minnesota T-wolves, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs, Toronto Raptors, Utah Jazz and Washington Wizards) and defensive teams characterized by low shooting ability and high personal fouls (AtlantaHawks, Charlotte Hornets, Chicago Bulls, Dallas Mavericks, Detroit Pistons, LA Lakers, Memphis Grizzlies, Brooklyn Nets, New Orleans Pelicans, NY Knicks, Orlando Magic, Philadelphia 76ers, Phoenix Suns and Oklahoma City Thunder). The second block

[2]Freely available at `http://www.dougstats.com/16-17RD.Team.Opp.txt`.

is composed of seven features (field goal attempted, 3 points attempted, total rebounds, assists, steals, turnovers, and points) and permits to identify two teams (GS Warriors and Houston Rockets) that obtain better statistics for general performances (see parameters presented in Table 5.3). Finally, the last block is composed of six features detected as irrelevant for clustering.

		fmgr	3pmr	pf
offensive teams ($\pi_{11} = 0.57$)	mean	0.468	0.371	1628.042
	sd	0.009	0.010	76.138
defensive teams ($\pi_{12} = 0.43$)	mean	0.446	0.342	1635.503
	sd	0.005	0.010	173.701

Table 5.2: Parameters of block 1 of the model selected by BIC for the NBA team data.

			fga	3pa	tr	as	st	to	pts
GS Warriors/	($\pi_{11} = 0.07$)	mean	7144	2934	3642	2281	728	1186	9481
Houston Rockets		sd	3	372	3	211	59	3	22
Other	($\pi_{11} = 0.93$)	mean	6992	2162	3564	1825	626	1091	8600
teams		sd	183	262	141	131	45	105	259

Table 5.3: Parameters of block 2 of the model selected by BIC for the NBA team data.

Model selection with MICL If MICL is considered for model selection, then the proposed approach leads to a variable selection for clustering. Indeed, if one block is considered then the selected model has only one component (MICL = -2288.564). If two blocks are considered, then the MICL of the selected model is -2287.664 (estimation is done in about 1 min). For this model, the first block groups four variables (total minutes played, turnovers, personal fouls, and points) modeled by a bi-component mixture while a second block groups the other variables which are detected as irrelevant for clustering. The best model selected by MICL provides an unbalanced partition because component 1 ($\pi_{11} = 0.09$) contains only three teams (Denver Nuggets, GS Warriors, and Houston Rockets). These two classes can be interpreted by using the parameters presented in Table 5.4. Note that even if more than two blocks are considered, then the selected model considers only one block of discriminative variables and one block of irrelevant variables.

5.2.5.2 Wine data

Data description Data are twenty-seven chemical and physical properties of three types of wine (Barolo, Grignolino, Barbera) from the Piedmont region of

			min	to	pf	pts
Denver Nuggets/ GS	($\pi_{11} = 0.09$)	mean	19769	1186	1595	9374
Warriors / Houston Rockets		sd	15	2	28	152
Other	($\pi_{12} = 0.91$)	mean	19812	1088	1636	8579
teams		sd	54	105	138	240

Table 5.4: Parameters of block 1 of the model selected by MICL for the NBA team data.

Italy. These data were introduced by Forina et al. (1986) and are available in the R package *pgmm* (McNicholas, ElSherbiny, et al., 2018). Information about the type of wine is used to validate the clustering results so the variable is not used during clustering. Moreover, note that the wines in this study were collected throughout 1970–1979. Model-based clustering of this data set has been conducted by McNicholas and Murphy (2008).

Model selection with BIC The model selected by the BIC considers four blocks of variables. Moreover, if more than four blocks are allowed, then the model selected by the BIC only fills four blocks while the other blocks are empty. The models selected by the BIC for different numbers of blocks are presented in Table 5.5.

B	BIC	Time	Block	$\boldsymbol{g}$	ARI
1	-6025.00	30	1	4	0.78
2	-5947.88	280	1	3	0.87
			2	4	0.16
3	-5921.42	1590	1	4	0.74
			2	4	0.20
			3	2	0.02
4	-5918.06	6065	1	4	0.75
			2	2	0.21
			3	3	0.02
			4	2	0.00

Table 5.5: Models selected by the BIC for different numbers of blocks for the wine data: BIC (BIC), computing time in seconds (Time), number of components (G) and ARI between the estimated partitions and the type of wine (ARI).

The model selected by the BIC provides four partitions. Block 1 is composed of 19 variables (Alcohol, Sugar-free Extract, Tartaric Acid, Uronic Acids, Alkalinity of Ash, Calcium, Magnesium, Phosphate, Total Phenols, Flavanoids, Non-flavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of Diluted Wines, OD280/OD315 of Flavanoids, Glycerol, 2-3-Butanediol, Proline). The partition provided by block 1 is related to the type of wines (ARI=0.75, see also Table 5.6).

Its ARI is close to the ARI obtained by variable selection in clustering (ARI=0.78) but is outperformed by parsimonious Gaussian mixture (ARI=0.98; see McNicholas and Murphy (2008)). The second block of variables is composed of four variables (Fixed Acidity, Malic Acid, pH, Total Nitrogen). Its information is mainly about the year of production (see Table 5.7) because class 1 (respectively class 2) mainly groups the wines harvested before (respectively after) 1974.

	Barolo	Grignolino	Barbera
Class 1	0	45	0
Class 2	0	5	48
Class 3	58	1	0
Class 4	1	20	0

Table 5.6: Classification table for partition provided by block 1 of the model selected by BIC for the wine data

	Year								
	1970	1971	1972	1973	1974	1975	1976	1978	1979
Class 1	8	25	4	27	31	3	1	0	0
Class 2	1	3	3	2	14	6	16	29	5

Table 5.7: Classification table for partition provided by block 2 of the model selected by BIC for the wine data

Model selection with MICL The model selected by the MICL considers three blocks of variables. Moreover, if more than three blocks are allowed, then the model selected by the MICL only fills four blocks while the other blocks are empty. The models selected by the MICL for different numbers of blocks are presented in Table 5.8.

B	BIC	Time	Block	$\boldsymbol{g}$	ARI
1	-7012.15	464	1	1	0.00
2	-6114.31	7210	1	3	0.80
			2	3	0.18
3	-6102.89	18880	1	3	0.88
			2	3	0.19
			3	1	0.00

Table 5.8: Models selected by the MICL for different numbers of blocks for the wine data: MICL (MICL), computing time in seconds (Time), number of components ($\boldsymbol{g}$) and ARI between the estimated partitions and the type of wine (ARI).

The MICL selects three blocks (if more than three blocks are considered, then only three non-empty blocks are returned by the algorithm). The first block is com-

posed of twenty variables (Alcohol, Sugar-free Extract, Tartaric Acid, Uronic Acids, Ash, Alkalinity of Ash, Calcium, Magnesium, Phosphate, Total Phenols, Flavanoids, Non-flavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of Diluted Wines, OD280/OD315 of Flavanoids, Glycerol, 2-3-Butanediol, Proline). This block considers three components which mainly correspond to the types of the wines (ARI=0.88, see also Table 5.9). Like for the model selected by the BIC, the block 2 is related to the year of the wines by separating the wines harvested before and after 1974 (ARI=0.16, see also Table 5.10). Block 3 groups the variables which are detected as irrelevant for clustering.

	Barolo	Grignolino	Barbera
Class 1	0	64	0
Class 2	0	4	48
Class 3	59	3	0

Table 5.9: Classification table for partition provided by block 1 of the model selected by MICL for the wine data

	1970	1971	1972	1973	1974	1975	1976	1978	1979
Class 1	1	3	3	2	14	5	15	29	5
Class 2	5	3	0	4	1	1	1	0	0
Class 3	3	22	4	23	30	3	1	0	0

Table 5.10: Classification table for partition provided by block 2 of the model selected by MICL for the wine data

5.2.6 Conclusion

In this section, we have presented a method for performing clustering with multiple partitions. The proposed model is easily interpretable and permits also to associate each produced partition with a subset of variables generating it. Thus, allowing to perform a clustering of variables of eventually different kinds as a by-product. Such kind of model allows in some sense to limit the subjectivity of the choice of variables in clustering and allows to find several potentially interesting structures in the data without imposing that all the variables define the same clustering. The strength of the proposed approach is to use a simpler model, *i.e.* conditional independence assumption, than the state of the art methods. Thus, the challenging problem of model selection can be circumvented, even for a large number of variables. Indeed, model selection can be done efficiently by maximizing classical information criteria (BIC or MICL). The proposed method is available as the MGMM an R package available on R-forge[3].

[3] http://r-forge.r-project.org/projects/mgmm

The proposed method could be improved by proposing additional initialization strategies, and more efficient ways to explore the large space of possible models $\mathcal{M}$ according to B and $\boldsymbol{g}$. This could be performed for instance by aggregating the clusterings defined for each variable and then exploring different possible cuttings of this hierarchy as possible initializations of the algorithm.
The proposed method offers many possible extensions. On the first hand, since it performs the clustering of the individuals and of the variables simultaneously, it can be in some sense interpreted as a co-clustering method. However, to fit with the standard formulation of co-clustering with only one partition for the individuals, an additional modeling layer should be added to summarize the multi-partition by only a single partition. On the other hand, it would also be interesting in the quantitative setting to derive some k-means type approximation of the proposed method to deal with the very high dimensional setting as Witten and Tibshirani (2010) in the variable selection framework by including LASSO type penalty.
Finally, model (5.1) can be extended by relaxing the assumption of within component independence. Classical results about identifiability of mixtures can be used to show the identifiability of the resulting model. However, model selection becomes more complex. Indeed, the two algorithms introduced in this paper cannot be used directly due to the within component dependencies. The M-step of the EM algorithm optimizing BIC and the Model step of the iterative algorithm optimizing MICL are not explicit.

5.3 Multiple partition by linear projections

In this section, we present another possibility to perform multiple partition clustering, by using linear projections. The outputs are now linear projections of the data each one conveying a particular clustering. Compared with the previous approach it is limited to the continuous setting with a moderate number of variables. The combinatorial problem of finding the best partition of the variables is now replaced with the problem of finding the best clustering projections.

5.3.1 Introduction

In exploratory data analysis, the statistician often uses clustering and visualization to improve his knowledge of the data. In visualization he looks for some principal components explaining some major characteristics of the data. For example in principal component analysis (PCA) the goal is to find a linear combination of the variables explaining the major variability of the data. In cluster analysis, the goal is to find some clusters explaining the major heterogeneity of the data. Here, we suppose that the data can contain several clustering latent variables, that is, we are in the multiple partition setting, and we are simultaneously looking for clustering subspaces, that is, linear projections of the data each one related to some clustering latent variable, thus the developed model is later called multi-partitions subspace clustering. A solution to perform multi-partition subspace clustering is

to use a probabilistic model on the data such as a mixture model (McLachlan and Peel, 2004), it allows to perform the estimation of the parameters, and model selection such as the choice of the number of subspaces and the number of clusters per subspace using standard model choice criteria such as BIC (Schwarz, 1978). Thus the main fields related to our work are model-based subspace clustering and multi-partitions clustering.

In the model-based subspace clustering framework, let first notice that PCA can be re-interpreted in a probabilistic way by considering a parsimonious version of a multivariate Gaussian distribution (Tipping and Bishop, 1999) and that the k-means algorithm can be re-interpreted as a particular parsimonious Gaussian mixture model estimated using a classification EM algorithm (Celeux and Govaert, 1995). A re-interpretation of the probabilistic PCA has also been used in clustering by Bouveyron, Girard, and Schmid (2007) to cluster high-dimensional data. Although the proposed high dimensional mixture does not perform dimension reduction, it rather operates a class per class dimension reduction which does not allow to have a global model-based data visualization. Thus Bouveyron and Brunet (2012) proposed the so-called Fisher-EM algorithm which simultaneously performs clustering and dimension reduction. This is performed through a modified version of the EM algorithm (Dempster, Laird, and Rubin, 1977) by including a Fisher step between the E and the M step. This approach allows the same projection to be applied to all data, but does not guarantee the increasing of the likelihood at each iteration of the algorithm.

In the context of multi-partitions clustering, Galimberti and Soffritti (2007) assumed that the variables can be partitioned into several independent blocks, each one following a full-covariance Gaussian mixture model. The model selection was done by maximizing the BIC criterion by a forward/backward approach. Then, Galimberti, Manisi, and Soffritti (2018) generalized their previous work by relaxing the assumption of block independence. The proposed extension takes into account three types of variables, classifying variables, redundant variables, and non-classifying variables. In this context, the choice of the model is difficult because several roles have to be taken into account for each variable, which requires a lot of calculations, even for the reallocation of only one variable. Poon et al. (2013) also took into account the multi-partition setting, called facet determination in their article. The model considered is similar to that of Galimberti and Soffritti (2007), but it also allows tree dependency between latent class variables, resulting in the Pouch Latent Tree Models (PLTM). Model selection is performed by a greed search to maximize the BIC criterion. The resulting model allows a broad understanding of the data, but the tree structure search makes estimation even more difficult as the number of variables increases. In the previous section, we have presented a tractable multi-partition clustering algorithm not limited to continuous data.

In this section, we still suppose that the data can contain *several* clustering latent variables. But contrary to the above section where it is assumed that variables are divided into blocks each one related to some clustering of the data, we are looking for *clustering subspaces*, i.e., linear projections of the data each one related to some

particular clustering latent variable thus replacing the combinatorial question of finding the partition of the variables in independent sub-vectors by the question of finding the coefficients of the linear combinations. The proposed approach can be related to the independent factor analysis (Attias, 1999) where the author deals with source separation, in our framework a source can be interpreted as some specific clustering subspace; however, their approach becomes intractable as the numbers of sources increases and does not allow to consider multivariate subspaces. Moreover, it is not invariant up to a rotation and rescaling of the data, where our proposed methodology is.

5.3.2 Multi-Partition Subspace Mixture Model

5.3.2.1 Presentation of the Model

It is supposed that n quantitative data in dimension d are available, the data number i will be denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$, where x_{ij} is the value of variable j of data i. The whole dataset will be denoted by $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$. Let suppose that we have H class variables $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ with $g_1, \ldots, g_H$ modalities. It is assumed that $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ are independent, with $p(z_{ik}^h = 1)$ denoted by π_k^h. Let also denote by $\boldsymbol{y}_i^h$ the latent continuous features variables related to the clustering variable $\boldsymbol{z}_i^h$ such that:

$$\boldsymbol{y}_i^h | z_{ik}^h = 1 \sim \mathcal{N}_{p_h}(\boldsymbol{\nu}_k^h, I_{p_h})$$

and that we will denote by $p_\bullet = \sum_{h=1}^H p_h$, and $\boldsymbol{\nu}_k^h \in \mathbb{R}^{p_h}$.
Let also assume that it exists a latent vector $\boldsymbol{u}_i$ of non clustering variables

$$\boldsymbol{u}_i \sim \mathcal{N}_{d-p_\bullet}(\gamma, I_{d-p_\bullet}).$$

with $\gamma \in \mathbb{R}^{d-p_\bullet}$.
Let us now assume that the observed data $\boldsymbol{x}_i$ is a linear combination of the clustering and non-clustering variables:

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i^1 \\ \vdots \\ \boldsymbol{y}_i^H \\ \boldsymbol{u}_i \end{pmatrix}.$$

with $\boldsymbol{V}_h \in \mathcal{M}_{p_h,d}$ for all h in $\{1, \ldots, H\}$ and $\boldsymbol{R} \in \mathcal{M}_{d-p_\bullet,d}$
The Figure 5.2 illustrates the model in the case of $H = 2$.

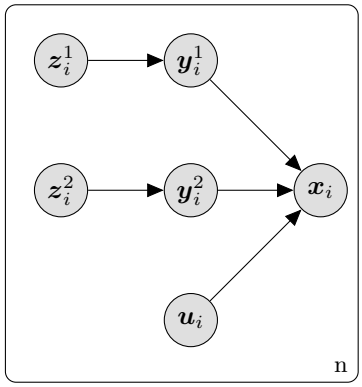


Figure 5.2: Bayesian dependency graph to the multi-partition setting, for $H = 2$ clustering variables.

Let us notice that this model allows us to visualize many clustering viewpoints in a low dimensional space since $\boldsymbol{x}_i$ can be summarized by $\boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^H$. For instance, one can assume that $p_1 = \ldots = p_H = 1$. In this case each clustering variable can be visualized on one component. We will denote by $\boldsymbol{\theta} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_H, \boldsymbol{R}, \boldsymbol{\gamma}, \boldsymbol{\nu}_1^1, ..., \boldsymbol{\nu}_{g_H}^H)$ the parameters of the model to be estimated.

5.3.2.2 Discussion about the Model

The Cartesian product of cluster spaces results in $\prod_{h=1}^{H} g_h$ clusters, which can be very large without needing many parameters. Thus the proposed model can be interpreted as being a very sparse Gaussian mixture model allowing to deal with a very large number of clusters, the resulting conditional means and covariances matrices are given in the following formulas:

$$\mathbb{E}(\boldsymbol{x}_i | z_{ik_1}^1 = 1, z_{ik_1}^2 = 1, \ldots, z_{ik_H}^H = 1) = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\nu}_{k_1}^1 \\ \vdots \\ \boldsymbol{\nu}_{k_H}^H \\ \boldsymbol{\gamma} \end{pmatrix},$$

and

$$\mathbb{V}(\boldsymbol{x}_i | z_{ik_1}^1 = 1, z_{ik_1}^2 = 2, \ldots, z_{ik_H}^H = 1) = (\boldsymbol{V}_1^{\mathrm{T}} \boldsymbol{V}_1 + \cdots + \boldsymbol{V}_H^{\mathrm{T}} \boldsymbol{V}_H + \boldsymbol{R}^{\mathrm{T}} \boldsymbol{R})^{-1}.$$

Thus, the expectation of $\boldsymbol{x}_i$ given in all the clusters is a linear combination of the cluster-specific means which can be referred to as a multiple-way MANOVA setting. On the one hand, as a particular homoscedastic Gaussian mixture, our model is more prone to model bias than free homoscedastic Gaussian mixture, and in the case when our model would be well-specified the homoscedastic Gaussian mixture would give a similar clustering for a large sample size (i.e., the same partitions with respect to the partition resulting from the product space of our multi-partitions model). On the other hand, our approach produces a factorized version of the partition space as well as the related clustering subspaces which is not a standard

output of clustering methods, and it can deal with a large number of clusters in a sparse way which can be particularly useful for a moderated sample size. In practice, the choice between our model and another mixture model can simply be performed through the BIC criterion.
In some sense, our model can be linked with the mixture of factor analyzers (Ghahramani, Hinton, et al., 1996). In mixture of factor analyzers the model is of the type:

$$\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{y}_i + \boldsymbol{u}_i,$$

where $\boldsymbol{A}$ is a low rank matrix. But here we have chosen a model of the type

$$\boldsymbol{x}_i = \boldsymbol{A}\begin{pmatrix}\boldsymbol{y}_i\\ \boldsymbol{u}_i\end{pmatrix},$$

which allows us to deal with the noise in a different way. Our model is invariant up to a bijective linear transformation of the data which is not the case for the mixtures of factor analyzers. On the other hand, our model can only deal with data with moderated dimension with respect to the number of statistical units; it assumes that the sources $\boldsymbol{y}_i$ can be recovered from the observed data $\boldsymbol{x}_i$.

5.3.3 Estimation of the Parameters of the Model

Likelihood in the supervised setting In the supervised setting the likelihood of the model can be written:

$$\begin{aligned}\ell(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) &= n\log\left|\det\begin{pmatrix}\boldsymbol{V}_1^T & \cdots & \boldsymbol{V}_H^T & \boldsymbol{R}^T\end{pmatrix}\right| - \sum_{i=1}^{n}\sum_{h=1}^{H}\sum_{k=1}^{g_h} z_{ik}^h \|\boldsymbol{V}_h^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\nu}_k^h\|^2\\ &\quad + \sum_{i=1}^{n}\sum_{h=1}^{H}\sum_{k=1}^{g_h} z_{ik}^h \log(\pi_k^h) - \sum_{i=1}^{n}\|\boldsymbol{R}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\gamma}\|^2 - \frac{n}{2}\log(2\pi).\end{aligned}$$

The likelihood cannot be maximized directly. However, in the case of $H = 1$, it reduces to the problem of Linear Discriminant Analysis (Campbell, 1984; Trevor Hastie, 1996). Let notice that if all the parameters are fixed except $\boldsymbol{V}_h$ and $\boldsymbol{R}$, $\boldsymbol{\nu}_k^h$ and $\boldsymbol{\gamma}$, the optimisation can be easily performed by constraining $\boldsymbol{V}_h^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}$ to be linear combinations of $\boldsymbol{V}_h^{(r)}$ and $\boldsymbol{R}^{(r)}$. Thus the likelihood will be optimized by using an alternate optimization algorithm. Let $\boldsymbol{M} \in \mathcal{M}_{d-p_\bullet+p_h,d-p_\bullet+p_h}(\mathbb{R})$ the matrix which allow to compute $\boldsymbol{V}_h^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}$ based on $\boldsymbol{V}_h^{(r)}$ and $\boldsymbol{R}^{(r)}$:

$$\begin{pmatrix}\boldsymbol{V}_h^{(r+1)}\\ \boldsymbol{R}^{(r+1)}\end{pmatrix} = \boldsymbol{M}\begin{pmatrix}\boldsymbol{V}_h^{(r)}\\ \boldsymbol{R}^{(r)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{M}_1\\ \boldsymbol{M}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_h^{(r)}\\ \boldsymbol{R}^{(r)}\end{pmatrix},$$

where $\boldsymbol{M}_1$ is the sub-matrix containing the p_h first rows of $\boldsymbol{M}$ and $\boldsymbol{M}_2$ the matrix containing the last $d - p_\bullet$ rows of $\boldsymbol{M}$. By denoting

$$\begin{pmatrix}\boldsymbol{y}_i^{h(r)}\\ \boldsymbol{u}_i{}^{(r)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{V}_h^{(r)}\\ \boldsymbol{R}^{(r)}\end{pmatrix}\boldsymbol{x}_i,$$

Thus $\boldsymbol{M}$ and the others parameters can be obtained by applying simple Linear Discriminant to previously projected data $(\boldsymbol{y}_i^{h(r)}, \boldsymbol{u}_i{}^{(r)})$.

Unsupervised setting: EM algorithm In practice $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ are unknown. Consequently, we will use an EM algorithm to "reconstitute the missing label" in order to maximize the likelihood, except that the data at each iteration are now weighted by $t_{ik}^{h\,(r+1)}$ instead of z_{ik}^h.
The algorithm is the following:

- Until convergence, for $h \in \{1, \ldots, H\}$ iterates the following steps:
  - E step: compute

$$t_{ik}^{h\,(r+1)} = \frac{\pi_k p(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_k^{h(r)}, \boldsymbol{I}_p)}{\sum_{k'=1}^{g} \pi_{k'} p(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_{k'}^{h\,(r)}, \boldsymbol{I}_p)}.$$

  - M step: compute $\pi_1^{h(r+1)}, \ldots, \pi_{g_h}^{h\,(r+1)}$, $\boldsymbol{V}_h^{(r+1)}$, $\boldsymbol{R}^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\nu}_k^{h(r+1)}$ based on formulas given in the supervised setting.

5.3.3.1 Model Choice

The proposed model needs the user to define the number of clustering subspaces H, the number of cluster in each clustering subspace $g_1, \ldots, g_H$, and the dimensionality $p_1, \ldots, p_H$ of each subspace. The constraints are that $H < d$, that $p_h \leq g_h - 1$ and $p_\bullet = p_1 + \cdots + p_H < d$. It is clear that the number of possible models can become very high. To limit the combinatorial aspect, one can impose $g_1 = \cdots = g_H = g$ and/or $p_1 = \cdots = p_h = p$. In practice the choice of $p = 1$ enforces to find clustering which could be visualized in one dimension, which can help the practitioner. Moreover, choosing $g = 2$ is the minimal requirement in order to investigate a clustering structure. However, if possible we recommend to explore the largest possible number of models and choosing the best one with the BIC. For a given model $\boldsymbol{m}$ the BIC is computed as:

$$BIC(\boldsymbol{m}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}; \mathbf{x}) - \frac{\nu_{\boldsymbol{m}}}{2} \log n,$$

where $\nu_{\boldsymbol{m}}$ is the number of parameters of the model. Thus the model choice consists of choosing the model maximizing the BIC. Let notice that in practice the user could be mainly interested by a low value of H, since even $H = 2$ can provide him with new insights about his data, focusing on finding several clustering view points.

5.3.4 Experiments on real data

Here results of the model are just presented on the crabs datasef (Campbell and Mahon, 1974), experiments on simulated data can be found in Vandewalle (2020). The crabs dataset consists of 200 crabs morphological data, each crab has two categorical (cluster) attributes—the species, orange or blue, and the sex, male or female. The dataset is composed of 50 males orange, 50 males blue, 50 females orange, 50 females blue for which 5 numerical attributes have been measured: the frontal lobe size, the rear width, the carapace length, the carapace width, and the body depth. We can see the PCA of the data in Figure 5.3. We see that component

two separates males and females well, whereas component three separates orange and blue subspecies. However, we will see that by applying our model we obtained a better separation of the clusters.
We will take $H = 2$, $p_1 = p_2 = 1$ and $g_1, g_2 \in \{1, \ldots, 5\}$. The resulting BIC tabular is given in Table 5.11, it suggests the choice of $g_1 = 3$ and $g_2 = 4$. The resulting visualization of the clustering variables in given Figure 5.4. Let us notice that Y_2 is divided into four clusters, however, we only see three since two of them have the same mean. We can see that even if the numbers of clusters do not correspond, the first clustering subspace finds the subspecies, whereas the second clustering subspace finds the sex. We could also look at the solution provided by $g_1 = g_2 = 2$ on Figure 5.5, this one has a lower BIC but seems more natural for the problem at hand. We see that the obtained map is in fact quite similar to the map obtained Figure 5.4; however, we notice that from a density approximation point of view we obtain a lower fit. In fact, if we look at the correlations between Y_1 Figure 5.4 and Y_2 Figure 5.5 we have a correlation of -0.97, and a similar correlation is obtained between Y_2 Figure 5.4 and Y_1 Figure 5.5. Thus, the produced subspace is finally quite similar.

Table 5.11: Value of the BIC criterion according to g_1 and g_2, for the choice of the number of clusters on the crabs dataset, best value in bold.

$g_1 \setminus g_2$	**1**	**2**	**3**	**4**	**5**
1	−62.66	0.41	10.40	5.11	0.80
2		17.82	16.57	18.88	0.49
3			3.75	**22.52**	17.44
4				−26.65	−26.64
5					12.06

5.3.5 Conclusions

We have presented a model that allows us to combine visualization and clustering with many clustering viewpoints. Moreover, we have shown the possibility of performing model choice by using the BIC criterion. The proposed model can provide new information on the structure present in the data by trying to reinterpret the cluster as a result of the Cartesian product of several clustering variables.
The proposed model is limited to the homoscedastic setting, which could be seen as a limitation; however, from our point of view this is more robust than the heteroscedastic setting, which is known to be jeopardized by the degeneracy issue (Biernacki and Chrétien, 2003). However, the extension of our work on the heteroscedastic setting can easily be performed from the modeling point of view; the main issue, in this case, would be the estimation of the parameters where an extension of the FDA to the heteroscedastic setting would be needed, as presented in Kumar and Andreou (1998). Another difficult issue is the choice of H, $g_1 \ldots, g_H$ and $p_1, \ldots, p_H$, which

is very combinatorial. Here we have proposed an estimation strategy for all these tuning parameters being fixed, and then performed a selection of the best tuning according to BIC. However, in future work, a model selection strategy to perform the model selection through a modified version of the EM algorithm will also be investigated as in Green (1990); it would thus limit the combinatorial aspect of the global model search through EM-wise local model searches.

5.4 Conclusion and perspectives

This chapter has presented two models for performing multiple partition clustering. The first model has the advantage to be very sparse a merits some additional study such as the choice of the number of blocks, and the number of clusters in each block, which are really important issues for the model to be used in practice. All these models could also include covariates which could be important from the interpretation point of view, where each partition could be particularly liked to some additional covariates. These perspectives are further detailed in Chapter 10.

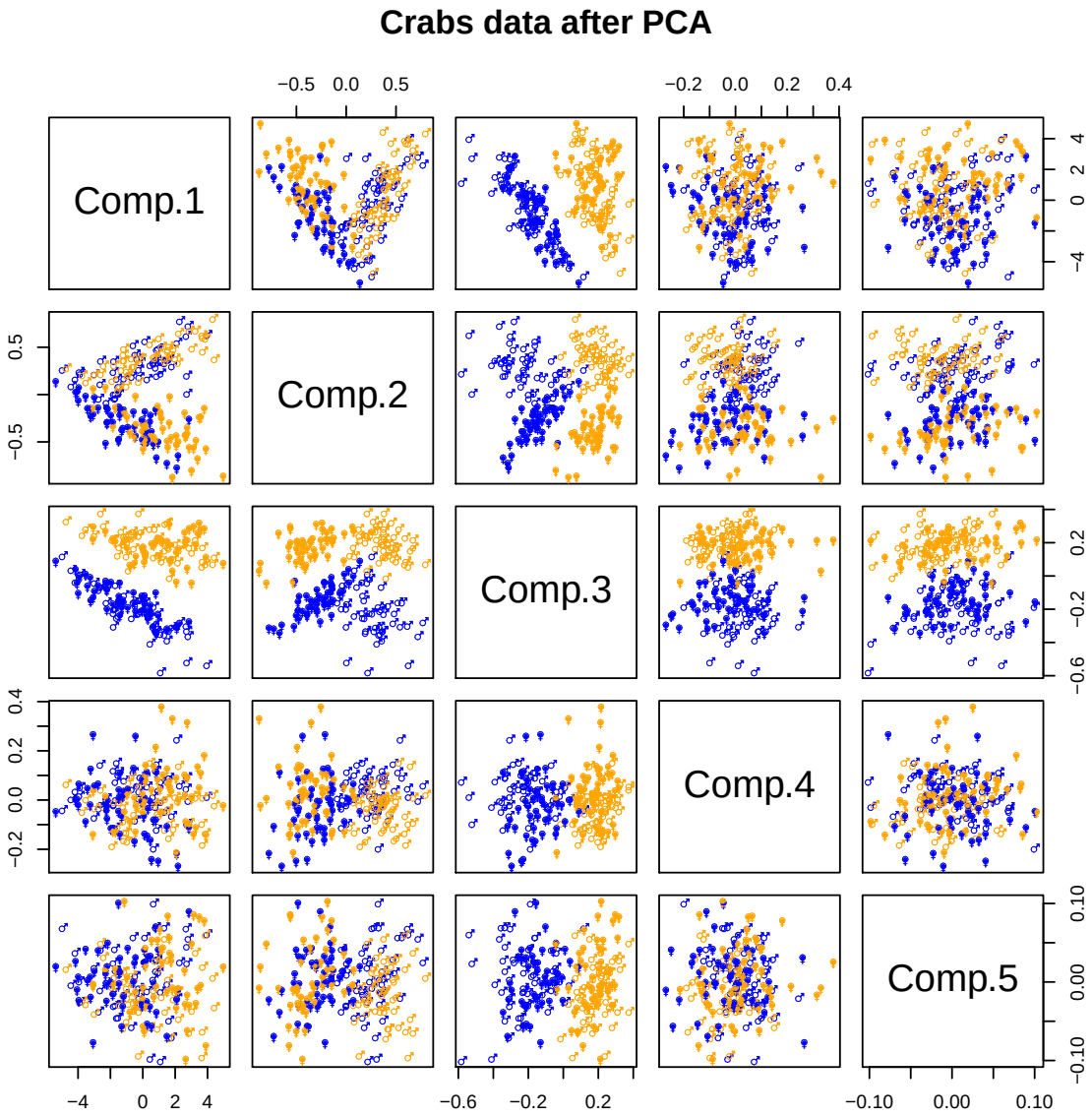


Figure 5.3: Scatter plots of the crabs data after PCA. Subspecies are represented according to their color, and sex is represented according to its symbol.

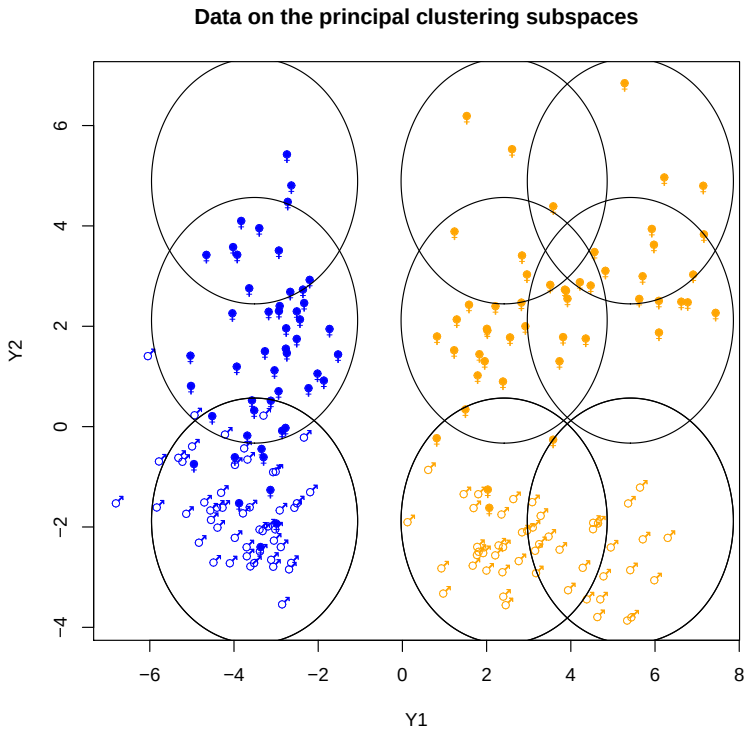


Figure 5.4: Scatter plots of the clustering subspace on the crabs data for $g_1 = 3$ and $g_2 = 4$, 95% isodensity is given for each component resulting of the Cartesian product.

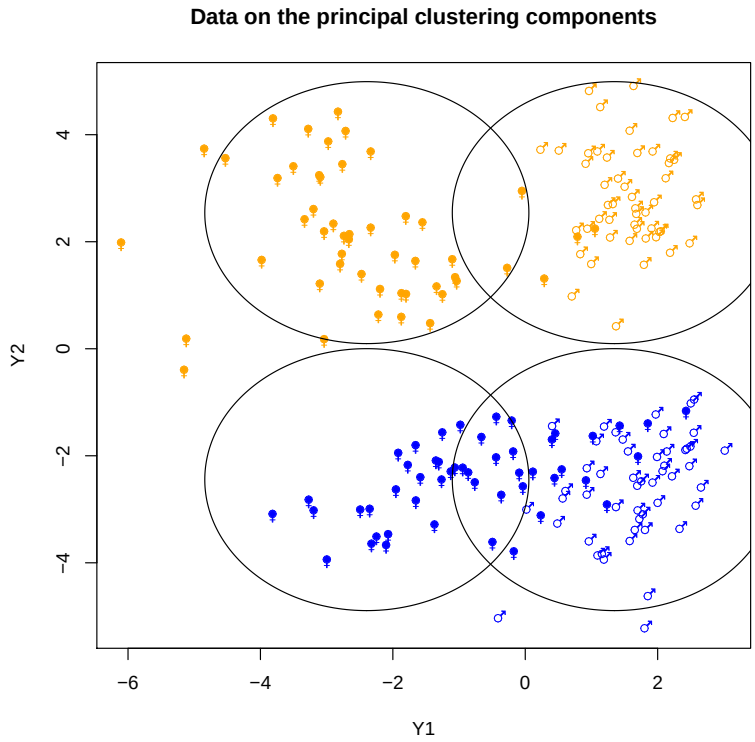


Figure 5.5: Scatter plots of the clustering subspace on the crabs data for $g_1 = g_2 = 2$, 95% isodensity is given for each component resulting of the Cartesian product.

Bibliography

Allman, E., Matias, C., and Rhodes, J. (2009). "Identifiability of parameters in latent structure models with many observed variables". In: *The Annals of Statistics* 37.6A, pp. 3099–3132.

Attias, H. (1999). "Independent factor analysis". In: *Neural computation* 11.4, pp. 803–851.

Biernacki, C., Celeux, G., and Govaert, G. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.

Biernacki, C. and Chrétien, S. (2003). "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM". In: *Statistics & probability letters* 61.4, pp. 373–382.

Bouveyron, C. and Brunet, C. (2012). "Simultaneous model-based clustering and visualization in the Fisher discriminative subspace". In: *Statistics and Computing* 22.1, pp. 301–324.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*. Vol. 50. Cambridge University Press.

Bouveyron, C., Girard, S., and Schmid, C. (2007). "High-dimensional data clustering". In: *Computational Statistics & Data Analysis* 52.1, pp. 502–519.

Campbell, N. A. (1984). "Canonical variate analysis - A general model formulation". In: *Australian Journal of Statistics* 26.1, pp. 86–96. ISSN: 1467-842X. DOI: 10.1111/j.1467-842X.1984.tb01271.x. URL: http://dx.doi.org/10.1111/j.1467-842X.1984.tb01271.x.

Campbell, N. A. and Mahon, R. J. (1974). "A multivariate study of variation in two species of rock crab of the genus Leptograpsus". In: *Australian Journal of Zoology* 22.3, pp. 417–425.

Celeux, G. and Govaert, G. (1995). "Gaussian parsimonious clustering models". In: *Pattern recognition* 28.5, pp. 781–793.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

Forina, M., Armanino, C., Castino, M., and Ubigli, M. (1986). "Multivariate data analysis as a discriminating method of the origin of wines". In: *Vitis* 25.3, pp. 189–201.

Galimberti, G., Manisi, A., and Soffritti, G. (2018). "Modelling the role of variables in model-based cluster analysis". In: *Statistics and Computing* 28.1, pp. 145–169.

Galimberti, G. and Soffritti, G. (2007). "Model-based methods to identify multiple cluster structures in a data set". In: *Computational Statistics & Data Analysis* 52.1, pp. 520–536. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.

2007.02.019. URL: http://www.sciencedirect.com/science/article/pii/S0167947307000758.

Ghahramani, Z., Hinton, G. E., et al. (1996). *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto.

Govaert, G. and Nadif, M. (2003). "Clustering with block mixture models". In: *Pattern Recognition* 36.2, pp. 463–473.

Green, P. J. (1990). "On use of the EM for penalized likelihood estimation". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 443–452.

Hand, D. and Keming, Y. (2001). "Idiot's Bayes, not so stupid after all?" In: *International statistical review* 69.3, pp. 385–398.

Hennig, C. (2015). "What are the true clusters?" In: *Pattern Recognition Letters* 64, pp. 53–62.

Kumar, N. and Andreou, A. G. (1998). "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition". In: *Speech communication* 26.4, pp. 283–297.

Luxburg, U. von, Williamson, R. C., and Guyon, I. (2012). "Clustering: Science or Art?" In: ed. by I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver. Vol. 27. Proceedings of Machine Learning Research. Bellevue, Washington, USA: JMLR Workshop and Conference Proceedings, pp. 65–79. URL: http://proceedings.mlr.press/v27/luxburg12a.html.

Marbac, M., Patin, E., and Sedki, M. (2018). "Variable selection for mixed data clustering: a model-based approach". In: *Journal of Classification* to appear.

Marbac, M. and Sedki, M. (2017). "Variable selection for model-based clustering using the integrated complete-data likelihood". In: *Statistics and Computing* 27.4, pp. 1049–1063.

Marbac, M. and Vandewalle, V. (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). "Variable selection for clustering with Gaussian mixture models". In: *Biometrics* 65.3, pp. 701–709.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm*. Wiley-Interscience, New York: Wiley Series in Probability, Statistics: Applied Probability, and Statistics.

McLachlan, G. J. and Peel, D. (2000). *Finite mixutre models*. Wiley-Interscience, New York: Wiley Series in Probability, Statistics: Apllied Probability, and Statistics.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McNicholas, P. D. (2016). *Mixture model-based classification*. CRC press.

McNicholas, P. D., ElSherbiny, A., McDaid, A., and Murphy, B. (2018). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.2. URL: https://CRAN.R-project.org/package=pgmm.

McNicholas, P. D. and Murphy, T. (2008). "Parsimonious Gaussian mixture models". In: *Statistics and Computing* 18.3, pp. 285–296. ISSN: 1573-1375. DOI: 10.1007/s11222-008-9056-0. URL: https://doi.org/10.1007/s11222-008-9056-0.

Moustaki, I. and Papageorgiou, I. (2005). "Latent class models for mixed variables with applications in Archaeometry". In: *Computational statistics & data analysis* 48.3, pp. 659–675.

Poon, L. K., Zhang, N. L., Liu, T., and Liu, A. H. (2013). "Model-based clustering of high-dimensional data: Variable selection versus facet determination". In: *International Journal of Approximate Reasoning* 54.1, pp. 196–215.

Raftery, A. E. and Dean, N. (2006). "Variable Selection for Model-Based Clustering". In: *Journal of the American Statistical Association* 101.473, pp. 168–178. DOI: 10.1198/016214506000000113. eprint: https://doi.org/10.1198/016214506000000113. URL: https://doi.org/10.1198/016214506000000113.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Teicher, H. (1963). "Identifiability of Finite Mixtures". In: *The Annals of Mathematical Statistics*, pp. 1265–1269.

— (1967). "Identifiability of mixtures of product measures". In: *Annals of Mathematical Statistics* 38, pp. 1300–1302.

Tipping, M. E. and Bishop, C. M. (1999). "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.

Trevor Hastie, R. T. (1996). "Discriminant Analysis by Gaussian Mixtures". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 155–176. ISSN: 00359246. URL: http://www.jstor.org/stable/2346171.

Vandewalle, V. (2020). "Multi-Partitions Subspace Clustering". In: *Mathematics* 8.4, p. 597.

Witten, D. M. and Tibshirani, R. (2010). "A framework for feature selection in clustering". In: *Journal of the American Statistical Association* 105.490, pp. 713–726.

Chapter 6

Contribution to general issues in model-based clustering

Contents

6.1 Introduction . . . 104
6.2 Dealing with the label switching . . . 104
6.2.1 State of the art . . . 105
6.2.2 Reminding of the label switching problem . . . 105
6.2.3 Our proposal: posterior distribution restricted by the partition 106
6.2.4 Sampling according to a Gibbs algorithm . . . 107
6.2.5 Conclusion . . . 108
6.3 Missing data . . . 108
6.3.1 Gaussian mixture with missing data . . . 108
6.3.2 Distance estimation with missing data . . . 109
6.3.3 Degeneracy in mixture . . . 109
6.4 Visualization in mixture . . . 113
6.4.1 Introduction . . . 113
6.4.2 Possible mapping strategies . . . 114
6.4.3 New proposed method: controlling the distribution family . . 116
6.4.4 Example . . . 117
6.4.5 Conclusion . . . 119
6.5 Conclusion . . . 119

Related scientific production

1. C. Biernacki and V. Vandewalle (2011). "Label switching in mixtures". In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401

2. E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki (2014). "Mixture of Gaussians for distance estimation with missing data". In: *Neurocomputing* 131, pp. 32–42

3. V. Vandewalle and C. Biernacki (2015). "An efficient SEM algorithm for Gaussian Mixtures with missing data". In: *8th International Conference of the ERCIM WG on Computational and Methodological Statistics*. Londres, United Kingdom. URL: https://hal.inria.fr/hal-01242588

4. C. Biernacki, M. Marbac, and V. Vandewalle (2020). "Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering". In: *Journal of Classification*. DOI: 10.1007/s00357-020-09369-y. URL: https://doi.org/10.1007/s00357-020-09369-y

6.1 Introduction

Most of my recent research production has focused on proposing models adapted to various data setting, such as categorical, functional, or for the multiple partition clustering. This is important in fact to develop many models to be able to manage clustering in various frameworks. However, there are also many general issues in mixture models that are of interest whatever the model used and that may have a great impact.

In this scope, I have worked on proposing a solution to the label switching issue which is generic in the inference of parameters of a mixture in a Bayesian setting (Biernacki and Vandewalle, 2011). In the framework of missing data, I have been also interested in distance estimation when some variables are missing based on a mixture model (Eirola et al., 2014), this has raised the issue of the degeneracy of the EM algorithm when considering missing data for which we have proposed some solutions (Vandewalle and Biernacki, 2015), but it is still a work in progress. Finally, I have been interested in the question of visualization of a mixture model, in this framework, we propose a method that can be applied for the visualization of any mixture model (Biernacki, Marbac, and Vandewalle, 2020).

6.2 Dealing with the label switching

In this section, I present and solution to the label switching issue that we have developed with Christophe Biernacki. This work has been presented in a conference with proceedings (Biernacki and Vandewalle, 2011). However, after some additional research, we found that a similar solution had also been proposed in Papastamoulis and Iliopoulos (2010), thus stopping our work for this solution. We have also initiated some work with Benjamin Guedj, on this issue. Our goal was to quantify the probability of label switching, according to the class separation, the number of iterations, the sample size, ... The main intuition being that this probability is often very low, thus justifying the current practice to ignore label switching.

6.2.1 State of the art

During the last fifteen years, there has been an increasing interest in using Bayesian methods in mixtures models. A reason for this success is the emergence of MCMC methods. However, one of the principal issues of these methods is the non-identifiability of components caused by symmetric prior, which makes the Gibbs outputs often useless for inference; this problem is known as label switching. The four main ones are now reminded. One can impose identifiability constraints on the parameters (Diebolt and Robert, 1994), use relabelling algorithms on the generated parameters (M. Stephens and Phil, 1997; Celeux, 1998), use loss invariants functions up to the parameters permutation (Celeux, Hurn, and Robert, 2000; Frühwirth-Schnatter, 2006) or use a probabilistic approach to take into account the uncertainty of the choice of the permutation of the parameters (Jasra, Holmes, and D. A. Stephens, 2005; Sperrin, Jaki, and Wit, 2010). These methods allow to partially solve the problem, however, they can be inefficient in practice or they do not take into account the uncertainty of the relabelling or it can be hard to define and optimize the invariant loss function or they can require a visual calibration. For more details see the state of the art made by Jasra, Holmes, and D. A. Stephens (2005).
We propose a posterior distribution for which the latent partition is restrained up to a particular numbering leading to the greatest separation with its permutations. A Gibbs algorithm that allows sampling easily according to this new distribution is detailed. The two main advantages of the proposed method, besides its real simple implementation, are on the first hand to take into account the uncertainty of the parameter permutation labeling and on an other hand, the needless of assumptions on the non switched posterior distribution.

6.2.2 Reminding of the label switching problem

Let a generic mixture of g distributions as defined in Chapter 2

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}|\boldsymbol{\alpha}_k)$$

where $\boldsymbol{\theta}_k = (\pi_k, \boldsymbol{\alpha}_k)$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$.
Starting from a n i.i.d. sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ coming from $p(\boldsymbol{x}|\boldsymbol{\theta})$ and a prior distribution $p(\boldsymbol{\theta})$, any Bayesian inference is based on the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Notice that $p(\boldsymbol{\theta}|\mathbf{x})$ is invariant up to a renumbering of the mixture components as soon as $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are invariant up to a renumbering. In other words, let $\mathcal{P}_g$ the set of the permutations of $\{1, \ldots, g\}$ and $\sigma(\boldsymbol{\theta}) = (\boldsymbol{\theta}_{\sigma(1)}, \ldots, \boldsymbol{\theta}_{\sigma(g)})$ be the parameter $\boldsymbol{\theta}$ permuted in index with $\sigma \in \mathcal{P}_g$, we have $p(\boldsymbol{\theta}|\mathbf{x}) = p(\sigma(\boldsymbol{\theta})|\mathbf{x})$ for every $\sigma \in \mathcal{P}_g$. This exact symmetry of the posterior distribution, also called label switching problem, makes meaningless direct computation of many usual punctual estimators as the posterior mean.

6.2.3 Our proposal: posterior distribution restricted by the partition

In order to solve the label switching problem, we propose to condition the posterior distribution by a particular numbering, not on the parameter $\boldsymbol{\theta}$ as it is usually done, but rather on a latent partition.
Let $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathcal{Z}$ the latent partition which has been used to generate $\mathbf{x}$. Let $\tilde{\mathcal{Z}} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_{g!}\}$ a stratification of the set of all the partitions $\mathcal{Z}$, each strata can be deduced from one another by a simple indexes permutation :

$$\forall h, h' \in \{1, \ldots, g!\}, \exists! \sigma \in \mathcal{P}_g \text{ such that } \mathbf{z} \in \mathcal{Z}_h \Leftrightarrow \sigma(\mathbf{z}) \in \mathcal{Z}_{h'}$$

with $\sigma(\mathbf{z}) = (\sigma(\boldsymbol{z}_1), \ldots, \sigma(\boldsymbol{z}_n))$ meaning that $\mathbf{z}$ is permuted in index for $\sigma \in \mathcal{P}_g$. In this framework, the standard posterior distribution can be written as a mixture of $g!$ posterior distributions given any particular numbering $\tilde{\mathcal{Z}}$ of partitions

$$p(\boldsymbol{\theta}|\mathbf{x}) = \sum_{h=1}^{g!} p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h) p(\mathcal{Z}_h|\mathbf{x}) = \frac{1}{g!} \sum_{h=1}^{g!} p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h).$$

Notice that, contrary to $p(\boldsymbol{\theta}|\mathbf{x})$, the distributions $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$ are not any more strictly invariant up to $\mathbf{z}$ renumbering. The conditioning on $\mathcal{Z}_h$ can be interpreted as the addition of the information that the partition $\mathbf{z}$ which has been used to generate $\mathbf{x}$ comes from a particular numbering. The importance of the asymmetry of $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$ clearly depends on the choice of $\tilde{\mathcal{Z}}$. In order to get the furthest from symmetry, the key idea of our proposal is then to choose a stratification $\tilde{\mathcal{Z}}$ which separates the best the distributions $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$ of this mixture and then to retain as the new posterior distribution any of these $g!$ distributions $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$, for instance $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_1)$, the choice of a particular h is arbitrary and without any consequence.
A first natural choice of $\tilde{\mathcal{Z}}$, noted $\tilde{\mathcal{Z}}^{KL}$, is the choice that leads to the largest Kullback-Leibler divergence between the components of the mixture on $\mathcal{Z}_h$ and can be written

$$\tilde{\mathcal{Z}}^{KL} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \int_{\Theta} p(\theta|x, \mathcal{Z}_1) \ln \left[\frac{p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_1)}{p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)} \right] d\theta.$$

This criterion is intractable even for very small sample sizes because of the combinatorial on the number of partitions, that is why we propose a simpler criterion, maximizing the difference between the distributions in a particular $\boldsymbol{\theta}$ instead of the whole parameters space Θ. We keep for this task the maximum a posteriori estimator $\boldsymbol{\theta}^{MAP}$, which gives a new optimal numbering noted $\tilde{\mathcal{Z}}^{MAP}$ defined by

$$\tilde{\mathcal{Z}}^{MAP} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \frac{p(\boldsymbol{\theta}^{MAP}|\mathbf{x}, \mathcal{Z}_1)}{p(\boldsymbol{\theta}^{MAP}|\mathbf{x}, \mathcal{Z}_h)}.$$

In practice, $\tilde{\mathcal{Z}}^{MAP}$ is straightforward to compute for any sample size since it consists in taking the most likely numbering for each statistical unit computed in θ^{MAP} :

$$\mathcal{Z}_1^{MAP} = \left\{ \mathbf{z} \in \mathcal{Z} / Id = \arg\max_{\sigma \in \mathcal{P}_g} p(\sigma(\mathbf{z})|\mathbf{x}, \theta^{MAP}) \right\},$$

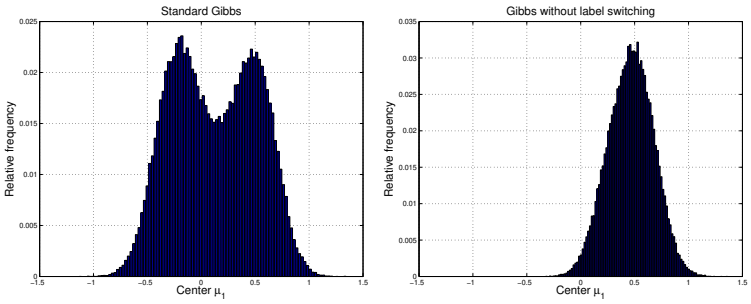


Figure 6.1: (a): usual posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$, (b) proposed posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_1^{MAP})$.

where Id stands for the identity permutation. θ^{MAP} can be interpreted as a reference parameter for the latent partition numbering.

6.2.4 Sampling according to a Gibbs algorithm

The standard Gibbs algorithm is slightly modified compared to Chapter 2 to take into account the conditioning on $\mathcal{Z}_1$ but is still easy (see Algorithm 5).

Algorithm 5 Gibbs sampling algorithm for mixture unswitched

start from $\boldsymbol{\theta}^{(0)}$
for $r = 1$ to $r_{\max}$ **do**
  for $i = 0$ to n **do**
    Sample $\boldsymbol{z}_i^{(r)}$ from $\boldsymbol{z}_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(r-1)}$
  permute $\mathbf{z}^{(r)}$ so that $\sigma(\mathbf{z}^{(r)}) \in {\mathcal{Z}_1}^{KL}$ or $\sigma(\mathbf{z}^{(r)}) \in {\mathcal{Z}_1}^{MAP}$ (according the selected criterion)
  Sample $\boldsymbol{\pi}^{(r)}$ from $\boldsymbol{\pi}|\mathbf{z}^{(r)}$
  for $k = 1$ to g **do**
    Sample $\boldsymbol{\alpha}_k^{(r)}$ from $\boldsymbol{\alpha}_k|\{\boldsymbol{x}_i/z_{ik}^{(r)} = 1\}$
return $(\mathbf{z}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\mathbf{z}^{(r)}, \boldsymbol{\theta}^{(r)})$

All the $g!$ permutations do not need to be investigated since the Hungarian algorithm can be used to solve the optimal affectation.

Figure 6.1 illustrates the posterior distribution obtained through a Gibbs sampling for the expectation of a mixture of Gaussian, with the standard output of the Gibbs from one part, and our proposed posterior distribution from the other part. We see that the proposed solution offers a good solution to the label switching issue by giving a relevant non-switched posterior distribution. However, such a solution requires to perform a maximum likelihood estimation of the parameters before it can be used.

6.2.5 Conclusion

We have proposed a solution to solve the label switching among others. However, this problem is still often ignored in practice since in many cases of interest when classes are well separated, and enough data are available, the switching probability is very low.

6.3 Missing data

Mixture model can easily deal with missing data (Ghahramani and Jordan, 1995; Hunt and Jorgensen, 2003). This adds a second level of missing data in addition to the class label. Then the obtained clustering can be interpreted by itself, but it also gives access to the expectation of the missing values given the observed values. In the next section, we present the related EM algorithm in the multivariate Gaussian setting.

6.3.1 Gaussian mixture with missing data

Let consider the full sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ coming from a mixture of g Gaussian component in dimension d. Let denote by $O_i \subseteq \{1, \ldots, d\}$ the set of observed variables for the individual i and by M_i the complementary set for the missing variables. Let $\boldsymbol{x}_i^o$ denote the observed variables for unit i, and $\mathbf{x}^o$ the observed dataset.
Let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ be the mean and variance covariance matrix of the class k. Let $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\boldsymbol{\theta}$ be the global parameter of the mixture. Finally $\phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian density of expectation $\boldsymbol{\mu}_k$ and variance covariance matrix $\boldsymbol{\Sigma}_k$. $\boldsymbol{\mu}_{ik}^o$ is the sub-vector of $\boldsymbol{\mu}_k$ associated to the index O_i (idem for M_i). $\boldsymbol{\Sigma}_{ik}^{om}$ is the sub-matrix of $\boldsymbol{\Sigma}_k$ associated to the rows O_i and columns M_i (as the same for any combination $\boldsymbol{\Sigma}_{ik}^{oo}$, ...).

EM algorithm with missing data The EM algorithm allowing to take into account missing data is the following, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^+$ are respectively parameters at two steps (*idem* for missing data):
E step: missing data

$$t_{ik}^+ = \frac{\pi_k \phi(\boldsymbol{x}_i^o; \boldsymbol{\alpha}_k)}{\sum_{k'=1}^g \pi_{k'} \phi(\boldsymbol{x}_i^o; \boldsymbol{\alpha}_{k'})} \quad \text{and} \quad \boldsymbol{x}_{ik}^{m+} = \boldsymbol{\mu}_{ik}^m + \boldsymbol{\Sigma}_{ik}^{mo} \left(\boldsymbol{\Sigma}_{ik}^{oo}\right)^{-1} (\boldsymbol{x}_i^o - \boldsymbol{\mu}_{ik}^o).$$

$\boldsymbol{x}_{ik}^{m+}$ can be interpreted as an imputation of missing data given the class and the observed variables.
M step: parameters

$$\pi_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+, \ \boldsymbol{\mu}_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+ \boldsymbol{x}_{ik}^+ \quad \text{and} \quad \boldsymbol{\Sigma}_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+ \left[(\boldsymbol{x}_{ik}^+ - \boldsymbol{\mu}_k^+)(\boldsymbol{x}_{ik}^+ - \boldsymbol{\mu}_k^+)' + \boldsymbol{\Sigma}_{ik}^+ \right]$$

where $n_k^+ = \sum_{i=1}^n t_{ik}^+$, $\boldsymbol{x}_{ik}^+ = \begin{pmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{x}_{ik}^{m+} \end{pmatrix}$ and $\boldsymbol{\Sigma}_{ik}^+ = \begin{pmatrix} 0_i^o & 0_i^{om} \\ 0_i^{mo} & \boldsymbol{\Sigma}_{ik}^{m+} \end{pmatrix}$ with 0 being the null matrix $d \times d$ and $\boldsymbol{\Sigma}_{ik}^{m+} = \boldsymbol{\Sigma}_{ik}^{mm} - \boldsymbol{\Sigma}_{ik}^{mo} \left(\boldsymbol{\Sigma}_{ik}^{oo}\right)^{-1} \boldsymbol{\Sigma}_{ik}^{om}$. The term $\boldsymbol{\Sigma}_{ik}^{m+}$ can be interpreted as a variance correction to take into account the underestimation of the variance caused by the "missing data imputation".
The iteration of these two steps leads to an increase in the likelihood at each iteration. Let notice that the algorithm is very near to the standard EM algorithm in Gaussian mixture with complete data, the differences being the imputation of missing values, and the variance correction. Let also notice here, that the value of the parameter at the next step depends on the parameters at the previous step not only through the weight but also in the update formulas. This possibly explains the slow degeneracy that can be observed with missing data.

6.3.2 Distance estimation with missing data

In the article Eirola et al. (2014), by using a mixture of Gaussian with missing data we can compute the expected distance between two individuals, which is particularly useful for supervised learning algorithms such as kernel methods. The proposed approach has the advantage to avoid any prior imputation to estimate these distances. It is needed to compute $\mathbb{E}[\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 | \boldsymbol{x}_i^o, \boldsymbol{x}_{i'}^o]$ which is straightforward in the mixture Gaussian setting framework, since it only requires to compute conditional expectations and variances given the observed variables in the Gaussian setting. In Eirola et al. (2014), we also propose to use sparse Gaussian mixture models (Bouveyron, Girard, and Schmid, 2007) which is important when the number of individuals is low compared with the number of variables.
Given this experience, we have observed that this context of missing data in the mixture setting could point out the problem of degeneracy in mixtures (convergence to a degenerate solution), and that this problem was hard to detect, since in this case we have observed that convergence is very slow thus risking to consider as valid a degenerated solution, for which we have proposed solutions in conferences (Vandewalle and Biernacki, 2015; Biernacki, Castellan, et al., 2016) but not yet published article. In the next section, we briefly discuss this issue, however, it is still a work in progress.

6.3.3 Degeneracy in mixture

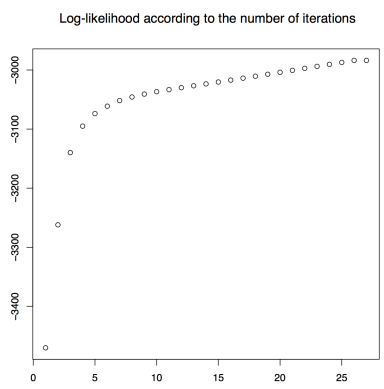
6.3.3.1 State of the art

In heteroscedastic Gaussian mixtures with complete data, the likelihood is known to be unbounded. For instance, in the univariate framework, a degenerated solution can be reached by setting the mean of one component equal to an observed data and letting tend its variance to zero (Day, 1969). These degenerated solutions on the border of the parameters space are out of interest, and a root of the gradient of the likelihood is searched because it is known that one of them is consistent (Redner and Walker, 1984). In practice, the EM algorithm is used to estimate the parameters.

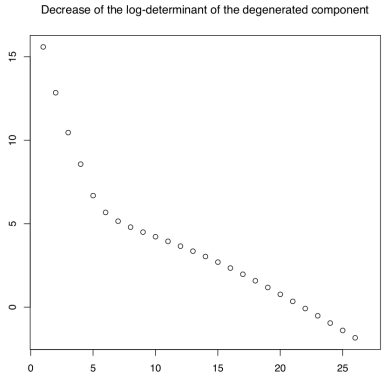
When it encounters a degenerated solution the likelihood goes to infinity, which is a symptom that the parameters are close to the border.
In the univariate (Biernacki and Chrétien, 2003) and multivariate (Ingrassia and Rocci, 2007) complete data settings, the degeneracy is very fast. In univariate Gaussian mixture models with binned data, the likelihood stays bounded however the degeneracy problem remains. In fact, when all the non-empty intervals are small enough, the global maximum of the likelihood is located on the border of the parameters space (Biernacki, 2007). In this case, the EM algorithm can be trapped by a degenerated solution with a very slow speed.
The solutions to avoid the problem of degeneracy are either to modify the estimator of the parameters or to detect it through the dynamic of the algorithm EM. On the one hand, the parameters estimator can be regularized through a prior in the Bayesian framework or by adding constraints to avoid the border of the parameters space (Snoussi and Mohammad-Djafari, 2001). On the other hand, the degeneracy can be detected through the dynamic of the EM algorithm. In the complete data case, the degeneracy is so fast that it is always detected, and it is just needed to restart the algorithm. In the binned data case, it can be detected by using the bound developed in Biernacki (2007).
Very few results are available for the degeneracy of Gaussian mixtures with some missing data. However, with the increasing of the number of available variables, the risk that data contain missing values also increases. The framework of missing data is halfway from the multivariate framework with complete data and the univariate binned data framework. The likelihood stays unbounded, as with complete data, but missing data can be interpreted as intervals of infinite size.
In this section, we are interested in the degeneracy in the missing data framework and characterize its dynamic.

6.3.3.2 Illustration of degeneracy

To illustrate the problem degeneracy let consider the dataset breast cancer tissue of the UCI database repository. It is composed of 106 statistical units and 9 variables. We have artificially hidden 10% of the data completely at random. Then we have adjusted a mixture model with 4 clusters.
The evolution of the log-likelihood for a degenerated solution is given in Figure 6.2a. We can see that the growth of the log-likelihood seems to be linear. If we look at the evolution of the log-determinant of the component with the lowest variance Figure 6.2b we also see that it seems to decrease linearly as the number of iterations increases. Moreover, if we look at the data related to the degenerated solution Table 6.1 we see that the number of data in this component is equal to 14, thus greater than the number of variables, but the number of complete data is equal to five which is lower than the number of variables.
Thus we see that the convergence toward a degenerated component is relatively slow (log-likelihood linear according to the number of iterations) and that the number of points of the degenerated solution is greater than the space dimension d (but the



(a) Evolution of the log-likelihood according to the number of iterations of a degeneracy path



(b) Evolution of the log-determinent according to the number of iterations of a degeneracy path

Figure 6.2: Evolution of the log-likelihood and of the log-determinent according to the number of iterations.

number of complete points is lower than d).

6.3.3.3 Degeneracy speed on a toy example

In order to guess the dynamic of the EM algorithm, let start with a univariate framework without mixture and with only one datum that we will denote by x. In this framework the estimated mean is equal to x and the estimated variance is equal to 0 which leads to an infinite likelihood. Let imagine that in addition to this observed datum, $n-1$ data have not been observed. Then, it is possible to perform an EM algorithm (useless here) which will converge toward the expected degenerated solution. In this oversimplified instance, it is possible to explicitly express the algorithm pathway, updating formulas at the M step become:

$$\mu^+ = \frac{(n-1)\mu + x}{n} \quad \text{et} \quad \Sigma^+ = \frac{(n-1)\Sigma + (x-\mu^+)^2}{n}.$$

The convergence speed of the log-likelihood ℓ at iteration q is written:

$$\ell(\boldsymbol{\theta}^{(q)}; x) \sim -0.5q \ln \frac{n-1}{n}$$

and geometrical convergence rate towards 0 for the variance:

$$\Sigma^{(q)} \sim \Sigma^{(0)} \left(\frac{n-1}{n}\right)^q.$$

By comparison with degeneracy with mixtures of complete data, it is seen that the new speed is now much slower. Let also remark that the convergence speed decreases as the rate of missing data increases. It can then be expected that the EM algorithm is trapped by degenerated solutions with a slower dynamic.

	1	2	3	4	5	6	7	8	9
1	211.00		0.09	30.75	151.98	4.94	14.27	27.24	217.13
2	196.86	0.02	0.09	28.59	82.06	2.87	7.97	27.66	200.75
3	144.00	0.12	0.05	19.65	70.43	3.58		7.57	160.37
4	172.52	0.13	0.04		192.22	5.12	19.32	32.19	174.93
5	121.00	0.17	0.09	24.44	144.47	5.91	22.02	10.59	141.77
6	223.00	0.12	0.08	33.10	197.01	5.95	30.45	12.96	252.48
7		0.17	0.23	34.22	94.35	2.76	31.28	13.88	180.61
8	303.00	0.06	0.04	22.57		4.54	21.83	5.72	321.65
9	250.00	0.09	0.09	29.64	180.76	6.10	26.14	13.96	280.12
10	391.00	0.06	0.01	35.78		7.41	22.13	28.11	400.99
11	176.00	0.09	0.08	20.59	79.71		18.23	9.58	191.99
12	145.00		0.11	21.22	82.46	3.89	20.30	6.17	162.51
13	124.13	0.13	0.11	20.59			18.46	9.12	134.89
14	103.00	0.16	0.29	23.75	78.26	3.29	22.32	8.12	124.98

Table 6.1: Data belonging to the degenerated component.

6.3.3.4 Influence of the missing data rate

Let take again the instance on the breast cancer tissue of the UCI database repository, and make vary the rate of missing data. In our results presented Table 6.2 we see that as the rate of missing data increases, the rate of degeneracy increases, and the number of iterations before degeneracy decreases.

% missing data	0	5	10	15	20	25	30
% deg.	16	4	12	11	46	51	100
Average nb of iterations before deg.	2	13	13	82	304	138	215

Table 6.2: Frequency and speed of degeneracy (deg.) according to the rate of missing data on the breast cancer data set.

6.3.3.5 Solution to the degeneracy

Existing strategies for avoiding degeneracy try artificially to add some supplementary information, typically on the parameters (which typically has not invariance-scale property). The first possible solution is to constrain the covariance matrices (Tanaka and Takemura, 2006) (*e.g.* numeric tolerance)

$$\forall k, |\Sigma_k| \geq \alpha_{(n)} > 0.$$

Another possibility is to impose relative constraints between covariance matrices (Hathaway, 1985; Ingrassia and Rocci, 2007; García-Escudero et al., 2015)

$$\forall k \neq j, |\Sigma_k| \geq \beta |\Sigma_j| \quad (0 < \beta \leq 1).$$

Instead of changing the parameters space, one can regularize the estimation of the parameters through a Bayesian approach with a well-behaved prior γ (Snoussi and Mohammad-Djafari, 2001; Ciuperca, Ridolfi, and Idier, 2003), and then maximize

$$\ln \ell(\boldsymbol{\theta}; \boldsymbol{x}) + \ln \gamma(\boldsymbol{\theta}).$$

The common difficulty to all these methods is that they require the additional information on α, β, or γ which is difficult to set in practice.

6.3.3.6 Discussion for further work

Thus empirically we have observed linear degeneracy of the log-likelihood, and that the phenomenon tends to occur more often when increasing the rate of missing value. This rate of converge has been proved on a toy example, however, extending it to a multivariate mixture setting is still challenging. Indeed, we are working on a solution to avoid degeneracy by adding some information on the latent partition, such as imposing a minimal number of points in each cluster.

6.4 Visualization in mixture

6.4.1 Introduction

The exploratory field of multivariate statistics essentially encompasses the clustering and visualization tasks. Both are often jointly involved: either visualization is performed in the hope of revealing the "graphical evidence" of a cluster structure in the dataset, or clustering is performed first and the visualization task follows in the hope of providing a better understanding of the estimated cluster structure. We are primarily interested in the second scenario.

The framework of model based clustering allows for the analysis of different types of data by "simply" adapting the related cluster distribution: continuous data (Banfield and Raftery, 1993; Celeux and Govaert, 1995; McNicholas and Murphy, 2008), categorical data (Goodman, 1974; Celeux and Govaert, 1991; Gollini and Murphy, 2014; Marbac, Biernacki, and Vandewalle, 2016), mixed data (Kosmidis and Karlis, 2015; McParland and Gormley, 2016; Punzo and Ingrassia, 2016; Marbac, Biernacki, and Vandewalle, 2017; Mazo, 2017), functional data (Samé et al., 2011; Bouveyron and Jacques, 2011; Jacques and Preda, 2014), networks data (Daudin, Picard, and Robin, 2008; Zanghi, Ambroise, and Miele, 2008; Ambroise and Matias, 2012).

Visualization is designed to express, in a user-friendly manner, the estimated clustering structure. Its general principle is to design a mapping of the data, or of other related statistical results such as the cluster shape, within a "friendly" space

(generally $\mathbb{R}^2$) while maintaining some properties that the data, or the related statistical results, have in their native space. The vast majority of proposed mapping relies on different variants of factorial analysis or other distance-based methods (like multidimensional scaling). However, all standard mappings waste most clustering information that is conveyed by the probabilistic approach, except Scrucca (2010) which uses the full model-based approach for the mapping. However, this approach is limited to continuous data.
The main steps of the method we propose are:

1. select a model-based clustering technique for data at hand;
2. extract the whole distribution of the posterior classification probabilities from the fitted model;
3. fit a multivariate spherical Gaussian mixture respecting as close as possible to the distribution of the previous classification probabilities;
4. (a) draw the spherical Gaussian mixture pdf on the most discriminative bivariate map;
   (b) draw a "pseudo" bivariate scatter plot representing the individual classification probabilities on the most discriminative bivariate map.

6.4.2 Possible mapping strategies

Individual mapping The clustering visualization task is probably thought as firstly as visualizing simultaneously the data set $\mathbf{x}$ and its estimated partition $\hat{\mathbf{z}}$. Typically, the corresponding mapping, designated below by M^{ind}, transforms the data set $\mathbf{x}$, defined on $\mathcal{X}$, into a new data set $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, defined on a new space $\mathcal{Y}$, as follows:

$$M^{\text{ind}} \in \mathcal{M}^{\text{ind}} : \ \mathbf{x} \in \mathcal{X}^n \mapsto \mathbf{y} = M^{\text{ind}}(\mathbf{x}) \in \mathcal{Y}^n. \tag{6.1}$$

Here $\mathcal{M}^{\text{ind}}$ denotes a particular mapping family. This family varies according to the type of data involved in $\mathcal{X}$ and also depending on whether they use only data $\mathbf{x}$ or additional clustering information $\hat{\mathbf{z}}$ or $\mathbf{t}(\hat{f})$ (where $\hat{f}$ is the estimated mixture adapted to the kind of data and $\mathbf{t}(\hat{f})$ the resulting posterior probabilities) .
Methods relying on data $\mathbf{x}$ (thus discarding clustering information) are certainly the most frequent. In terms of continuous data, *principal component analysis* (PCA; Josse, Pagès, and Husson (2011), Verbanck, Josse, and Husson (2015), and Audigier, Husson, and Josse (2016b)) serves to represent the data on a map by focusing on their dispersion. Similarly, categorical data can be visualized using *multiple correspondence analysis* (MCA; Van der Heijden and Escofier (2003), Josse, Chavent, et al. (2012), and Greenacre (2017)), a mix of continuous and categorical data can be visualized using *mixed factorial analysis* (MFA; Chavent, Kuentz-Simonet, and Saracco (2012) and Audigier, Husson, and Josse (2016a)) and functional data can be

visualized using *functional principal component analysis* (FPCA; Ramsay and Silverman (2005), Zhou and Pan (2014), and Chen and Lei (2015)). *Multidimensional scaling* (MDS; Young (1987) and T. Cox and M. Cox (2001)) is more general since it can be used to deal with any type of data. It relies on dissimilarities between pairs of individuals for inputs $\mathbf{x}$ and also for outputs $\mathbf{y}$, the resulting coordinate matrix $\hat{\mathbf{y}}$ being obtained by minimizing a loss function. However, dissimilarities have to be defined specifically with respect to the type of data under consideration. To just illustrating this point, the Euclidean distance is frequent for continuous data whereas the Hamming distance is more suitably for binary data.
Methods taking into account additional clustering information $\hat{\mathbf{z}}$ or $\mathbf{t}(\hat{f})$ are less common and are mostly restricted to continuous data. We can cite *linear discriminant analysis* (LDA; Fisher (1936) and Xanthopoulos, Pardalos, and Trafalis (2013)) which takes into account cluster separation by defining the mapping through particular factorial analysis of the cluster means. Also, in the specific case of continuous data, Hennig (2004), Scrucca (2010), and Morris, McNicholas, and Scrucca (2013) defined a specific linear mapping between $\mathcal{X}$ and $\mathcal{Y}$. In that case, the distribution of $\mathbf{y}$ is itself a (less-dimensional) Gaussian mixture or a multivariate t-mixture, with the same number of components and the same proportions, which can be expressed as $h = \sum_k \pi_k h_k$. Finally, their method aims to preserve the related conditional membership probabilities $\mathbf{t}(\hat{f})$ and $\mathbf{t}(h)$, namely the classification probabilities of $\mathbf{x}$ with $\hat{f}$ and the classification probabilities of $\mathbf{y}$ with h, respectively. In other words, the aim is to find a linear mapping that preserves as far as possible, through the mapping mixture h, the cluster separation occurring in the original mixture f. Somewhat the method we proposed in this paper is related to this idea but it is not restricted to continuous distributions in the mixture and it does not rely on a linear mapping.

Pdf mapping Many visualizations are in practice overlaid by additional information relating to the corresponding mapping distribution. This mapping transforms the initial mixture $f = \sum_k \pi_k f_k$, defined on the distributional space $\mathcal{F}$, into a new mixture $h = \sum_k \pi_k h_k$, defined on the distributional space $\mathcal{H}$. It can be expressed as the following mapping, designated here by M^{pdf}:

$$M^{\text{pdf}} \in \mathcal{M}^{\text{pdf}} : \ f \in \mathcal{F} \mapsto h = M^{\text{pdf}}(f) \in \mathcal{H}, \tag{6.2}$$

where $\mathcal{M}^{\text{pdf}}$ denotes a particular mapping family. It is important to note that the pdf mapping M^{pdf} is rarely defined "from scratch" since it can be obtained as a "simple" by-product from the previous individual mapping M^{ind}. However, in practice, the resulting mixture h can be particularly tedious to calculate (possibly no closed-form solution available outside linear mappings), which can be partially overcome by displaying the empirical mapping of a very large sample. But the resulting pdf can also have a non-conventional isodensity shape per cluster (for instance clusters with disconnected parts), undermining somewhat all the user-friendliness that is expected when using pdf visualization.

Traditional way: controlling the mapping family The cornerstone of all traditional pdf visualization procedures is based on defining the mapping family $\mathcal{M}^{\text{pdf}}$ (or more exactly $\mathcal{M}^{\text{ind}}$ from which $\mathcal{M}^{\text{pdf}}$ is almost always deduced). As just an example, the reader can have in mind the classical linear mapping for the continuous case. Then, the pdf family $\mathcal{H}$ of h is a simple by-product of $\mathcal{M}^{\text{pdf}}$, and thus can be denoted by $\mathcal{H}(\mathcal{M}^{\text{pdf}})$. Using the general mapping expression (6.2), $\mathcal{H}(\mathcal{M}^{\text{pdf}})$ is naturally expressed as follows:

$$\mathcal{H}(\mathcal{M}^{\text{pdf}}) = \{h : h = M^{\text{pdf}}(f), f \in \mathcal{F}, M^{\text{pdf}} \in \mathcal{M}^{\text{pdf}}\}\,. \tag{6.3}$$

As an immediate consequence, the nature of $\mathcal{H}$ can depend to a great extent on the choice of $\mathcal{M}^{\text{pdf}}$, leading potentially to very different cluster shapes. Arguments that lead to traditional $\mathcal{M}^{\text{pdf}}$ (or $\mathcal{M}^{\text{ind}}$) rely essentially on a combination of user-friendly and easy-to-compute properties. For instance, in the continuous case, linear mappings are often retained (like for PCA). In the categorical case, a continuous space $\mathcal{Y}$ is often targeted (like for MCA). It is a similar situation for functional data with FPCA or also for mixed data with MFA or MDS, even if MDS is a somewhat more complex procedure since it is not always defined in closed-form. However, such choices may vary significantly from one statistician to another one. For instance, MDS relies on defining dissimilarities both inside spaces $\mathcal{X}$ and $\mathcal{Y}$ and changing them could significantly affect the resulting mapping.

6.4.3 New proposed method: controlling the distribution family

Alternatively, the general mapping expression (6.2) can be seen as indexed by the distribution family $\mathcal{H}$, the mapping $\mathcal{M}^{\text{pdf}}$ being now obtained as a by-product, and thus now denoted by $\mathcal{M}^{\text{pdf}}(\mathcal{H})$. This new point of view is straightforwardly expressed as:

$$\mathcal{M}^{\text{pdf}}(\mathcal{H}) = \{M^{\text{pdf}} : h = M^{\text{pdf}}(f), f \in \mathcal{F}, h \in \mathcal{H}\}\,. \tag{6.4}$$

It corresponds to the reversed situation of (6.3) where $\mathcal{H}$ has to be defined instead of M^{pdf}. This new freedom indeed provides an opportunity to directly force $\mathcal{H}$ to be a user-friendly mixture family.

Constrained spherical Gaussians as matching candidates One of the most simple candidate belonging to the "user-friendly mixture family" is probably the spherical Gaussian mixture defined on $\mathcal{Y} = \mathbb{R}^{d_Y}$. Its pdf is defined for any $\boldsymbol{y} \in \mathbb{R}^{d_Y}$ by

$$h(\boldsymbol{y}; \boldsymbol{\nu}) = \sum_{k=1}^{g} \pi_k \phi_{d_Y}(\boldsymbol{y}; \boldsymbol{\nu}_k, \boldsymbol{I}), \tag{6.5}$$

where $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_g)$ and $\phi_{d_Y}(.; \boldsymbol{\nu}_k, \boldsymbol{I})$ is the pdf of the Gaussian distribution with mean $\boldsymbol{\nu}_k = (\mu_{k1}, \ldots, \mu_{kd_Y}) \in \mathbb{R}^{d_Y}$ and covariance matrix equal to identity $\boldsymbol{I}$.
Because clustering visualization is the central task of this work, it is expected to require that both mixtures f and $h(\cdot; \boldsymbol{\nu})$ have the most similar clustering information. This information is measured by the posterior probabilities of classification.

The main idea of the proposal is thus to find $\boldsymbol{\nu}$ such that the distribution of posterior class membership probabilities resulting from mixture $h(\cdot;\boldsymbol{\nu})$ be as close as possible distribution of posterior class membership probabilities from mixture f in the initial space. This is measured in terms of Kullback-Leibler divergence between these posterior class membership probabilities distribution which have the advantage to be defined on the same space (the simplex) contrarily to density $h(\cdot;\boldsymbol{\nu})$ and f, thus allowing a generic approach. For sake of simplicity, it is not detailed here but can be found in Biernacki, Marbac, and Vandewalle (2020).

Final visualization as bivariate spherical Gaussians Because h is defined on $\mathbb{R}^{g-1}$, it is inconvenient to draw this distribution if $g \geq 4$. Therefore, we apply an LDA to h to represent this distribution on its most discriminative map (*i.e.,* eigen value decomposition of the covariance matrix computed on the centers $\hat{\boldsymbol{\nu}}$ by considering the mixture proportions $\boldsymbol{\pi}$), leading to the following bivariate spherical Gaussian mixture $\tilde{h}$:

$$\tilde{h}(\tilde{\boldsymbol{y}};\tilde{\boldsymbol{\nu}}) = \sum_{k=1}^{g} \pi_k \phi_2(\tilde{\boldsymbol{y}};\tilde{\boldsymbol{\nu}}_k, \boldsymbol{I}), \tag{6.6}$$

where $\tilde{\boldsymbol{y}} \in \mathbb{R}^2$, $\tilde{\boldsymbol{\nu}} = (\tilde{\boldsymbol{\nu}}_1, \ldots, \tilde{\boldsymbol{\nu}}_g)$ and $\tilde{\boldsymbol{\nu}}_k \in \mathbb{R}^2$. The (standard) percentage of inertia of LDA serves to measure the quality of the mapping from h to $\tilde{h}$. In addition, the accuracy of the mapping from the initial mixture f to the final "ready-to-be-drawn" mixture $\tilde{h}$ can be easily compared through the following difference between the normalized (theoretical) entropy of the partition related to f and the normalized (theoretical) entropy of the partition related to $\tilde{h}$.

Remarks When the initial data set $\mathbf{x}$ is in the continuous space $\mathcal{X} = \mathbb{R}^d$ and also when the initial clustering relies on a Gaussian mixture f whose covariance matrices are identical, then the proposed mapping is strictly equivalent to applying an LDA to the centers of f. We have presented the approach for the pdf mapping. It is also possible to obtained and individual mapping as a by-product, it is not presented here since it is not the main focus of the proposal, the main focus being to visualize distribution in a Gaussian-like way.

6.4.4 Example

Presentation of the example We consider the data set of Schlimmer (1987). It is composed of votes for each of the $n = 435$ U.S. House of Representatives Congressmen on $d_X = 16$ key votes. For each vote, three levels are considered: yea, nay, or unknown disposition. Data are clustered by a mixture of products of multinomial distributions (Goodman, 1974). Parameter estimation is performed by maximum likelihood and model selection is done by the BIC (Schwarz, 1978), which selects $g = 4$ components. The R package Rmixmod (Lebret et al., 2015) is used for inference.
As an output of this estimation step, the user is provided with a partition and a parameter. It may be not convenient to have a detailed look at the partition of 435

individuals. Regarding the parameters, the mixing proportions can be suitable for a quick, but partial, understanding of the clustering result. However, going further into the clustering understanding by analyzing the multinomial parameters can be very laborious since it entails $192 = 16 \times 3 \times 4$ values to be observed and compared. It is also possible to analyze the clustering results graphically in a conventional way. Figure 6.3 presents the scatter plot of the Congressmen and their partition on the first map of the MCA, obtained by the R package FactoMineR (Lê, Josse, Husson, et al., 2008). It appears that the scatter plot provided by MCA is quite hard to read. Firstly, it is well-known that total inertia is hard to interpret, and consequently, the information about a possible relative positioning of clusters can be questionable. Secondly, even if faithful, the overlap between components is not fully visible and thus does not allow for a straightforward interpretation of f.

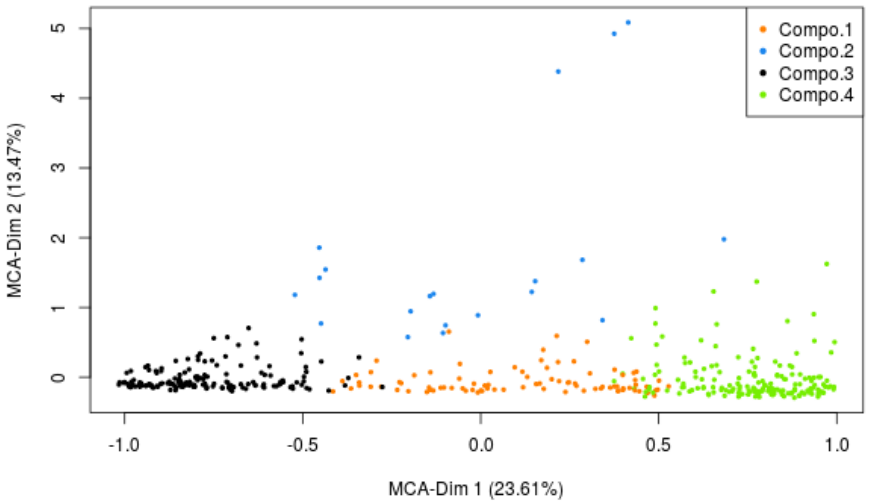


Figure 6.3: Scatter plot of the Congressmen and their partition on the first MCA map.

Visualization proposal on the example We now illustrate the previous visualization proposition on the running example. Figure 6.4 is the component interpretation graph obtained for the congressional voting records. It presents the Gaussian-like component overlap on the most discriminative map. In this way, it provides more visually than a traditional confusion table the overlap information of the initial mixture f. We also graphically observe the ranking between the different cluster spreads, indicating some variety in mixing proportions (numerically speaking, we have $\hat{\pi}_1 = 0.21$, $\hat{\pi}_2 = 0.05$, $\hat{\pi}_3 = 0.35$ and $\hat{\pi}_4 = 0.39$). Note that the mapping of f on this graph is accurate because the difference between entropies is almost zero (*i.e.,* $\delta_{\mathrm{E}}(f, \tilde{h}) = -0.08$). For instance, this figure also shows that the

components with most observations (*i.e.,* components three and four) are composed of strongly different Congressmen. Indeed, the overlap between these components is almost zero. Moreover, component one contains Congressmen which are more moderate than Congressmen of components three and four.

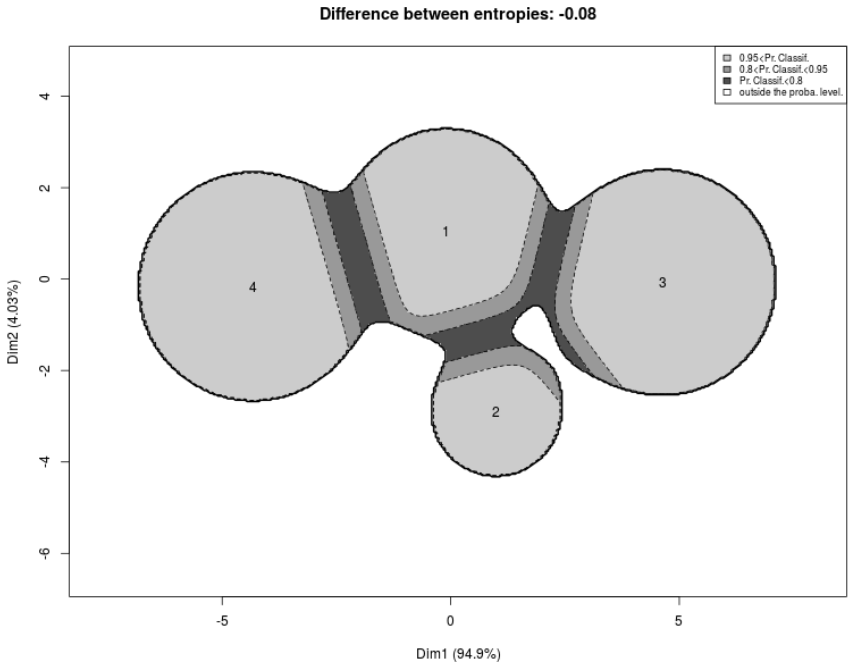


Figure 6.4: Component interpretation graph of the congressional voting records.

6.4.5 Conclusion

We have presented a generic method for visualizing the results of a model-based clustering in a "Gaussian way". This method allows for visualization of any model-based clustering made on any type of data, because it is only based on the distribution of classification probabilities. It permits to interpret the results of a model-based clustering but not to select the best clustering method (choosing a clustering method has to be performed before through a classical model selection process). In this way, it is not an exploratory visualization method, as such methods are often dedicated to. The developed method is available as an R package; **ClusVis** available on the CRAN.

6.5 Conclusion

This Chapter has presented contributions to model-based clustering to the general issue. I discussed the label-switching problem and some solutions that we have proposed. Developing tools, such as visualization, helping the interpretation of clustering is very important in practice to help the user to appropriate results for

the potentially complex model on complex data spaces. The study of the degeneracy problem is still a work in progress.

Bibliography

Ambroise, C. and Matias, C. (2012). "New consistent and asymptotically normal parameter estimates for random-graph mixture models". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 74.1, pp. 3–35. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2011.01009.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2011.01009.x.

Audigier, V., Husson, F., and Josse, J. (2016a). "A principal component method to impute missing values for mixed data". In: *Advances in Data Analysis and Classification* 10.1, pp. 5–26.

— (2016b). "Multiple imputation for continuous variables using a Bayesian principal component analysis". In: *Journal of Statistical Computation and Simulation* 86.11, pp. 2140–2156.

Banfield, J. and Raftery, A. (1993). "Model-based Gaussian and non-Gaussian clustering". In: *Biometrics* 49.3, pp. 803–821. ISSN: 0006-341X. DOI: 10.2307/2532201. URL: http://dx.doi.org/10.2307/2532201.

Biernacki, C. (2007). "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures for grouped data and behaviour of the EM algorithm". In: *Scandinavian Journal of Statistics* 34.3, pp. 569–586.

Biernacki, C., Castellan, G., Chretien, S., Guedj, B., and Vandewalle, V. (2016). "Pitfalls in Mixtures from the Clustering Angle". In: *Working Group on Model-Based Clustering Summer Session*. Paris, France. URL: https://hal.archives-ouvertes.fr/hal-01419755.

Biernacki, C. and Chrétien, S. (2003). "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM". In: *Statistics & probability letters* 61.4, pp. 373–382.

Biernacki, C., Marbac, M., and Vandewalle, V. (2020). "Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering". In: *Journal of Classification*. DOI: 10.1007/s00357-020-09369-y. URL: https://doi.org/10.1007/s00357-020-09369-y.

Biernacki, C. and Vandewalle, V. (2011). "Label switching in mixtures". In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401.

Bouveyron, C., Girard, S., and Schmid, C. (2007). "High-dimensional data clustering". In: *Computational Statistics & Data Analysis* 52.1, pp. 502–519. ISSN: 0167-9473. DOI: 10.1016/j.csda.2007.02.009. URL: http://www.sciencedirect.com/science/article/pii/S0167947307000692.

Bouveyron, C. and Jacques, J. (2011). "Model-based clustering of time series in group-specific functional subspaces". In: *Advances in Data Analysis and Classification* 5.4, pp. 281–300.

Celeux, G. (1998). "Bayesian inference for mixture: The label switching problem". In: *Compstat*. Springer, pp. 227–232.

Celeux, G. and Govaert, G. (1991). "Clustering criteria for discrete data and latent class models". In: *Journal of Classification* 8.2, pp. 157–176. ISSN: 1432-1343. DOI: 10.1007/BF02616237. URL: https://doi.org/10.1007/BF02616237.

— (1995). "Gaussian parsimonious clustering models". In: *Pattern recognition* 28.5, pp. 781–793.

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and inferential difficulties with mixture posterior distributions". In: *Journal of the American Statistical Association* 95.451, pp. 957–970.

Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012). "Orthogonal rotation in PCAMIX". In: *Advances in Data Analysis and Classification* 6.2, pp. 131–146.

Chen, K. and Lei, J. (2015). "Localized functional principal component analysis". In: *J. Amer. Statist. Assoc.* 110.511, pp. 1266–1275. ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1016225. URL: http://dx.doi.org/10.1080/01621459.2015.1016225.

Ciuperca, G., Ridolfi, A., and Idier, J. (2003). "Penalized maximum likelihood estimator for normal mixtures". In: *Scandinavian Journal of Statistics* 30.1, pp. 45–59.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling*. Chapman and Hall.

Daudin, J.-J., Picard, F., and Robin, S. (2008). "A mixture model for random graphs". In: *Stat. Comput.* 18.2, pp. 173–183. ISSN: 0960-3174. DOI: 10.1007/s11222-007-9046-7. URL: http://dx.doi.org/10.1007/s11222-007-9046-7.

Day, N. E. (1969). "Estimating the components of a mixture of normal distributions". In: *Biometrika* 56.3, pp. 463–474.

Diebolt, J. and Robert, C. P. (1994). "Estimation of finite mixture distributions through Bayesian sampling". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.2, pp. 363–375.

Eirola, E., Lendasse, A., Vandewalle, V., and Biernacki, C. (2014). "Mixture of Gaussians for distance estimation with missing data". In: *Neurocomputing* 131, pp. 32–42.

Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2015). "Avoiding spurious local maximizers in mixture modeling". In: *Statistics and Computing* 25.3, pp. 619–633.

Ghahramani, Z. and Jordan, M. (1995). *Learning From Incomplete Data*. Tech. rep. Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab.

Gollini, I. and Murphy, T. (2014). "Mixture of latent trait analyzers for model-based clustering of categorical data". In: *Statistics and Computing* 24.4, pp. 569–588.

Goodman, L. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". In: *Biometrika* 61.2, pp. 215–231.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

Hathaway, R. J. (1985). "A constrained formulation of maximum-likelihood estimation for normal mixture distributions". In: *The Annals of Statistics*, pp. 795–800.

Hennig, C. (2004). "Asymmetric linear dimension reduction for classification". In: *Journal of Computational and Graphical Statistics* 13.4, pp. 930–945.

Hunt, L. and Jorgensen, M. (2003). "Mixture model clustering for mixed data with missing information". In: *Computational Statistics & Data Analysis* 41.3–4, pp. 429–440. ISSN: 0167–9473. DOI: 10.1016/S0167-9473(02)00190-1. URL: http://www.sciencedirect.com/science/article/pii/S0167947302001901.

Ingrassia, S. and Rocci, R. (2007). "Constrained monotone EM algorithms for finite mixture of multivariate Gaussians". In: *Computational Statistics & Data Analysis* 51.11, pp. 5339–5351.

Jacques, J. and Preda, C. (2014). "Model-based clustering for multivariate functional data". In: *Computational Statistics and Data Analysis* 71, pp. 92–106.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling". In: *Statistical Science*, pp. 50–67.

Josse, J., Chavent, M., Liquet, B., and Husson, F. (2012). "Handling missing values with regularized iterative multiple correspondence analysis". In: *Journal of classification* 29.1, pp. 91–116.

Josse, J., Pagès, J., and Husson, F. (2011). "Multiple imputation in principal component analysis". In: *Advances in data analysis and classification* 5.3, pp. 231–246.

Kosmidis, I. and Karlis, D. (2015). "Model-based clustering using copulas with applications". English. In: *Statistics and Computing*, pp. 1–21. ISSN: 0960-3174. DOI: 10.1007/s11222-015-9590-5. URL: http://dx.doi.org/10.1007/s11222-015-9590-5.

Lê, S., Josse, J., Husson, F., et al. (2008). "FactoMineR: an R package for multivariate analysis". In: *Journal of statistical software* 25.1, pp. 1–18.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2015). "Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library". In: *Journal of Statistical Software* 67.6, pp. 241–270.

Marbac, M., Biernacki, C., and Vandewalle, V. (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207.

— (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656.

Mazo, G. (2017). "A semiparametric and location-shift copula-based mixture model". In: *Journal of Classification* 34.3, pp. 444–464.

McNicholas, P. D. and Murphy, T. (2008). "Parsimonious Gaussian mixture models". In: *Stat. Comput.* 18.3, pp. 285–296. ISSN: 0960-3174. DOI: 10.1007/s11222-008-9056-0. URL: http://dx.doi.org/10.1007/s11222-008-9056-0.

McParland, D. and Gormley, I. C. (2016). "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2, pp. 155–169.

Morris, K., McNicholas, P. D., and Scrucca, L. (2013). "Dimension reduction for model-based clustering via mixtures of multivariate t-distributions". In: 7, pp. 321–338.

Papastamoulis, P. and Iliopoulos, G. (2010). "An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions". In: *Journal of Computational and Graphical Statistics* 19.2, pp. 313–331.

Punzo, A. and Ingrassia, S. (2016). "Clustering bivariate mixed-type data via the cluster-weighted model". In: *Computational Statistics* 31.3, pp. 989–1013.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Second. Springer Series in Statistics. Springer, New York, pp. xx+426. ISBN: 978-0387-40080-8; 0-387-40080-X.

Redner, R. A. and Walker, H. F. (1984). "Mixture densities, maximum likelihood and the EM algorithm". In: *SIAM review* 26.2, pp. 195–239.

Samé, A., Chamroukhi, F., Govert, G., and Aknin, P. (2011). "Model-based clustering and segmentation of time series with changes in regime". In: *Advances in Data Analysis Classification* 5, pp. 301–321.

Schlimmer, J. (1987). "Concept acquisition through representational adjustment". PhD thesis. Department of Information and Computer Science, University of California.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Scrucca, L. (2010). "Dimension reduction for model-based clustering". In: *Statistics and Computing* 20.4, pp. 471–484. ISSN: 1573-1375. DOI: 10.1007/s11222-009-9138-7. URL: http://dx.doi.org/10.1007/s11222-009-9138-7.

Snoussi, H. and Mohammad-Djafari, A. (2001). "Penalized maximum likelihood for multivariate gaussian mixture". In: *arXiv preprint physics/0111007*.

Sperrin, M., Jaki, T., and Wit, E. (2010). "Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models". In: *Statistics and Computing* 20.3, pp. 357–366.

Stephens, M. and Phil, D. (1997). "Bayesian methods for mixtures of normal distributions". In:

Tanaka, K. and Takemura, A. (2006). "Strong Consistency of the Maximum Likelihood Estimator for Finite Mixtures of Location: Scale Distributions When the Scale Parameters Are Exponentially Small". In: *Bernoulli*, pp. 1003–1017.

Van der Heijden, P. and Escofier, B. (2003). "Multiple correspondence analysis with missing data". In: *Analyse des correspondances. Recherches au cżur de l'analyse des donnees*, pp. 152–170.

Vandewalle, V. and Biernacki, C. (2015). "An efficient SEM algorithm for Gaussian Mixtures with missing data". In: *8th International Conference of the ERCIM WG on Computational and Methodological Statistics*. Londres, United Kingdom. URL: https://hal.inria.fr/hal-01242588.

Verbanck, M., Josse, J., and Husson, F. (2015). "Regularised PCA to denoise and visualise data". In: *Statistics and Computing* 25.2, pp. 471–486.

Xanthopoulos, P., Pardalos, P. M., and Trafalis, T. B. (2013). "Linear Discriminant Analysis". In: *Robust Data Mining*, pp. 27–33.

Young, F. W. (1987). *Multidimensional scaling: History, theory, and applications*. Lawrence Erlbaum Associates.

Zanghi, H., Ambroise, C., and Miele, V. (2008). "Fast online graph clustering via Erdős–Rényi mixture". In: *Pattern Recognition* 41.12, pp. 3592–3599. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2008.06.019. URL: http://www.sciencedirect.com/science/article/pii/S0031320308002483.

Zhou, L. and Pan, H. (2014). "Principal component analysis of two-dimensional functional data". In: *J. Comput. Graph. Statist.* 23.3, pp. 779–801. ISSN: 1061-8600. DOI: 10.1080/10618600.2013.827986. URL: http://dx.doi.org/10.1080/10618600.2013.827986.

Part II

Contribution to some applications

Chapter 7

Contribution to the analysis of usability study in medicine

Contents

7.1 Introduction 129
7.2 Bayesian modeling of the discovery matrix 130
7.2.1 Motivation from a practical point of view 130
7.2.2 Proposed solution 131
7.2.3 Performance of the method 134
7.2.4 Discussion 134
7.3 Number of additional subjects needed 134
7.3.1 Validation study in usability testing 135
7.3.2 Modeling the economic consequences of undetected problems 135
7.4 Conclusion and perspectives 137

Related article

1. V. Vandewalle, A. Caron, C. Delettrez, R. Périchon, S. Pelayo, A. Duhamel, and B. Dervaux (2020). "Estimating the number of usability problems affecting medical devices: modelling the discovery matrix". In: *BMC Medical Research Methodology* 20.234

7.1 Introduction

This chapter is the result of work with Alexandre Caron and Benoît Dervaux, both members of the METRICS team. It is in the framework of a usability study of a medical device, where one of the goals is to determine the number of possible problems linked to the use of this device as well as their respective occurrence probabilities. Estimating this number and the different probabilities is essential to determine whether or not an additional usability study should be conducted, and the number of users to be included in the study to maximize the expected benefits from the economical point of view. This work has been conducted in the

scope of the USEVAL-DM ANR project[1], one case study motivating our theoretical developments was usability data from the Zeneo insulin needless injector pen [2] (see Figure 7.1).



(a) (b)

Figure 7.1: Zeneo needless injection pump

The discovery process can be modeled by a binary matrix, a matrix whose number of columns depends on the number of defects discovered by users. In this framework, we have proposed probabilistic modeling of this matrix. We have included this modeling in a Bayesian context where the number of problems and the probabilities of discovery are considered as random variables. In this framework, the article Vandewalle et al. (2020) shows the interest of the approach. This approach beyond point estimation also makes it possible to obtain the distribution of the number of problems and their respective probabilities given the discovery matrix. The proposed model allows us to implement an approach aiming at measuring the value of additional information related to the discovery process. In this framework, we are currently finishing a second paper and developing an R package that should help practitioners in the field of usability to better sizing their studies.
In Section 7.2, I present the medical issue and the modeling of the discovery matrix of Vandewalle et al. (2020). In Section 7.3, I present how it can be used to compute the number of needed users for a second usability study to limit the number of final users encountering not yet discovered problems with respect to the cost of the second usability study.

7.2 Bayesian modeling of the discovery matrix

7.2.1 Motivation from a practical point of view

Usability testing is a cornerstone of medical device development, and proof of usability is mandatory for market access in both the European Union and the United States (US-FDA, 2016). The overall objective of a usability assessment is to ensure that a medical device is designed and optimized for use by the intended users in the

[1] https://anr.fr/Project-ANR-15-CE36-0007
[2] https://www.crossject.com/fr/notre-technologie/technologie-sans-aiguille

environment in which the device is likely to be used (UK-MHRA, 2017). The goal is to identify problems (called "use errors") that could cause harm to the user or impair medical treatment (*e.g.* an inappropriate number of inhalations, finger injection with an adrenaline pen, ...) (US-FDA, 2012). The detection of usability problems must be as comprehensive as possible because medical devices are safety-critical systems (Borsci, Macredie, et al., 2013). However, the total number of usability problems is never known in advance. The main challenge during the usability testing is thus to estimate this number, to assess the completeness of the problem discovery process (US-FDA, 2012). In practice, participants are placed under actual conditions of use (real or simulated), and usability problems are observed and listed by human factor engineers. The experimental conditions are defined in a risk analysis that gathers together possible usability problems. Throughout the usability testing, problems are discovered and added to a discovery matrix - a binary matrix with the participants as the rows and the problems as the columns. The current approach involves estimating the total number of problems as the usability testing progresses, starting from the first sessions. The number is estimated iteratively as the sample size increases until the objective of completeness has been achieved (Lewis, 1994).

7.2.2 Proposed solution

From a statistical perspective, the current estimation procedure is based on a model of how the usability problems are detected; this is considered to be a binomial process. The literature suggests that the total number of usability problems can be estimated from the discovery matrix's problem margin (the sum of the columns)(Kanis, 2011; Lewis, 2001; Hertzum and Jacobsen, 2003; Schmettow, 2012; Borsci, Londei, and Federici, 2011). However, this estimation is complicated by (i) the small sample size usually encountered in usability studies of medical devices (Faulkner, 2003) and (ii) as-yet unobserved problems that truncate the margin and bias estimates (Lewis, 2000; Sauro and Lewis, 2016; Thomas and Gart, 1971). Let also notice that this problem has also been investigated in ecology while determining the number of species in a population(Chao, 1984; Klingwort, Buelens, and Schnell, 2019), however, it what follows we focus on the extension of the approach of Schmettow (2008) related to heterogenous probabilities of detection.

Data available: the discovery matrix The human factor engineer collects the results of the usability testing in a problem-discovery matrix $\mathbf{d}$. Each row corresponds to a participant, and each column corresponds to a usability problem. The result is 1 if the participant discovered the problem and 0 otherwise. Considering that after the inclusion of n participants, j problems have been discovered, a $n \times j$ matrix is built. By way of an example, the discovery matrix obtained after $n = 8$ participants (in rows) might be the one presented below:

$$
\begin{array}{l}
\\
n \text{ patients}
\end{array}
\overbrace{\left\{\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{pmatrix}\right.}^{j \text{ discovered problems}} = \mathbf{d}
$$

In this example, $j = 10$ different problems (in columns) have been detected so far. The first participant discovered only one problem (column 1), whereas the second discovered two new problems (columns 2 and 3), etc. For instance, if we consider an insulin pump, it can be that the user did not succeed in opening the insulin pump, or that he used it in the wrong way and picks his thumb.

At this stage, some problems might not have been detected, and the total number of usability problems (m) is unknown. It should be noted that by definition, $m \geq j$ and $m-j$ problems remain undetected. Indeed, $\mathbf{d}$ comes from a complete but unobserved matrix of dimensions $n \times m$. This matrix is denoted as $\mathbf{x}$. Thus, the "observed" matrix $\mathbf{d}$ is a truncated version of the "complete" matrix $\mathbf{x}$ where the columns have been ordered according to the discovery order of the problems. Hereafter, we use the following notation: $\mathbf{x} = (x_{il})_{1 \leq i \leq n, 1 \leq l \leq m}$ where $x_{il} = 1$ if the participant i experiences the problem l, and $x_{il} = 0$ otherwise.

The human factor engineer's goal is to estimate the total number of problems m from the discovery matrix $\mathbf{d}$ and thus deduce the number of problems that have not been detected ($m - j$). He is also interested in the probabilities of detection of each problem $p_1, \ldots, p_m$, since it is also an important parameter when evaluating the probability that a user encounters some problem in real life.

The matrix-based method The details of the state of the art methods are not presented here but can be found in Vandewalle et al. (2020). Here we present a Bayesian inference approach based on matrix $\mathbf{d}$ that we proposed. For sake of simplicity we first assume a model on $\mathbf{x}$ the complete discovery matrix, then deduce a model on the observed discovery matrix $\mathbf{d}$.

We assume that the probability of detection is specific to each problem $x_{il} \sim \mathcal{B}(p_l)$ and that each p_l are independent and follow a logit normal distribution $\text{logit}(p_l) \sim \mathcal{N}(\mu, \sigma)$. Thus $p_1, \ldots, p_m$ are latent variables, and the likelihood of the complete discovery matrix $\mathbf{x}$ given the parameters μ, σ can be obtained by integrating over the latent variables

$$
P(\mathbf{x}|\mu, \sigma) = \int_0^1 \ldots \int_0^1 P(\mathbf{x}|p_1, \ldots, p_m) f(p_1, \ldots, p_m|\mu, \sigma) dp_1 \ldots dp_m
$$

where $f(p_1, \ldots, p_m|\mu, \sigma)$ is the probability density function of $p_1, p_2, \ldots, p_m$, that

simplifies to $\prod_{k=1}^{m} f(p_j|\mu,\sigma)$ since $p_1,\ldots,p_m$ are assumed to be independent given μ and σ.
In a Bayesian framework (Robert, 2007), we assumed the prior distribution on (μ,σ,m). Moreover, we add the prior independence assumption of μ, σ and m:

$$P(\mu,\sigma,m) = P(\mu)P(\sigma)P(m).$$

Each prior distribution is defined as follows:

- $\mu \sim \mathcal{N}(0;A)$: a Gaussian distribution with variance $A = 1.5$,
- $\sigma^2 \sim \text{inv}-\chi^2_\nu$: an inverse chi-squared distribution with $\nu = 1$ degrees of freedom.
- $P(m) = \frac{1}{M} \forall m \in \{1,\ldots,M\}$: a uniform distribution with M being a predetermined upper boundary for m.

Such choice on μ and σ^2 was made in order to get flat prior on p_l, near from the uniform distribution.
Thus, the integrated likelihood $P(\mathbf{x})$ can be obtained by integrating over μ and σ

$$P(\mathbf{x}) = \int_0^{+\infty} \int_{-\infty}^{+\infty} P(\mathbf{x}|\mu,\sigma)P(\mu)P(\sigma)d\mu d\sigma$$

This integral can be approximated with Markov Chain Monte Carlo (MCMC) techniques. We sample from $P(\mu,\sigma|\mathbf{x})$. The parameters are sampled using the parameter space augmented by $p_1,\ldots,p_m$ (i.e. from $\mu,\sigma,p_1,\ldots,p_m|\mathbf{x}$) using an adaptive Hamiltonian Monte Carlo algorithm (Stan Development Team, 2020). Then, a numerical approximation of the integrated likelihood $P(\mathbf{x})$ is obtained via bridge sampling (Meng and Wong, 1996).
However $\mathbf{x}$ is not observed, thus Bayesian inference is performed based on $\mathbf{d}$. Since the columns of $\mathbf{x}$ are exchangeable (the integrated likelihood of $\mathbf{x}$ is the same for any permutation of its columns) we have:

$$P(\mathbf{d}|m) = \frac{1}{j_1!\ldots j_r!} \times A_m^j \times P(\hat{\mathbf{x}}^m)$$

where $\hat{\mathbf{x}}^m$ is the complete discovery matrix obtained based on $\mathbf{d}$ for the value m of the number of problems, it can be thought as the matrix $\mathbf{d}$ padded with $m-j$ null columns.
To estimate the number of problems we focused on $P(m|\mathbf{d})$, which is obtained using Bayes' theorem:

$$P(m|\mathbf{d}) = \frac{P(m) \times P(\mathbf{d}|m)}{\sum_{m'=1}^{M} P(m') \times P(\mathbf{d}|m')}.$$

Thus in practice it is needed to compute $P(\hat{\mathbf{x}}^m)$ for each m in $\{1,\ldots,M\}$. This also allows us to obtain credibility intervals to investigate the completeness of the discovery process.

7.2.3 Performance of the method

We compared the performance of five methods (naïve, GT, double-deflation, LNBzt, and matrix-based methods) first in a simulation study and then using literature data from actual usability studies in Vandewalle et al. (2020). This is not detailed in this manuscript but the main conclusions are the following. The simulations show that as expected accuracy of the estimation of the number of problems increases with the sample size for all estimates. The matrix-based method shows less bias overall. The matrix-based method gave the lowest RMSE in all settings, especially when the number of "rare" problems is high. As expected since the simulations are performed in the heterogeneous framework the estimators assuming homogeneity systematically underestimate the number of undiscovered problems. In practice, the homogeneity assumption does not hold. Thus from the human factor engineer's point of view, the matrix-based approach and the logit normal binomial zero truncated approach (Schmettow, 2008) are the only reliable ones: they gave a good coverage probability in almost any sample size.

7.2.4 Discussion

For the probability of problem discovery $p_1, \ldots, p_m$; we used a logit-normal distribution as a plugin to model the uncertainty. The choice of this distribution was convenient in that it allowed us to compare our method with the only published model that accounts for heterogeneity (Schmettow, 2008) in the usability framework. However, there are no data for confirming the validity of this choice. Nevertheless, this limitation could be easily overcome by replacing the logit-normal with another distribution (such as beta or gamma) if it proves to be more appropriate. This choice could be made using model choice criteria (e.g. the Akaike information criterion or the Bayesian information criterion), or Bayesian model averaging could be performed. However, it should be borne in mind that for small sample size, fitting for both incompleteness and heterogeneity is complex and inevitably leads to a high degree of uncertainty.

One interest of our proposed full Bayesian approach is that it will make it possible to assess the relevance of conducting some additional usability tests to find not yet discovered problems and precise the values of probabilities of problems not yet detected. Thus facilitating the decision-making process for both regulators and device manufacturers, which is discussed in the next section.

7.3 Number of additional subjects needed

This section results from a work that is about to be submitted to Health Economics. It is in direct line with previous work applied to risk assessment from a medico-economic point of view.

7.3.1 Validation study in usability testing

Usability testing is performed iteratively during the development of the medical device. This continuous back-and-forth between the development team and the human factor engineers is called "formative" assessment as usability problems are detected and corrected as it goes.
Once the design of the medical device is mature, a "validation" study is performed as a part of the premarket submission. This section focuses on the latter. In the rest of the manuscript the term "usability testing" will refer to the study performed during the validation step.
In Section 7.2, we introduced the matrix-based method as a Bayesian approach for the estimation of the number of problems affecting medical devices. In this section, we present a framework for sample size estimation applying a Bayesian decision-theoretic approach from the manufacturer's perspective.

7.3.2 Modeling the economic consequences of undetected problems

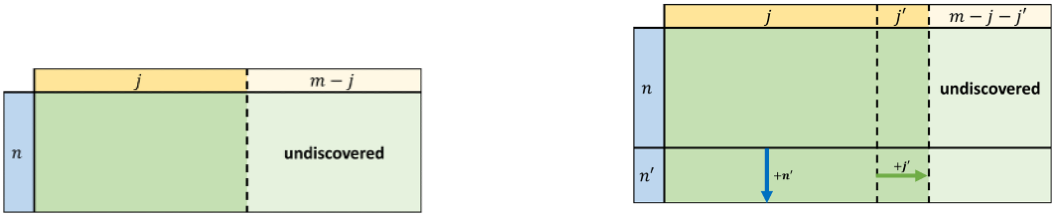
From the manufacturer's perspective, the economic consequences of a usability problem that remains undetected after the usability study are better modeled according to its severity. The latter has been well studied in the HFE literature and is a combination of three factors (Nielsen, 1994): the frequency of the problem, its impact, and its persistence. These three dimensions are usually synthesized in a single scale with usability problems being classified as cosmetic, minor, major, or catastrophic. For sake of simplicity, we will only consider non-critical problems. For such problems, a redesign is not deemed necessary and the problem is considered solved with a mention in the user manual. Critical problems are also taken into consideration in the submitted article but are not detailed here.
In our case study, we model the consequences as follows. If the end-user undergoes a non-critical problem, he will ask the manufacturer for reimbursement or will not use the medical device, which means that the manufacturer will make no profit. Notwithstanding, we are well aware that estimating the costs of undetected usability problems is a task that is way more complex and will normally require to perform a risk analysis. However, the scenario described above is inspired by the real adrenaline pens recalls and we consider it relevant for our study case.
One key assumption that we make is that problems mentioned in the user manual are not eligible for reimbursement of the device. Thus we are interested in y the number of users that encounter at least one new problem when considering N final user (N supposed to be known). We are interested in the distribution of $y|\mathbf{d}$ which can be sampled from Algorithm 6
Thus it is possible to deduce an approximation of $\mathbb{E}[y|\mathbf{d}]$, the average number of end-users that encounter at least one new problem given the discovery matrix $\mathbf{d}$.

Bringing in n' new users before going to market We now assume that n' additional users test the device before going to market. We will denote by $\mathbf{d}_n$ the

Algorithm 6 Sampling of $y|\mathbf{d}$

1. Sample $m|\mathbf{d}$ et $p_1, \ldots, p_m|\mathbf{d}, m$ (MCMC),
2. Sample $y|m, p_1, \ldots, p_m, \mathbf{d} \sim \mathcal{B}\left(N, 1 - \prod_{k=j+1}^{m}(1-p_k)\right)$.

initial discovery matrix and $\mathbf{d}_{n+n'}$ the augmented discovery matrix (see Figure 7.2).



(a) Initial (b) Augmented

Figure 7.2: Initial (a) and augmented (b) matrices (discovery matrices in dark green)

These n' new users may allow to find new usability problems and confirm low discovery probabilities of not yet discovered problems. Thus we would expect that:

$$\mathbb{E}\left[y|\mathbf{d}_n\right] - \mathbb{E}\left[y|\mathbf{d}_{n+n'}\right] \geq 0.$$

At this stage, $\boldsymbol{d}_{n+n'}$ in not observed thus it is necessary to consider the expectation of $\mathbb{E}\left[y|\mathbf{d}_{n+n'}\right]$ given the observed discovery matrix $\mathbf{d}_n$ which is a well-posed problem in the Bayesian framework.
Moreover from an economic point of view it is looked for a trade-off between the cost of the test on n' new users and the expected benefits. The optimal number of needed subjects $n'^{\star}$ can be expressed as follows:

$$n'^{\star} = \arg\max_{n'} \underbrace{c_{pb} \times \left(\mathbb{E}\left[y|\mathbf{d}_n\right] - \mathbb{E}\left[\mathbb{E}\left[y|\mathbf{d}_{n+n'}\right]|\mathbf{d}_n\right]\right)}_{\text{Expected value of sample information}} - \underbrace{c_{test} \times n'}_{\text{Costs}}$$

where c_{pb} is the cost of a problem and c_{test} is the cost for performing an additional test when assuming linear costs.
One key element is the computation of the double expectation $\mathbb{E}\left[\mathbb{E}\left[y|\mathbf{d}_{n+n'}\right]|\mathbf{d}_n\right]$ that can be performed as presented in Algorithm 7. This sampling can be very expensive since a new MCMC would be needed for each new augmented matrix $\mathbf{d}_{n+n'}$. However, as a first approximation it possible to froze some part of the parameters generated to sample $\mathbf{d}_{n+n'}$ (such as m, μ, and σ) then run a simple univariate update independently for each p_l of undiscovered problems. For sake of simplicity, this point is not detailed here.
In order to illustrate the proposed approach, let consider a simple example of a device after 40 initial usability tests. Figure 7.3 shows the trade-off that we obtain

Algorithm 7 Approximation of $\mathbb{E}\left[\mathbb{E}\left[y|\mathbf{d}_{n+n'}\right]|\mathbf{d}_n\right]$

For $b \in \{1, \ldots, B\}$

1. Sample $\mathbf{d}_{n+n'}|\mathbf{d}_n$:
   (a) Sample $m|\mathbf{d}_n$
   (b) Sample $\mu, \sigma, p_1, \ldots, p_m|m, \mathbf{d}_n$
   (c) Sample $\mathbf{d}_{n+n'}|p_1, \ldots, p_m, \mathbf{d}_n$
2. Sample $m|\mathbf{d}_{n+n'}$ and $p_1, \ldots, p_m|\mathbf{d}_{n+n'}, m$
3. Sample of y_b according to $y|m, p_1, \ldots, p_m, \mathbf{d}_{n+n'}$

Return $\frac{1}{B}\sum_{b=1}^{B} y_b$.

between the expected value of sample information and the costs. Thus allowing the practitioner to choose the most relevant sample size given the available information.

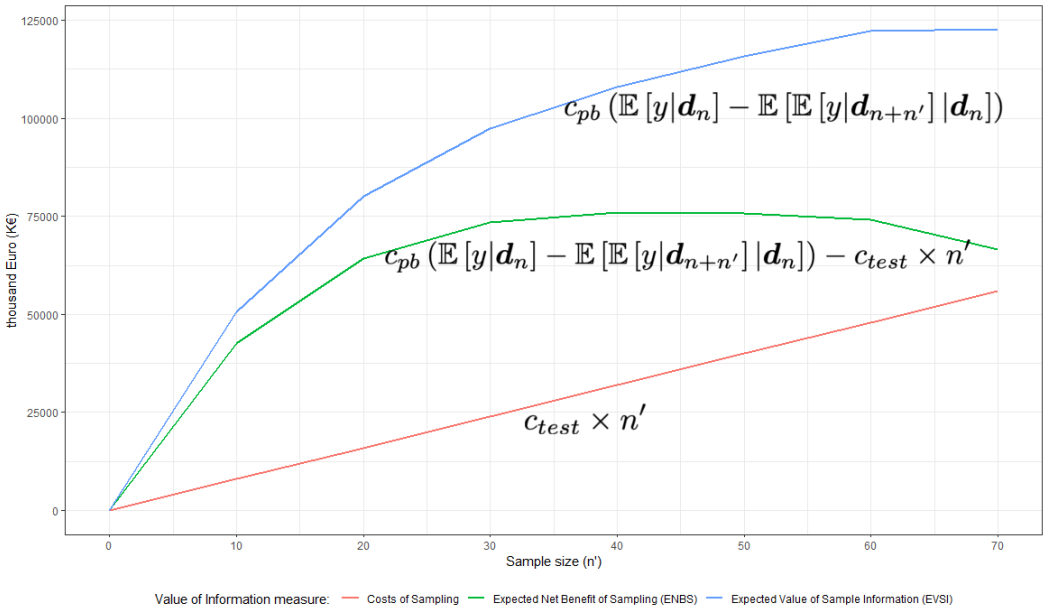


Figure 7.3: Expected value of sample information and costs according to the number of subjects.

7.4 Conclusion and perspectives

In this section, we have shown how we have made use of Bayesian statistic tools in the field of usability, in particular to the question of the number of needed subjects. The approach has been implemented in the R package `useval` which will be available

soon. It makes use of the `rstan` and `bridgesampling` packages.
This section has shown how it has been possible, starting from the initial issue of usability test to formalize the problem from the statistical point of view and propose a quite innovative solution. Such questions also refer to the missing data issues where some variable would be totally missing. In fact, in practice there are always "missing variables" however making assumptions of the way they could be totally missing is hard to answer from the general point of view.
Let notice that a continuous latent discovery variable has been considered, but it would also have been possible to consider a latent class variable leading to the clustering framework. Let also notice that only heterogeneity in columns has been considered, but heterogeneity in rows could also be considered. In that case, it could result in a particular co-clustering (see Govaert and Nadif, 2013) framework with truncation over some variables.

Bibliography

Borsci, S., Londei, A., and Federici, S. (2011). "The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation". In: *Cogn Process* 12.1. Edition: 2010/11/04, pp. 23–31. ISSN: 1612-4790 (Electronic) 1612-4782 (Linking). DOI: 10.1007/s10339-010-0376-6. URL: https://www.ncbi.nlm.nih.gov/pubmed/21046191.

Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J., and Young, T. (2013). "Reviewing and extending the five-user assumption: a grounded procedure for interaction evaluation". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 20.5, pp. 1–23. ISSN: 1073-0516.

Chao, A. (1984). "Nonparametric estimation of the number of classes in a population". In: *Scandinavian Journal of statistics*, pp. 265–270.

Faulkner, L. (2003). "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing". In: *Behavior Research Methods, Instruments, & Computers* 35.3, pp. 379–383. ISSN: 0743-3808.

US-FDA (2012). "Medical device recall report FY2003 to FY2012". In: *Center for Devices and Radiological Health*.

— (2016). "Applying human factors and usability engineering to medical devices: Guidance for industry and Food and Drug Administration staff". In: *Washington, DC: FDA*.

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Hertzum, M. and Jacobsen, N. E. (2003). "The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods". In: *International Journal of Human-Computer Interaction* 15.1. Section: 183, pp. 183–204. ISSN: 1044-7318 1532-7590. DOI: 10.1207/s15327590ijhc1501_14.

Kanis, H. (2011). "Estimating the number of usability problems". In: *Appl Ergon* 42.2. Edition: 2010/10/01, pp. 337–47. ISSN: 1872-9126 (Electronic) 0003-6870 (Linking). DOI: 10.1016/j.apergo.2010.08.004. URL: https://www.ncbi.nlm.nih.gov/pubmed/20880514.

Klingwort, J., Buelens, B., and Schnell, R. (2019). "Capture–Recapture Techniques for Transport Survey Estimate Adjustment Using Permanently Installed Highway-Sensors". In: *Social Science Computer Review*, p. 0894439319874684.

Lewis, J. R. (1994). "Sample sizes for usability studies: Additional considerations". In: *Human factors* 36.2, pp. 368–378. ISSN: 0018-7208.

— (2000). *Using discounting methods to reduce overestimation of p in problem discovery usability studies*. Tech. rep. Citeseer.

— (2001). "Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples". In: *International Journal of Human-Computer Interaction* 13.4. Section: 445, pp. 445–479. ISSN: 1044-7318 1532-7590. DOI: 10.1207/s15327590ijhc1304_06.

Meng, X. .-L. and Wong, W. H. (1996). *Simulating ratios of normalizing constants via a simple identity: a theoretical exploration*. Statistica Sinica, pp. 831–860.

UK-MHRA (2017). *Human Factors and Usability Engineering – Guidance for Medical Devices Including Drug-Device Combination Products*. Tech. rep. URL: https://www.gov.uk/government/publications/guidance-on-applying-human-factors-to-medical-devices.

Nielsen, J. (1994). *Usability inspection methods*. Paper presented at the Conference companion on Human factors in computing systems.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation: Springer Science & Business Media*.

Sauro, J. and Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann. ISBN: 0-12-802548-4.

Schmettow, M. (2008). "Heterogeneity in the usability evaluation process". In: *People and Computers XXII Culture, Creativity, Interaction 22*, pp. 89–98.

— (2012). "Sample size in usability studies". In: *Communications of the ACM* 55.4, pp. 64–70. ISSN: 0001-0782.

Stan Development Team (2020). *The Stan Core Library*. Version 2.21.2. URL: http://mc-stan.org/.

Thomas, D. G. and Gart, J. J. (1971). "Small sample performance of some estimators of the truncated binomial distribution". In: *Journal of the American Statistical Association* 66.333, pp. 169–177. ISSN: 0162-1459.

Vandewalle, V., Caron, A., Delettrez, C., Périchon, R., Pelayo, S., Duhamel, A., and Dervaux, B. (2020). "Estimating the number of usability problems affecting medical devices: modelling the discovery matrix". In: *BMC Medical Research Methodology* 20.234.

Chapter 8

Artificial intelligence for aviation

Contents

8.1 Introduction 141
8.2 Flight data and aviation practices 142
8.2.1 Flight data 142
8.2.2 Aviation practices 142
8.3 Inferring performance tables from flight data 144
8.4 Optimizing flight trajectory through a functional approach 146
8.4.1 General trajectory optimization problem 146
8.4.2 A model for instantaneous consumption 147
8.4.3 Decomposition of the trajectory in a basis and resulting quadratic optimization problem 148
8.4.4 Satisfying constraints 148
8.5 Conclusion 149

Related article

1. F. Dewez, B. Guedj, and V. Vandewalle (2020). "From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning". In: *Data-Centric Engineering* 1, e11. DOI: 10.1017/dce.2020.12

8.1 Introduction

This Chapter presents some work related to the PERF-AI European project where I had been involved for two years. The PERF-AI European Project[1] has been built based on a collaboration between SafetyLine, a start-up working on big-data for aviation, and Inria. From the Inria part, Benjamin Guedj and I have mounted the project. The project started in November 2018 and finishes at the end of October 2020. During the project, we hired a post-doctoral researcher, Florent Dewez, for 18 months, and an engineer, Arthur Talpaert, for 12 months. Schematically, SafetyLine was in charge of bringing data, his expertise, and real problematic. Inria was in

[1] https://cordis.europa.eu/project/id/815914

charge of proposing mathematical modeling and developing a prototype then industrialized by SafetyLine. The main challenge of the project is to make use of flight recorder data to update performance tables and to propose optimized trajectories. One particular interest is to reduce fuel consumption and thus the emission of CO_2. This results in a first article (Dewez, Guedj, and Vandewalle, 2020), a second article about to be submitted, and the development of Python library, Pyrotor, available soon.
In Section 8.2, I present the available data and discuss some of the practices in the aviation domain. In Section 8.3, I present how performance tables can be updated using the available data and taking into account aviation constraints. In Section 8.4, I present the strategy that we have proposed for optimizing flight trajectory based on a data-driven consumption model.

8.2 Flight data and aviation practices

8.2.1 Flight data

For security reasons, flight data are recorded all along the flight (see Figure 8.1). It may for instance help to understand the causes of an air crash, but these data are also easily available for flight companies that pay more and more attention to use it to optimize flight performances. These data are recorded according to some norms and are available through the Quick Access Recorder (QAR). Thus, for each second of the flight it is possible to get directly or to compute the following variables among others: angle of attack (α), path angle (γ), true airspeed (V), Mach number (M), Altitude (h), Mass (m), Fuel flow (FF), Static air temperature (SAT), Air density (ρ), engine thrust ($N1$). From a statistical point of view, all these available variables recorded according to the time can be viewed as multivariate functional data (Ramsay and Silverman, 2005). Based on these available data one goal could be for instance to predict the total fuel consumption based on the observation of all the other state variables along the time. There also could be some online optimization perspective such as adapting some control variables according to some feedback from the consumption or other viewpoints. However, before advancing on what can be performed from a statistical point of view based on these data it is important to understand how a flight is planned in practice and what are the actual degrees of freedom concerning these practices in the scope of the PERF-AI project.

8.2.2 Aviation practices

For security and certification reasons, aviation devices and practices have not changed so much for 40 years. When planning his flight the pilot enters some basic required information in the flight monitoring systems (FMS) (see Figure 8.2). These data are, for instance, control points for the flight trajectory, piece-wise constant control parameters such as the engine thrust, the altitude, ... Based on these basics information the FMS returns some outputs which are mainly the time of the flight



Figure 8.1: Flight data recorder

and the fuel consumption. At this stage, it is important to notice that FMS has very limited computation power, and thus performs very limited calculus. These computations are mainly performed based on performance tables which indicate for a given engine thrust, altitude, angle of attack, ... what is the consumption, and what is speed. These tables are given by the aircraft manufacturer at the beginning of the aircraft life based on a few real flight tests and measures in wind tunnels. During the entire life of the aircraft, the performance tables are generally not updated, however, it is known that these performances can vary during the life of the aircraft (Airbus, 2002). In practice based on the difference between the observed and predicted consumption the pilot has only the possibility to update the computation by a multiplicative coefficient (the so-called perf-factor), affecting similarly all the phases of the flight. However, the data from the QAR would allow a finer update. Thus, one of the first tasks in the project was to propose a strategy to update the performance tables based on the data of the QAR, leading to the article Dewez, Guedj, and Vandewalle, 2020.



Figure 8.2: Flight Monitoring System

8.3 Inferring performance tables from flight data

I now present how it is possible to update performance tables based on flight data. First notice that if performance tables would give the expected time and fuel consumption according to the flight variables, all the variables being recorded during the flight, it would be easy to learn a regression model to update performance tables. However, performance tables give among others drag and lift coefficients which are not directly recorded, but allow the FMS to deduce the speed, the fuel consumption, ... Moreover, we do not have access to the computations which are performed inside the FMS for industrial reasons.

A previous study has been performed on this topic by Cedric Rommel a former Ph.D. student at SafetyLine Rommel, 2018. However, this approach has identifiability issues, making it impossible to recover relevant estimated coefficients. He proposed some solution to remove this identifiability problem by adding some penalty to enforce proximity between the obtained coefficient and some expected value. But it was not sufficient to guarantee accurate estimation.

To solve this problem we have proposed to place ourselves in a framework where point mechanics equations can be simplified (basically in stationary flight), thus making it possible to deduce drag and lift coefficients. In Figure 8.3 we present the main forces which apply to the plane in stationary flight, making it possible to apply Newton's second law. More details about these computations can be found in Dewez, Guedj, and Vandewalle (2020). Then, based on the approximation of these coefficients, standard regression models such as linear, polynomial, or gradient tree boosting models can be used to fill the regression tables. With the state of the art giving us access to the relevance of the physical approximation, we were able to derive upper bounds on the expectation of the absolute error between the predicted value of the variable and the true (unknown) value of the variable.

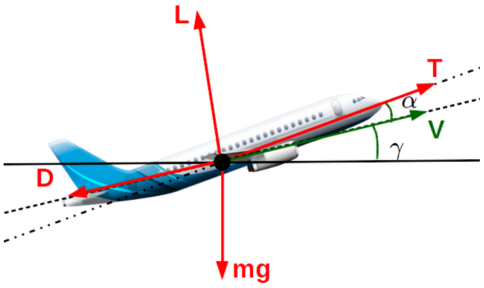


Figure 8.3: Main forces in stationary flight: the thrust force (T), the lift force (L), the drag force (D), the weight (mg), the speed vector (V), the angle of attack (α), the path angle (γ).

This point is not deeply detailed here but some pre-processing of the data coming from the QAR needs to be performed before it can be used. First, even if latter the temporal structure of the data is not used in the regression, it can be used for

smoothing the data, which is particularly important for at least two reasons: (i) the QAR records the variables with limited precision, thus making a potential abrupt change in the signal, (ii) like every sensor there can be some measure noise that can be limited thanks to smoothing. Second, it was needed to filter the data to only keep stationary flight data (constant speed, without turn, ...). Third for the learning also to be reliable, on a relatively small range of variation for the flight parameters we limited ourselves to particular flight phases (climb, cruise, descent). In the end, a data set as presented in Table 8.1 is obtained by considering the concatenation of several flights of a plane for instance, or of several planes if not enough data from a particular plane is available.

Observation	ρ	V	α	FF	...	m	γ
1	0.3224	234.5	0.0324	0.6716	...	62,519	0.0139
2	0.3704	236.8	0.0224	0.6503	...	64,960	0.0198
3	0.3224	234.8	0.0305	0.6637	...	66,974	0.0159
⋮	⋮	⋮	⋮	⋮		⋮	⋮
164,054	0.3433	232.9	0.0332	0.6642	...	66,673	0.0150

Table 8.1: Example of a preprocessed data set.

Figures 8.4 allows to visualize the tendencies of estimators of drag and lift coefficients concerning the Mach number (ratio of flow velocity past a boundary to the local speed of sound) for different fixed values of the angle of attack (α). Now we mention that 90% of Mach number data are between 0.77 and 0.80 and 90% of the angle of attack data are between 1.9° and 2.9°. Then we observe that both predicted C_D and C_L globally increase when the Mach number or the angle of attack increases. This global tendency is expected in this small range of values according to Anderson (1999, Part 1, Chap. 2): the larger the angle of attack or the Mach number, the larger the drag and lift coefficients. Nevertheless, this natural tendency for the lift coefficient is not verified by the estimators when α is too large, namely $\alpha = 2.75°$ or $\alpha = 3°$. This unexpected behavior can be explained by the approximated nature of the variable C_L. Indeed it may behave in a way that is different from the true value of C_L in certain regions of the cruise domain. In this case, any estimator for C_L is likely to inherit this unexpected behavior and we believe refined aeronautics-supported approximations would bring a solution.

An internal to the PERF-AI project Python package has been developed and is now available to SafetyLine to deeply test it and industrialize it. One of the main interests of this work from a practical point of view is that it makes it possible to be used in the short term since it does not changes current practices, just needing some update of the performance tables.

From an optimization point of view, adding the constraint to pass through performance tables expressed in terms of drag and lift coefficients that are never observed in practice could be seen as an unnecessary step. Moreover, discretizing the performance model on the grid of the performance tables may also be unnecessary.

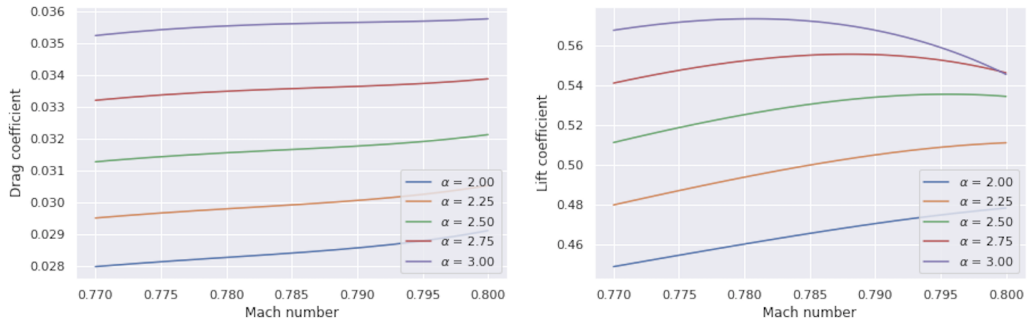


Figure 8.4: Predictions of drift (C_D) and lift (C_L) coefficients from polynomial models

However, a method not complying with these constraints would also take more time before being used in real life. In Section 8.4, I present how optimization can be considered by directly using the available flight data removing the constraint that the calculation passed through the FMS. The article related to this work is about to be submitted.

8.4 Optimizing flight trajectory through a functional approach

In this section, I show our proposal to optimize the flight trajectory. It is mainly based on a decomposition of the coordinates of the trajectory according to some basis, thus making it possible to replace the initial optimization on a multivariate functional space with optimization on the coefficients of the decomposition. Additional elements are also added to take into account constraints on the trajectory, or eventual multi-colinearity of the coefficients which may be observed on real flights. One advantage of the proposed approach is that it is very fast, which is an important requirement since the pilot has only 30 minutes between his flight plan and takeoff.

8.4.1 General trajectory optimization problem

First of all, it is needed to define a trajectory and endpoint conditions. We will look for trajectories that last a time T.

Definition 1 (Trajectory) *Let $T > 0$ be a real number and let $D \geqslant 1$ be an integer. Any continuous $\mathbb{R}^D$-valued map y defined on $[0,T]$,* i.e. *$y \in \mathcal{C}\big([0,T],\mathbb{R}^D\big)$, is called a* trajectory *over the time interval $[0,T]$. The d-th component of a trajectory y will be denoted by $y^{(d)}$.*

Here we are looking for trajectories that start and end at certain given states. The space of such trajectories is introduced below.

Definition 2 (Endpoint conditions) *Let $y_0, y_T \in \mathbb{R}^D$. We define the set $\mathcal{D}(y_0, y_T) \subset \mathcal{C}\big([0,T], \mathbb{R}^D\big)$ as the set of trajectories over $[0,T]$ being in the initial state y_0 and in the final state y_T,* i.e.

$$y \in \mathcal{D}(y_0, y_T) \qquad \Longleftrightarrow \qquad \begin{cases} y(0) = y_0 \\ y(T) = y_T \end{cases}$$

Let now define the cost function F related to the total fuel consumption of the flight $F : \mathcal{C}\big([0,T], \mathbb{R}^D\big) \longrightarrow \mathbb{R}$ then some optimal trajectory $y^\star$ with respect to F would be

$$y^\star \in \underset{y \in \mathcal{D}(y_0, y_T)}{\arg\min} \; F(y).$$

Thus the initial optimization problem consists of optimizing a real-valued function over a multivariate functional space. This initial problem can be related to the field of optimal control theory and some solutions could eventually be found. However, let notice at this stage that before the optimization, the function F needs to be learned based on flight data presented in Section 8.2. This can be a hard issue since even with a high number of data we are dealing with the problem of multivariate functional regression where the considered trajectories for learning do not systematically have the same length.

8.4.2 A model for instantaneous consumption

A first assumption that will be made is that the total consumption of the flight is the integral of instantaneous consumption, and that instantaneous consumption at time t can be explained by the state variables at time t denoted by $y(t)$. Let denote by $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ this instantaneous consumption function we have:

$$F(y) := \int_0^T f\big(y(t)\big)\, dt \; .$$

Such formulation makes the problem easier to learn from a statistical point of view since the recorded database gives us at each second of the flight the state variables and the instantaneous consumption (fuel flow). Thus before some smoothing of the data, we were able to constitute a learning database with the instantaneous consumption and the state variables. Merging all these data points we were able to learn an accurate model for instantaneous consumption. Then making it easy to predict the total fuel consumption.
Without additional assumption, the optimization problem stays quite general and we would recommend using optimal control tools. In fact, in this work, the optimization is performed phase per phase, and we have first focused on the climbing phase. In this phase, it is possible to assume a quadratic model for the instantaneous fuel consumption given the other variables.

8.4.3 Decomposition of the trajectory in a basis and resulting quadratic optimization problem

We will now make assumptions more precise to keep a tractable global optimization problem. To do this we will assume (i) that each component of y can be decomposed on a functional basis (ii) that f is a quadratic function. This will make the computation particularly efficient, by summarizing the initial optimization problem as a quadratic optimization problem according to the coefficients. This optimization is very fast since it only involves the integrals of the products of base functions, which is only needed to be performed once and can be stored for further use. Let denote by $\check{F}$ the function that takes the coefficients c of the decomposition of the coordinates of the trajectory in some bases and returns the total consumption:

$$\check{F}(c) = c^T Q c + w^T c + rT$$

where Q is a square matrix involving the coefficient of the quadratic regression model and the integral of the product of basis functions, w is a vector of the same dimension as c and r some constant. Thus obtaining the optimal trajectory through a standard quadratic programming algorithm.

8.4.4 Satisfying constraints

In practice, the trajectory can be submitted to flight constraints which can be expressed as follows.

Definition 3 (Additional constraints) *For $l = 1, \dots, L$, let g_l be a real-valued function defined on $\mathbb{R}^D$. We define the set $\mathcal{G} \subset \mathcal{C}\big([0,T], \mathbb{R}^D\big)$ as the set of trajectories over $[0,T]$ satisfying the following L inequality constraints given by the functions g_l,* i.e.

$$y \in \mathcal{G} \qquad \Longleftrightarrow \qquad \forall\, l = 1, \dots, L \quad \forall\, t \in [0,T] \qquad g_l\big(y(t)\big) \leqslant 0 \; .$$

Satisfying these constraints could be difficult inside the algorithm, thus we propose to considers some reference trajectories satisfying these constraints and consider the optimization as a trade-off between total fuel consumption and the distance with reference trajectories, the problem is still expressed according to the decomposition in the bases. Thus we are looking for $c^\star$ in the following way.

$$c^\star \in \arg\min_c \check{F}(c) + \kappa \sum_{i=1}^{I} \big(c - c_{R_i}\big)^T \Sigma^\dagger \big(c - c_{R_i}\big) \; ,$$

where c_{R_i} are the coefficients of the reference trajectory number i and $\Sigma^\dagger$ some metric. In practice we propose to estimate this metric as the pseudo-inverse of the covariance matrix estimated based on many observed trajectories. We recommend to choose the regularization parameter κ as the lowest value such as the constraints are satisfied, this search can be performed in an iterative way. The penalty can also

be reinterpreted in the Bayesian framework has a Gaussian multivariate prior on c centered around reference trajectories and with $\kappa\Sigma^{\dagger}$ as inverse covariance matrix. For instance, a plot of the optimized trajectory obtained is given in Figure 8.5. We observe that the optimized trajectory seeks to reach the maximum altitude in the minimum amount of time; this is in accordance with the existing literature (see for instance Codina and Menéndez (2014) and references therein). In particular, the duration is equal to $1,048$ seconds which is slightly shorter than the reference duration. We note also that the optimized Mach number shares a very similar pattern with the references. On the other hand, the optimized engine's rotational speed tends to slowly decrease until the cruise regime before reaching the top of the climb. This is not the case for the reference engine's speed which falls to the cruise regime just after reaching the final altitude. Most of the savings seem to be achieved in these last moments of the climb. To finish we emphasize that the optimized trajectory presents a realistic pattern inherited from the reference trajectories.

8.5 Conclusion

Working on the PERF-AI project was a very valuable experience. It allowed me to follows a project from end to end. Starting from the industrial issue, mounting the project, hiring a postdoctoral researcher and an engineer, and proposing some accurate solutions from the company and academic points of view.
This field which consists of merging some prior physic model and statistical data is very promising, and asks interesting issues. For instance how it is possible to merge different kinds of paradigms (physical and data-based one) to take the best of both worlds. The proposed approach could be applied to a wide range of problems such as sailing for instance.

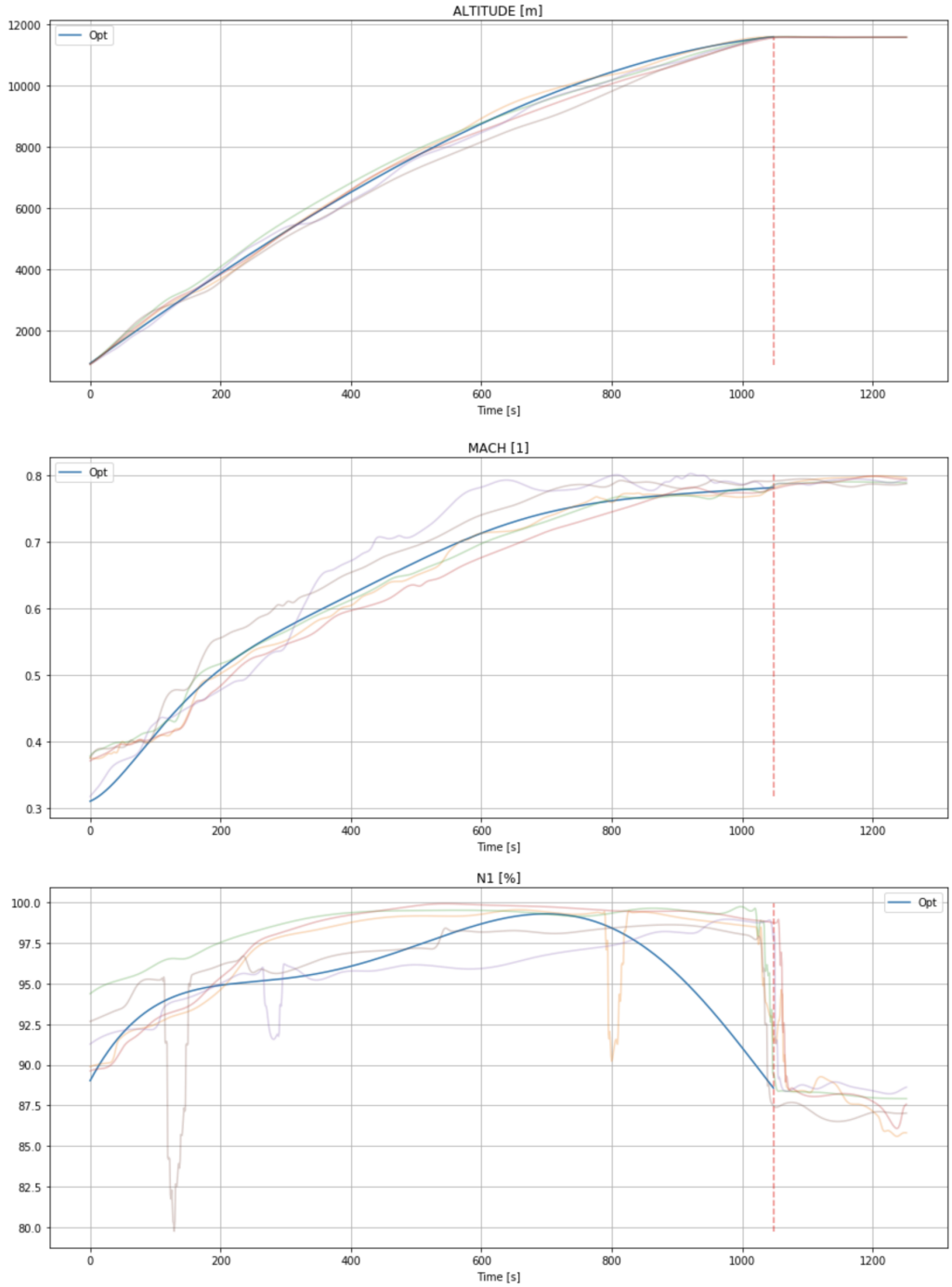


Figure 8.5: Optimised and reference altitudes, Mach numbers and engines rotational speeds – The optimised trajectory is represented by the blue curves.

Bibliography

Airbus (2002). *Getting to Grips with Aircraft Performance Monitoring*. Tech. rep. http://www.smartcockpit.com/docs/Getting_to_Grips_With_Aircraft_Performance.pdf. Airbus.

Anderson, J. (1999). *Aircraft performance and design*. McGraw-Hill international editions: Aerospace science/technology series. WCB/McGraw-Hill.

Codina, R. D. and Menéndez, X. P. (2014). "How much fuel and time can be saved in a perfect flight trajectory ? Continuous cruise climbs vs. conventional operations". In: *Proceedings of the 6th International Congress on Research in Air Transportation (ICRAT)*.

Dewez, F., Guedj, B., and Vandewalle, V. (2020). "From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning". In: *Data-Centric Engineering* 1, e11. DOI: 10.1017/dce.2020.12.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Second. Springer Series in Statistics. Springer, New York, pp. xx+426. ISBN: 978-0387-40080-8; 0-387-40080-X.

Rommel, C. (2018). "Exploration de données pour l'optimisation de trajectoires aériennes". PhD thesis. École Polytechnique.

Chapter 9

Contributions to Credit scoring

Contents

9.1 Introduction . . . 154
9.2 Rational review of reject inference methods . . . 154
9.2.1 Problem presentation . . . 154
9.2.2 General parametric model . . . 155
9.2.3 General EM principle . . . 156
9.2.4 Several reintegration algorithm . . . 156
9.2.5 Concluding remarks . . . 158
9.3 Model-embedded feature quantization . . . 158
9.3.1 Motivation . . . 158
9.3.2 Quantization as a combinatorial challenge . . . 159
9.3.3 State of the art . . . 160
9.3.4 Proposed criterion . . . 161
9.3.5 A relaxation of the optimization problem . . . 162
9.3.6 A neural network-based estimation strategy . . . 165
9.3.7 Conclusion . . . 167
9.4 Conclusion and perspectives . . . 167

Related production

1. A. Ehrhardt, V. Vandewalle, C. Biernacki, and P. Heinrich (2018). "Supervised multivariate discretization and levels merging for logistic regression". In: *23rd International Conference on Computational Statistics*. Iasi, Romania. URL: https://hal.archives-ouvertes.fr/hal-01949128

2. A. Ehrhardt, C. Biernacki, V. Vandewalle, P. Heinrich, and S. Beben (2017). "Réintégration des refusés en Credit Scoring". In: *49e Journées de Statistique*. Avignon , France. URL: https://hal.archives-ouvertes.fr/hal-01653767

3. A. Ehrhardt, C. Biernacki, V. Vandewalle, and P. Heinrich (2019). "Feature quantization for parsimonious and interpretable predictive models". working paper or preprint. URL: https://hal.archives-ouvertes.fr/hal-01949135

9.1 Introduction

From 2016 to 2019, I have co-supervised Adrien Ehrhardt's thesis with Christophe Biernacki and Philippe Heinrich. This thesis was carried out by a CIFRE contact at the company CA-CF (a company specialized in consumer loans). In this chapter, I present the main points of Adrien Ehrhardt's thesis Ehrhardt (2019). These works have been presented in conferences and are the subjects of articles submitted or in the process of being submitted.
Score construction is an important practice in many fields, such as credit granting (credit scoring), medical diagnosis (survival score), marketing (appetite score), ... The thesis of Adrien has focused on credit scoring but the obtained results also apply on many others fields. In this framework, we have mainly focused on logistic regression which is still the most widely used method in credit scoring. This mainly because it is easy to interpret which is mandatory for regulatory reasons and because it produces good results in practice. The major contribution of Adrien's thesis is a rational review of reject inference methods and the proposal of a model-embedded feature quantization approach. The first one considers how the data from the non-financed clients may be useful/used to improve the relevance of a score. The second one considers embedding the quantization process (grouping modalities of categorical variables or discretizing quantitative variables) into the scoring model building rather than used as a pre-processing step.
In Section 9.2, I present the rationale review that we have proposed on reject inference methods, trying to highlight the often hidden assumptions that are made when using these methods. In Section 9.3, I present the embedded quantization approach that we have proposed relying on a smoothed relaxation of the initial combinatorial optimization problem.

9.2 Rational review of reject inference methods

9.2.1 Problem presentation

An important issue in credit scoring is that the score used for future clients is developed based on financed clients only. Thus, inducing a potential drift between the population on which the model is learned and the population on which the model is used. Depending on the selection mechanism and of the model used this can have consequences on the scorecard's relevance. Reject inference methods consist of using the data of non-financed clients (their answers in the questionnaire) to update the scorecards. For a review of such methods see Viennet, Soulié, and Rognier (2006), Guizani et al. (2013), Banasik and Crook (2007), and Nguyen (2016) among others. Most of the reintegration methods are quite empirical and are used without explicitly giving the assumptions that are made.
From a general point of view reject inference is a particular case on semi-supervised learning (Chapelle, Schlkopf, and Zien, 2010) since it consists of learning from both labeled data (financed clients) and unlabeled data (non-financed clients). For in-

stance Feelders (2000) investigated this issue by using mixture models. In semi-supervised learning, it is often assumed that labeled and unlabeled data come from the same distributions, whereas in the particular case of credit scoring this assumption does not hold due to the selection of credit applicants. Moreover contrary to the semi-supervised framework where unlabeled data are more numerous than labeled data, in credit scoring the number of financed clients and the number of non-financed clients are often balanced.

In our particular study, we have focused on logistic regression. Since this model is a local model as defined in Zadrozny (2004), as a model directly modeling $p(y|\boldsymbol{x})$, the probability of y ($y \in \{0,1\}$, 1 is the client refunded his loan and 0 otherwise) given the available features $\boldsymbol{x}$, under missing at random assumption (MAR) is immune to biasedness of $\boldsymbol{x}$, which is not the case of generative models for instance. The accuracy of the reject inference methods according to sampling and model assumptions are discussed.

9.2.2 General parametric model

Firstly, it is both convenient and realistic to assume that triplets in the complete sample $\mathcal{D}_c = \{\boldsymbol{x}_i, y_i, z_i\}_{1 \leq i \leq n}$ are all independent and identically distributed (i.i.d), including the unknown values of y_i when $i \in$ NF (NF representing the set of non-financed clients and F the set of financed clients), z_i indicating whether or not the client is financed ($z_i \in \{\text{f}, \text{nf}\}$). Secondly, it is usual and convenient to assume that the unknown distribution $p(y|\boldsymbol{x})$ belongs to a given parametric family $\{p_{\boldsymbol{\theta}}(y|\boldsymbol{x})\}_{\boldsymbol{\theta} \in \Theta}$, where Θ is the parameter space. For instance, logistic regression is often considered in practice.

As in any missing data situation (here z indicates if y is observed or not), the relative modeling process, namely $p(z|\boldsymbol{x}, y)$, has also to be clarified. For convenience, we can also consider a parametric family $\{p_{\boldsymbol{\phi}}(z|\boldsymbol{x}, y)\}_{\boldsymbol{\phi} \in \Phi}$, where $\boldsymbol{\phi}$ denotes the parameter and Φ the associated parameter space of the financing mechanism. Note that we consider here the most general missing data situation, namely a Missing Not At Random (MNAR) mechanism (see Little and Rubin, 2014). It means that z can be stochastically dependent on some missing data y, *i.e.* $p(z|\boldsymbol{x}, y) \neq p(z|\boldsymbol{x})$.

Finally, combining both previous distributions $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ and $p_{\boldsymbol{\phi}}(z|\boldsymbol{x}, y)$ leads to express the joint distribution of (y, z) conditionally to $\boldsymbol{x}$ as:

$$p_{\boldsymbol{\gamma}}(y, z|\boldsymbol{x}) = p_{\boldsymbol{\phi}(\boldsymbol{\gamma})}(z|y, \boldsymbol{x}) p_{\boldsymbol{\theta}(\boldsymbol{\gamma})}(y|\boldsymbol{x}) \tag{9.1}$$

where $\{p_{\boldsymbol{\gamma}}(y, z|\boldsymbol{x})\}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}}$ denotes a distribution family indexed by a parameter $\boldsymbol{\gamma}$ evolving in a space $\boldsymbol{\Gamma}$. Here it is clearly expressed that both parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can depend on $\boldsymbol{\gamma}$, even if in the following we will note shortly $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\gamma})$ and $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\gamma})$. In this very general missing data situation, the missing process is said to be *non-ignorable*, meaning that parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be functionally dependent (thus $\boldsymbol{\gamma} \neq (\boldsymbol{\phi}, \boldsymbol{\theta})$).

9.2.3 General EM principle

Mixing previous model and data, the maximum likelihood (ML) principle can be invoked for estimating the whole parameter $\boldsymbol{\gamma}$, thus yielding as a by-product an estimate of the parameter $\boldsymbol{\theta}$. Indeed, $\boldsymbol{\theta}$ is of particular interest, the goal of the financial institutions being solely to obtain an estimate of $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. The observed log-likelihood can be written as:

$$\ell(\boldsymbol{\gamma};\mathcal{D}) = \sum_{i\in \mathrm{F}} \ln p_{\boldsymbol{\gamma}}(y_i, \mathrm{f}|\boldsymbol{x}_i) + \sum_{i'\in \mathrm{NF}} \ln \left[\sum_{y\in\{0,1\}} p_{\boldsymbol{\gamma}}(y, \mathrm{nf}|\boldsymbol{x}_{i'}) \right], \tag{9.2}$$

where $\mathcal{D}$ is the observed data (y_i unknown for $z_i = \mathrm{nf}$. Within this missing data paradigm, the Expectation-Maximization (EM) algorithm (see Dempster, Laird, and Rubin, 1977) can be used: it aims at maximizing the expectation of the complete likelihood $\ell_c(\boldsymbol{\gamma};\mathcal{T}_c)$ (defined hereafter) over the missing labels. Starting from an initial value $\boldsymbol{\gamma}^{(0)}$, iteration (s) of the algorithm is decomposed into the following two classical steps:

E-step compute the conditional probabilities of missing y_i values ($i \in \mathrm{NF}$):

$$y_i^{(s)} = p_{\boldsymbol{\theta}(\boldsymbol{\gamma}^{(s-1)})}(1|\boldsymbol{x}_i, \mathrm{nf}) = \frac{p_{\boldsymbol{\gamma}^{(s-1)}}(1, \mathrm{nf}|\boldsymbol{x}_i)}{\sum_{y'\in\{0,1\}} p_{\boldsymbol{\gamma}^{(s-1)}}(y', \mathrm{nf}|\boldsymbol{x}_i)}; \tag{9.3}$$

M-step maximize the conditional expectation of the complete log-likelihood:

$$\ell_c(\boldsymbol{\gamma};\mathcal{D}_c) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\gamma}}(y_i, z_i|\boldsymbol{x}_i) = \sum_{i\in \mathrm{F}} \ln p_{\boldsymbol{\gamma}}(y_i, \mathrm{f}|\boldsymbol{x}_i) + \sum_{i\in \mathrm{NF}} \ln p_{\boldsymbol{\gamma}}(y_{i'}, \mathrm{nf}|\boldsymbol{x}_{i'}), \tag{9.4}$$

leading to:

$$\begin{aligned}\boldsymbol{\gamma}^{(s)} &= \arg\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \mathbb{E}_{\mathbf{y}_{\mathrm{nf}}}[\ell_c(\boldsymbol{\gamma};\mathcal{D}_c)|\mathcal{T}, \boldsymbol{\gamma}^{(s-1)}] \\ &= \arg\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{i\in \mathrm{F}} \ln p_{\boldsymbol{\gamma}}(y_i, \mathrm{f}|\boldsymbol{x}_i) + \sum_{i'\in \mathrm{NF}} \sum_{y\in\{0,1\}} y_{i'}^{(s)} \ln p_{\boldsymbol{\gamma}}(y, \mathrm{nf}|\boldsymbol{x}_{i'}).\end{aligned}$$

Usually, stopping rules rely either on a predefined number of iterations, or on a predefined stability criterion of the observed log-likelihood.
Most of reject inference methods try to mimic this EM algorithm without making explicit assumptions on $p_{\boldsymbol{\theta}(\boldsymbol{\gamma})}(1|\boldsymbol{x}_i, \mathrm{nf})$ which is a difficult task since y is not known for non financed clients.

9.2.4 Several reintegration algorithm

In this section, we discuss some of the main reject inference methods. More details can be found in Ehrhardt (2019).

Strategy 1: Ignoring non-financed clients The simplest reject inference strategy is to ignore non-financed clients for estimating $\boldsymbol{\theta}$. Thus it consists in estimating $\boldsymbol{\theta}$ by maximizing the log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{f}})$.
This strategy leads trivially to a consistent estimator in the case of missing completely at random (MCAR). Under well specified model hypothesis and missing at random assumption this strategy leads to consistent estimates (Zadrozny, 2004). In other cases, it is not possible to draw definitive conclusions.

Strategy 2: Fuzzy Augmentation This strategy can be found in Nguyen, 2016. It corresponds to an algorithm which is starting with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$ (see previous section). Then, all $\{y_i\}_{i \in \mathrm{NF}}$ are imputed by their expected value given by: $\hat{y}_i^{(1)} = p_{\hat{\boldsymbol{\theta}}^{(0)}}(1|\boldsymbol{x}_i)$ (notice that these imputed values are not in $\{0, 1\}$ but in $]0, 1[$). However, this does not modified the obtained estimator of strategy 1.

Strategy 3: Reclassification This strategy corresponds to an algorithm which is starting with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$. Then, all $\{y_i\}_{i \in \mathrm{NF}}$ are imputed by the *maximum a posteriori* (MAP) principle given by: $\hat{y}_i^{(1)} = \arg\max_{y \in \{0,1\}} p_{\hat{\boldsymbol{\theta}}^{(0)}}(y|\boldsymbol{x}_i)$. This solution consist in a CEM (Celeux and Govaert, 1992) algorithm which known to produce biased solutions.

Strategy 4: Augmentation Augmentation can be found in Viennet, Soulié, and Rognier, 2006. It is also documented as a "Re-Weighting method" by Guizani et al., 2013; Banasik and Crook, 2007; Nguyen, 2016. This technique is directly influenced by the importance of sampling literature (see works from Zadrozny, 2004 for an introduction in a similar context as here). Indeed, intuitively, as for all selection mechanisms such as survey respondents, observations should be weighted according to their inverse probability of being in the sample w.r.t. the whole population, *i.e.* by the inverse of $p(z|\boldsymbol{x}, y)$. By assuming implicitly a MAR and ignorable missingness mechanism, we get $p(z|\boldsymbol{x}, y) = p(z|\boldsymbol{x})$. Compared with strategy 1, this strategy has the advantage to produce the best logistic parameters in the MAR setting, even when the model is miss-specified.

Strategy 5: Twins This reject inference method is documented internally at CACF. It consists of combining two logistic regression-based scorecards: one predicting y learned on financed clients (denoted by $\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ as previously), the other predicting z learned on all applicants (denoted by $\hat{\boldsymbol{\phi}}$), before learning the final scorecard using the predictions made by both previous scorecards on financed clients. The detailed procedure is provided in Adrien Erhardt's thesis. This procedure does not modify the obtained results by strategy 1.

Strategy 6: Parcelling The parcelling method can be found in works from Guizani et al., 2013; Banasik and Crook, 2007; Viennet, Soulié, and Rognier, 2006. This method aims to correct the log-likelihood estimation in the MNAR case by making

further assumptions on $p(y|\boldsymbol{x}, z)$. It is a little deviation from the Fuzzy Augmentation method in a MNAR setting, where the payment status $\hat{y}_i^{(1)}$ for non-financed clients ($i \in \text{NF}$) is estimated by a quantity now differing from this one associated to financed clients (which was namely $p_{\hat{\boldsymbol{\theta}}^{(0)}}(1|\boldsymbol{x}_i, \text{f})$, with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\text{f}}$). The core idea is to propose an estimate $\hat{y}_i^{(1)} = \hat{p}(1|\boldsymbol{x}_i, \text{nf}) = 1 - \hat{p}(0|\boldsymbol{x}_i, \text{nf})$, for $i \in \text{NF}$, with

$$\hat{p}(0|\boldsymbol{x}_i, \text{nf}) \propto \varepsilon_{k(\boldsymbol{x}_i)} p_{\hat{\boldsymbol{\theta}}^{(0)}}(0|\boldsymbol{x}_i, \text{f}),$$

where $k(\boldsymbol{x})$ is the scoreband index among K equal-length scorebands $B_1, \ldots, B_K$ and $\varepsilon_1, \ldots, \varepsilon_K$ are so-called "prudence factors". These latter are generally such that $1 < \varepsilon_1 < \cdots < \varepsilon_K$, and they aim to counterbalance the fact that non-financed low refunding probability clients are considered way riskier, all other things being equal, than their financed counterparts. All these ε_k values have to be fixed by the practitioner. The method is thereafter strictly equivalent to Fuzzy Reclassification by maximizing over $\boldsymbol{\theta}$ the complete log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{D}_{\text{c}}^{(1)})$ with $\mathcal{D}_{\text{c}}^{(1)} = \mathcal{D} \cup \hat{\mathbf{y}}_{\text{nf}}^{(1)}$ and $\hat{\mathbf{y}}_{\text{nf}}^{(1)} = \{\hat{y}_i^{(1)}\}_{i \in \text{NF}}$. It yields a final parameter estimate $\hat{\boldsymbol{\theta}}^{(1)}$. This method can adjust the obtained results, however the prudence factors cannot be estimated from the data nor tested and is consequently a matter of unverifiable expert knowledge.

9.2.5 Concluding remarks

For years, the necessity of reject inference at CACF and other institutions (as it seems from the large literature coverage this research area has had) has been a question of personal belief. Moreover, there even exist contradictory findings in this area.

By formalizing the reject inference problem, we were able to pinpoint in which cases the current scorecard construction methodology, using only financed clients' data, could be unsatisfactory: under a MNAR missingness mechanism and/or a misspecified model. We concluded that no current reject inference method could enhance the current scorecard construction methodology: only the Augmentation method (Strategy 4) and the Parcelling method (Strategy 6) had theoretical justifications but introduce other estimation procedures.

In light of those limitations, adding to the fact that implementing those methods is a non-negligible time-consuming task, we recommend credit modelers to work only with financed loans' data unless there is significant information available on either rejected applicants or on the acceptance mechanism $\boldsymbol{\phi}$ in the MNAR setting.

9.3 Model-embedded feature quantization

9.3.1 Motivation

The second main contribution is to embed the quantization step in the model parameters estimation process. The idea is that in many applications continuous predictors are discretized to produce a scorecard, *i.e.* a table assigning a grade to

an applicant in credit scoring depending on its predictors given in a certain interval. This can have two major advantages, (i) it makes the score easier to interpret in practice, (ii) it may allow more flexibility when the true relationship is non-linear. In this section, we also consider the issue of merging modalities of a categorical variable which limits the number of parameters to estimate thus potentially achieving a better bias-variance trade-off when considering a limited amount of data. The term quantization will stand for both discretization of continuous features as levels' grouping of categorical ones. This question of quantization in practice is often used as a prior pre-processing of the data based on chi-square related criteria. In practice, these pre-processing can require a lot of human time and may vary from a practitioner to another. Thus we have proposed a strategy in the scope of logistic regression to automate this quantization step, the R package `glmdisc` to automate this process is available on the CRAN. For sake of simplicity, I will only present the solution that we have proposed based on a continuous relaxation of the initial combinatorial problem without considering interactions between variables. An approach considering latent discretization variables and allowing to include interactions has also been developed but it is not presented here. For more details see Ehrhardt (2019).

9.3.2 Quantization as a combinatorial challenge

The quantization procedure consists in turning a d-dimensional raw vector of continuous and/or categorical features $\boldsymbol{x} = (x_1, \ldots, x_d)$ into a d-dimensional categorical vector *via* a component-wise mapping $\boldsymbol{q} = (\boldsymbol{q}_j)_1^d$:

$$\boldsymbol{q}(\boldsymbol{x}) = (\boldsymbol{q}_1(x_1), \ldots, \boldsymbol{q}_d(x_d)).$$

Each of the univariate quantizations $\boldsymbol{q}_j(x_j) = (q_{j,1}(x_j), \ldots, q_{j,m_j}(x_j))$ is a vector of m_j dummies:

$$q_{j,h}(x_j) = 1 \text{ if } x_j \in C_{j,h}, 0 \text{ otherwise, } 1 \leq h \leq m_j, \tag{9.5}$$

where m_j is an integer, denoting the number of intervals/groups to which x_j is mapped and the sets $C_{j,h}$ are defined with respect to each feature type as is described just below.

Raw continuous features If x_j is a continuous component of $\boldsymbol{x}$, quantization $\boldsymbol{q}_j$ has to perform a discretization of x_j and the $C_{j,h}$'s, $1 \leq h \leq m_j$, are contiguous intervals:

$$C_{j,h} = (c_{j,h-1}, c_{j,h}], \tag{9.6}$$

where $c_{j,1}, \ldots, c_{j,m_j-1}$ are increasing real numbers called cutpoints, $c_{j,0} = -\infty$, $c_{j,m_j} = \infty$. Discretization is visually exemplified on Figure 9.1.

Raw categorical features If x_j is a categorical component of $\boldsymbol{x}$, quantization $\boldsymbol{q}_j$ consists in grouping levels of x_j thus the $C_{j,h}$'s form a partition of the set $\{1, \dots, l_j\}$ s.t. $\bigcup_{h=1}^{m_j} C_{j,h} = \{1, \dots, l_j\}$ and $\forall h, h' \neq h$, $C_{j,h} \cap C_{j,h'} = \varnothing$. Note that it is assumed that there are no empty buckets, *i.e.* $\nexists j, h$ s.t. $C_{j,h} = \varnothing$. Grouping is visually exemplified in Figure 9.2.

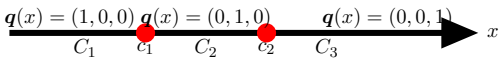


Figure 9.1: Quantization (discretization) of a continuous feature.

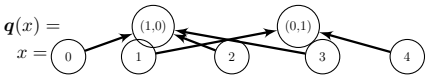


Figure 9.2: Quantization (factor levels merging) of categorical feature.

In both continuous and categorical cases, keep in mind that m_j is the dimension of $\boldsymbol{q}_j$. For notational convenience, the (global) order of the quantization $\boldsymbol{q}$ is set as

$$|\boldsymbol{q}| = \sum_{j=1}^{d} m_j.$$

The space where quantizations $\boldsymbol{q}$ live (resp. $\boldsymbol{q}_j$) will be denoted by $\boldsymbol{Q}_{\boldsymbol{m}}$ in the sequel (resp. $\boldsymbol{Q}_{j,m_j}$), when the number of levels $\boldsymbol{m} = (m_j)_1^d$ is fixed. Since it is not known, the full model space is $\boldsymbol{Q} = \cup_{\boldsymbol{m} \in \mathbb{N}_\star^d} \boldsymbol{Q}_{\boldsymbol{m}}$ where $\mathbb{N}_\star^d = (\mathbb{N} \setminus \{0\})^d$.
The space of quantization may be very large making it hard to investigate from a brute force search, moreover since dealing with piece-wise constant function, tools such are optimization on a continuous space are not available.

9.3.3 State of the art

The state of the art in quantization consists of optimizing a heuristic criterion, often totally unrelated (unsupervised methods) or at least explicitly (supervised methods) to prediction, and mostly univariate (each feature is quantized irrespective of other features' values).
Many algorithms have thus been designed and a review of approximately 200 discretization strategies, gathering both criteria and related algorithms, can be found in Ramírez-Gallego et al., 2016, preceded by other review articles such as Dougherty, Kohavi, and Sahami, 1995; H. Liu, Hussain, et al., 2002. They classify discretization methods by distinguishing, among other criteria and as said previously, unsupervised and supervised methods (y is used to discretize $\boldsymbol{x}$), for which model-specific (assumptions on the predictive model to be used after quantization) or model-free approaches are distinguished, univariate and multivariate methods (features $\boldsymbol{x}_{-\{j\}} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ may influence the quantization scheme of x_j) and other criteria.
For factor levels grouping, we found no such taxonomy, but some discretization methods, *e.g.* χ^2 independence test-based methods can be naturally extended to this type of quantization, which is for example what the CHAID algorithm, proposed by Kass, 1980 and applied to each categorical feature, relies on. A simple idea is

also to use Group LASSO (proposed by Meier, Van De Geer, and Bühlmann, 2008) which attempts to shrink to zero all coefficients of a categorical feature to avoid situations where a few levels enter the model, which is arguably less interpretable. Another idea would be to use Fused LASSO (proposed by Tibshirani et al., 2005), which seeks to shrink the pairwise absolute difference of selected coefficients and apply it to all pairs of levels: the levels for which the difference would be shrunk to zero would be grouped. A combination of both approaches would allow both selection and grouping.
For benchmarking purposes, and following results found in the taxonomy of Ramírez-Gallego et al., 2016, we used the MDLP discretization method (proposed by Fayyad and Irani, 1993), which is a popular supervised univariate discretization method, and we implemented an extension of the discretization method ChiMerge (proposed by Kerber, 1992) to categorical features, performing pairwise χ^2 independence tests rather than only pairs of contiguous intervals. Note that various refinements of ChiMerge have been proposed in the literature, Chi2 by H. Liu and Setiono, 1995, ConMerge by Wang and B. Liu, 1998, ModifiedChi2 by Tay and Shen, 2002, and ExtendedChi2 by Su and Hsu, 2005, which seek to correct for multiple hypothesis testing (see Shaffer, 1995 for an overview of this problem) and automize the choice of the confidence parameter α in the χ^2 tests, but adapting them to categorical features for benchmarking purposes would have been too time-consuming.

9.3.4 Proposed criterion

Focus is now given to logistic regression since it is a requirement in many industries, including Credit Scoring and in particular for CACF. Nevertheless, subsequent results apply to any other supervised classification model.

Logistic regression on quantized data Quantization is a widespread preprocessing step to perform a learning task consisting in predicting, say, a binary variable $y \in \{0, 1\}$, from a quantized predictor $\boldsymbol{q}(\boldsymbol{x})$, through, say, a parametric conditional distribution $p_{\boldsymbol{\theta}}(y|\boldsymbol{q}(\boldsymbol{x}))$ like logistic regression; the whole process can be visually represented as a dependence structure among $\boldsymbol{x}$, its quantization $\boldsymbol{q}(\boldsymbol{x})$ and the target y on Figure 9.3. Considering quantized data instead of raw data has a double benefit. First, the quantization order $|\boldsymbol{q}|$ acts as a tuning parameter for controlling the model's flexibility and thus the bias/variance trade-off of the estimate of the parameter $\boldsymbol{\theta}$ (or of its predictive accuracy) for a given dataset. This claim becomes clearer with the example of logistic regression we focus on, as a still very popular model for many practitioners:

$$\ln\left(\frac{p_{\boldsymbol{\theta}}(1|\boldsymbol{q}(\boldsymbol{x}))}{1-p_{\boldsymbol{\theta}}(1|\boldsymbol{q}(\boldsymbol{x}))}\right) = \theta_0 + \sum_{j=1}^{d} \boldsymbol{q}_j(x_j)'\boldsymbol{\theta}_j, \tag{9.7}$$

where $\boldsymbol{\theta} = (\theta_0, (\boldsymbol{\theta}_j)_1^d) \in \mathbb{R}^{|\boldsymbol{q}|+1}$ and $\boldsymbol{\theta}_j = (\theta_j^1, \ldots, \theta_j^{m_j})$ with $\theta_j^{m_j} = 0$, $1 \leq j \leq d$, for identifiability reasons. Second, at the practitioner level, the previous tuning of

$|\boldsymbol{q}|$ through each feature's quantization order m_j, especially when it is quite low, allows an easier interpretation of the most important predictor values involved in the predictive process. The log-likelihood

$$\ell_{\boldsymbol{q}}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{q}(\boldsymbol{x}_i)) \tag{9.8}$$

provides a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{q}}$ of $\boldsymbol{\theta}$ for a given quantization $\boldsymbol{q}$.

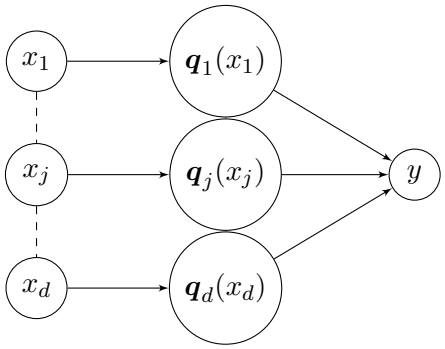


Figure 9.3: Dependence structure between x_j, $\boldsymbol{q}_j$ and y.

Quantization as a model selection problem As dicussed in the previous section, and emphasized in the literature review, quantization is often a preprocessing step; however, quantization can be embedded directly in the predictive model. Continuing our logistic example, a standard information criterion such as the BIC (proposed by **bic**) can be used to select the best quantization:

$$\hat{\boldsymbol{q}} = \underset{\boldsymbol{q} \in \boldsymbol{Q}}{\arg\min}\, \mathrm{BIC}(\hat{\boldsymbol{\theta}}_{\boldsymbol{q}}) = \underset{\boldsymbol{q} \in \boldsymbol{Q}}{\arg\min}\, -2\ell_{\boldsymbol{q}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{q}}; \mathcal{D}) + (|\boldsymbol{q}| - d + 1) \ln n. \tag{9.9}$$

where $|\boldsymbol{q}|-d+1$ is the dimension of the parameters space resulting from quantization $\boldsymbol{q}$. It allows to perform a trade-off between the quantization order $|\boldsymbol{q}|$ and the data fit $\ell_{\boldsymbol{q}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{q}}; \mathcal{D})$. Moreover it enjoys good theoretical properties such as consistency in many frameworks, *e.g.* multiple regression (Nishii, 1984), in the exponential family setting (Poskitt, 1987; Haughton, 1988), for selecting the order in Markov models , for selecting the number of components in a mixture (Keribin, 1998).

9.3.5 A relaxation of the optimization problem

In this section, we propose to relax the constraints on $\boldsymbol{q}_j$ to simplify the search of $\hat{\boldsymbol{q}}$. Indeed, the derivatives of $\boldsymbol{q}_j$ are zero almost everywhere and consequently, gradient descent cannot be directly applied to find an optimal quantization.

Smooth approximation of the quantization mapping A classical approach consists in replacing the binary functions $q_{j,h}$ (see Equation (9.5)) by smooth parametric ones with a simplex condition, namely with $\boldsymbol{\alpha}_j = (\boldsymbol{\alpha}_{j,1}, \ldots, \boldsymbol{\alpha}_{j,m_j})$:

$$\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot) = \left(q_{\boldsymbol{\alpha}_{j,h}}(\cdot)\right)_{h=1}^{m_j} \text{ with } \sum_{h=1}^{m_j} q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = 1 \text{ and } 0 \leq q_{\boldsymbol{\alpha}_{j,h}}(\cdot) \leq 1,$$

where functions $q_{\boldsymbol{\alpha}_{j,h}}(\cdot)$, properly defined hereafter for both continuous and categorical features, represent a fuzzy quantization in that, here, each level h is weighted by $q_{\boldsymbol{\alpha}_{j,h}}(\cdot)$ instead of being selected once and for all as in Equation (9.5). The resulting fuzzy quantization for all components depends on the global parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d)$ and is denoted by $\boldsymbol{q}_{\boldsymbol{\alpha}}(\cdot) = \left(\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot)\right)_{j=1}^{d}$.

For continuous features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha_{j,h}^0, \alpha_{j,h}^1) \in \mathbb{R}^2$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp(\alpha_{j,h}^0 + \alpha_{j,h}^1 \cdot)}{\sum_{g=1}^{m_j} \exp(\alpha_{j,g}^0 + \alpha_{j,g}^1 \cdot)} \tag{9.10}$$

where $\boldsymbol{\alpha}_{j,m_j}$ is set to $(0, 0)$ for identifiability reasons.

For categorical features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha_{j,h}(1), \ldots, \alpha_{j,h}(l_j)) \in \mathbb{R}^{l_j}$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp\left(\alpha_{j,h}(\cdot)\right)}{\sum_{h'=1}^{m_j} \exp\left(\alpha_{j,h'}(\cdot)\right)}$$

where l_j is the number of levels of the categorical feature x_j.

Parameter estimation With this new fuzzy quantization, the logistic regression for the predictive task is then expressed as

$$\ln\left(\frac{p_{\boldsymbol{\theta}}(1|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}))}{1 - p_{\boldsymbol{\theta}}(1|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}))}\right) = \theta_0 + \sum_{j=1}^{d} \boldsymbol{q}_{\boldsymbol{\alpha}_j}(x_j)' \boldsymbol{\theta}_j, \tag{9.11}$$

where $\boldsymbol{q}$ has been replaced by $\boldsymbol{q}_{\boldsymbol{\alpha}}$ from Equation (9.7). Note that as $\boldsymbol{q}_{\boldsymbol{\alpha}}$ is a sound approximation of $\boldsymbol{q}$ (see above), this logistic regression in $\boldsymbol{q}_{\boldsymbol{\alpha}}$ is consequently a good approximation of the logistic regression in $\boldsymbol{q}$ from Equation (9.7). The relevant log-likelihood is here

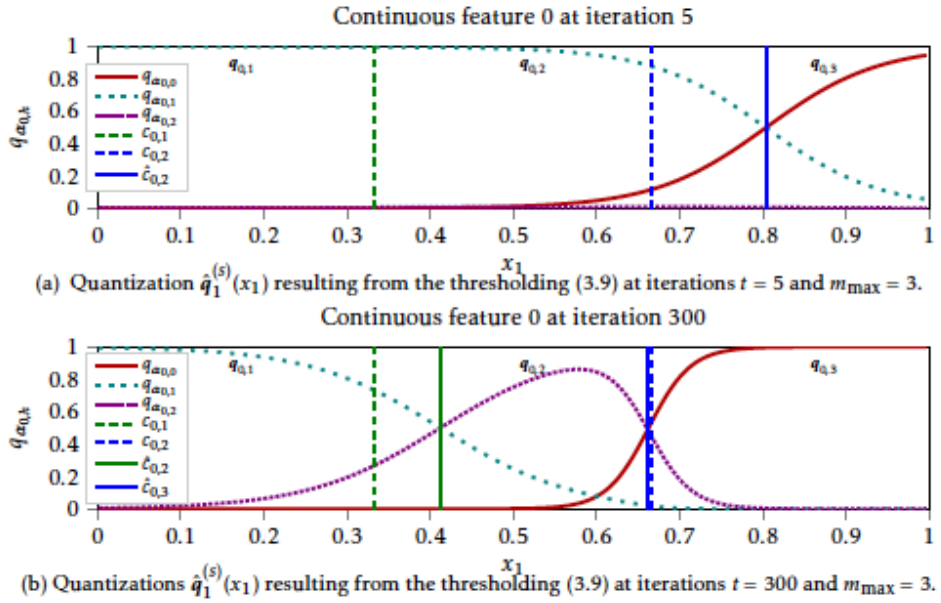
$$\ell_{\boldsymbol{q}_{\boldsymbol{\alpha}}}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)) \tag{9.12}$$

and can be used as a tractable substitute for (9.8) to solve the original optimization problem (9.9), where now both $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ have to be estimated, which is discussed in the next section.

Deducing quantizations from the relaxed problem We wish to maximize the log-likelihood (9.11) which would yield parameters $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$; these are consistent if the model is well-specified (*i.e.* there is a "true" quantization under classical regularity conditions). Denoting by A the space of $\boldsymbol{\alpha}$ and $\boldsymbol{Q}_A$ the space of $\boldsymbol{q}_{\boldsymbol{\alpha}}$, to "push" $\boldsymbol{Q}_A$ further into $\boldsymbol{Q}$, $\hat{\boldsymbol{q}}$ is deduced from a *maximum a posteriori* procedure applied to $\boldsymbol{q}_{\hat{\boldsymbol{\alpha}}}$:

$$\hat{q}_{j,h}(x_j) = 1 \text{ if } h = \underset{1 \leq h' \leq m_j}{\arg\max}\, q_{\hat{\boldsymbol{\alpha}}_{j,h'}}(x_j), 0 \text{ otherwise.} \tag{9.13}$$

If there are several levels h that satisfy (9.13), we simply take the level that corresponds to smaller values of x_j to be in accordance with the definition of $C_{j,h}$ in Equation (9.6). This *maximum a posteriori* principle will be exemplified in Figure 9.4 on simulated data. These approximations are justified by the following arguments.



(a) Quantization $\hat{\boldsymbol{q}}_1^{(s)}(x_1)$ resulting from the thresholding (3.9) at iterations $t = 5$ and $m_{\max} = 3$.

(b) Quantizations $\hat{\boldsymbol{q}}_1^{(s)}(x_1)$ resulting from the thresholding (3.9) at iterations $t = 300$ and $m_{\max} = 3$.

Figure 9.4: Principle for deducing quantization from the relaxed problem.

Rationale From a deterministic point of view, we have $\boldsymbol{Q} \subset \boldsymbol{Q}_A$: First, the *maximum a posteriori* step (9.13) produces contiguous intervals (*i.e.* there exists $C_{j,h}$; $1 \leq j \leq d$, $1 \leq h \leq m_j$, s.t. $\hat{\boldsymbol{q}}$ can be written as in Equation (9.5)) (see Samé et al., 2011). Second, in the continuous case, the higher $\alpha^1_{j,h}$, the less smooth the transition from one quantization h to its "neighbor" $h+1$, whereas $\dfrac{\alpha^0_{j,h}}{\alpha^1_{j,h}}$ controls the point in $\mathbb{R}$ where the transition occurs (see Chamroukhi et al., 2009). Concerning the categorical case, the rationale is even simpler as $q_{\lambda\boldsymbol{\alpha}_{j,h}}(x_j) \to 1$ if $h = \arg\max_{h'} q_{\alpha_{j,h'}}(x_j)$

as $\lambda \to +\infty$ (see Reverdy and Leonard, 2016).
From a statistical point of view, provided the model is well-specified, *i.e.*:

$$\exists \boldsymbol{q}^\star, \boldsymbol{\theta}^\star, \forall \boldsymbol{x}, y,\ p(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}^\star}(y|\boldsymbol{q}^\star(\boldsymbol{x})); \tag{9.14}$$

and under standard regularity conditions and with a suitable estimation procedure (see later for the proposed estimation procedure), the maximum likelihood framework would ensure the consistency of $(\boldsymbol{q}_{\hat{\boldsymbol{\alpha}}}, \hat{\boldsymbol{\theta}})$ towards $(\boldsymbol{q}^\star, \boldsymbol{\theta}^\star)$ if $\boldsymbol{\alpha}^\star$ s.t. $\boldsymbol{q}_{\boldsymbol{\alpha}^\star} = \boldsymbol{q}^\star$ was an interior point of the parameter space A. However, as emphasized in the previous paragraph, "$\boldsymbol{\alpha}^\star = +\infty$" such that the maximum likelihood parameter is on the edge of the parameter space which hinders asymptotic properties (*e.g.* normality) in some settings (see Self and Liang, 1987), but not "convergence" on which we focus here.
For the continuous case, the intuition is that the only consistent solution is the convergence of the smooth approximation to a binary function. We did not investigate this issue further since numerical experiments showed consistency: from an empirical point of view, we see in Figure 9.4, that the smooth approximation $\boldsymbol{q}_{\hat{\boldsymbol{\alpha}}}$ converges towards "hard" quantizations[1] $\boldsymbol{q}$.
However, and as is usual, the log-likelihood $\ell_{\boldsymbol{q}_{\boldsymbol{\alpha}}}(\boldsymbol{\theta}, \mathcal{D})$ cannot be directly maximized w.r.t. $(\boldsymbol{\alpha}, \boldsymbol{\theta})$, so that we need an iterative procedure. To this end, the next section introduces a neural network of suitable architecture.

9.3.6 A neural network-based estimation strategy

Neural network architecture To estimate parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ in the model (9.11), a particular neural network architecture can be used. We shall insist that this network is only a way to use common deep learning frameworks, namely Tensorflow (see Martín Abadi et al., 2015) through the high-level API Keras (see Chollet et al., 2015) instead of building a gradient descent algorithm from scratch to optimize (9.12). The most obvious part is the output layer that must produce $p_{\boldsymbol{\theta}}(1|\boldsymbol{q}_{\boldsymbol{\alpha}}(\boldsymbol{x}))$ which is equivalent to a densely connected layer with a sigmoid activation (the reciprocal function of logit).
For a continuous feature x_j of $\boldsymbol{x}$, the combined use of m_j neurons including affine transformations and softmax activation obviously yields $\boldsymbol{q}_{\boldsymbol{\alpha}_j}(x_j)$. Similarly, an input categorical feature x_j with l_j levels is equivalent to l_j binary input neurons (presence or absence of the factor level). These l_j neurons are densely connected to m_j neurons without any bias term and a softmax activation. The softmax outputs are next aggregated via the summation in model (9.11), say $\Sigma_{\boldsymbol{\theta}}$ for short, and then the sigmoid function σ gives the final output. All in all, the proposed model is straightforward to optimize with a simple neural network, as shown in Figure 9.5.

Stochastic gradient descent as a quantization provider By relying on stochastic gradient descent, the smoothed likelihood (9.12) can be maximized over $(\boldsymbol{\alpha}, \boldsymbol{\theta})$.

[1] Up to a permutation on the labels $h = 1 \dots m_j$ to recover the ordering in $C_{j,h}$ (see Equation (9.6)).

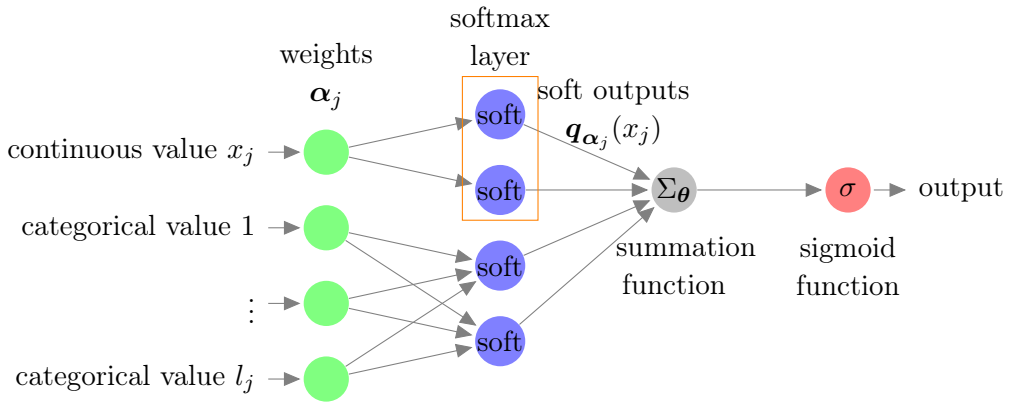


Figure 9.5: Proposed shallow architecture to maximize (9.12).

Due to its convergence properties (see Bottou, 2010), the results should be close to the maximizers of the original likelihood (9.8) if the model is well-specified, when there is a true underlying quantization. However, in the misspecified model case, there is no such guarantee. Therefore, to be more conservative, we evaluate at each training epoch (s) the quantization $\hat{\boldsymbol{q}}^{(s)}$ resulting from the *maximum a posteriori* procedure explicited in Equation (9.13), then classically estimate the logistic regression parameter *via* maximum likelihood, as done in Equation (9.8):

$$\hat{\boldsymbol{\theta}}^{(s)} = \arg\max_{\boldsymbol{\theta}} \ell_{\hat{q}^{(s)}}(\boldsymbol{\theta}; \mathcal{D}) \tag{9.15}$$

and the resulting $\mathrm{BIC}(\hat{\boldsymbol{\theta}}^{(s)})$ as in (9.9). If S is a given maximum number of iterations of the stochastic gradient descent algorithm, the quantization retained at the end is then determined by the optimal epoch

$$s^\star = \arg\min_{s\in\{1,\dots,S\}} \mathrm{BIC}(\hat{\boldsymbol{\theta}}^{(s)}). \tag{9.16}$$

S can be seen as a computational budget: contrary to classical early stopping rules (*e.g.* based on validation loss) used in neural network fitting, this network only acts as a stochastic quantization provider for (9.16) which will naturally prevent overfitting. We reiterate that, in (9.16), the BIC can be swapped for the user's favourite model choice criterion.

Choosing an appropriate number of levels Concerning now the number of intervals or factor levels $\boldsymbol{m} = (m_j)_1^d$, they have also to be estimated since in practice they are unknown. Looping over all candidates $\boldsymbol{m}$ is intractable. But in practice, by relying on the *maximum a posteriori* procedure developed in Equation (9.13), a lot of unseen factor levels might be dropped. Indeed, for a given level h, all training observations $x_{i,j}$ in $\mathcal{T}$ and all other levels h', if $q_{\boldsymbol{\alpha}_{j,h}}(x_{i,j}) < q_{\boldsymbol{\alpha}_{j,h'}}(x_{i,j})$, then the level h "vanishes". In practice, we recommend to start with a user-chosen $\boldsymbol{m} = \boldsymbol{m}_{\max}$, then the proposed approach is able to explore small values of $\boldsymbol{m}$ and to select a value $\hat{\boldsymbol{m}}$ drastically smaller than $\boldsymbol{m}_{\max}$.

9.3.7 Conclusion

Numerical results of the approach are not presented here, see Ehrhardt (2019) for more details. The main conclusion is that it produces better results than state of the art quantization methods such as MDLP. Such an approach is of great interest in practice since it can drastically reduce the human time needed to develop the score while producing optimized quantization for the final used predictive model. The presentation here has focused on the quantization in the logistic regression framework, however, it is not limited to this framework, and the simple steps proposed could be embedded in many other models.

9.4 Conclusion and perspectives

In this chapter, I have presented some of the main contributions of Adrien Erhardt's thesis. Even if motivated by the industrial context of credit scoring, with some limitations resulting from the use of logistic regression, there are still many developments needed. One important challenge is to reduce the gap between the available data and the final goal. Thus, each development such as the embedded quantization approach may have a great impact on practice by limiting arbitrary pre-processing.
Some great perspective in credit scoring would also be to adapt models to new types of data available such as navigation of the website. Such functional data can be challenging covariates to deal with. Among several possible solutions, clustering/dimension reduction for such navigation data can be considered. Moreover, in an embedded framework one could imagine embedding this clustering in the predictive model in such a way that it can be linked with predictive clustering issues.

Bibliography

Banasik, J. and Crook, J. (2007). "Reject inference, augmentation, and sample selection". In: *European Journal of Operational Research* 183.3, pp. 1582–1594. URL: http://www.sciencedirect.com/science/article/pii/S0377221706011969 (visited on 08/25/2016).

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010*. Springer, pp. 177–186.

Celeux, G. and Govaert, G. (1992). "A classification EM algorithm for clustering and two stochastic versions". In: *Computational statistics & Data analysis* 14.3, pp. 315–332.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). "A regression model with a hidden logistic process for feature extraction from time series". In: *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, pp. 489–496.

Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. 1st. The MIT Press. ISBN: 0262514125, 9780262514125.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). "Supervised and unsupervised discretization of continuous features". In: *Machine Learning Proceedings 1995*. Elsevier, pp. 194–202.

Ehrhardt, A. (2019). "Formalisation et étude de problématiques de scoring en risque de crédit: inférence de rejet, discrétisation de variables et interactions, arbres de régression logistique". PhD thesis. Lille 1.

Ehrhardt, A., Biernacki, C., Vandewalle, V., and Heinrich, P. (2019). "Feature quantization for parsimonious and interpretable predictive models". working paper or preprint. URL: https://hal.archives-ouvertes.fr/hal-01949135.

Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., and Beben, S. (2017). "Réintégration des refusés en Credit Scoring". In: *49e Journées de Statistique*. Avignon , France. URL: https://hal.archives-ouvertes.fr/hal-01653767.

Ehrhardt, A., Vandewalle, V., Biernacki, C., and Heinrich, P. (2018). "Supervised multivariate discretization and levels merging for logistic regression". In: *23rd International Conference on Computational Statistics*. Iasi, Romania. URL: https://hal.archives-ouvertes.fr/hal-01949128.

Fayyad, U. and Irani, K. (1993). "Multi-interval discretization of continuous-valued attributes for classification learning". In: *13th International Joint Conference on Artificial Intelligence*, pp. 1022–1029.

Feelders, A. (2000). "Credit scoring and reject inference with mixture models". In: *International Journal of Intelligent Systems in Accounting, Finance & Manage-*

ment 9.1, pp. 1–8. URL: http://www.ingentaconnect.com/content/jws/isaf/2000/00000009/00000001/art00177 (visited on 08/25/2016).

Guizani, A., Souissi, B., Ammou, S. B., and Saporta, G. (2013). "Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit". In: *45 emes Journées de statistique*. URL: http://cedric.cnam.fr/fichiers/art_2753.pdf (visited on 08/25/2016).

Haughton, D. M. A. (1988). "On the Choice of a Model to Fit Data from an Exponential Family". In: *The Annals of Statistics* 16.1, pp. 342–355. ISSN: 00905364. URL: http://www.jstor.org/stable/2241441.

Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data". In: *Applied statistics*, pp. 119–127.

Kerber, R. (1992). "Chimerge: Discretization of numeric attributes". In: *Proceedings of the tenth national conference on Artificial intelligence*. Aaai Press, pp. 123–128.

Keribin, C. (1998). "Consistent estimate of the order of mixture models". In: *Comptes Rendus De L Academie Des Sciences Serie I-Mathematique* 326.2, pp. 243–248.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002). "Discretization: An enabling technique". In: *Data mining and knowledge discovery* 6.4, pp. 393–423.

Liu, H. and Setiono, R. (1995). "Chi2: Feature selection and discretization of numeric attributes". In: *Tools with artificial intelligence, 1995. proceedings., seventh international conference on*. IEEE, pp. 388–391.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). "The group lasso for logistic regression". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 53–71.

Nguyen, H. T. (2016). *Reject inference in application scorecards: evidence from France*. Tech. rep. University of Paris West-Nanterre la Défense, EconomiX. URL: http://economix.fr/pdf/dt/2016/WP_EcoX_2016-10.pdf (visited on 08/25/2016).

Nishii, R. (1984). "Asymptotic properties of criteria for selection of variables in multiple regression". In: *The Annals of Statistics*, pp. 758–765.

Poskitt, D. S. (1987). "Precision, Complexity and Bayesian Model Determination". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 49.2, pp. 199–208. ISSN: 00359246. URL: http://www.jstor.org/stable/2345420.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2016). "Data discretization: taxonomy and big data challenge". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6.1, pp. 5–21.

Reverdy, P. and Leonard, N. E. (2016). "Parameter estimation in softmax decision-making models with linear objective functions". In: *IEEE Transactions on Automation Science and Engineering* 13.1, pp. 54–67.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). "Model-based clustering and segmentation of time series with changes in regime". In: *Advances in Data Analysis and Classification* 5.4, pp. 301–321.

Self, S. G. and Liang, K.-Y. (1987). "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions". In: *Journal of the American Statistical Association* 82.398, pp. 605–610. ISSN: 01621459. URL: http://www.jstor.org/stable/2289471.

Shaffer, J. P. (1995). "Multiple hypothesis testing". In: *Annual review of psychology* 46.1, pp. 561–584.

Su, C.-T. and Hsu, J.-H. (2005). "An extended chi2 algorithm for discretization of real value attributes". In: *IEEE transactions on knowledge and data engineering* 17.3, pp. 437–441.

Tay, F. E. and Shen, L. (2002). "A modified chi2 algorithm for discretization". In: *IEEE Transactions on Knowledge & Data Engineering* 3, pp. 666–670.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 91–108.

Viennet, E., Soulié, F. F., and Rognier, B. (2006). "Evaluation de Techniques de Traitement des Refusés pour l'Octroi de Crédit". In: *arXiv preprint cs/0607048*. URL: http://arxiv.org/abs/cs/0607048 (visited on 08/25/2016).

Wang, K. and Liu, B. (1998). "Concurrent discretization of multiple attributes". In: *Pacific Rim International Conference on Artificial Intelligence*. Springer, pp. 250–259.

Zadrozny, B. (2004). "Learning and evaluating classifiers under sample selection bias". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 114.

Part III

Research Project

Chapter 10

Research project

Contents

10.1 Projects in the scope of model-based clustering 175
10.1.1 Multiple partition clustering dissemination 175
10.1.2 Predictive clustering 176
10.1.3 Clustering of recurrent event data and integration of covariates 176
10.2 Perspectives motivated by applications 177
10.2.1 Application to medical data 177
10.2.2 Application to the industry/retail 178
10.3 Reducing the gap between data and end-use 179

Based on my previous work, I will continue to work on model-based clustering, from both model proposal/inference point of view, but also on more generic problems such as the degeneracy in mixtures. I will continue to feed my methodological research by applied problems especially in medicine but also in the industry/retail. I would finally like to investigate further global approaches starting from the data and going all the way to the end-use, such a global approach needing to be carried out in the scope of collaborative projects.

10.1 Projects in the scope of model-based clustering

10.1.1 Multiple partition clustering dissemination

Among my recent research work, I feel that multiple partition clustering is one of the most promising topics. The model proposed in Marbac and Vandewalle, 2019 while being very simple, allows to deal with mixed variables (thanks to conditional independence assumption), missing values, and gives some rich interpretation of the data at hand by grouping variables according to their clustering behavior with respect to the individuals. Using such a model could be very valuable for the practitioner, he can apply to his dataset without any prior variable selection and analyze the resulting grouping of variables into blocks and the clustering which results from each group of variables. The proposed model is only available as an R package on R-forge (`http://r-forge.r-project.org/projects/mgmm/`), but I think that there is potential for further software integration. Moreover, there is some need for user interaction with the output of the model, this can be by developing a user-friendly interface this interface can for instance visualizing results as presented in Chapter 6.

The main lock for using the multiple partition model is that it needs the choice of the number of blocks and the number of clusters in each block. Thus I project to work on strategies to avoid exhaustive search in a huge model collection. This could be done by searching efficiently for the number of clusters in each block inside the algorithm. For instance, Bayesian approaches with some efficient Gibbs sampler algorithms could be investigated. Once some accurate strategy can be found to solve this problem efficiently, the question of additional software development and dissemination will be investigated.
From a practical point of view, k-means is still the most used clustering method (Jain, 2010). Thus, it would be interesting in practice to derive k-means-like extensions of our multiple partition model. Applying this idea is straightforward by using the interpretation of k-means as a particular Gaussian mixture estimated using the Classification EM (CEM) algorithm (Celeux and Govaert, 1995), however, this would require further investigation to select accurate numbers of blocks and cluster per block and define relevant indicator related to the obtained results. This could lead to a popular "Multiple k-means" algorithm.

10.1.2 Predictive clustering

I am currently working with Matthieu Marbac, Christophe Biernacki, and Mohammed Sedki on the question of predictive clustering. This starts from the idea that in epidemiology, for instance, clustering is often performed based on some variables (eating habits), then the obtained clusters (good or bad eating habits) are used as a feature variable in a predictive model (predict obesity for instance). From a statistical point of view, the question is first to study the possible theoretical guarantees of this two steps approach. The second question is to produce clustering potentially more related to the prediction. Using mixture models this question can be easily answered from a practical point of view since the clustering and the predictive model can be estimated simultaneously thus leading to a one-step approach. However, even if this solution can produce a clustering more related to the outcome, it is also more prone to over-fitting thus asking the question of finding a good trade-off between both points of view.
This question of predictive clustering could be linked with the question of multiple partition clustering. Indeed, one could see multiple partition as providing several potentially useful summaries of the variables. Thus one or several of these summaries could be related to the outcome variable, multiple partition resulting as an unsupervised dimension reduction step.

10.1.3 Clustering of recurrent event data and integration of covariates

I have work in progress with Génia Babykina (Assistant Professor in statistics) and Jean-Baptiste Beuscart (Professor and hospital practitioner), both from METRICS teams, on the analysis of recurrent event data. We are interested in the

study of re-hospitalization data, the goal is to cluster patients according to their re-hospitalization profile. Thus mixture models will be a particularly useful tool for this issue.
In this scope, I would also like to consider the question of using covariates in clustering. More precisely, when performing clustering in practice this can lead to obtaining some obvious clustering, for instance, old people have more recurrent events than young people. In this case, the added value of clustering compared to expert knowledge could be poor. Thus one possibility to bring new assumptions by clustering is to include these obvious variables as covariates in the model (Leisch, 2004; Chiquet, Mariadassou, and Robin, 2018), then proposing a new explanation of the remaining heterogeneity. This cycle could continue until some remaining heterogeneity is found.

10.2 Perspectives motivated by applications

10.2.1 Application to medical data

10.2.1.1 Taking into account temporal structure in the statistical analysis of high-throughput proteomic data

As part of a collaboration with Guillemette Marot, Florence Pinet, and Christophe Bauters over the past few years, we have been interested in the use of protein concentrations in plasma to predict various clinical events such as death or ventricular remodeling following myocardial infarction. This work has led to a published article (Cuvelliez et al., 2019), in which we have used LASSO-type approaches to predict patient death and have identified 6 markers that are possibly predictive of death.
As part of this collaboration, since September 2019, I am co-supervising Wilfried Heyse's thesis with Christophe Bauters and Guillemette Marot. Wilfried Heyse's thesis focuses on temporal structures in the statistical analysis of high throughput proteomic data. In this context, we are currently developing models that allow grouping together proteins with similar temporal evolution over patients. The objective is then to correlate these different groups of proteins with a clinical phenomenon such as ventricular remodeling. In addition to the obvious interest it represents from a clinical point of view, this problem also highlights a statistical lock which consists in the clustering of multivariate high dimensional temporal data. In this scope, we will investigate, among others, the mixture of mixed models which will be particularly well suited for such type of data (James and Sugar, 2003).

10.2.1.2 Patient path at hospital

Working on the ClinMine project has revealed to me the real challenge of studying the patient path at hospital. It cannot be summarized by considering only one categorical variable over time, but also needs to consider a huge number of variables varying according to time. Thus I will continue to work on this issue with other members of the MODAL and METRICS teams. One of the main goals will be to

define what is a patient path and to develop statistical models well suited for the study of these data. This falls within the scope of multivariate functional data, a kind of data on which I have already worked with Cristian Preda (MODAL) and Sophie Dabo (MODAL).
For being useful such an approach could be built in collaboration with physicists of the METRICS team to produce a relevant and useful analysis of the available data, starting with the study of some specific pathology. It would also be interesting to introduce a cost-sensitive way to link the statistical model with the decision-making process as discussed in Chapter 7.

10.2.1.3 Linking different kinds of omics data through a model-based clustering approach

I have work in progress with Guillemette Marot and Camille Ternynck on linking differents kinds of omics data such as microarray (continuous) and RNAseq (count) data for instance. Preliminary work has been presented in a conference (Vandewalle, Ternynck, and Marot, 2019). The idea is to link microarray and RNAseq variables through a mixture model, assuming that measures from the same gene whatever the considered technology (microarray or RNAseq) come from the same cluster. This is a work in progress, one important fact resulting from this work is that one kind of variables tends to dominate the other when performing the clustering, thus also raising more theoretical issues.

10.2.2 Application to the industry/retail

10.2.2.1 Sales predictions by grouping low turnover products

The Ph.D. thesis of Axel Potier in collaboration with the ADEO company (parent company of Leroy merlin, specialized in do-it-yourself equipment) will start in the next weeks. I will co-supervise the thesis of Axel Potier with Christophe Biernacki and Matthieu Marbac. The thesis aims at proposing a specific estimation of sales forecasts for references with low turnover. The proposed originality is essentially based on the estimation of sales of groups of products, this grouping being done either by grouping different individual products or by grouping identical products but from different stores. In both cases, the estimation of sales volume becomes mechanically more accurate and can be based on standard methods that have already proven themselves for “classic” sales volumes.
From a mathematical point of view, it consists of constructing an oriented graph based on sales data, where the nodes are the products and the probabilities of substitutability are the edges, the substitutability groups can be obtained by unsupervised classification of large graphs oriented by probabilistic approaches of the Stochastic Block Model type (Nowicki and Snijders, 2001; Côme and Latouche, 2015). Finally, it will remain to define rigorously, as a by-product of the groups obtained and of the internal substitutability forces (the edge probabilities), the use of this classified

graph to improve the predictions of effective sales by product groups and also the ideal distribution of individual products within each group.
The other question is to group identical products and estimate their global sale over several stores. The objective is then to position the right stock level of a product with low turnover in a warehouse that can serve several stores with given flexibility. The warehouses are located at different geographical granularities, for example at the regional or national level, the so-called local level being the store itself. It is therefore a question of optimizing the supply chain associated with a given product. The objective is then to optimize in the end a trade-off between the probability of a customer to accept a waiting time for a given product and the accuracy of the stock for the company.

10.2.2.2 Data driven trajectory optimization

In Chapter 8, I have presented the work in the scope of the PERF-AI project. This raises perspectives for data-driven trajectory optimization applied to other fields than aviation. For instance, we are investigating applications to sailing where similar tools could be used. Considering reference trajectories coming from data could help in many setting to propose relevant optimized trajectories.

10.3 Reducing the gap between data and end-use

A question that goes beyond the result of the model (for instance a produced clustering) is the question of its end-use in human organization. This is perhaps not directly to the statistician to address this question, but it seems to me that approaches that try to see the problem globally, as statistical models partially do, can help to reduce the gap between data and end-use. This can be for instance by integrating some pre-processing in the model (for instance in Chapter 9), it also possible by including information from the problem at hand inside the model for a relevant cost-sensitive evaluation (as discussed in Chapter 7). For instance, mixture models can be a good building block, that can be embedded in a more global approach.
In IFCS 2019 meeting, I had the chance to assist at a presentation of David Hand where he emphasizes that statistics and in particularly clustering have consequences. He was, in particular, talking about the AUC criterion in prediction that he presented as nonsensical because it is an average over many possible decision thresholds whereas when used, the score only makes use of one particular threshold which is chosen according to the misclassifications costs which are *properties of the problem*, and not of the classifier (Hand, 2009). Another important point of his talk was that classification (both supervised and unsupervised) has consequences in daily lives. This is obvious in credit scoring where the acceptance of the credit depends on the score, but such consequence also exists in unsupervised learning. However, the potential consequences of some particular clustering can be difficult to assess in advance.

More precisely in unsupervised learning, groups are made based on the least possible prior information, only generated based on a set of variables hopefully related to the problem for which clustering is desired. Recently when discussing with a physicist working about some disease, he explained the way by which patients are clustered for instance in two or six clusters, and how it can have consequences on their treatment or the priorities on treatments developed by the pharmacological industry. In this case, such basic hierarchical clustering helped to identify six subgroups where before only two groups were considered by the state of the art. This question is highly linked with personalized medicine where ideally the treatment should be personalized for each patient. However, due to the clinical trial setting, it is not possible to consider as many possible treatments as patients, and it is necessary to reduce the possible number of clusters with respect to the capacity of management of clinical trials.
Thus clustering which is often described as a pre-processing step in practice has consequences and it is often hard to come back from a clustering that has been admitted as a gold standard. It allows building a simplified and understandable version of the world, as in the medical example each group can benefit from its personalized treatment. One "Graal" would be to chose such clusters that enable the best treatment for each homogeneous group of patients, however bringing such continuity between these steps is often impossible and we are limited to deal with sub-optimal steps. At this stage, it is important to notice that *any development that could help in reducing the gap between the data and the final use can have a huge impact on practice*. This remark can be linked with Chapter 7, where we evaluate the number of needed subjects in a cost-sensitive way. Thus collaborations with physicists and experts in the health economy would be very useful to reduce this gap to improve patient care.

Bibliography

Celeux, G. and Govaert, G. (1995). "Gaussian parsimonious clustering models". In: *Pattern recognition* 28.5, pp. 781–793.

Chiquet, J., Mariadassou, M., and Robin, S. (2018). "Variational inference for probabilistic Poisson PCA". In: *Ann. Appl. Stat.* 12.4, pp. 2674–2698. DOI: 10.1214/18-AOAS1177. URL: https://doi.org/10.1214/18-AOAS1177.

Côme, E. and Latouche, P. (2015). "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood". In: *Statistical Modelling* 15.6, pp. 564–589.

Cuvelliez, M., Vandewalle, V., Brunin, M., Beseme, O., Hulot, A., Groote, P. de, Amouyel, P., Bauters, C., Marot, G., and Pinet, F. (2019). "Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction". In: *Scientific reports* 9.1, pp. 1–12.

Hand, D. J. (2009). "Measuring classifier performance: a coherent alternative to the area under the ROC curve". In: *Machine learning* 77.1, pp. 103–123.

Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2009.09.011. URL: http://www.sciencedirect.com/science/article/pii/S0167865509002323.

James, G. M. and Sugar, C. A. (2003). "Clustering for sparsely sampled functional data". In: *Journal of the American Statistical Association* 98.462, pp. 397–408.

Leisch, F. (2004). "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R". In: *Journal of Statistical Software, Articles* 11.8, pp. 1–18. ISSN: 1548-7660. DOI: 10.18637/jss.v011.i08. URL: https://www.jstatsoft.org/v011/i08.

Marbac, M. and Vandewalle, V. (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179.

Nowicki, K. and Snijders, T. A. B. (2001). "Estimation and prediction for stochastic blockstructures". In: *Journal of the American statistical association* 96.455, pp. 1077–1087.

Vandewalle, V., Ternynck, C., and Marot, G. (2019). "Linking different kinds of Omics data through a model-based clustering approach". In: *IFCS 2019, Thessaloniki, Greece*.

Appendices

Appendix A

Publications and scientific activities

Contents

A.1 Publications . . . 185
A.1.1 Articles published in peer-reviewed journals . . . 185
A.1.2 Book chapter . . . 187
A.1.3 Talks in conferences . . . 187
A.2 Synthesis of scientific activities . . . 190
A.2.1 Participation in research projects . . . 190
A.2.2 Doctoral and scientific supervision . . . 191
A.2.3 Participation in thesis juries . . . 192
A.2.4 Participation in selection committees . . . 192
A.2.5 Contracts with companies . . . 193
A.3 Scientific responsibilities . . . 193
A.3.1 Participation in scientific societies . . . 193
A.3.2 Organization of national and international conferences . . . 193
A.3.3 Reviewing activities . . . 194

A.1 Publications

A.1.1 Articles published in peer-reviewed journals

Post-thesis articles

1. V. Vandewalle (2020). "Multi-Partitions Subspace Clustering". In: *Mathematics* 8.4, p. 597

2. C. Biernacki, M. Marbac, and V. Vandewalle (2020). "Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering". In: *Journal of Classification*. DOI: 10.1007/s00357-020-09369-y. URL: https://doi.org/10.1007/s00357-020-09369-y

3. F. Dewez, B. Guedj, and V. Vandewalle (2020). "From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning". In: *Data-Centric Engineering* 1, e11. DOI: 10.1017/dce.2020.12

4. V. Vandewalle, A. Caron, C. Delettrez, R. Périchon, S. Pelayo, A. Duhamel, and B. Dervaux (2020). "Estimating the number of usability problems affecting medical devices: modelling the discovery matrix". In: *BMC Medical Research Methodology* 20.234

5. M. Cuvelliez, V. Vandewalle, M. Brunin, O. Beseme, A. Hulot, P. de Groote, P. Amouyel, C. Bauters, G. Marot, and F. Pinet (2019). "Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction". In: *Scientific reports* 9.1, pp. 1–12

6. M. Marbac and V. Vandewalle (2019). "A tractable multi-partitions clustering". In: *Computational Statistics & Data Analysis* 132, pp. 167–179

7. C. Dhaenens, J. Jacques, V. Vandewalle, M. Vandromme, E. Chazard, C. Preda, A. Amarioarei, P. Chaiwuttisak, C. Cozma, G. Ficheur, et al. (2018). "ClinMine: Optimizing the management of patients in hospital". In: *IRBM* 39.2, pp. 83–92

8. M. Marbac, C. Biernacki, and V. Vandewalle (2017). "Model-based clustering of Gaussian copulas for mixed data". In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656

9. M. Marbac, C. Biernacki, and V. Vandewalle (2016). "Latent class model with conditional dependency per modes to cluster categorical data". In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207

10. M. Marbac, C. Biernacki, and V. Vandewalle (2015b). "Model-based clustering for conditionally correlated categorical data". In: *Journal of Classification* 32.2, pp. 145–175

11. E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki (2014). "Mixture of Gaussians for distance estimation with missing data". In: *Neurocomputing* 131, pp. 32–42

Articles from thesis

1. V. Vandewalle (2009a). "Les modèles de mélange, un outil utile pour la classification semi-supervisée." In: *Monde des Util. Anal. Données* 40, pp. 121–145

2. V. Vandewalle, C. Biernacki, G. Celeux, and G. Govaert (2013). "A predictive deviance criterion for selecting a generative model in semi-supervised classification". In: *Computational Statistics & Data Analysis* 64, pp. 220–236

Pre-thesis article

1. S. Robin, S. Schbath, and V. Vandewalle (2007). "Statistical tests to compare motif count exceptionalities". In: *BMC bioinformatics* 8.1, p. 84

A.1.2 Book chapter

1. V. Vandewalle, C. Preda, and S. Dabo (2020). "Clustering spatial functional data". In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons

A.1.3 Talks in conferences

Conferences with proceedings

1. C. Biernacki and V. Vandewalle (2011b). "Label switching in mixtures". In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401

Invited talks

1. V. Vandewalle and M. Marbac (2018). "A tractable multi-partitions clustering". COMPSTAT 2018 - 23rd International Conference on Computational Statistics, Iasi, Romania. (invité)

2. V. Vandewalle, T. Mottet, and M. Marbac (2017). "Model-based variable clustering". CMStatistics/ERCIM 2017 - 10th International Conference of the ERCIM WG on Computational and Methodological Statistics, London, United Kingdom. (invité)

3. V. Vandewalle (2017). "Simultaneous dimension reduction and multi-objective clustering". IFCS Meeting, Tokyo, August 8th (invité)

4. V. Vandewalle and C. Biernacki (2017a). "Dealing with missing data through mixture models". *154th ICB Seminar on "Statistics and clinical practice"* Warsaw May 11 (invité)

5. V. Vandewalle and C. Biernacki (2017b). "Survival analysis with complex covariates: a model-based clustering preprocessing step". IEEE PHM, Dallas June 19th (invité)

6. V. Vandewalle and C. Preda (2017). "Clustering categorical functional data: Application to medical discharge letters". 20th conference of the society of probability and statistics of Roumania, Brasov (Roumania), April 28 (invité)

7. V. Vandewalle (2016). "Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis". CMStatistics 2016 Sevilla, Spain. (invité)

8. V. Vandewalle and C. Biernacki (2015). "An efficient SEM algorithm for Gaussian Mixtures with missing data". In: *8th International Conference of the ERCIM WG on Computational and Methodological Statistics*. Londres, United Kingdom. URL: https://hal.inria.fr/hal-01242588

9. M. Marbac, C. Biernacki, and V. Vandewalle (2015c). "Model-based clustering of categorical data by relaxing conditional independence." Classification Society Meeting, Hamilton, Ontario, Canada (invité)

Others talks

1. V. Vandewalle, C. Ternynck, and G. Marot (2019). "Linking different kinds of Omics data through a model-based clustering approach". In: *IFCS 2019, Thessaloniki, Greece*

2. V. Vandewalle, C. Preda, and S. Dabo (2018). "Clustering spatial functional data". ERCIM 2018, Pise, Italy

3. M. Marbac, C. Biernacki, M. Sedki, and V. Vandewalle (2018). "A targeted multi-partitions clustering". In: *The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)*. Pise, Italy. URL: https://hal.archives-ouvertes.fr/hal-01949111

4. C. Biernacki, V. Vandewalle, and M. Marbac (2018). "Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering". 23rd International Conference on Computational Statistics, Iasi, Romania

5. A. Ehrhardt, V. Vandewalle, C. Biernacki, and P. Heinrich (2018). "Supervised multivariate discretization and levels merging for logistic regression". In: *23rd International Conference on Computational Statistics*. Iasi, Romania. URL: https://hal.archives-ouvertes.fr/hal-01949128

6. V. Vandewalle, C. Cozma, and C. Preda (2015). "Clustering categorical functional data Application to medical discharge letters". 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, Dec 2015, Londres, United Kingdom

7. M. Marbac, C. Biernacki, and V. Vandewalle (2015a). "Classification de données mixtes par un modèle de mélange de copules gaussiennes". 46èmes Journées de la SFDS, Rennes

8. S. Iovleff, C. Biernacki, and V. Vandewalle. (2014). "Visualisation des Modèles de Mélange". Big Data Mining and Visualization, Journées communes aux Groupes de Travail EGC et AFIHM, Lille

9. F. Letué, E. Gabriel, and V. Vandewalle (2014). "Table ronde STID-groupe Enseignement de la statistique de la SFdS : comment s'appuyer sur nos réseaux d'anciens étudiants pour mieux promouvoir nos formations en statistique". 46èmes Journées de la SFDS, Rennes

10. M. Marbac, C. Biernacki, and V. Vandewalle (2014b). "Model-based clustering of Gaussian copulas for mixed data". Working meeting "Handling categorical and continuous data" of GdR MASCOT-NUM. IHP, Paris

11. M. Marbac, C. Biernacki, and V. Vandewalle (2014a). "Classification de données mixtes par un mélange de copules Gaussiennes". 1ère journée YSP. IHP, Paris

12. V. Vandewalle (2013). "Quel est le bagage statistique de nos futurs étudiants ?" 45èmes Journées de la SFDS, Toulouse

13. V. Vandewalle and C. Biernacki (2013). "Mise en garde sur l'utilisation des mélanges gaussiens avec données manquantes". 45èmes Journées de la SFDS, Toulouse

14. E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki (2012). "Mixture of Gaussians for Distances Estimations with Missing Data". Workshop New Challenges in Neural Computation

15. M. Marbac, C. Biernacki, and V. Vandewalle (2012). "Modèle de mélange pour classifier des données qualitatives conditionnellement corrélées". 44èmes Journées de la SFDS, Bruxelles

16. C. Biernacki and V. Vandewalle (2011a). "Label swicthing dans les mélanges". 43èmes Journées de la SFDS, Tunis

17. V. Vandewalle (2010). "How to take into account the discrete parameters in the BIC criterion?" COMPSTAT 2010

18. V. Vandewalle (2009b). "Sélection prédictive d'un modèle génératif par le critère AICp". *41èmes Journées de Statistique*, Bordeaux, France

19. V. Vandewalle, C. Biernacki, G. Celeux, and G. Govaert (2008). "Are unlabeled data useful in semi-supervised model-based classification? Combining hypothesis testing and model choice". In: *proceedings of SFC-CLADAG meeting*. Caserta, Italy, pp. 433–436

Posters

1. V. Vandewalle and C. Preda (2016). "Clustering categorical functional data Application to medical discharge letters Medical discharge letters". Working Group on Model-Based Clustering Summer Session: Paris, July 17-23, 2016 Paris, France

A.2 Synthesis of scientific activities

In terms of my scientific activities in the broadest sense, I have had the opportunity to participate in four research projects (two ANR projects, one European project and a CNRS Mastodons challenge project). I have also participated in the supervision of two theses, plus another thesis in progress and one about to start. I have supervised six M2 internships. I have participated in three thesis juries. I have participated in three selection committees for lecturer positions. I have also taken various responsibilities in scientific societies, been a reviewer for several journals, and participated in the organisation of sessions at various international conferences. Finally, as part of my activities within the MODAL team, I have participated in research contracts with companies.

A.2.1 Participation in research projects

ANR Smiles Project: Statistical Modeling and Inference for unsupervised Learning at LargE-Scale (2018-2022) Since November 2018, I have been participating in the ANR Smiles project which focuses on the development of methods for clustering in large sample sizes. In particular in the context of mixture models, the aim is to adapt the estimation algorithms to large numbers of data, for example in order to detect small classes that are not very visible in smaller samples.

European project PERF-AI: Enhance Aircraft Performance and Optimisation through utilisation of Artificial Intelligence (2018-2020) . Since November 2018, I have been participating in the European PERF-AI project in partnership with the company Safety Line. In particular, based on data collected during flights, it involves developing Machine Learning models to optimize the aircraft's trajectory in relation to fuel consumption, for example.

ANR ClinMine ANR project: Optimizing Patient Care in Hospitals (2013-2017) . In this ANR project I worked on the development of mixture models for clustering of categorical functional data. Indeed, we have a lot of hospital data on categorical variables over time (type of pathology, state of mail processing, ...), but few models exist for their clustering, especially when the period over which the observation is made varies from one individual to another.

Projet CloHe du challenge Mastodons CNRS " Big data and data quality " (2016-2018) From 2016 to 2018 I participated in the CloHe project of the Mastodons CNRS challenge " Big data and data quality ". In this project we worked on the design of supervised and unsupervised classification algorithms for the classification of observed satellite images in multi-spectral resolution over time. It is essentially a matter of processing multivariate functional data with missing values. A new approach for this problem is under development, notably through data modeling by a Gaussian process.

A.2.2 Doctoral and scientific supervision

Theses supervision

1. In progress since September 2019: Wilfried Heyse, *PhD thesis from the University of Lille, Biostatistics Specialization. Title :* " Taking into account time structure in the statistical analysis of high throughput proteomic data. "
   Home laboratory : INSERM Lille and Inria Lille-Nord Europe (MODAL team)
   Financing : INSERM
   Framework : Co-supervision with Christophe Bauters and Guillemette Marot

2. 2016-2019 : Adrien Ehrhardt, *PhD thesis from the University of Lille, Specialized in Statistics. Title :* " Predictive models for large and biased data. Application to the improvement of credit risk scoring. "
   Home laboratory : Paul Painlevé Mathematics Laboratory (CNRS 8524) and Inria Lille-Nord Europe (MODAL team)
   Financing : CIFRE within CACF
   Framework : Co-supervision with Christophe Biernacki and Philippe Heinrich

3. 2011-2014: Matthieu Marbac-Lourdelle, *PhD thesis from the University of Lille, Specialized in Statistics. Title :* " Mixture models for clustering of categorical and mixed data " defended September 23, 2014.
   Home laboratory : Paul Painlevé Mathematics Laboratory (CNRS 8524) and Inria Lille-Nord Europe (MODAL team)
   Financing : DGA & Inria
   Framework : Co-supervision with Christophe Biernacki

Master 2 internship supervision

1. March-August 2019 : Wilfried Heyses, M2 ISN internship, University of Lille, France, Statistical analysis of high-throughput proteomic data

2. April-September 2018 : Souane Ibrahima, M2 Mathematical Engineering internship : University of Nice, " Study and implementation of a large scale multi-partition clustering model "

3. April-September 2017 : Thierry Mottet, M2 statistical modeling internship in Besançon, France, Study and implementation of a model for clustering variables according to their grouping behavior.

4. May-September 2016: Hamza Cherkaoui, M2 research internship in Applied Mathematics, University Lille 1 / Diploma of Engineering of Centrale Lille option DAD, " Implementation of a generative model for clustering qualitative functional data "

5. April-September 2014: Komi Nagbe, M2 Stochastic and Computer Science Methods for Decision-making at the University of Pau, " Development of methods for the visualization of probabilistic models in classification ".

6. April-September 2011 : Matthieu Marbac, M2 Digital Systems Engineering internship at the University of Lille 1, " Clustering of conditionally correlated data "

A.2.3 Participation in thesis juries

1. Adrien Ehrhardt, *PhD thesis from the University of Lille, Specialized in Statistics. Title :*, " Formalization and study of credit risk scoring problematics, defended on September 30, 2019. "
   Thesis supervisors: Christophe Biernacki, Philippe Heinrich, Vincent Vandewalle
   Reviewers: François Husson, Jean-Michel Loubes
   Examiner: Camelia Goga

2. Florence Loigneville, *PhD thesis from the University of Lille, Specialty Statistics. Title :*, " Hierarchical generalized linear model Gamma-Poisson for quality control in microbiology, defended January 22, 2016."
   Thesis supervisors:. Julien Jacques, Cristian Preda
   Reviewers: Ali Gannoun, Enachescu Denis
   Examiners: Filipe Marques, Vincent Vandewalle

3. Matthieu Marbac: *PhD thesis from the University of Lille, Specialized in Statistics. Title :* " Mixing models for unsupervised classification of qualitative and mixed data " defended September 23, 2014."
   Thesis supervisors: Christophe Biernacki, Vincent Vandewalle
   Reviewers: Dimitris Karlis, Jean-Michel Marin
   Examiners: Gilles Celeux, Nicolas Wicker

A.2.4 Participation in selection committees

1. 2020: Position of Assistant Professor in Statistics for Data Science, University of Avignon

2. 2019 : Position of Assistant Professor in Statistics and Applications, University of Versailles Saint Quentin

3. 2018 : Position of Assistant Professor in Mathematical Modeling for Medical Data Analysis, University of Lille

A.2.5 Contracts with companies

Since November 2018, I participate in the European project PERF-AI with transfers in the field of aeronautics.
In January 2018 I participated in the mathematics/enterprise week organized by AMIES in Lille as a scientific referent for one of the proposed subjects.
From April 2016 to September 2019 I co-supervised Adrien Ehrhardt's CIFRE thesis at CACF.
Within the framework of the MODAL team of Inria I participated in contracts with companies through the supervision of engineers. I have participated in contracts with the following companies:

- Cyland (2015-2016): Sale prediction
- Auchan (2014-2015): Variable selection in regression
- Rouge gorge (2014): Customer segmentation
- Natural security (2011): Risk calculation in the context of new payment technology

A.3 Scientific responsibilities

A.3.1 Participation in scientific societies

- **2016 to 2018 :** Member of the scientific animation cell of the bioinformatics platform of Lille.
- **2016 :** Member of the steering committee of the school Cimpa-Sénégal, Statistical methods for the evaluation of extreme risks.
- **2013 to today :** Organization of the STID-SFdS session at the JdS.
- **2013 to today :** Member of the jury for the best STID-SFdS course.
- **2013 to 2016 :** Member of the board of the STID France association.

A.3.2 Organization of national and international conferences

- Organization of a "Model-based and multivariate functional data" session at the CRoNoS & MDA conference, Cyprus, 2019.

- Organization of a session "Advances in model-based clustering" at the 11th ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018) Conference, Pisa, 2018.
- Organization of a session "Functional spatio-temporal data and applications" at the 2nd Satellite CRoNoS Workshop on Functional Data Analysis, Iasi, 2018.
- Chairman of a session "Statistical modelling" at COMPSTAT 2018 Conference, Iasi, 2018.
- Organization of a "Model-based clustering" session at the 10th ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017) Conference, London, 2017.
- Participation in the JOBIM 2017 program committee (https://project.inria.fr/jobim2017/en/).
- Participation in the organization of the conference "Learning with functional data" in Lille on October 7, 2016 (https://functional-data.univ-lille1.fr)
- Since April 2016, member of the scientific animation unit of the bilille platform, in this context I have participated in the organization of five thematic days (https://wikis.univ-lille1.fr/bilille/animation).

A.3.3 Reviewing activities

I have been a reviewer for the following journals: Journal of Classification, Neurocomputing, Pattern Recognition, Journal of SFdS, Statistics and computing, International Journal of Computer Mathematics, Advances in Data Analysis and Classification, Journal of Computational and Graphical Statistics, Methodology & Computing in Applied Probability, Spatial Statistics.

Appendix B

Teaching activities and responsibilities

Contents

B.1 Teaching activities 195
B.1.1 Teachings 195
B.1.2 Internship and project monitoring 196
B.2 Teaching responsibilities 196
B.2.1 Head of STID Department (2012-2015) 196
B.2.2 Promotional of STID department (2013-2014) 197
B.2.3 Head of Tutored Projects (2014-2015) 198
B.2.4 Director of Studies (2011-2012) 198

B.1 Teaching activities

B.1.1 Teachings

I am doing most of my teachings to undergraduate students in the first and second year of DUT STID (Statistic and business intelligence). I am also involved in the related professional license SID (L3) at IUT C, as well as in the GIS department of Polytech'Lille, and in the Master Data Science for health at Institute of Lille for health (ILIS).
The teaching that I do on a recurring basis in IUT C:

- DUT STID
  - First year (L1): probabilities (elementary probabilities, notion of random variable), statistical programming (introduction to the R software)
  - Second year (L2): cases study in statistics, supervised classification (logistic regression and classification trees), mathematics option (additional analysis)
- Licence SID (L3): Machine learning

The other teachings in which I intervene punctually in other components :

- Polytech'Lille, department GIS (Computer Engineering and Statistics)
  - M1: Supervised classification
  - M2: Biostatistics; survival analysis
- ILIS, master data-science for health data
  - M2: Introduction to Bayesian statistics

B.1.2 Internship and project monitoring

The DUT STID is a professional training. As such, internships and projects play an important role. Every year since my recruitment, I have supervised projects and internships.

- Project monitoring (3 groups of students per year)
  - Regular updates on the progress of the project with the students
  - Participation in the intermediate defense
  - Reading and evaluating reports
  - Evaluation of the final defense
- Internship follow-up (4 groups of students per year)
  - Regular updates on the progress of the internship with students
  - Meeting with the professional tutor in the company
  - Reading and evaluating reports
  - Evaluation of the final defense

I set up the professionalization contract in the second year of DUT STID in 2013.
During the academic year 2013-2014, I participated in the setting up of a transversal statistical-computing project in the framework of the implementation of the new national pedagogical program.
During the confinement in spring 2020, I participated in the building of a professional case study aiming at replacing the impossible internship that year.
In the building of the new program of DUT STID, I am currently involved in the working group on statistical modeling.

B.2 Teaching responsibilities

B.2.1 Head of STID Department (2012-2015)

As Head of the STID department, I have carried out the following missions related to the management of the diploma:

- Organization of student recruitment, reception, information, and orientation.

- Elaboration and adaptation of the timetable.
- Distribution of services among the teachers under the control of the selection committee and the Director of the Institute.
- Redesign of pedagogical models.
- Organization of pedagogical meetings.
- Monitoring of cohorts.
- Implementation of actions for professional integration.
- Validation of internship offers.
- Participation in the board of directors of the IUT.
- Participation in STID department heads' meetings.
- Recruitment of temporary teachers.

I also had to work specifically on the following files, which required a very particular investment:

- Participation in the reform of the national educational program for which I was coordinator of the statistical part.
- Implementation of the second year of the DUT STID in a professionalization contract and initiation of the file to make it qualify as an apprenticeship.
- Setting up of the four-year assessment file of the DUT STID and defending it with the experts.
- Implementation of transversal management tools (change of schedule management software).

This responsibility has been very interesting. However, I have chosen not to apply for a second mandate in order to devote more time to my research work.

B.2.2 Promotional of STID department (2013-2014)

Throughout my mandate, as head of the STID department, I have participated in the promotion of the STID department. This mission gave rise to a specific responsibility during the 2013-2014 academic year.

- **Setting up a partnership with the Academic Inspectorate of Mathematics :** Within the framework of the reform of the high school mathematics curriculum, I came into contact with academic inspectors of mathematics, for the organization of :

  - A half-day conference: " Computer Science and Statistics, the two pillars of decision support " (participants: 130 mathematics teachers)
  - Two half-days of training for mathematics teachers

- **Organization of high school visits through lectures, as well as mini-sessions of practical work :**.
  - Lycée André Malraux, Béthune
  - Lycée Saint Rémi, Roubaix

B.2.3 Head of Tutored Projects (2014-2015)

During the academic year 2014-2015, I was responsible for the tutored projects. These projects intervene for semesters 2, 3, and 4 of the DUT:

- In semester 2: a transversal statistical / IT project given to students by the teaching team
- In semesters 3 and 4: students must find a company for which they carry out a project and are supervised by a teaching tutor.

As the person in charge of the tutored projects for semester 2, I coordinated the transversal project: research of data, organization of meetings with the pedagogical team, explanation of the project organization to the students. This activity proved to be particularly complex since it took place in the context of the implementation of the new pedagogical program (first cross-cutting project between statistics and computer science contributors) requiring new coordination.
For semesters 3 and 4 I carried out the following missions:

- Validation of student missions
- Assignment of students to tutors
- Project monitoring management
- Organization of support

B.2.4 Director of Studies (2011-2012)

I took the responsibility of director of studies (first and second year) one year after my recruitment as Assistant Professor in the STID department of the IUT of Université Lille 2. On this occasion, I carried out the following missions:

- Attendance control.
- Organization of supervised homework weeks.
- Management of the collection of grades.

- Preparation and participation in jurys.
- Meeting with students following jury decisions.
- Meeting with students to discuss the various difficulties encountered.
- Writing notices for student prosecution files.

At the end of this first experience, I wanted to get more involved in the department by taking the responsibility of Head of the STID department.

Contribution to model-based clustering of heterogeneous data

Abstract: I am an Assistant Professor at the University of Lille since 2010. In the continuity of my thesis defended in 2009 on semi-supervised classification, my research work has focused on model-based clustering. These models are particularly useful tools that permit to perform clustering of data through the inference of a probabilistic model. It can cluster any kind of data as soon as some model is available for this kind of data, enabling the use of standard statical tools such as model selection to rationale some choices such as the number of clusters.

In the scope of model-based clustering, I have worked on the proposal of models allowing the clustering of data containing different kinds of variables, in particular taking into account categorical, mixed, or functional variables. I have worked on the problem of multiple partition clustering which consists of searching for several latent class variables, then allowing several clustering points of view to be revealed by the model. Finally, I have also been interested in more general model-based clustering issues such as the label switching problem, taking into account missing data, or visualizing the output of a mixture.

In another part of my research, I have been worked on application-driven issues. In this scope, I have worked on the usability study of medical devices through the modeling of the discovery matrix modeling thus enabling the manufacturer to access the completeness of the discovery process and evaluate its performances. I have worked on credit scoring through the automation of pre-processing by embedding it in the scoring model estimation. Finally, I have worked on aircraft flight data to propose a data-driven optimization of the plane trajectory.

These different research works have been carried out within the framework of research projects and have led to publications in peer review journals. All along these works, I have participated in the supervision of internships, thesis, post-doctoral students, and engineers. In this manuscript, I present a synthesis of my research works as well as the resulting research projects.

Keywords: Clustering, mixture models, heterogeneous data, Bayesian inference, model choice, visualization, missing data, functional data, medical applications