Université Rennes I

Ecole Doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication (MathSTIC)

Mémoire d'habilitation présenté par

Matthieu Marbac

en vue de l'obtention du diplôme d'Habilitation à Diriger des Recherches de l'Université de Rennes I

Contributions to Model-Based Clustering

Mémoire soutenu publiquement le 06 septembre 2022 devant le jury:

Charles Bouveyron	Université Côte d'Azur	Examinateur
Julien Chiquet	MIA Paris-Saclay AgroParisTech/INRAE	Rapporteur
David Hunter	Penn State University	Rapporteur
Valérie Monbet	Université de Rennes 1	Présidente
Stéphane Robin	Sorbonne Université	Examinateur
Cinzia Viroli	University of Bologna	Rapporteur

Remerciements

First of all, I would like to thank the three referees of this manuscript, Julien Chiquet, David Hunter and Cinzia Viroli, who kindly accepted to take a large part of their time to write a report on my past and recent work. I would also like to thank Charles Bouveyron, Valérie Monbet and Stéphane Robin for accepting to participate in my defense as examiners. 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.

Je souhaite remercier les enseignants qui m'ont formé lors de mon parcours universitaire à Lille 1. À ce titre, j'ai une pensée particulière pour mes premiers enseignants en probabilité et statistiques: Charles Suquet (dont les polycopiés de cours restent pour moi une référence), Gwénaëlle Castellan et Laurence Marsalle. Je remercie également Tran Viet Chi pour son encadrement lors de mon mémoire de master 1 qui m'a donné envie de poursuivre dans le monde académique. Enfin, je souhaite remercier particulièrement Bernhard Beckermann qui reste pour moi un modèle de rigueur et d'implication auprès des étudiants.

Je remercie chaleureusement mes deux directeurs de thèse Christophe Biernacki et Vincent Vandewalle avec qui j'ai eu la chance d'échanger depuis plusieurs années. Leur bienveillance, leur disponibilité et leur passion pour la recherche m'ont particulière marqué. Cela a toujours été un plaisir de travailler avec eux, et j'espère poursuivre nos collaborations. Plusieurs autres membres de l'équipe Modal m'ont guidé lors de mon doctorat, et je souhaite profiter de cette occasion pour les remercier: Julien Jacques et Critian Preda (qui m'ont notamment aidé pour mes premiers enseignements...mais pas seulement) ainsi que Alain Célisse, Serge Iovleff, Guillemette Marot et Sandrine Meilen.

Je remercie fortement Mohammed Sedki pour son encadrement lors de mon post-doctorat. C'est toujours un plaisir d'échanger avec lui grâce à sa bonne humeur et à son franc parler.

Je remercie l'ensemble des membres (passés et actuels) de l'Ensai avec qui j'ai pris plaisir à travailler. Je remercie notamment Valentin Patilea pour son accueil lors de mon arrivée et pour m'avoir permis de m'ouvrir à d'autres thématiques ainsi que Hong-Phuong Dang et Salima El Kolei pour leur bonne humeur. Enfin, je remercie particulièrement Marie Du Roy de Chaumaray avec qui c'est un plaisir de travailler chaque jour et avec qui je souhaite collaborer encore longtemps.

Enfin, je souhaite remercier l'ensemble des personnes avec qui j'ai eu la chance de collaborer: Mohamed Ahmed, Christophe Biernacki, Claire Boyer, Gilles Celeux, Amay Cheam, Patricia

Dargent, Marie Du Roy de Chaumaray, Orianne Dumas, Salima El Kolei, Marc Fredette, Julie Josse, Michael Genin, Mohammed Sedki, Paul McNicholas, Fabien Navarro, Valentin Patilea, Etienne Patin, Cécilia Saldanha Gomes, Aude Sportisse, Gilles Stupfler, Vincent Vandewalle.

Contents

Chapter 1

Introduction

Contents

1.1 Summary of my contributions

1.1.1 Presentation

I was appointed Assistant Professor in Statistics at ENSAI/CREST (Bruz, France) in September 2017, after holding postdoctoral fellows at Institut national de la santé et de la recherche médicale (Villejuif, France) then at the University of McMaster (Hamilton, Ontario, Canada) then holding an engineer position at Institut national de recherche en sciences et technologies du numérique (Lille, France). I was awarded my PhD in October 2014 at the University of Lille.

I have been working on mathematical aspects in empirical likelihood, mathematical and methodological aspects in clustering and applications of these methods for biostatistics. My work has resulted in the publication/acceptance of 19 papers in international, peer-reviewed journals in fields of mathematical statistics (*The Annals of Statistics*), computational and methodological statistics (*Computational Statistics & Data Analysis*, *Journal of Computational and Graphical Statistics*, *Statistics and Computing*), applied statistics (*The Annals of Applied Statistics*) and epidemiology (*The Annals of Epidemiology*). Results of my work have been presented in many international conferences (*Bernoulli-IMS One World Symposium*, *CMstatistics*, *CompStat*, *EcoSta*, *ICML-workshop*, *SDSS*) and have been implemented in 8 R packages available on *CRAN*. I have etablished international and national collaborations with different departments of statistics (*CREST*, *HEC Montreal*, *Inria*, *University of McMaster*, *Université de Lille*, *UPMC*) and epidemiology (*Institut Pasteur Paris*, *INSERM*).

I am an Associated Editor in *Computational Statistics & Data Analysis* (since April 2021), I belongs to the Scientific Program Committee of the international conference EcoSta 2022 and I have organized different invited sessions for international conferences (EcoSta 2021, Canadian Conference in Applied Statistics 2021, CMstatistics 2022). At a national level, I am a member of the scientific commitee of evaluation of the French Agence of Research for the topic "mathematics, numerical science, bioloy and health" (ANR-AAPG-2022 CES 45) and I gave lectures at the summer school *19èmes Journées d'Étude en Statistique* organized by the french statistical society (SFDS). Finally, I have supervised 8 MSc internships and I am supervizing 1 PhD student.

1.1.2 Research Interests

My developments in clustering focus on the study of *mixture models* for complex data (*i.e.,* categorical data (Marbac, Biernacki, and Vandewalle (2016) and Marbac, Biernacki, and Vandewalle (2015)), mixed-type data (Marbac, Biernacki, and Vandewalle (2017)), high-dimensional data (Marbac and Sedki (2017a) and its R package Marbac and Sedki (2016b), and Marbac and McNicholas (2016)).

We have developed two methods for selecting the variables in clustering, by considering parametric mixture models. The first method allows for selecting the model without estimating the model parameters (Marbac and Sedki (2017b)). The second method simultaneously performs the selection of the relevant variables and the parameter estimation (Marbac, Sedki, and Patin (2020)). Both of these methods are implemented in the R package VarSelLCM (Marbac and Sedki (2018) and Marbac and Sedki (2020)) and have been extended to the case of multiple partitions (Marbac and Vandewalle (2019)). Recently, in Du Roy de Chaumaray and Marbac (2021a), we propose an approach for selecting the subset of relevant variables and the number of components in a semi-parametric mixture model.

Some of my developments previously described have been used in epidemiology (Dumas et al. (2021), Saldanha Gomes et al. (2020), Marbac et al. (2018)). Moreover, these collaborations raise new methodological problems such as the use of a clustering results in a predictive model. Such an approach is classical in epidemiology but produces biased results. Hence, in Marbac et al. (2022) and its companion R package (Marbac et al. (2021)), we circumvent this issue by simultanously estimating the clustering and the prediction models. Discussions with epidemiologists encouraged me to develop a visualization method for the clustering output (Biernacki, Marbac, and Vandewalle (2021) and the companion R package Biernacki, Marbac, and Vandewalle (2019)) but also different methods to cluster data with missingness having a non-ignorable mechanism (Biernacki et al. (2021) and Du Roy de Chaumaray and Marbac (2020) and the compagnion R package Du Roy de Chaumaray and Marbac (2021b))

I have been involved in different projects in *biostatistics* that have led me to extend some statistical methods for high-dimensional data (Marbac, Tubert-Bitter, and Sedki (2016) and the companion R package Marbac and Sedki (2016a)) or to develop new methodologies for clustering functional data (Cheam, Marbac, and McNicholas (2017), Du Roy de Chaumaray, Marbac, and Navarro (2020) and Cheam et al. (2020) and their companion R packages Cheam, Marbac, and McNicholas (2020) and Du Roy de Chaumaray, Marbac, and Navarro (2019)). Thus, I have worked on a method, based on mixture models of hidden Markov chains, allowing data collected by accelerometers to be analyzed without considering arbitrary thresholds (Du Roy de Chaumaray, Marbac, and Navarro (2020)). Moreover, I was involved in a project investigating the geographical disparities of the COVID-19 deaths. This approach analyses the daily-reports of COVID-19 deaths using wavelet decomposition, semi-parametric regression and mixture models (Cheam et al. (2020)). Moreover, with colleagues from the department of medicine of the Uni-

versity of Lille, I am currently working on *spatial scan statistics*. We have proposed a method for considering functional covariates in scan statistics (Frévent et al. (2021)). Moreover, we are working on a method that allows for multiple cluster detection by avoiding the exhaustive search for clusters detection. Finally, we are conducting research to obtain an explicit form of the asymptotic distribution of the scan statistics. The works on scan statistics permit interactions with other statisticians but also two Professors of Medicine.

Finally, I was interested in *empirical likelihood*. Thus, in Du Roy de Chaumaray, Marbac, and Patilea (2021), we consider parameter inference for a semi-parametric regression model with weakly dependent data (α-mixing).

1.1.3 Publications

Submitted to peer-reviewed journals

Biernacki, C. et al. (2021). *Model-based Clustering with Missing Not At Random Data*. URL: `https://arxiv.org/abs/2112.10425`.

Cheam, A.M.S. et al. (2020). "Translation-invariant functional clustering on COVID-19 deaths adjusted on population risk factors". URL: `https://arxiv.org/abs/2012.10629`.

Du Roy de Chaumaray, M. and M. Marbac (2020). "Clustering Data with nonignorable Missingness using Semi-Parametric Mixture Models." URL: `https://arxiv.org/abs/2009.07662`.

Du Roy de Chaumaray, M. and M. Marbac (2021a). "Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models". URL: `https://arxiv.org/abs/2112.05684`.

Book chapters

Marbac, M. (2022a). *Introduction à une étude statistique avec données manquantes, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Marbac, M. (2022b). *Méthodes basées sur la vraisemblance pour données manquantes ayant un mécanisme ignorable, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Marbac, M. (2022c). *Méthodes de pondération pour données manquantes, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Peer-reviewed journals (Statistics)

Biernacki, C., M. Marbac, and V. Vandewalle (2021). "Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering". *Journal of Classification* 38, pp. 129–157. URL: `https://link.springer.com/article/10.1007/s00357-020-09369-y`.

Cheam, A.S.M., M. Marbac, and P.D. McNicholas (2017). "Model-based clustering for spatiotemporal data on air quality monitoring". *Environmetrics* 28(3), e2437. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2437`.

Du Roy de Chaumaray, M., M. Marbac, and F. Navarro (2020). "Mixture of hidden Markov models for accelerometer data". *The Annals of Applied Statistics* 14(4), pp. 1834–1855. URL: https://projecteuclid.org/euclid.aoas/1608346901.

Du Roy de Chaumaray, M., M. Marbac, and V. Patilea (2021). "Wilks' theorem for semiparametric regressions with weakly dependent data". *The Annals of Statistics* 49(6), pp. 3228 –3254. URL: https://doi.org/10.1214/21-AOS2081.

Frévent, C. et al. (2021). "Detecting spatial clusters on functional data: a parametric scan statistic approach". *Spatial Statistics* 46, p. 100550. ISSN: 2211-6753. URL: https://www.sciencedirect.com/science/article/pii/S2211675321000609.

Marbac, M., C. Biernacki, and V. Vandewalle (2016). "Latent class model with conditional dependency per modes to cluster categorical data". *Advances in Data Analysis and Classification* 10(2), pp. 183–207. URL: https://link.springer.com/article/10.1007/s11634-016-0250-1.

Marbac, M., C. Biernacki, and V. Vandewalle (2015). "Model-based clustering for conditionally correlated categorical data". *Journal of Classification* 32(2), pp. 145–175. URL: https://link.springer.com/article/10.1007/s00357-015-9180-4.

Marbac, M., C. Biernacki, and V. Vandewalle (2017). "Model-based clustering of Gaussian copulas for mixed data". *Communications in Statistics-Theory and Methods* 46(23), pp. 11635–11656. URL: https://www.tandfonline.com/doi/abs/10.1080/03610926.2016.1277753.

Marbac, M. and P.D. McNicholas (2016). "Dimension Reduction in Clustering". *Wiley StatsRef: Statistics Reference Online*, pp. 1–7. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat07846.

Marbac, M. and M. Sedki (2017a). "A family of block-wise one-factor distributions for modeling high-dimensional binary data". *Computational Statistics & Data Analysis* 114, pp. 130–145. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167947317300932.

Marbac, M. and M. Sedki (2017b). "Variable selection for model-based clustering using the integrated complete-data likelihood". *Statistics and Computing* 27(4), pp. 1049–1063. URL: https://link.springer.com/article/10.1007/s11222-016-9670-1.

Marbac, M. and M. Sedki (2018). "VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values". *Bioinformatics* 35(7), pp. 1255–1257. URL: https://academic.oup.com/bioinformatics/article/35/7/1255/5091183?login=true.

Marbac, M., M. Sedki, and E. Patin (2020). "Variable selection for mixed data clustering: Application in human population genomics". *Journal of Classification*, pp. 1–19. URL: https://link.springer.com/article/10.1007%2Fs00357-018-9301-y.

Marbac, M., P. Tubert-Bitter, and M. Sedki (2016). "Bayesian model selection in logistic regression for the detection of adverse drug reactions". *Biometrical Journal* 58(6), pp. 1376–1389. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201500098.

Marbac, M. and V. Vandewalle (2019). "A tractable multi-partitions clustering". *Computational Statistics & Data Analysis* 132, pp. 167–179. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167947318301592.

Marbac, M. et al. (2022). "Simultaneous semi-parametric estimation of clustering and regression." *Journal of Computational and Graphical Statistics* forthcoming, pp. 1–9. URL: https://doi.org/10.1080/10618600.2021.2000872.

Peer-reviewed journals (Epidemiology)

Dumas, O. et al. (2021). "Household cleaning and poor asthma control among elderly women". *The Journal of Allergy and Clinical Immunology: In Practice*. URL: https://www.sciencedirect.com/science/article/abs/pii/S2213219821002026?via%3Dihub.

Marbac, M. et al. (2018). "Patterns of cleaning product exposures using a novel clustering approach for data with correlated variables". *The Annals of Epidemiology* 28(8), pp. 563–569. URL: https://www.sciencedirect.com/science/article/abs/pii/S104727971630504X.

Saldanha Gomes, C. et al. (2020). "Clusters of diet, physical activity, television exposure and sleep habits and their association with adiposity in preschool children: the EDEN mother-child cohort." *International Journal of Behavioral Nutrition and Physical Activity* 17(1). URL: https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-020-00927-6#citeas.

R packages

Biernacki, C., M. Marbac, and V. Vandewalle (2019). *ClusVis: Gaussian-Based Visualization of Gaussian and Non-Gaussian Model-Based Clustering*. R package version 1.2.0. URL: https://CRAN.R-project.org/package=ClusVis.

Cheam, A.M.S., M. Marbac, and P.D. McNicholas (2020). *SpaTimeClus: Model-Based Clustering of Spatio-Temporal Data*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=SpaTimeClus.

Du Roy de Chaumaray, M. and M. Marbac (2021b). *MNARclust: Clustering Data with Non-Ignorable Missingness using Semi-Parametric Mixture Models*. R package version 1.0.0. URL: https://CRAN.R-project.org/package=MNARclust.

Du Roy de Chaumaray, M., M. Marbac, and F. Navarro (2019). *MHMM: Mixture of hidden Markov models for accelerometer data*. R package version 1.0. URL: https://cran.rstudio.com/web/packages/MHMM/index.html.

Marbac, M and M. Sedki (2016a). *MHTrajectoryR: Bayesian Model Selection in Logistic Regression for the Detection of Adverse Drug Reactions*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=MHTrajectoryR.

Marbac, M and M. Sedki (2016b). *MvBinary: Modelling Multivariate Binary Data with Blocks of Specific One-Factor Distribution*. R package version 1.1. URL: https://CRAN.R-project.org/package=MvBinary.

Marbac, M and M. Sedki (2020). *VarSelLCM: Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values*. R package version 2.1.3.1. URL: `https://cran.r-project.org/web/packages/VarSelLCM/index.html`.

Marbac, M. et al. (2021). *ClusPred: Simultaneous Semi-Parametric Estimation of Clustering and Regression*. R package version 1.0.0. URL: `https://CRAN.R-project.org/package=ClusPred`.

1.2 Outline of the manuscript

The manuscript is organized as follows. The rest of this chapter presents a brief introduction to model-based clustering. The next five chapters describe my contributions on subjects I wish to continue working on. Chapter 2 presents my contributions in variable selection for model-based clustering. Chapter 3 presents my contributions to model-based clustering with missingness under a non-ignorable mechanism. Chapter 4 details a method to consider a clustering output in a prediction model. Chapter 5 presents two of my applications in biostatistics of functional data clustering. Chapter 6 presents my contribution to empirical likelihood for semi-parametric regression model with dependent data. Chapter 7 details my future research projects.

1.3 Brief introduction to model-based clustering

1.3.1 Mixture models

Clustering aims to summarize data sets composed by many subjects with few homogeneous classes. Indeed, it assesses a partition among the subjects by grouping similar subjects into the same class while two classes are composed of strongly different subjects. We can split the clustering methods into two families: distance-based methods and model-based methods. Distance-based methods use a distance or a similarity between subjects to define a notion of homogeneity within class. Among the the distance-based methods, the two standard approaches are the K-means clustering (Lloyd (1982), Arthur and Vassilvitskii (2006) and Lu and Zhou (2016)) and the hierarchical ascendent classification (Ward (1963), Szekely and Rizzo (2005) and Gao, Bien, and Witten (2020)). Model-based methods (McLachlan and Peel (2000), McNicholas (2016) and Fruhwirth-Schnatter, Celeux, and Robert (2019)) do not explicitly define a distance between subjects (despite the fact that they implicitly define some distances; for instance Gaussian mixture models consider a Mahalanobis distance). Indeed, model-based methods achieve the clustering aim by modelling the distribution of the observed variables. This manuscript focuses on model-based clustering methods. Thus, we consider data to cluster $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ that are composed of n independent observations $\boldsymbol{x}_i \in \mathcal{X}$ where $\mathcal{X}$ depends on the type of variables. We suppose that each observation arises from the same mixture model with K components defined by the probability distribution function (pdf)

$$f(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i), \tag{1.1}$$

where $\boldsymbol{m}$ defines the model (*i.e.,* the number of components, the family of the components,...), $\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}$ groups all the model parameters including the vector of proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ where $0 < \pi_k$ and $\sum_{k=1}^{K} \pi_k = 1$, $\Theta_{\boldsymbol{m}}$ being the space of the parameters of model $\boldsymbol{m}$. Mixture

models can be considered in a parametric framework that assumes that the component belongs to a parametric distribution family such that we have

$$f(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k), \tag{1.2}$$

where $\boldsymbol{\alpha}_k$ groups the parameters of components k, leading to $\boldsymbol{\theta} = (\boldsymbol{\pi}^\top, \boldsymbol{\alpha}^\top)^\top$ where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_K^\top)^\top$ groups the parameters of the mixture components. Mixture models can also be considered in a semiparametric framework (Lindsay and Lesperance (1995), Hunter, Richards, and Rosenberger (2011) and Xiang, Yao, and Yang (2019)) that does not assume that the components belong to some parametric distribution families. In this case, $\boldsymbol{\theta}$ groups the finite dimensional parameters $\boldsymbol{\pi}$ and the infinite dimensional parameters (*i.e.,* the distribution $f_1, \ldots, f_K$).

From a mixture model, clustering can be achieved by computing the posterior probabilities of classification. Indeed, from the observed data $\mathbf{x}$, clustering aims to estimate the partition $\mathbf{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$ where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})^\top$, $z_{ik} = 1$ if subject i belongs to cluster k and $z_{ik} = 0$ otherwise. Thus, the posterior probabilities of classification are defined by

$$\mathbb{P}(Z_{ik} = 1 \mid \mathbf{X}_i = \boldsymbol{x}_i, \boldsymbol{m}, \boldsymbol{\theta}) = \frac{\pi_k f_k(\boldsymbol{x}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}_i)}. \tag{1.3}$$

Note that, in this manuscript, we consider that one class groups the subjects arising from the same component. Some extension allows some mixture components to be merged to define a class but these approaches are beyond the scope of this manuscript (see Hennig (2010) and Baudry et al. (2010) for some extensions and Hennig (2015) for a discussion on the definition of the clusters). The posterior probabilities of classification permit the incertainty of the class assignment to be taken into account. Moreover, from these probabilities, a classification rule permits an hard assignment of the subjects into class to be obtained. Thus, it is standard to apply the rule of the maximum *a posteriori* that affects a subject of the class maximizing its posterior probability of classification.

The choice of considering a parametric or semiparametric framework does not affect the use of mixture models for clustering because both frameworks allow for computing the posterior probabilities of classification. Semiparametric mixtures limit the assumptions made on the data distribution and thus the bias obtained by the parametric mixture models when their parametric assumptions are violated. Thus, one could be surprised to see that both approaches continue to be used but there are some technical reasons. First, one can note that some important properties were first stated for parametric mixtures and were not (until recently) proven for semiparametric mixtures. Among these properties we have in mind, one can cite the following properties that we detailled below: identifiability of the parameters, implementation of an algorithm having the monotonicity property, availability of an approach for model selection (especially for the estimation of the number of clusters) and properties of the estimators. Note that the three first properties are now stated for some semiparametric mixtures but there are properties of some estimators in semiparametric mixtures (bias and variance of the estimators, ideal bandwidth, *etc*) that still need to be etablished. One other argument in favour of the parametric mixture models is their easiness for interpretation. Indeed, clusters could be described by (few) parameters while descriptive statistics must be computed from the resulting partition when clustering is achieved by semiparametric mixtures. For these reasons, we believe that both families are of interest despite the fact that our recent research is more focused on semiparametric approaches. In this manuscript, we try to compare both families with numerical experiments, when it is possible (see Chapters 2, 3 and 4). However, these comparison consider continuous data, so are in favour of the semiparametric mixtures because the developments of these models mainly concerns continuous

data sets. Note that parametric mixture models have been developed to deal with different types of variables: contiuous (Banfield and Raftery (1993), Celeux and Govaert (1995) and McNicholas and Murphy (2008)), categorical (Goodman (1974), Celeux and Govaert (1991) and Gollini and Murphy (2014)), mixed-type (Hunt and Jorgensen (2011), Kosmidis and Karlis (2016) and Marbac, Biernacki, and Vandewalle (2017)), functional (Jacques and Preda (2014a), Jacques and Preda (2014b) and Cheam and Fredette (2020)) or network data (Nowicki and Snijders (2001), Daudin, Picard, and Robin (2008) and Tabouy, Barbillon, and Chiquet (2020)).

1.3.2 Identifiability of the model parameters

The unicity of the partition, up to a relabelling of the components, is an imperative property of a clustering method. This property holds true for mixture model whose the parameters are identifiable. Thus, the parameters of a fixed model $\boldsymbol{m}$ are identifiable when

$$\forall \boldsymbol{x}_1 \in \mathcal{X},\ f(\boldsymbol{x}_1; \boldsymbol{m}, \boldsymbol{\theta}) = f(\boldsymbol{x}_1; \boldsymbol{m}, \widetilde{\boldsymbol{\theta}}) \Leftrightarrow \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}. \tag{1.4}$$

Considering parametric components, the first results of identifiability for mixture models are stated by Teicher (1963), Teicher (1967) and Yakowitz and Spragins (1968). When the data are categorical, the usual model assumes that the variables are independent within each component. This model, called the latent class model, does not satisfy the property of identifiability of the parameters. Thus, a less restrictive property is to consider the generic identifiability of the parameters that considers that the space where the parameters do not statisfy (1.4) has a null measure. Allman, Matias, and Rhodes (2009) show that, under a mild relation between the number of components and the number of modalities of the variables, the latent class model for categorical data is generically identifiable. Considering the semi-parametric mixture models, constraints on the component distributions must be made to obtain identifiability. This leads to two important families of mixtures that are identifiable: a generalization of the latent class model that assumes that the density of a component is defined as a product of univariate densities or the location scale mixture models. In this manuscript, especially in Chapter 2, we focus on the generalization of the latent class model.

1.3.3 Parameter estimation

Parametric mixture models and the EM algorithm To be able to compute the posterior classification probabilities (1.3) associated with a parametric mixture model (1.2), the model parameters need to be estimated from the observed sample. Independence between the subjects implies that, the observed log-likelihood function is defined by

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

The maximum likelihood estimate (MLE), defined by $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}} = \arg\max_{\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}} \ell(\boldsymbol{\theta}; \mathbf{x}, \boldsymbol{m})$, cannot be obtained directly for mixture models and thus requires the use of optimization algorithms. The popular approach to obtain the MLE is to use the Expectation-Maximization algorithm (EM algorithm; Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977), Wu (1983) and McLachlan and Krishnan (2007)). The use of this algorithm is natural because the EM algorithm is devoted to the case of inference with missing values (in clustering, the partition is also interpreted as a missing value). The main idea of the EM algorithm is to consider a second optimization problem defined from all the data. Thus, we can define the completed data

log-likelihood (log-likelihood function computed over all the data, including the missing values) by

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \boldsymbol{m}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right).$$

The EM algorithm is an iterative algorithm that alternates between a step of data completion (E-step) that consists of the computation of the conditional expectation of the complete-data log-likelihood given the parameters defined by $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]}) := \mathbb{E}\left[\ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{m}) \mid \boldsymbol{X} = \mathbf{x}, \boldsymbol{\theta}^{[s]}, \boldsymbol{m} \right]$ and a step of parameter estimation (M-step) that consists of an updating of the parameters by maximizing this conditional expectation. Thus, starting from the initial value $\boldsymbol{\theta}^{[0]}$, its iteration s is defined by

- E-step: computation of the posterior probabilities of classification given the parameter $\boldsymbol{\theta}^{[s]}$

$$t_{ik}(\boldsymbol{\theta}^{[s]}) = \frac{\pi_k^{[s]} f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[s]})}{\sum_{\ell=1}^{K} \pi_\ell^{[s]} f_\ell(\boldsymbol{x}_i; \boldsymbol{\alpha}_\ell^{[s]})}.$$

- M-step: updating the parameters

$$\boldsymbol{\theta}^{[s+1]} = \underset{\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}}{\arg\max} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\boldsymbol{\theta}^{[s]}) \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right).$$

The EM algorithm is monotonic (*i.e.,* it increases the observe-data log-likelihood at each iteration) and thus converges into a local optimum of the objective function. Moreover, this algorithm is deterministic, meaning that the point of convergence only depends on the starting point $\boldsymbol{\theta}^{[0]}$. Hence, this algorithm needs to be run with different starting points (generally randomly sampled) to obtain the MLE. Note that Balakrishnan, Wainwright, and Yu (2017) developed a theoretical framework for quantifying when and how quickly EM-type iterates converge to a small neighborhood of a given global optimum of the population likelihood. Moreover, their approach allows for a characterization of the region of convergence of EM-type iterates to a given population fixed point, that is, the region of the parameter space over which convergence is guaranteed to a point within a small neighborhood of the specified population fixed point.

The sucess of the EM algorithm is due to its easy implementation explained by the use of the complete-data log-likelihood. However, for some models, the computation of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]})$ can be complex thus making the implementation of the E-step harder. For other models, the maximization of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]})$ with respect to $\boldsymbol{\theta}$ can lead to a problem with no closed-form thus making the implementation of the M-step harder. For these reasons, different extensions of the EM algorithm have been proposed. When the maximization implied by the M-step is not easy, the Generalized Expectation-Maximization algorithm (see Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977) and McLachlan and Krishnan (2007)) and the Expectation and Conditional Maximization algorithm (see Meng and Rubin (1993) and Liu and Rubin (1994)) can be used. When the computation of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]})$ is complex, simplifications can be obtained by using the Monte-Carlo Expectation-Maximization algorithm (see Wei and Tanner (1990), Chan and Ledolter (1995) and Booth and Hobert (1999)).

Parameter estimation can also be conducted in a Bayesian framework. Again, algorithms based on data augmentation such as the Gibbs sampler are generally used (Marin, Mengersen, and Robert (2005)).

Semi-parametric mixture models and the MM algorithm To estimate the finite and infinite dimensional parameters, the first approach was to use an EM-like algorithm Benaglia, Chauveau, and Hunter (2009). Despite the fact that this algorithm is very simple, it suffers from a lack of theoretical justification. Thus, Levine, Hunter, and Chauveau (2011) propose using a majorization-minimization algorithm (MM algorithm; see Hunter and Lange (2004) and Lange (2016)) to perform an estimation by maximizing the smoothed loglikelihood function defined by

$$\mathcal{S}\ell(\theta; \mathbf{x}, \boldsymbol{m}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N} f_k(\boldsymbol{x}_i) \right),$$

where $K(\boldsymbol{u}) = \prod_{j=1}^{J} K(u_j)$ is defined as a product of univariate kernels, $h > 0$ is a bandwidth, $K_h(\boldsymbol{x}_i - \boldsymbol{u}) = h^{-J} \prod_{j=1}^{J} K(h^{-1}u_j)$ is its rescale version and

$$\mathcal{N} f_k(\boldsymbol{x}_i) = \exp \left(\int K_h(\boldsymbol{x}_i - \boldsymbol{u}) \ln f_k(\boldsymbol{x}_i) d\boldsymbol{u} \right).$$

Similarily to the EM algorithm (in fact EM algorithm is a particular type of MM algorithm), the MM algorithm is an iterative algorithm that starts at the point $\theta^{[0]}$ and whose iteration $[s]$ is defined by

- Computation of the smoothed posterior probabilities of classification given the parameter $\boldsymbol{\theta}^{[s]}$

$$t_{ik}(\boldsymbol{\theta}^{[s]}) = \frac{\pi_k^{[s]} \mathcal{N} f_k^{[s]}(\boldsymbol{x}_i)}{\sum_{\ell=1}^{K} \pi_\ell^{[s]} \mathcal{N} f_\ell^{[s]}(\boldsymbol{x}_i;)}.$$

- Updating the parameters

$$\boldsymbol{\theta}^{[s+1]} = \underset{\boldsymbol{\theta} \in \Theta_{\boldsymbol{m}}}{\arg\max} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\boldsymbol{\theta}^{[s]}) \ln \left(\pi_k \mathcal{N} f_k^{[s]}(\boldsymbol{x}_i) \right).$$

The resulting algorithm has the monotonicity property that justifies its use.

1.3.4 Model selection

Parametric mixture models and information criteria Model selection is an important issue in model-based clustering. This task consists of finding the best model $\hat{\boldsymbol{m}}$ among a set of competing models $\mathcal{M}$. To determine such a model, it is usefull to consider an information criterion (IC) and thus we have

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m} \in \mathcal{M}}{\arg\max}\, \mathrm{IC}(\boldsymbol{m}).$$

Among the information criteria, the Bayesian Information Criterion (BIC; Schwarz (1978)) is generally used for model selection in mixture models. This criterion is defined by

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

where $\nu_{\boldsymbol{m}}$ is the number of parameters implied by model $\boldsymbol{m}$. Note that the BIC requires the estimation of the MLE. The consistency of this criterion is not straightforward because of a loss of identifiability of the parameters when the model is overfitted (*e.g.,* overestimation of the number of components) thus leading to an information matrix that is not invertible. To

circumvent this issue, locally conic-parametrization can be considered (Dacunha-Castelle and Gassiat (1997) and Dacunha-Castelle and Gassiat (1999)). The idea of this reparametrization is that a first positive and real parameter measures the "distance" to the true model and the second multivariate parameter defines some direction of approach to the true model. Thus, this parametrization allows the distribution of the likelihood ratio to be stated. Based on this result Keribin (2000) states the consistency the information criteria including the BIC.

If the consistency of the BIC is a blessing, it could also be a curse when its is used for selecting the model of parametric mixtures on real data. Indeed, in many cases, the parametric assumptions are not satisfied and thus, in this case, the BIC tends to overestimate the number of components. Moreover, the BIC does not take into account the objective of clustering: to provide homogeneous classes that are well-separated. To overcome these limitations, the integrated complete likelihood (Biernacki, and Govaert (2000) and Biernacki, C. and Celeux, G. and Govaert, G. (2010)) can be considered. The ICL is defined by

$$\mathrm{ICL}(\boldsymbol{m}) = \int_{\Theta_{\boldsymbol{m}}} \exp(\ell(\boldsymbol{\theta}; \mathbf{x}, \hat{\mathbf{z}}, \boldsymbol{m}))p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta})$ is the prior of $\boldsymbol{\theta}$ and $\hat{\mathbf{z}}$ is the partition given by the MAP rule when the posterior probabilities of classification are computed with the MLE. When the mixture components belong to the exponential family and when conjugate priors are used, then ICL has a closed form. Otherwise, one can consider the BIC-like approximation

$$\mathrm{ICL}_{\mathrm{approx}}(\boldsymbol{m}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}; \mathbf{x}, \boldsymbol{m}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln t_{ik}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}),$$

Semiparametric mixture models In semiparametric mixture models, the issue of model selection is generally restricted to the estimation of the number of components. Moreover, model selection is complex because there are no theoretical results on the estimator maximizing the smoothed log-likelihood. However, tools are available if the component densities are defined as a product of univariate densities (see Kasahara and Shimotsu (2014), Bonhomme, Jochmans, and Robin (2016b), Bonhomme, Jochmans, and Robin (2016a) and Kwon and Mbakop (2020)). Works of Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020) are based on the results of Allman, Matias, and Rhodes (2009) that state the identifiability result. Kasahara and Shimotsu (2014) provide an estimation of the lower bound of the number of components by considering a partition of the support of each variable (*e.g.,* using a decomposition into bins) and by using the identifiability of the latent class model (*i.e.,* mixture models where each component is a product of multinomial distributions). This discretization allows the tensor defining the probability of each event to considered, while the rank of this tensor permits a lower bound on K to derived. Note that previous works on non-parametric mixture models considered a bin decomposition (*i.e.,* a specific discretization method) to estimate non-parametric mixture models or to study their identifiability but not for model selection (see Hettmansperger and Thomas (2000), Cruz-Medina, Hettmansperger, and Thomas (2004) and Elmore, Hettmansperger, and Thomas (2004)). However, Kasahara and Shimotsu (2014) do not provide a method for selecting the discretization (*i.e.,* number of elements, locations of those elements). Thus, their method is only consistent for a lower bound of K (see Section 2.3 in Kwon and Mbakop (2020)). Alternatively, Kwon and Mbakop (2020) consider an integral operator, identified from the distribution of $\mathbf{X}_i$, that has a rank equal to K. Noting that the singular values of operators are stable under perturbations (to handle the fact that this operator is estimated from the observed sample), a thresholding rule, allowing the number of non-zero singular values to be counted, provides a consistent estimator of K. One advantage of the approach of Kwon and Mbakop (2020) is to avoid the use of

discretization, even if some connexions can be established with the approach of Kasahara and Shimotsu (2014) (see Section 2.3 in Kwon and Mbakop (2020)). One elegant property of the methods of Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020) is that both methods determine an estimator of K without performing the density estimation for different numbers of clusters and without determining ahead a maximum number of clusters (contrary to the use of the BIC for selection the number of components of a parametric mixture that requires fixing a maximum number of clusters in advance). Thus, those methods start with a step of model selection followed by the estimation of the parameters for the selected model. This is quite unusual. Indeed, when model selection is conducted for a parametric mixture model via an information criterion, parameter estimation needs to be first performed for each competing model in order to compute the information criterion. Note that the use of the identifiability results stated by Allman, Matias, and Rhodes (2009) is crucial for studying the rank of the objects considered by Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020). The approaches of Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020) are mainly based on the distribution of a pair of variables. Thus, if the number of variables J is large, computational issues can arise while considering all possible pairs of variables. It restricts the use of their methods to data sets composed of few variables. Moreover, the nature of the approach makes a variable selection impossible. In Chapter 2, we propose a new method for model selection that also permit variable selection. This method also relies on a discretization step.

Chapter 2

Feature selection in clustering

Contents

2.1 Introduction

2.1.1 State of the art

This chapter focuses on a full model selection for parametric and non-parametric mixture models (*i.e.,* estimating the number of components and the subset of the relevant variables) which is a crucial step in model-based clustering. Like any statistical method, the behavior of clustering methods can be deteriorated in high dimensions. To circumvent this issue, the analysis can be conducted by parsimonious approaches that add constraints on the parameter space. Inspired by the success of variable selection methods in regression, several authors have considered variable

selection for clustering. The main idea is to consider that only a small subset of the variables explain the true underlying clusters. Thus, selecting variables is very challenging in clustering because the role of a variable (relevant or irrelevant for clustering, see below for the definition of these roles) is defined from a variable that is not observed. Thus, the selection of the variables and the clustering need to be performed simultaneously. Note that selecting the variables in clustering has two strong benefits. First, it facilitates the interpretation of the different components as it only has to be done on the subset of discriminative variables. Second, it improves the accuracy of the estimators because it reduces the number of estimators to be considered. Indeed, by considering a n-sample arisen from a mixture of two isotropic Gaussian distributions of dimension J, Azizyan, Singh, and Wasserman (2013) show that the minimax expected loss (worst case expected loss for the best estimator) of the assignment function, is of order (ignoring constants and log terms) $\kappa^{-2}\sqrt{J/n}$ where κ is the signal to noise ratio (*i.e.,* the ratio of mean separation to standard deviation). However, if only s variables are relevant for clustering, with $s << J$, feature selection can improve the accuracy of clustering. Indeed, if the number s is known but the subset of the s relevant variables is estimated via principal component analysis, then Azizyan, Singh, and Wasserman (2013) show that that the minimax expected loss is upper bounded by $\kappa^{-1}(s^2(\ln J)/n)^{1/4}$. Thus, if s is sufficiently small, the accuracy of the estimated partition is improved by the variable selection. Note that the mixture of two isotropic Gaussian distributions makes variable selection easy but this is not the case for a general mixture (*e.g.,* it is no longer the case for a mixture of three isotropic Gaussian distributions of a mixture of two homoscedastic Gaussian distributions). Thus, feature selection in clustering requires the development of specific methods which achieve this aim via regularization methods or via model selection.

In model-based clustering, the issue of detection of the role of the variables has been introduced by Law, Figueiredo, and Jain (2004) and Tadesse, Sha, and Vannucci (2005). These authors consider that variables can be divided into two subsets: the *relevant* variables having different distributions for the mixture components and the *irrelevant* variables having the same distribution over the mixture components. In a parametric framework, a model $\boldsymbol{m}$ is defined by the number of components, by the family of the components and by the role of the variables. Since authors assume that irrelevant variables are independent of the relevant ones, the density of $\boldsymbol{x}_i$ is defined by

$$f(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \phi(\boldsymbol{x}_i^W; \boldsymbol{\gamma}, \boldsymbol{\tau}) \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i^S; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\boldsymbol{\theta}$ groups all the parameters, $\boldsymbol{x}_i^W$ corresponds to the set of irrelevant variables which follows a multivariate Gaussian distribution with mean $\boldsymbol{\gamma}$ and covariance matrix $\boldsymbol{\tau}$, and where $\boldsymbol{x}_i^S$ corresponds to the set of the relevant variables. Moreover, Law, Figueiredo, and Jain (2004) assume conditional independence, so covariance matrices $\boldsymbol{\Sigma}_k$ are diagonal, while Tadesse, Sha, and Vannucci (2005) do not impose such constraint. Considering the distribution defined by (2.1), parameter estimation can be easily achieved by maximizing the likelihood via an EM algorithm. However, model selection is challenging due to the large number of models. Indeed, considering that the maximum number of clusters is $K_{\max}$, then the number of competing models is $K_{\max}2^J$.

Raftery and Dean (2006) consider this independence assumption between the relevant and the irrelevant variables as too stringent. Thus, they introduce the notion of *redundant* variables denoted by $\boldsymbol{x}_i^U$. They consider that the redundant variables are conditionally independent of the class membership given the relevant variables. More precisely, the conditional distribution of the redundant variables follows a multivariate linear regression on all the relevant variables.

Therefore, the model density is

$$f(\boldsymbol{x}_i;\boldsymbol{m},\boldsymbol{\theta}) = \phi(\boldsymbol{x}_i^W;\boldsymbol{\gamma},\boldsymbol{\tau})\phi(\boldsymbol{x}_i^U;\mathbf{a}+\boldsymbol{x}_i^S\mathbf{b},\boldsymbol{\Omega})\sum_{k=1}^{K}\pi_k\phi(\boldsymbol{x}_i^S;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

where vector $\mathbf{a}$ corresponds to the intercepts of the linear regression, $\mathbf{b}$ is the matrix of the coefficients, and $\boldsymbol{\Omega}$ is the variance matrix of the regression. Raftery and Dean (2006) proposed performing model selection by optimizing the BIC via a deterministic procedure (backward-forward) that the authors admit being sub-optimal. This method is implemented in the R package **clustvarsel** (Scrucca and Raftery (2014)). Maugis, Celeux, and Martin-Magniette (2009a) note that the previous method involves many parameters since all the relevant variables are considered as linearly dependent of all the relevant ones. Hence, they propose detecting the predictor variables in each linear regression during the model selection. The difficult issue of the identifiability of such a method has been studied in Maugis, Celeux, and Martin-Magniette (2009a). The general form of their proposal, as presented in Maugis, Celeux, and Martin-Magniette (2009b), involves three possible roles for the variables: the relevant variables $\boldsymbol{x}_i^S$, the redundant variables $\boldsymbol{x}_i^U$ and the independent variables $\boldsymbol{x}_i^W$. Moreover, the redundant variables $\boldsymbol{x}_i^U$ are only explained by a subset $\boldsymbol{x}_i^R$ of the relevant variables, while the variables $\boldsymbol{x}_i^W$ are assumed to be independent of the relevant variables. Therefore, the model, called SRUW, has the following density

$$f(\boldsymbol{x}_i;\boldsymbol{m},\boldsymbol{\theta}) = \phi(\boldsymbol{x}_i^W;\boldsymbol{\gamma},\boldsymbol{\tau})\phi(\boldsymbol{x}_i^U;\mathbf{a}+\boldsymbol{x}_i^R\mathbf{b},\boldsymbol{\Omega})\sum_{k=1}^{K}\pi_k\phi(\boldsymbol{x}_i^S;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{2.2}$$

where the covariance matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\tau}$ can be spherical, diagonal or full. Their method improves the results of the variable selection but again complicates the difficult challenge of model selection. Since the SRUW model collection is large, two embedded backward or forward stepwise algorithms are used for model selection: one for the clustering and one for the linear regression. A backward algorithm allows one to start with all variables in order to take variable dependencies into account. A forward procedure, starting with an empty clustering variable set or a small variable subset, could be preferred for numerical reasons if the number of variables is large. Thus, the optimization procedure is deterministic and performs many model comparisons to optimize the BIC. Therefore, it can suffer from local optima and it requires many calls of EM algorithm that makes the model selection time consuming when J is large. This procedure is implemented in the C++ code called **SelvarClustIndep** (Maugis (2009)).

Regularization methods are efficient approaches for feature selection in clustering. Pan and Shen (2007) propose adapting the approach of the Lasso regression (Tibshirani (1996)) to the clustering. Thus, an ℓ_1 penalty is applied on the means of the Gaussian distributions. Obviously, this approach has to be applied on the centered variables $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}} = \frac{1}{n}(\sum_{i=1}^n x_{i1},\ldots,\sum_{i=1}^n x_{iJ})$. Moreover, the authors propose using a parsimonious Gaussian mixture model since they consider the homoscedastic model with diagonal covariance matrices: $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_K = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_J^2)$. For a fixed penalty $\lambda > 0$, the aim is to maximize the objective function $F(.;\lambda)$ with

$$F(\boldsymbol{\theta};\lambda) = \sum_{i=1}^{n}\ln\left(\sum_{k=1}^{K}\pi_k\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1)\right) - \lambda\sum_{k=1}^{K}||\boldsymbol{\mu}_k||_1.$$

This maximization is performed by an EM algorithm for different values of λ used. For each value of λ, one model is returned. Thus, the model selection is performed, among the models resulting from a specific penalty value, by the BIC. However, this criterion is not computed with

the MLE but with the estimate maximizing the objective function $F(.;\lambda)$. The approach of Pan and Shen (2007) has been extended to the heteroscedastic diagonal Gaussian mixture (Xie, Pan, and Shen (2008)), then to the general Gaussian mixture (Zhou, Pan, and Shen (2009)). For this general case, a penalty is also used for the covariance matrices. Thus, the objective function implies two penalties $\lambda_1 > 0$ and $\lambda_2 > 0$ and is defined by

$$F(\boldsymbol{\theta}; \lambda_1, \lambda_2) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) - \lambda_1 \sum_{k=1}^{K} ||\boldsymbol{\mu}_k||_1 - \lambda_2 \sum_{k=1}^{K} ||\boldsymbol{\Sigma}_k^{-1}||_1.$$

Note that the penalization is applied tp the inverse of the covariance matrices to permit the introduction of independence between variables. Meynet (2012) shows that the ℓ_1 penalty can lead to biased estimates. Thus, she proposed to use the ℓ_1 penalty method only for setting a filter among the competing models, since this procedure provides only few different models. Therefore, it is doable to perform a classical MLE inference on this model. Finally, the model selection is achieved with a BIC-like criterion where the penalty term is modified for taking the dimension of the model space into account. The R package **SelvarMix** uses the same approach to carry out the model selection when the model defined by (2.2) is considered. It performs the estimation of the model SRUW in three steps: determination of a subset of models with the ℓ_1 based method, maximum likelihood inference for each model retained by the previous step, selection of the best model with BIC criterion. Among the regularization methods, the sparse K-means (Witten and Tibshirani (2010)) is the most popular because it requires a small computational overhead and is able to manage very high-dimensional datasets. The approach uses a lasso-type penalty to select the set of variables which are relevant for the clustering. Thus, clustering and variable selection are simultaneously achieved by maximizing

$$F(\boldsymbol{\theta}, \boldsymbol{z}) = \sum_{j=1}^{J} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} (x_{ij} - x_{i'j})^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n} \sum_{i'=1}^{n} z_{ik} z_{i'k} (x_{ij} - x_{i'j})^2 \right),$$

subject to the constraints that

$$\|\boldsymbol{w}\|^2 \leq 1, \ \|\boldsymbol{w}\|_1 \leq s \text{ and } w_j \geq 0, \ \forall j.$$

Thus, the weights w_j define the impact of each variable on the partition. Sparsity is added on $\boldsymbol{w}$ by considering a suitable choice of the tuning parameter s. The authors proposed selecting a suitable value of s with an extension of the gap statistics (Tibshirani, Walther, and Hastie (2001)).

2.1.2 Framework of the chapter

In this chapter, we focus on a full model selection for parametric (see Section 2.2) and non-parametric (see Section 2.2) mixture models where no more assumption are made on the component distribution except to be defined as a product of univariate densities (see Chauveau, Hunter, and Levine (2015) for a review). Thus, we consider a sample composed of n independent observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})^\top \in \mathcal{X}$ is the vector composed of the J variables collected on subject i defined on the space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_J$ where each $\mathcal{X}_j$ is compact. Each $\mathbf{X}_i$ is identically distributed according to the mixture of K components defined by the density

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \eta_{kj}(x_{ij}), \tag{2.3}$$

where the univariate densities η_{kj} are first considered to be parametric, and then are considered to be non-parametric. Model (2.3) has been used in different fields such as statistics (Hall and

Zhou (2003), Kasahara and Shimotsu (2014), Chauveau, Hunter, and Levine (2015), Zheng and Wu (2020) and Kwon and Mbakop (2020)) but also in behavioral science Clogg (1995), econometry (Hu, McAdams, and Shum (2013) and Compiani and Kitamura (2016)) or sociology (Hagenaars and McCutcheon (2002)). One standard situation where the conditional independence assumption implied by (2.3) holds true, is in the framework of the standard repeated-measure random-effect model, where the subject-level effect is replaced by a component-level effect. Moreover, this assumption is relevant in the context of high-dimensional data that is the situation where the feature selection has the strongest impact. Note that redundant variables cannot be considered for model (2.3), as it requires modeling intra-component dependencies. In this chapter we consider only two types of variables: the relevant and the irrelevant for clustering. The model-based framework implies that the selection of the variables falls into the scope of model selection. Thus, variable j is said to be irrelevant for clustering if $\eta_{1j} = \ldots = \eta_{Kj}$. Note that in the case of parametric distributions, the equality between the η_{kj} is equivalent to the equality of their parameters. In this context, a model $\boldsymbol{m} = \{K, \boldsymbol{\Omega}\}$ is defined by the number of components K and the indexes of the relevant variables $\boldsymbol{\Omega} \subset \{1, \ldots, J\}$. Note that in the case of parametric components, the family of the distributions is supposed to be known. If different families are considered (or different parsimonious constraints), then the family of the distribution must be included in the definition of $\boldsymbol{m}$. Therefore, considering the task of full model selection in (2.3) implies that each $\mathbf{X}_i$ is identically distributed according to a non-parametric mixture of K components defined by the density

$$f(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \left(\prod_{j \in \bar{\boldsymbol{\Omega}}} \eta_{1j}(x_{ij}) \right) \left(\sum_{k=1}^{K} \pi_k \prod_{j \in \boldsymbol{\Omega}} \eta_{kj}(x_{ij}) \right), \tag{2.4}$$

where $\bar{\boldsymbol{\Omega}} = \{1, \ldots, J\} \setminus \boldsymbol{\Omega}$ contains the indices of the irrelevant variables for clustering, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{m}}$ groups all the parameters and $\boldsymbol{\Theta}_{\boldsymbol{m}}$ is the parameter space implied by $\boldsymbol{m}$. Thus, in this chapter, we define the objective of full model selection by the double objective of estimating the number of components K and the subset of relevant variables $\boldsymbol{\Omega}$, as well.

2.1.3 Contributions to variable selection in model-based clustering

In a parametric framework, variable selection in (2.4) can be performed via an information criterion but it leads to computational issues because the number of competing models is of order 2^J. As a first contribution, we propose in Marbac and Sedki (2017b), to use a new information criterion based on the integrated complete-data likelihood and named MICL. This criterion does not require the maximum likelihood estimate and its maximization appears to be simple and computationally efficient. The original contribution of our approach is to perform model selection without requiring any parameter estimation. Parameter inference is then needed only for the unique selected model. This approach is used for the variable selection of a Gaussian mixture model with conditional independence assumed. As a second contribution, we present in Marbac, Sedki, and Patin (2020), two approaches for performing variables selection in (2.4) for the context of mixed-type data. The first approach optimizes the BIC with a modified version of the standard EM algorithm that simultaneously performs a maximum likelihood estimation of the parameters and the estimation of the subset of discriminative variables, for a fixed number of components. The second method performs model selection without requiring parameter inference by maximizing the MICL criterion. As a third contribution, we implemented both approaches in the R package **VarSelLCM** (Marbac and Sedki (2020)) available on CRAN and described in Marbac and Sedki (2018). This package permits a collaboration in epidemiology (Saldanha Gomes et al. (2020)) to be established for identifying clusters of boys and girls based on diet,

sleep and activity-related behaviors and their family environment at 2 and 5 years of age, and to assess whether the clusters identified, varied across maternal education levels and were associated with body fat at age 5. Finally, our last contribution, in the parametric framework, is to extend in Marbac and Vandewalle (2019), the case of feature selection in clustering to the research of multiple partitions. Section 2.2 is devoted to the developments in the parametric context.

Note that all existing methods of variables selection in model-based clustering are restricted to parametric distributions. Thus, if the parametric assumptions are violated, bias can occur for the estimator of the number of components or on the subset of discriminative variables. In Du Roy de Chaumaray and Marbac (2021a), we address the problem of full model estimation for non-parametric finite mixture models. Section 2.3 is devoted to the developments in the non-parametric context.

2.2 Full model selection for parametric mixture models

This section presents our contribution in feature selection for model-based clustering in a parametric framework. Section 2.2.1 presents the parametric context for feature selection in clustering. Section 2.2.2 is devoted to the feature selection via the BIC presented in Marbac, Sedki, and Patin (2020). Section 2.2.3 illustrates the relevance of the procedure with the application in epidemiology that investigates the adiposity in preschool children and that was considered in Saldanha Gomes et al. (2020). Section 2.2.4 is devoted to the feature selection via the MICL presented in Marbac and Sedki (2017b). Section 2.2.5 illustrates the relevance of the procedure with application to the human population genomic considered in Marbac, Sedki, and Patin (2020) Section 2.2.6 presents the extension of the BIC and MICL approaches to cases of multiple partitions introduced in Marbac and Vandewalle (2019). Numerical applications are performed with the R package **VarSelLCM** presented in Marbac and Sedki (2018).

2.2.1 Model-based clustering for mixed-type data

Data to analyze consists of n independent observations $\mathbf{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$, where each observation $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ is defined over the space $\mathcal{X}_1 \times \ldots \times \mathcal{X}_J$, $\mathcal{X}_j$ depending on the nature of variable j. Hence, $\mathcal{X}_j = \mathbb{R}$ ($\mathbb{N}$, $\{1, \ldots, m_j\}$ respectively) if variable j is continuous (integer and categorical with m_j levels, respectively). Observations are assumed to arise independently from the parametric mixture model defined by (2.4). Thus, each distribution η_{kj} is considered to have a parametric form $f_{kj}(\cdot; \boldsymbol{\alpha}_{kj})$. The univariate marginal distribution of variable j depends on its definition space, therefore $f_{kj}(\cdot; \boldsymbol{\alpha}_{kj})$ is considered as the pdf of a Gaussian distribution $\mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$ (Poisson $\mathcal{P}(\alpha_{kj})$ and multinomial $\mathcal{M}(\alpha_{kj1}, \ldots, \alpha_{kjm_j})$) if variable j is continuous (integer and categorical, respectively) with $\boldsymbol{\alpha}_{kj} = (\mu_{kj}, \sigma_{kj})$ ($\boldsymbol{\alpha}_{kj} = \alpha_{kj}$ and $\boldsymbol{\alpha}_{kj} = (\alpha_{kj1}, \ldots, \alpha_{kjm_j})^\top$, respectively). The probability distribution function (pdf) for model $\boldsymbol{m} = \{K, \boldsymbol{\omega}\}$ and parameters $\boldsymbol{\theta}$ is defined by

$$f(\boldsymbol{x}_i; \boldsymbol{m}, \boldsymbol{\theta}) = \left(\prod_{j \in \boldsymbol{\Omega}^c} f_{1j}(x_{ij}; \boldsymbol{\alpha}_{1j}) \right) \left(\sum_{k=1}^{K} \pi_k \prod_{j \in \boldsymbol{\Omega}} f_{kj}(x_{ij}; \boldsymbol{\alpha}_{kj}) \right), \tag{2.5}$$

where $\boldsymbol{\theta}$ groups all the model parameters: *i.e.,* π_k is the proportion of component k such that $0 < \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and the parameters of component k denoted by $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1}^\top, \ldots, \boldsymbol{\alpha}_{kJ}^\top)^\top$. Considering the equalities between the parameters defined by $\boldsymbol{m}$, the observed-data log-likelihood

of model $\boldsymbol{m}$ is defined by

$$\ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x}) = \sum_{j\in\boldsymbol{\Omega}^c}\sum_{i=1}^{n}\ln f_{1j}(x_{ij};\boldsymbol{\alpha}_{1j}) + \sum_{i=1}^{n}\ln\left(\sum_{k=1}^{K}\pi_k\prod_{j\in\boldsymbol{\Omega}}f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj})\right).$$

The MLE of the parameters corresponding to the irrelevant variables are explicit, but not those of the proportions and the relevant variables. Thus, it is standard to use an EM algorithm to maximize the observed-data log-likelihood. Here, the partition among the observations is unobserved. We denote this partition by $\mathbf{z} = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)$ with $\boldsymbol{z}_i = (z_{i1},\ldots,z_{iK})$, where $z_{ik} = 1$ if observation i arises from component k and $z_{ik} = 0$ otherwise. Hence, the complete-data likelihood of model $\boldsymbol{m}$ (log-likelihood computed on the observed and unobserved variables) is defined by

$$\ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x},\mathbf{z}) = \sum_{j\in\boldsymbol{\Omega}^c}\sum_{i=1}^{n}\ln f_{1j}(x_{ij};\boldsymbol{\alpha}_{1j}) + \sum_{k=1}^{K}\sum_{i=1}^{n}z_{ik}\ln\pi_k + \sum_{j\in\boldsymbol{\Omega}}\sum_{k=1}^{K}\sum_{i=1}^{n}z_{ik}\ln f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj}).$$

The EM algorithm that gives the MLE starts from the initial value $\boldsymbol{\theta}^{[0]}$ randomly sampled and its iteration $[r]$ is defined by

E-step Computation of the fuzzy partition $t_{ik}^{[r]} := \mathbb{E}[Z_{ik}|\boldsymbol{x}_i,\boldsymbol{m},\boldsymbol{\theta}^{[r-1]}]$, hence

$$t_{ik}^{[r]} := \frac{\pi_k^{[r-1]}\prod_{j=1}^{J}f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^{K}\pi_\ell^{[r-1]}\prod_{j=1}^{J}f_{\ell j}(x_{ij};\boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

M-step Maximization of the expected value of the complete-data log-likelihood over the parameters,

$$\pi_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ and } \boldsymbol{\alpha}_{kj}^{[r]} = \begin{cases} \boldsymbol{\alpha}_{jk}^{\star[r]} & \text{if } j\in\boldsymbol{\Omega} \\ \tilde{\boldsymbol{\alpha}}_{1j} & \text{otherwise} \end{cases},$$

where $n_k^{[r]} = \sum_{i=1}^{n}t_{ik}^{[r]}$, $\tilde{\boldsymbol{\alpha}}_{1j} = \arg\max_{\boldsymbol{\alpha}_{1j}}\sum_{i=1}^{n}\ln f_{1j}(x_{ij};\boldsymbol{\alpha}_{1j})$ is the MLE for an irrelevant variable, and $\boldsymbol{\alpha}_{jk}^{\star[r]} = \arg\max_{\boldsymbol{\alpha}_{kj}}\sum_{i=1}^{n}t_{ik}^{[r]}\ln f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj})$ is the estimate for an relevant variable. This algorithm converges to a local optimum of the observed-data log-likelihood. Thus, the MLE for the model $\boldsymbol{m}$, denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$, is obtained by performing many different random initializations of $\boldsymbol{\theta}^{[0]}$.

Model selection generally aims to find the model $\hat{\boldsymbol{m}}$ which maximizes a criterion among a collection of competing models $\mathcal{M}$. The number of components of the competing models is usually bounded by a fixed value $K_{\max}$. Thus, we can define the set of the competing models by

$$\mathcal{M} = \{\boldsymbol{m} = \{K,\boldsymbol{\Omega}\} :\ K\in\{1,\ldots,K_{\max}\} \text{ and } \boldsymbol{\Omega}\subseteq\{1,\ldots,J\}\}.$$

Due to the cardinality of $\mathcal{M}$, it is no possible to use an exhaustive approach for determining the best model (*i.e.,* computation of the information criterion for each competing model). Thus, standard model-based approaches perform model selection according to the BIC via a deterministic algorithm (*e.g.,* stepwise algorithm). In the next section, we show that, for model (2.5), a specific EM algorithm can simultaneously perform variable selection according to the BIC and maximum likelihood inference, thus avoiding the issues of suboptimality and computational time of the stepwise algorithms. Then, we propose another procedure that performs variable selection without performing parameter estimation according to a criterion that considers the task of clustering.

2.2.2 Full model selection via BIC

In a Bayesian framework, the best model is the model having the greatest probability *a posteriori*. Thus, by assuming uniformity for the prior distribution of $\boldsymbol{m}$, a natural estimator of $\boldsymbol{m}$ is the model $\hat{\boldsymbol{m}}$ defined by

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}\in\mathcal{M}} p(\mathbf{x}|\boldsymbol{m}),$$

where $p(\mathbf{x}|\boldsymbol{m})$ is the integrated likelihood defined by

$$p(\mathbf{x}|\boldsymbol{m}) = \int_{\boldsymbol{\Theta}_{\boldsymbol{m}}} p(\mathbf{x}|\boldsymbol{m},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta},$$

where $p(\mathbf{x}|\boldsymbol{m},\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{m},\boldsymbol{\theta})$ is the likelihood function, and $p(\boldsymbol{\theta}|\boldsymbol{m})$ is the pdf of the prior distribution of the parameters. Unfortunately, the integrated likelihood is intractable, but many methods permit approximations to its value (Friel and Wyse (2012)). The most popular approach consists of using the BIC, which approximates the logarithm of the integrated likelihood by a Laplace approximation, and thus requires MLE. The BIC is defined by

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

where $\nu_{\boldsymbol{m}}$ is the number of independent parameters required by $\boldsymbol{m}$.

For a fixed number of components K, selecting the variables necessitates the comparison of 2^J models. Therefore, an exhaustive approach approximating the integrated likelihood for each competing model is not feasible. Instead, Raftery and Dean (2006) carry out model selection via deterministic algorithms (like a *stepwise* method). However, this approach cannot ensure that the model maximizing the BIC is obtained. Moreover, it can be computationally expensive if many variables are observed. In Marbac and Sedki (2017b) and in Marbac, Sedki, and Patin (2020), model selection is an easier problem, because the model assumes within-component independence. This assumption permits the direct maximization of any penalized log-likelihood function defined by

$$\ell_{\mathrm{pen}}(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x}) = \ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x}) - \nu_{\boldsymbol{m}}c,$$

for any constant c. This function is maximized by using a modified version of the EM algorithm Green (1990). Hence, we introduce the penalized complete-data log-likelihood function

$$\ell_{\mathrm{pen}}(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x},\mathbf{z}) = \ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x},\mathbf{z}) - (K-1)c - cK\sum_{j\in\boldsymbol{\Omega}}\nu_j - c\sum_{j\in\boldsymbol{\Omega}^c}\nu_j,$$

where ν_j is the number of parameters for one univariate marginal distribution of variable j (*i.e.,* $\nu_j = 2$ if the variable is continuous, $\nu_j = 1$ if the variable is integer and $\nu_j = m_j - 1$ if the variable is categorical with m_j levels). This modified version of the EM algorithm finds the model maximizing the penalized log-likelihood for a fixed number of components. It starts at an initial point $\{\boldsymbol{m}^{[0]},\boldsymbol{\theta}^{[0]}\}$ randomly sampled with $\boldsymbol{m}^{[0]} = \{K,\boldsymbol{\Omega}^{[0]}\}$, and its iteration $[r]$ is composed of two steps:

E-step Computation of the fuzzy partition

$$t_{ik}^{[r]} := \frac{\pi_k^{[r-1]}\prod_{j\in\boldsymbol{\Omega}^{[s]}} f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^{K}\pi_\ell^{[r-1]}\prod_{j\in\boldsymbol{\Omega}^{[s]}} f_{\ell j}(x_{ij};\boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

M-step Maximization of the expectation of the penalized complete-data log-likelihood over $\{\boldsymbol{\Omega},\boldsymbol{\theta}\}$, hence $\boldsymbol{m}^{[r]} = \{K,\boldsymbol{\omega}^{[r]}\}$ with

$$\boldsymbol{\Omega}^{[r]} = \left\{j:\ \Delta_j^{[r]} > 0\right\},\ \pi_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ and } \boldsymbol{\alpha}_{jk}^{[r]} = \begin{cases} \boldsymbol{\alpha}_{kj}^{\star[r]} & \text{if } j\in\boldsymbol{\Omega}^{[r]} \\ \tilde{\boldsymbol{\alpha}}_{kj} & \text{otherwise} \end{cases},$$

where Δ_j is the difference between the maximum of the expected value of the penalized complete-data log-likelihood obtained when variable j is relevant and when it is irrelevant such that

$$\Delta_j = \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{[r]}\big(\ln f_{kj}(x_{ij};\boldsymbol{\alpha}_{kj}^{\star[r]}) - \ln f_{1j}(x_{ij};\tilde{\boldsymbol{\alpha}}_{1j})\big) - (K-1)\nu_j c,$$

where $\boldsymbol{\alpha}_{kj}^{\star[r]}$ and $\tilde{\boldsymbol{\alpha}}_{1j}$ are defined in 2.2.1. The resulting algorithm keeps the property of monotonicity such that for any iteration $[r]$

$$\ell_{\text{pen}}(\boldsymbol{\theta}^{[r]}|\boldsymbol{m}^{[r]},\mathbf{x},\mathbf{z}) \geq \ell_{\text{pen}}(\boldsymbol{\theta}^{[r-1]}|\boldsymbol{m}^{[r-1]},\mathbf{x},\mathbf{z}).$$

Note that this algorithm can be implemented because of the assumption of independence within components. Indeed, this assumption defines the penalized log-likelihood function as a sum of independent functions which only depend on the the partition and on a single variable. Thus, its optimization can be achieved by J independent optimizations. In the case of categorical variables, identifiability requires considering at least three relevant variables (the exact relation between the maximum number of components and the number of levels of the variables is stated in Theorem 4 in Allman, Matias, and Rhodes (2009)). This constraint can be easily considered at the M-step to ensure that the estimated model is identifiable. Thus, in the case of categorical variables, we suggest defining

$$\boldsymbol{\Omega}^{[r]} = \Big\{j : \Delta_j^{[r]} > 0\Big\} \cup \Delta_{(3)}^{[r]},$$

where $\Delta_{(3)}^{[r]}$ contain the three values of $j \in \{1,\ldots,J\}$ which lead to the three greatest values of $\Delta_j^{[r]}$. To obtain the pair $\{\boldsymbol{\Omega},\boldsymbol{\theta}\}$ maximizing the penalized observed-data log-likelihood, for a fixed number of components, many random initializations of this algorithm should be performed. Hence, the pair $\{\boldsymbol{m},\boldsymbol{\theta}\}$ maximizing the penalized observed-data log-likelihood is obtained by performing this algorithm for every values of K between one and $K_{\max}$. By considering $c = (\ln n)/2$, this algorithm carries out the model selection according to the BIC. Moreover, other criteria can also be considered like the AIC by setting $c = 1$.

2.2.3 Application to investigating adiposity in preschool children

Background: Despite the growing interest in the relation between adiposity in children and different lifestyle clusters, few studies have used a longitudinal design to examine a large range of behaviors in various contexts, in particular eating-related and sleep-related routines, and few studies have examined these factors in young children. The objectives of this study were to identify clusters of boys and girls based on diet, sleep and activity-related behaviors and their family environment at 2 and 5 years of age, and to assess whether the clusters identified varied across maternal education levels and were associated with body fat at age 5.

Methods: The EDEN mother-child study is a prospective cohort designed to assess prenatal and postnatal determinants of child health and development. This cohort is composed of 2002 pregnant women (less than 24 weeks of gestation) aged 18–44 years recruited between 2003 and 2006 in two university hospitals located in Nancy and Poitiers, France. Exclusion criteria were multiple pregnancies, history of diabetes, inability to speak or read French and any plan to move out of the region within the next 3 years. A total of 1903 children were born alive and then followed up periodically by postal questionnaires and clinical examinations. At age 2 and 5, respectively 1436 and 1195 parents (usually the mother) completed a postal questionnaire, the

data of which were used to construct the obesity-related behavior clusters of children. At 5 years old, 1101 children had a full clinical examination including anthropometric measurements and bioelectrical impedance analysis (BIA). Clustering is achieved with model (2.5) by considering a full model selection is performed via the BIC according to the approach described in Section 2.2.2. Each subject is described by 44 variables (2 continuous) related to the child's diet, PA (physical activity), TV use and sleep at age 2 and 40 variables (4 continuous) variables, at age 5 (see Saldanha Gomes et al. (2020) for more details).

One analysis is performed per gender. For each gender, two independent clusterings are performed. One clustering considers the variables measured at age 2 and the second clustering considers the variables measured at age 5. To examine how clusters of children evolved from the ages of 2 to 5 years, we cross-classified them according to their cluster membership at both ages and present the proportion of children from cluster at age 2 that moved to each cluster at age 5. Based on the cross-classification, each child was assigned to a given cluster evolution path from 2 to 5 years of age. Gender-stratified linear regression models were then used to assess the association between body fat percentage at 5 years and cluster membership at age 2 and at age 5 as well as the cluster evolution path. The reference in each case was the most favorable (a priori) cluster/ evolution path. These analyses were conducted in two steps. Model 1 was adjusted for study center, exact age at the 5-year clinical examination, and predicted BMI at age 2 (longitudinal and cluster evolution path analyses only). Model 2 was further adjusted for maternal education.

Results: At age 2, the selected model contains two clusters with 15 relevant variables for boys and 17 for girls. In both genders, the most discriminant variables corresponded to intake of energy-dense nutrient-poor (EDNP) food items such as soft drinks and processed and fast foods (e.g., processed meat, pizzas, French fries and potato chips). Because the two clusters were essentially characterized by opposite eating habits, cluster 1 was labeled 'unhealthy eating' and cluster 2 'healthy eating'. Among girls, the two clusters also had contrasting TV exposure (TV watching time and TV on during meals), PA (physical activity) and sleeping habits, with low TV/PA and regular sleeping routines clustered positively with healthy eating habits. The probability of belonging to the assigned cluster exceeded 80% for 88% of the boys and 80% of the girls. Children whose mothers had a lower educational level were more likely to belong to the 'unhealthy eating' cluster (p-value $< 10^{-4}$).

At age 5, the selected model contains 2 clusters and 14 relevant variables for the boys, and 4 clusters and only 5 relevant variables for the girls. In both genders, TV exposure variables were the most discriminant. In boys, the two clusters differed mainly regarding their TV exposure and eating habits (intake of EDNP food, snacking, soft drinks at mealtimes), with high TV exposure clustered positively with unhealthy eating habits (cluster 1 was therefore labeled 'high TV–unhealthy eating' and cluster 2 labeled 'moderate TV–healthy eating'). The two clusters also differed regarding types of PA: boys in the 'high TV–unhealthy eating' cluster spent more time walking, while boys in the 'moderate TV–healthy eating' cluster were more likely to participate in organized sports activities. Membership probabilities were greater than 80% for 85% of boys. Boys whose mothers had a lower educational level were more likely to belong to the 'high TV–unhealthy eating' cluster (p-value $< 10^{-4}$). About girls, the four clusters that emerged were mainly characterized by different activity (TV/PA) patterns, with TV viewing time being by far the most discriminant variable (mean TV time across clusters ranged from 35 to 174 min). The clustering of TV exposure and PA (outdoor playing/walking) was complex, with all possible combinations of favorable and unfavorable behaviors observed; cluster 1 was labeled 'low TV– low outdoor PA', cluster 2 'moderate TV–rather high outdoor PA', cluster 3 'high TV–low outdoor PA' and cluster 4 'very high TV–high outdoor PA'. TV during meals and sweetened beverages

at mealtimes clustered positively with overall TV time. Membership probabilities exceeded 80% for more than 60% of the girls assigned to clusters 1 to 3 and for more than 80% of those assigned to cluster 4. Girls whose mothers had a low educational level were more likely to belong to the 'very high TV–high outdoor PA' cluster whereas girls with more highly educated mothers were more likely to belong to the 'low TV–low outdoor PA' cluster (p-value $< 10^{-4}$).

Figure 2.1 shows how children evolved from each cluster at age 2 into clusters at age 5. In both genders, a higher proportion of mothers of children from the 'unhealthy eating' versus the 'healthy eating' age 2 cluster did not complete the 5-year-questionnaire (29% vs. 20%; p-value $< 10^{-4}$). The mothers who did not respond had a higher rate of 'no diploma' (31% vs. 22%; p-value $< 10^{-3}$). Although the clusters differed at age 2 and age 5, children who belonged to the 'unhealthy eating' age 2 cluster were more likely to move to the age 5 clusters characterized by unhealthy eating habits and/or higher TV exposure. There was also a relatively high cross-over between predominantly favorable and unfavorable (based on eating habits and TV) clusters. For example, of the girls in the 'healthy eating' cluster at age 2, the same proportions moved to the age 5 clusters with high and with low TV exposure. The associations between cluster membership and body fat percentage at 5 years are presented in Saldanha Gomes et al. (2020). The clusters at age 2 were not significantly associated with body fat percentage at 5 years for either gender. Cross-sectional analysis at 5 years showed a significant association between cluster membership and body fat percentage only in girls, with a trend towards increasing body fat percentage that increases with increasing TV exposure across clusters. With the 'moderate TV–rather high outdoor PA' cluster as the reference, girls belonging to the 'very high TV–high outdoor PA' cluster had significantly higher body fat percentage, even after adjustment for maternal education level (+ 1.53%). Examination of the evolution of the clusters from 2 to 5 years of age showed that, compared with the girls who moved from the 'healthy eating habits' cluster at age 2 to the 'moderate TV–rather high outdoor PA' cluster at age 5 (reference group), those who moved from the 'unhealthy eating habits' cluster at age 2 to the 'very high TV–high outdoor PA' cluster at age 5 had a significantly higher body fat percentage (+1.82%) at age 5 for a given BMI at age 2, and even after adjustment for maternal education level. There was no significant association between the cluster evolution path and body fat percentage in boys.

Conclusions: At 2 years of age, two clusters emerged that were essentially characterized by opposite eating habits. At 5 years of age, TV exposure was the most distinguishing feature, but the numbers and types of clusters differed by gender. An association between cluster membership and body fat was found only in girls at 5 years of age, with girls in the cluster defined by very high TV exposure and unfavorable mealtime habits (despite high outdoor playing and walking time) having the highest body fat. Girls whose mother had low educational attainment were more likely to be in this high-risk cluster. Girls who were on a cluster evolution path corresponding to the highest TV viewing time and the least favorable mealtime habits from 2 to 5 years of age had higher body fat at 5 years of age. Efforts to decrease TV time and improve mealtime routines may hold promise for preventing overweight in young children, especially girls growing up in disadvantaged families. These preventive efforts should start as early in life as possible, ideally before the age of two, and should be sustained over the preschool years.

2.2.4 Full model selection via MICL

Although the BIC has good properties of consistency, it does not focus the clustering goal that is to provided homogeneous clusters which are well-separated. Thus, criteria based on the complete-data likelihood have been introduced such as the integrated complete-data likelihood (ICL; Biernacki, and Govaert (2000)). The ICL can be analyzed as the integrated likelihood

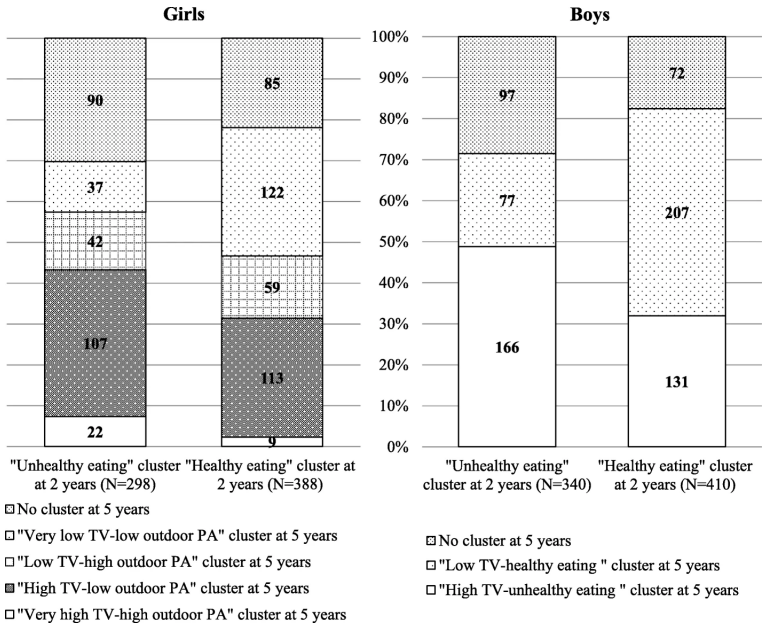


Figure 2.1: Distribution of children in clusters at 5 years among clusters at 2 years per gender.

plus a penalty term that reflects the entropy between clusters. Such criterion does not have the property of consistency (see Baudry (2015) to investigate the consistency of ICL and a comparison between BIC and ICL). However, ICL performs well in practice because it generally returns less clusters that the BIC and seems to be more robust to the misspecification of the components. Note that the BIC involves an approximation in terms that is asymptotically negligible but that can deteriorate the performances of BIC for a finite sample, especially when n is small or when $\mathcal{M}$ is large. ICL permits this issue to be circumvented, because it is an exact criteria Biernacki, C. and Celeux, G. and Govaert, G. (2010). The *integrated complete-data likelihood* is defined by

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int_{\boldsymbol{\Theta}_{\boldsymbol{m}}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}.$$

where $p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)]^{z_{ik}}$ is the complete-data likelihood. When conjugate prior distributions are used, the integrated complete-data likelihood has a closed form. Thus, we assume independence between the prior distributions, such that

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\tau}|\boldsymbol{m}) \prod_{j=1}^{J} p(\boldsymbol{\alpha}_{\bullet j}|\boldsymbol{m}),$$

with

$$p(\boldsymbol{\alpha}_{\bullet j}|K, \boldsymbol{m}) = \begin{cases} \prod_{k=1}^{K} p(\boldsymbol{\alpha}_{kj}) & \text{if } j \in \boldsymbol{\Omega} \\ p(\boldsymbol{\alpha}_{1j}) \prod_{k=1}^{K} \mathbb{1}_{\{\boldsymbol{\alpha}_{kj} = \boldsymbol{\alpha}_{1j}\}} & \text{if } j \in \boldsymbol{\Omega}^c \end{cases},$$

where $\boldsymbol{\alpha}_{\bullet j} = (\boldsymbol{\alpha}_{1j}, \ldots, \boldsymbol{\alpha}_{Kj})$. To obtain a closed-form of the integrated complete-data likelihood, we use conjugate prior distributions. Thus, $\boldsymbol{\tau}|\boldsymbol{m}$ follows a Dirichlet distribution $\mathcal{D}_g(u, \ldots, u)$. If variable j is continuous, $p(\boldsymbol{\alpha}_{kj}) = p(\sigma^2_{kj}) p(\mu_{kj}|\sigma^2_{kj})$ where σ^2_{kj} follows an Inverse-Gamma

distribution $\mathcal{IG}(a_j/2, b_j^2/2)$ and $\mu_{kj}|\boldsymbol{m}, \sigma_{kj}^2$ follows a Gaussian distribution $\mathcal{N}(c_j, \sigma_{kj}^2/d_j)$. If variable j is integer, then $\boldsymbol{\alpha}_{kj}$ follows a Gamma distribution $\mathcal{G}a(a_j, b_j)$ while $\boldsymbol{\alpha}_{kj}$ follows a Dirichlet distribution $\mathcal{D}_{m_j}(a_j, \ldots, a_j)$ if variable j is categorical with m_j levels. If there is no information *a priori* on the parameters, we use the Jeffreys non-informative prior distributions for the proportions (*i.e.,* $u_k = 1/2$) and for the hyper-parameters of a categorical variable (*i.e.,* $a_{jk} = 1/2$). Such prior distributions do not exist for the parameters of the Gaussian and Poisson distributions, so we use flat prior distributions. The conjugate prior distributions imply the following closed-form of the integrated complete-data likelihood

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^K} \frac{\prod_{k=1}^K \Gamma\left(n_k + \frac{1}{2}\right)}{\Gamma\left(n + \frac{K}{2}\right)} \prod_{j=1}^J p(\mathbf{x}_{\bullet j}|K, \omega_j, \mathbf{z}),$$

where $\mathbf{x}_{\bullet j} = (x_{ij}; i = 1, \ldots, n)$, $n_k = \sum_{i=1}^n z_{ik}$ and

$$p(\mathbf{x}_{\bullet j}|K, \omega_j, \mathbf{z}) = \int p(\boldsymbol{\alpha}_{\bullet j}|K, \omega_j) \prod_{k=1}^K \prod_{i=1}^n f_{kj}(x_{ij}; \boldsymbol{\alpha}_{kj})^{z_{ik}} d\boldsymbol{\alpha}_{\bullet j}. \tag{2.6}$$

The conjugate priors provide a closed-form of the integral defined by (2.6) and thus of the integrated complete-data likelihood (see Marbac, Sedki, and Patin (2020) for details). The value of the integrated complete-data likelihood depends whether or not j belongs to $\boldsymbol{\Omega}$. In Marbac and Sedki (2017b), we introduce the maximum integrated complete-data likelihood criterion (MICL) as the greatest value of the integrated complete-data likelihood among all the possible partitions. Thus, the MICL is defined by

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ with } \mathbf{z}^\star_{\boldsymbol{m}} = \arg\max_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}).$$

Obviously, this criterion is quite similar to the ICL because it is based on the integrated complete-data likelihood and inherits its main properties. In particular, it is less sensitive to model misspecification than the BIC. However ICL considers the partitions given by the MAP rule computed with the MLE while MICL considers the partition maximizing integrated complete-data likelihood. Thus, unlike the ICL and the BIC, MICL does not require computing the MLE for each competing model and thus avoids the multiple calls to the EM algorithm. Because $\boldsymbol{\Omega}$ does not impact the dimension of $\mathbf{z}$, we can maximize the integrated complete-data likelihood over $\{\boldsymbol{\Omega}, \mathbf{z}\}$, and thus the best model according the MICL can be obtained, for a fixed number of components by an iterative algorithm. Thus, this optimization algorithm is used for finding the model maximizing the MICL, for a fixed number of components. Starting at the initial point $\{\mathbf{z}^{[0]}, \boldsymbol{m}^{[0]}\}$ with $\boldsymbol{m}^{[0]} = \{K, \boldsymbol{\Omega}^{[0]}\}$, the algorithm alternates between two optimizations of the integrated complete-data likelihood: optimization over $\mathbf{z}$ given $\{\mathbf{x}, \boldsymbol{m}\}$, and maximization over $\boldsymbol{\Omega}$ given $\{\mathbf{x}, \mathbf{z}\}$. The algorithm is initialized as follows: each j belongs independently to $\boldsymbol{\Omega}^{[0]}$ with probability 1/2 then $\mathbf{z}^{[0]} = \hat{\mathbf{z}}_{\boldsymbol{m}^{[0]}}$ is defined as the partition provided by a MAP rule associated with model $\boldsymbol{m}^{[0]}$ and with its MLE $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}^{[0]}}$. Iteration $[r]$ of the algorithm is written as
Partition step: find $\mathbf{z}^{[r]}$ such that

$$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{m}^{[r]}).$$

Model step: find $\boldsymbol{m}^{[r+1]} = \arg\max_{\boldsymbol{\Omega}} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m})$ such that $\boldsymbol{m}^{[r+1]} = \{K, \boldsymbol{\Omega}^{[r+1]}\}$ with

$$\boldsymbol{\Omega}^{[r+1]} = \{j : p(\mathbf{x}_{\bullet j}|K, \omega_j = 1, \mathbf{z}^{[r]}) > p(\mathbf{x}_{\bullet j}|K, \omega_j = 0, \mathbf{z}^{[r]})\}.$$

At iteration $[r]$, the model step consists of finding the vector $\boldsymbol{m}^{[r+1]}$ maximizing the integrated completed-data likelihood, for the current partition $\mathbf{z}^{[r]}$. This optimization can be performed

independently for each variable $j \in \{1, \ldots, J\}$, due to the within component independence assumption. The partition step is more complex, hence $\mathbf{z}^{[r]}$ is defined as a partition increasing the value of the integrated complete-data likelihood for the current model. It is obtained by an iterative method initialized at the partition $\mathbf{z}^{[r-1]}$. Each iteration consists of uniformly sampling an individual which is affiliated with the component maximizing the integrated complete-data likelihood, while the other component memberships are unchanged (details are given in Marbac and Sedki (2017b)). Like the EM algorithm, the proposed algorithm converges to a local optimum of $\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m})$, so many different initializations should be performed. In a moderate computing time, the algorithm can manage datasets with a large number of variables and a relatively large number of individuals. However, the procedure of model selection is time-consuming if a huge number of individuals is observed. Thus, we recommend using this approach for samples composed of few observations and to use the modified EM algorithm presented previously for the large sample size.

2.2.5 Application to human population genomics

Based on the seminal work of Menozzi, Piazza, and Cavalli-Sforza (1978), principal component analysis (PCA) is widely used in population genetics to construct low-dimensional projections that summarize genetic variations across populations (see Patterson, Price, and Reich (2006), Price et al. (2006), Novembre et al. (2008) and Francois et al. (2010)). The PCA framework provides a first formal test for the presence of genetic structure in a sample of populations. Unfortunately, PCA does not attempt to classify individuals into populations and does not allow the subset of markers that contain the classification information to be selected. More importantly, PCA does not provide a satisfactory framework for estimating the true number of populations present in a dataset. Pritchard, Stephens, and Donnelly (2000) are the first to describe a model-based clustering method for using multi-locus genotype data to infer population structure and assign individuals to populations where the software implementation is given in the **STRUCTURE** software. Alexander, Novembre, and Lange (2009) developed the **ADMIXTURE** approach, which allows the parameters of a mixture model to be estimated without using the EM algorithm. **ADMIXTURE** was applied to a dataset with 13298 markers and 324 individuals and the problem of population choice was mentioned in the discussion Alexander, Novembre, and Lange (2009).

In this section, we study the genomic diversity of $n = 1318$ individuals from 35 populations of Western Central Africa Patin et al. (2017), genotyped at 690739 genetic markers (categorical variables having three levels). We restrict the analysis to $J = 160470$ independent markers ($r^2 < 0.1$, using **PLINK 1.9** Chang et al. (2015)). Clustering is achieved by considering a full model selection in (2.4) with a maximum number of clusters equals to 8 clusters. Because of the data dimension, we choose to use the MICL criterion. Indeed, it is known that BIC poorly performs when $n < d$ (see also the results on the Golub dataset in Marbac and Sedki (2017b)). For each possible number of populations, the algorithm optimizing the MICL was randomly initialized at 100 starting points. Algorithm implemented in **VarSelLCM** is parallelized and the calculations are performed on 48-(Intel(R) Xeon(R) CPU E7-8857 v2 @ 3.00GHz) cores. The full model selection procedure required about five days of computation.

Two clusters of populations were selected by MICL, separating rainforest hunter-gathering groups (RHG, derogatively called 'pygmies') from sedentary Bantu-speaking farmers (BSP). This partition is consistent with **ADMIXTURE** results (Patin et al. (2017)). A notable exception was the Bongo, a RHG population that was clustered with BSPs by the proposed method. Note that this group is known to be heavily admixed with farmers (Patin et al. (2017)). The method detected 58954 discriminative markers (37% of the observed markers). Markers were ordered

based on their discriminative power, defined as the difference between $p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}^\star)$ considering that the variable is relevant or irrelevant The discriminating power permits the most important variables for the clustering results to be detected. Table 2.1 presents the ten most discriminative markers.

Marker	Power	Chr.	Gene	Allele	$\hat{p}_{RHG}$	$\hat{p}_{BSP}$	F_{ST}	F_{ST} rank
rs2073933	117	9	*ADAMTS13*	C	0.219	0.023	0.281	2
rs10957505	108	8	*SLCO5A1*	A	0.509	0.165	0.271	4
rs1535842	107	9	*SMARCA2*	C	0.488	0.195	0.198	57
rs12440787	99	15	*MCTP2*	A	0.435	0.132	0.248	9
rs1146634	98	1	*RABGGTB*	C	0.657	0.298	0.232	13
rs916811	97	11	*CD6*	A	0.095	0.011	0.129	654
rs1352380	94	6	*TSG1*	T	0.298	0.082	0.189	82
rs7964862	94	12	*MMP17*	T	0.423	0.212	0.108	1325
rs2955032	94	15	*SPATA8*	C	0.702	0.353	0.212	32
rs675443	91	5	*CEP120*	A	0.392	0.124	0.212	33

Table 2.1: The ten genetic markers with the highest discriminating power, together with their chromosome, closest gene, their estimated allele frequency $\hat{p}$ in RHG and BSP, their population differentiation index F_{ST} and its genome-wide rank.

The power of markers to discriminate RHG from BSP was strongly correlated with F_{ST}, the population differentiation index ($r = 0.815$), a classical measure in population genetics. Nevertheless, some interesting exceptions were observed, for markers whose allele frequency (i.e., the proportion of chromosomes carrying one form of a genetic marker in the population) was systematically different in RHG populations relative to BSP, but the magnitude of this difference was typically low. Such markers are of interest for forensic sciences such as the ancestry-informative markers (Phillips (2012)) or the detection of polygenic selection between populations (Pritchard, Pickrell, and Coop (2010)).

2.2.6 Extension to the multiple partitions

We consider the problem of multivariate clustering that extends the usual framework of clustering to the case of multiple partitions. Classical clustering methods assume that the considered variables explain a single partition among the observations. However, the available data could convey more that one partition of the data. For instance, one can imagine that different blocks of variables describing a customer (variables about work, variables about leisure, variables about family, *etc*) can give different clustering/partitioning of the dataset at hand. In the absence of prior knowledge on how to group the variables into blocks, a challenging question for the statistician is to find these blocks of variables based on the data. Note that the application described in Section 2.2.3 considers a clustering with multiple partitions (one based on the variables measured at 2 years and one based on the variables measured at 5 years). However, in this case, the repartition of the variables is known and is given by the age of the subject.

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. 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. They then proposed a forward/backward search to perform model selection based on the maximization of the BIC. More recently, Galimberti, Manisi, and Soffritti (2018) have proposed an extension of their previous work 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 extension of the models proposed by Raftery and Dean (2006) and Maugis, Celeux, and Martin-Magniette (2009a), in the framework of variable selection in clustering (see model (2.2)). 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. This implies much 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 comparison of the BIC between two models. Thus, they perform many calls of the EM algorithm. Hence, these approach only can deal with a limited number of variables.

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, and the resulting models are 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 for a rich interpretation of the data, however the huge number of possibilities due to the tree structure search, makes 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, we proposed in Marbac and Vandewalle (2019), 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 within this block). This model is an extension of the approaches proposed by Marbac and Sedki (2017b) and Marbac, Sedki, and Patin (2020) 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 us to deal with categorical data while this is not possible in Galimberti, Manisi, and Soffritti (2018). The simplicity of the model allows us to estimate the repartition of the variables into blocks and the mixture parameters simultaneously as in Marbac and Sedki (2017b) and Marbac, Sedki, and Patin (2020). We present a procedure for performing model selection (choice of the number of blocks, the number of clusters inside each block and the repartition of variables into blocks) with the BIC or the MICL. The BIC enjoys consistency properties and does not require prior distributions to be defined. 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 (2017b) have proposed the MICL criterion derived from the ICL criterion. This criterion takes into account the classification purpose by computing the maximum integrated completed likelihood. Moreover, it is expected to behave well 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 clustering with multiple partitions, it is possible to simultaneously perform parameter estimation (partition estimation, respectively) and model selection with the BIC (MICL, respectively) criterion as in Marbac and Sedki (2017b) and Marbac, Sedki, and Patin (2020), thus avoiding running EM algorithms for each repartition of variables into blocks. Note that the proposed model allows mixed-data to be

dealt with as in Marbac, Sedki, and Patin (2020), and it also includes the variable selection as a special case. Moreover, the proposed model can give an answer to 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 us notice that the proposed framework has similarities with the biclustering framework, and in particular 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 as in the block clustering, our approach considers several partitions in rows. Moreover, block clustering is limited to deal with variables of the same kind assuming a homogenous distribution in each block while our approach allows us to deal with heterogeneous data.

The model

Data to cluster $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are composed of n observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ described by J variables potentially of different types (*i.e.,* each variable can be continuous, binary, count or categorical). Observations are assumed to arise independently from a multiple partitions model (MPM) which considers that variables are grouped into B independent blocks. The blocks of variables are defined by $\boldsymbol{\Omega} = \{\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_B\}$ where $\boldsymbol{\Omega}_b$ groups the indexes of variables belonging to block b. Moreover, MPM considers that variables of block b follow a K_b-component mixture assuming within-component independence. Thus, for a model $\boldsymbol{m} = \{B, \boldsymbol{K}, \boldsymbol{\Omega}\}$ with $\boldsymbol{K} = (K_1, \ldots, K_B)$, the probability distribution function (pdf) of $\boldsymbol{x}_i$ is

$$f(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{b=1}^{B} f(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}) \text{ with } f(\boldsymbol{x}_{i\{b\}}|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{K_b} \pi_{bk} \prod_{j \in \boldsymbol{\Omega}_b} f(x_{ij}|\boldsymbol{\alpha}_{jk}), \tag{2.7}$$

where , $\boldsymbol{x}_{i\{b\}} = (x_{ij}; j \in \boldsymbol{\Omega}_b)$ is the vector of observed variables of block b, $\boldsymbol{\theta} = (\boldsymbol{\pi}^\top, \boldsymbol{\alpha}^\top)^\top$ groups the model parameters, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_B^\top)^\top$ groups the proportions with $\boldsymbol{\pi}_b = (\pi_{b1}, \ldots, \pi_{bG_b})^\top$, $\pi_{bg} > 0$ and $\sum_{g=1}^{G_b} \pi_{bg} = 1$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_J^\top)^\top$ and $\boldsymbol{\alpha}_j = (\boldsymbol{\alpha}_{j1}, \ldots, \boldsymbol{\alpha}_{jK_{\omega_j}})^\top$ where ω_j indicates the block of variable j such that $\omega_j = b$ means that $j \in \boldsymbol{\Omega}_b$. The univariate margin of a component for a continuous (respectively binary, count and categorical), denoted by $f(x_{ij}|\boldsymbol{\alpha}_{jK})$, is a Gaussian (Bernoulli, Poisson and multinomial) distribution with parameters $\boldsymbol{\alpha}_{jK}$. Model (2.7) provides B partitions among the observations (one partition per block of variables). The partition of block b is denoted by $\mathbf{z}_b = (\boldsymbol{z}_{1b}, \ldots, \boldsymbol{z}_{nb}) \in \mathcal{Z}_{K_b}$, where $\mathcal{Z}_{K_b}$ is the set of the partitions of n elements in K_b clusters, and $\boldsymbol{z}_{ib} = (z_{ib1}, \ldots, z_{ibK_b})$ with $z_{ibg} = 1$ 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 $\boldsymbol{m}$ belong to $\boldsymbol{\mathcal{Z}}_{\boldsymbol{m}} = \mathcal{Z}_{K_1} \times \ldots \times \mathcal{Z}_{K_B}$.

Example 2.1. We consider $J = 4$ continuous variables arisen from MPM with $B = 2$ blocks of two variables with $\boldsymbol{\Omega}_1 = \{1, 2\}$ and $\boldsymbol{\Omega}_2 = \{3, 4\}$ (*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.,* $K_1 = K_2 = 2$) with equal proportions (*i.e.,* $\pi_{bK} = 1/2$), mean $\mu_{j1} = 4$, $\mu_{j2} = -4$ and variance $\sigma_{j1}^2 = \sigma_{j2}^2 = 1$. Figure 2.2 gives the bivariate scatter-plot of the observations. Colors and symbols indicate the component memberships of block 1 and 2 respectively.

Standard methods of clustering consider that the observed variables explain a single partition among the observations. However, if this assumption is violated, model (2.7) can circumvent this limit because it considers different partitions explained by different subsets of variables.

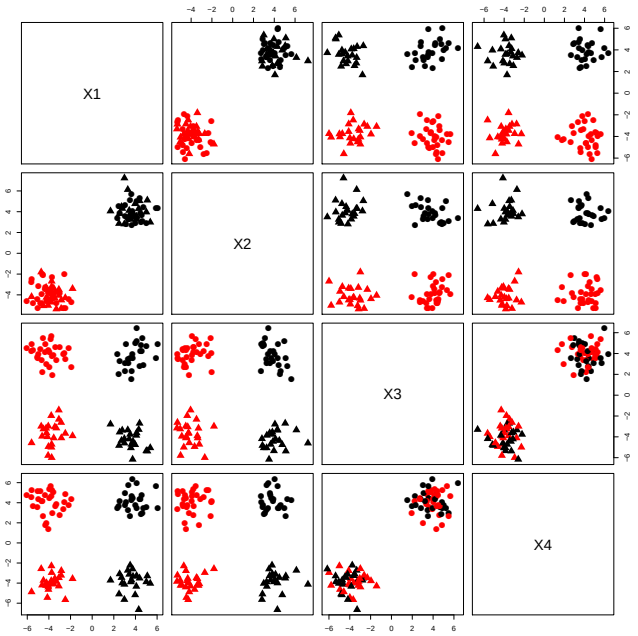


Figure 2.2: Sample generated from MPM where colors and symbols indicate the component memberships of block 1 and 2 respectively.

Moreover, (2.7) generalizes approaches used for variable selection in model-based clustering. Indeed, if $B = 2$ and $K_1 = 1$ then variables belonging to block 1 are irrelevant for the clustering, while variables belonging to block 2 are relevant. The model (2.7) 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 (2.7). Identifiability holds for blocks containing at least one continuous or integer variable (see Teicher (1963) and Teicher (1967)). Finally, identifiability holds for blocks only composed of categorical variables under mild conditions (Allman, Matias, and Rhodes (2009)).

Contrary to Galimberti and Soffritti (2007) who assume a full Gaussian covariance matrices, model (2.7) assumes that variables are independent within components. This assumption is quite standard for clustering categorical or mixed-type data (see Hand and Keming (2001) and Moustaki and Papageorgiou (2005)), and it limits the number of parameters. Hence, model (2.7) has $\nu_{\boldsymbol{m}} = \sum_{b=1}^{B}(K_b - 1) + \sum_{j \in \boldsymbol{\Omega}_b} \nu_j K_b$ 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 provides efficient approaches for model selection (see Sections 2.2.2 and 2.2.4).

Maximum likelihood inference

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

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

The complete-data log-likelihood is

$$\ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x},\mathbf{z}) = \sum_{b=1}^{B}\left(\ln p(\mathbf{z}_b|K_b,\boldsymbol{\pi}_b) + \sum_{j\in\Omega_b}\ln p(\mathbf{x}_j|K_b,\mathbf{z}_b,\boldsymbol{\alpha}_j)\right),$$

with

$$p(\mathbf{z}_b|K_b,\boldsymbol{\pi}_b) = \prod_{i=1}^{n}\prod_{k=1}^{K_b}\pi_{bk}^{z_{ibk}},$$

ands

$$p(\mathbf{x}_j|K_b,\mathbf{z}_b,\boldsymbol{\alpha}_j) = \prod_{i=1}^{n}\prod_{k=1}^{K_b}p(x_{ij};\boldsymbol{\alpha}_{jk})^{z_{ibk}},$$

where $\mathbf{x}_j = (x_{1j},\ldots,x_{nj})^\top$. The maximum likelihood estimates (MLE) can be obtained by an EM algorithm. Independence between the B blocks of variables permits the observed-data log-likelihood on each block to be maximized independently. However, here we present an EM algorithm performing the maximization of the full observed-data likelihood, because we modify later this algorithm, to simultaneously estimate $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$ in the spirit of the modified EM algorithm used to perform feature selection according to the BIC presented in Section 2.2.2. Starting from the initial value $\boldsymbol{\theta}^{[0]}$, the iteration $[r]$ of the EM algorithm maximizing the observed-data log-likelihood is composed of two steps:

E-step Computation of the fuzzy partitions $t_{ibk}^{[r]} := \mathbb{E}[Z_{ibg}|\boldsymbol{x}_{i\{b\}},\boldsymbol{m},\boldsymbol{\theta}^{[r-1]}]$, hence for $b = 1,\ldots,B$, for $k = 1,\ldots,K_b$, for $i = 1,\ldots,n$

$$t_{ibk}^{[r]} = \frac{\pi_{bk}^{[r-1]}\prod_{j\in\boldsymbol{\Omega}_b}f(x_{ij};\boldsymbol{\alpha}_{jk}^{[r-1]})}{\sum_{k=1}^{K_b}\pi_{bk}^{[r-1]}\prod_{j\in\boldsymbol{\Omega}_b}f(x_{ij};\boldsymbol{\alpha}_{jk}^{[r-1]})},$$

M-step Maximization of the expected value of the complete-data log-likelihood over $\boldsymbol{\theta}$,

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

where $Q(\boldsymbol{\alpha}_{jk};\mathbf{x}_j,\mathbf{t}_{bk}) = \sum_{i=1}^{n}t_{ibk}\ln f(x_{ij};\boldsymbol{\alpha}_{jk})$ and $n_{bk}^{[r]} = \sum_{i=1}^{n}t_{ibk}^{[r]}$.

Model selection with the BIC

The model has to be assessed from the data among a set of competing models $\mathcal{M}$ defined by

$$\begin{aligned}\mathcal{M} = \{&\boldsymbol{m} = \{B,\boldsymbol{K},\boldsymbol{\Omega}\};\\ &1\leq B\leq B_{\max},\ 1\leq K_b\leq K_{\max},\cup_{b=1}^{B}\boldsymbol{\Omega}_b = \{1,\ldots,J\},\boldsymbol{\Omega}_b\cap\boldsymbol{\Omega}_{b'} = \emptyset, 1\leq b,b'\leq B, b\neq b'\},\end{aligned}$$

where $B_{\max}$ is the maximum number of blocks and $K_{\max}$ is the maximum number of components within the block. The number of competing models is $\text{card}(\mathcal{M}) = \sum_{B=1}^{B_{\max}}S(J,B)K_{\max}^{B}$ where $S(J,B)$ denotes the Stirling number of the second kind. Model selection with the BIC consists of maximizing this criterion with respect to $\boldsymbol{m}$. Obviously, this is equivalent to maximizing the penalized likelihood on the pair $\{\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{m}}\}$. Thus, model and parameter inference leads to the search

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

Due to the number of competing models, an exhaustive approach which consists of computing BIC for each competing models, is not doable. However, holding $\{B, \boldsymbol{K}\}$ fixed, model selection with BIC and maximum likelihood inference implies maximizing the penalized likelihood with respect to $\{\boldsymbol{\Omega}, \boldsymbol{\theta}\}$. Similarily to the approach of Section 2.2.2, 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 (such as a 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 $\{B, \boldsymbol{K}\}$ allowed by $\mathcal{M}$. Therefore, for less than $\sum_{B=1}^{B_{\max}} K_{\max}^B$ different EM algorithms should be used. To implement this modified EM algorithm, we introduce the penalized complete-data likelihood

$$\begin{aligned}\ell_{pen}(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) &= \ell(\boldsymbol{\theta}_{\boldsymbol{m}}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) - \frac{\boldsymbol{\nu}_{\boldsymbol{m}}}{2} \log n \\ &= \sum_{b=1}^{B} \left(\ln p(\mathbf{z}_b|\boldsymbol{\pi}_b) - \frac{K_b - 1}{2} \ln n + \sum_{j \in \boldsymbol{\Omega}_b} \ln p(\mathbf{x}_j; \mathbf{z}_b, \boldsymbol{\alpha}_j) - \frac{\nu_j K_b}{2} \ln n \right).\end{aligned}$$

Holding $\{B, \boldsymbol{K}\}$ 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 $k = 1, \ldots, K_b$, for $i = 1, \ldots, n$

$$t_{ibk}^{[r]} = \frac{\pi_{bk}^{[r-1]} \prod_{j \in \boldsymbol{\Omega}_b^{[r-1]}} f(x_{ij}; \boldsymbol{\alpha}_{jk}^{[r-1]})}{\sum_{k=1}^{K_b} \pi_{bk}^{[r-1]} \prod_{j \in \boldsymbol{\Omega}_b^{[r-1]}} f(x_{ij}; \boldsymbol{\alpha}_{jk}^{[r-1]})},$$

M-step1 Updating the assignment of the variables to blocks

$$\boldsymbol{\Omega}_b^{[r]} = \left\{ j : b = \underset{b' \in \{1, \ldots, B\}}{\arg\max} \left(\sum_{k=1}^{G_{b'}} \max_{\boldsymbol{\alpha}_{jk} \in \Theta_j} Q(\boldsymbol{\alpha}_{jk}|\mathbf{x}_j, \mathbf{t}_{b'k}^{[r]}) - \frac{\nu_j G_{b'}}{2} \ln n \right) \right\},$$

M-step2 Updating the model parameters

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

where $\omega_j^{[r]} = b$ if $j \in \boldsymbol{\Omega}^{[r]}$. As for the standard EM algorithm, the objective function (*i.e.,* the penalized complete-data likelihood) 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.

Integrated complete-data likelihood The integrated complete-data likelihood is defined by

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

We assume independence between the prior distributions, so

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = \prod_{b=1}^{B} p(\boldsymbol{\pi}_b|K_b) \prod_{j=1}^{J} p(\boldsymbol{\alpha}_j|\boldsymbol{K}, \omega_j) \text{ where } p(\boldsymbol{\alpha}_j|\boldsymbol{K}, \omega_j) = \prod_{k=1}^{K_{\omega_j}} p(\boldsymbol{\alpha}_{jk}),$$

where $\omega_j = b$ if $j \in \boldsymbol{\Omega}_b$. Thus, the integrated complete-data likelihood has the form defined by

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

where $p(\mathbf{z}_b|K_b) = \int_{\mathcal{S}(K_b)} p(\mathbf{z}_b|K_b, \boldsymbol{\pi}_b) p(\boldsymbol{\pi}_b|K_b) d\boldsymbol{\pi}_b$, $\mathcal{S}(K_b)$ denotes the simplex of dimension K_b and $p(\mathbf{x}_j|\mathbf{z}_b, K_b) = \int_{\Theta_j^{K_b}} p(\mathbf{x}_j|\mathbf{z}_b, \boldsymbol{\alpha}_j) p(\boldsymbol{\alpha}_j|\boldsymbol{K}, \omega_j) d\boldsymbol{\alpha}_j$. We use conjugate prior distributions, thus integrals $p(\mathbf{z}_b|K_b)$ and $p(\mathbf{x}_j|\mathbf{z}_{\omega_j}, K_{\omega_j})$ have closed forms (see Marbac and Vandewalle (2019) for details). 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

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

Model selection with MICL consists of finding $\boldsymbol{m}^\star = \arg\max_{\boldsymbol{m} \in \mathcal{M}} \text{MICL}(\boldsymbol{m})$. Holding $\{B, \boldsymbol{K}\}$ fixed, maximizing MICL corresponds to maximizing the integrated complete-data likelihood with respect to the assignment of the variables into blocks $\boldsymbol{\Omega}$ and to the partition $\mathbf{z}$. Starting at the initial value $\boldsymbol{\Omega}^{[0]}$ where each variable is independently assigned to each block with the same probability $1/B$, the algorithm alternates between two steps defined at iteration $[r]$ by

Partition step: Updating the partition $\mathbf{z}_b^{[r]}$ for each block $b = 1, \ldots, B$

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

Model step: Updating the assignment of the variables to blocks

$$\boldsymbol{\Omega}_b^{[r]} = \left\{ j : b = \underset{b' \in \{1,\ldots,B\}}{\arg\max} \; p(\mathbf{x}_j|\mathbf{z}_{b'}^{[r]}, K_{b'}) \right\}.$$

Optimization at the Partition step is not obvious, despite the fact that it is done on each block independently. So, the partition $\mathbf{z}_b^{[r]}$ is defined as a partition which 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 remains (see Marbac and Vandewalle (2019) for more details). 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.

2.3 Full model selection for nonparametric mixture models

This section addresses the problem of full model estimation for non-parametric finite mixture models. It presents an approach for selecting the number of components and the subset of

discriminating variables (*i.e.,* the subset of variables having different distributions among the mixture components). The proposed approach considers a discretization of each variable into B bins and a penalization of the resulting log-likelihood. Considering that the number of bins tends to infinity as the sample size tends to infinite, we prove that our estimator of the model (number of components and subset of relevant variables for clustering) is consistent under a suitable choice of the penalty term.

2.3.1 Introduction

This section focuses on a full model selection (*i.e.,* estimation of the number of components and detection of the subset of the relevant variables for clustering) for non-parametric mixture models where no assumptions are made on the component distribution except to be defined as a product of univariate densities (see Chauveau, Hunter, and Levine (2015) for a review). Thus, we consider a sample composed of n independent observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})^\top \in \mathcal{X}$ is the vector composed of the J variables collected on subject i defined over the space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_J$ where each $\mathcal{X}_j$ is compact. Each $\mathbf{X}_i$ is identically distributed according to the non-parametric version of (2.3). Thus, the model is a mixture of K components defined by the density

$$g(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \eta_{kj}(x_{ij}), \tag{2.8}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ is a finite dimensional parameter belonging to the simplex of size K, $\mathcal{S}_K = \{\boldsymbol{u} \in [0,1]^K : \sum_{k=1}^{K} u_k = 1\}$ and where the univariate densities η_{kj} constitute infinite dimensional parameters. Among the recent developments related to (2.8), one can cite the papers of Hall and Zhou (2003), Hall et al. (2005) and Allman, Matias, and Rhodes (2009) who studied the model identifiability, while Benaglia, Chauveau, and Hunter (2009), Levine, Hunter, and Chauveau (2011) and Zheng and Wu (2020) proposed an algorithm for estimating the parameters when the number of components K is known.

In Du Roy de Chaumaray and Marbac (2021a), we addressed the issue of full model selection (*i.e.,* double objective of estimating the number of components K and the subset of relevant variables $\boldsymbol{\Omega}$ as well) in non-parametric mixture models defined by (2.8). To the best of our knowledge, this paper presents the first method that permits a full-model selection (*i.e.,* estimation of K and $\boldsymbol{\Omega}$) for non-parametric multivariate mixture models. Moreover, it allows many variables to be managed, which makes it a complementary work to Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020), even in the case where all the variables are considered to be relevant and only the number of components needs to be estimated. As proposed by Tadesse, Sha, and Vannucci (2005), we consider two types of variables: the relevant variables and the irrelevant variables for clustering. Thus, variable j is said to be irrelevant for clustering if $\eta_{1j} = \ldots = \eta_{Kj}$ and a model $\boldsymbol{m} = \{K, \boldsymbol{\Omega}\}$ is defined by the number of components K and the indices of the relevant variables $\boldsymbol{\Omega} \subset \{1, \ldots, J\}$. Therefore, considering the task of full model selection in (2.8) implies that each $\mathbf{X}_i$ is identically distributed according to a non-parametric mixture of K components defined by the density

$$g_{\boldsymbol{m},\boldsymbol{\psi}}(\boldsymbol{x}_i) = \left(\prod_{j \in \bar{\boldsymbol{\Omega}}} \eta_{1j}(x_j) \right) \left(\sum_{k=1}^{K} \pi_k \prod_{j \in \boldsymbol{\Omega}} \eta_{kj}(x_j) \right), \tag{2.9}$$

where $\bar{\boldsymbol{\Omega}} = \{1, \ldots, J\} \setminus \boldsymbol{\Omega}$ contains the indices of the irrelevant variables for clustering and $\boldsymbol{\psi} \in \Psi_{\boldsymbol{m}}$ groups the finite dimensional parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top \in \mathcal{S}_K$ and the infinite

dimensional parameters composed of the univariate densities η_{kj}. To achieve the full model selection, we use a discretization of each continuous variable into B bins. The number of bins tends to infinite with the sample size to ensure the consistency of the approach. Indeed, if B were fixed, the estimated model could be a sub-model of the true model (*i.e.,* the number of components and the subset of the discriminative variables could be underestimated). The distribution of the resulting discretized variables follows a latent class model where each component is a product of multinomial distributions (Goodman (1974)). This discretization is convenient, because model selection can then be achieved, for the latent class model, by using the penalized likelihood (*e.g.,* BIC) whose consistency has been proven for mixture models (Keribin (2000)). Moreover, in this framework, a specific EM algorithm optimizing the penalized likelihood can be used for simultaneously detecting the subset of the relevant variables and estimating the model parameters (see Marbac, Sedki, and Patin (2020)), for a known number of components. Thus, by considering an upper-bound of the number of components, the full-model selection can be achieved. Unlike in Kasahara and Shimotsu (2014), the procedure provides a consistent estimation of the univariate densities of the components η_{kj} from the discretized data. Therefore, we prove the consistency of the procedure for a wide range of number of bins B, at an appropriate rate that we detail. The consistency of the procedure cannot be proven by using the consistency of information criterion for parametric mixture models (see Keribin (2000)) because the parameters space depends on B and thus increases with sample size. The growth rate of B is mainly driven to avoid underestimation of the model while the range of the penalty is mainly driven to avoid overestimation of the model. The case of model underestimation is analyzed by extending the proof of Keribin (2000) in order to deal with the increasing dimension of the parameters space. In the case of model overestimation, the asymptotic distribution of the likelihood ratio is investigated by performing a locally conic parametrization (see Dacunha-Castelle and Gassiat (1997) and Dacunha-Castelle and Gassiat (1999)) of the model obtained on the discretized data. An upper bound of the likelihood ratio is obtained by controlling, on the one hand, the deviation of the likelihood ratio from its asymptotic distribution by using results on empirical processes stated in Chernozhukov, Chetverikov, and Kato (2014) and, on the other hand, the supremum of the asymptotic distribution by applying deviation results on Gaussian processes (see Dudley (2014)).

The method proposed in Du Roy de Chaumaray and Marbac (2021a) uses a discretization that provides an estimator of the densities of the components. However, we advice to use the proposed approach only for model estimation. When the model has been selected, we advice to use a kernel-based method for density estimation. Indeed, the bin-density estimate that are known to be outperformed by kernel-based estimators. Thus, for a real data analysis, we advice to use the proposed approach for model selection then, for the selected model, to perform density estimation with a EM-like algorithm (Benaglia, Chauveau, and Hunter (2009)) or by maximizing the smoothed log-likelihood (Levine, Hunter, and Chauveau (2011)). The final partition is thus computed from the model selected by the proposed methods and the densities estimated via a kernel method.

2.3.2 Model selection by bin estimation and penalized log-likelihood

This section considers the estimation of the number of components for model (2.9), from an n-sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X}_i \in \mathcal{X}$ with $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_J$, J being fixed. The method used for selecting the number of components discretizes each variable into B non-overlapping bins $I_{Bj1}, \ldots, I_{BjB}$ such that $\cup_{b=1}^{B} I_{Bjb} = \mathcal{X}_j$ and for any (b, b') with $b \neq b'$, $I_{Bjb} \cap I_{Bjb'} = \emptyset$. Thus, we consider the function σ_{Bjb} with $b \in \{1, \ldots, B\}$, such that $\sigma_{Bjb}(x_{ij}) = 1$ if $x_{ij} \in I_{jb}$ and $\sigma_{Bjb}(x_{ij}) = 0$ if $x_{ij} \notin I_{Bjb}$, and we denote by l_{Bjb} the size of the bin I_{Bjb}. The discretized

variables follow a latent class model where each component is a product of J multinomial distributions each having B levels. Therefore, the pdf of the discretized subject i is

$$f_{\boldsymbol{m},B,\boldsymbol{\theta}}(\boldsymbol{x}_i) = \prod_{j\in\boldsymbol{\Omega}}\prod_{b=1}^{B}\left(\frac{\alpha_{B1jb}}{l_{Bjb}}\right)^{\sigma_{Bjb}(x_{ij})}\left(\sum_{k=1}^{K}\pi_k\prod_{j\in\boldsymbol{\Omega}}\prod_{b=1}^{B}\left(\frac{\alpha_{Bkjb}}{l_{Bjb}}\right)^{\sigma_{Bjb}(x_{ij})}\right), \tag{2.10}$$

where $\boldsymbol{\theta}$ groups the component proportions π_k and the probabilities α_{Bkjb} that one subject arisen from component k takes level b for the variable j when this variable is discretized into B bins. The parameter space is given by the simplexes $S_K \times S_B^{K|\boldsymbol{\Omega}|+(J-|\boldsymbol{\Omega}|)}$, where $|\boldsymbol{\Omega}|$ denotes the cardinal of the set of discriminative variables $\boldsymbol{\Omega}$. Note that $f_{\boldsymbol{m},B,\boldsymbol{\theta}}$ is an approximation of $g_{\boldsymbol{m},\boldsymbol{\psi}}$ and that this approximation becomes more accurate when B tends to infinity.

The probabilities α_{Bkjb} are unknown and must be estimated from the observed sample. This estimation can be achieved by maximizing the log-likelihood defined by

$$\ell_n(f_{\boldsymbol{m},B,\boldsymbol{\theta}}) = \sum_{i=1}^{n}\ln f_{\boldsymbol{m},B,\boldsymbol{\theta}}(\boldsymbol{x}_i).$$

The maximum likelihood statistics for a model with K components and B bins per variable is

$$T_{n,\boldsymbol{m},B} = \sup_{\boldsymbol{\theta}\in\Theta_{\boldsymbol{m},B,\varepsilon}}\ell_n(f_{\boldsymbol{m},B,\boldsymbol{\theta}}),$$

where, in order to avoid numerical issues, we introduced a threshold ε such that the parameter space becomes $\Theta_{\boldsymbol{m},B,\varepsilon} = S_{K,\varepsilon} \times S_{B,\varepsilon}^{K|\boldsymbol{\Omega}|+(J-|\boldsymbol{\Omega}|)}$, with $\varepsilon > 0$ being the minimal value of all the elements defined in the simplexes, *i.e.* $S_{B,\varepsilon} = \{\boldsymbol{u} \in \mathbb{R}^B : \ u_b > \varepsilon, \ \sum_{b=1}^{B} u_b = 1\}$. Under the condition that $B\varepsilon$ tends to zero as B goes to infinity and ε to zero, the parameter space $\Theta_{\boldsymbol{m},B,\varepsilon}$ converges to the whole parameter space. Note that, due to the growth rate of B which will be stated by Assumption 2.4(i) in the next section, it is sufficient to set $\varepsilon^{-1} = O(n^{\alpha+1})$ for some $\alpha > 0$. This maximization can be achieved via an EM algorithm.

The penalized likelihood is defined by substracting from the maximum likelihood statistics a penalty term $a_{n,\boldsymbol{m},B}$ which takes into account the sample size and the complexity of model (2.10). Thus, we obtain the following information criterion

$$W_{n,\boldsymbol{m},B} = T_{n,\boldsymbol{m},B} - a_{n,\boldsymbol{m},B}. \tag{2.11}$$

Depending on the choice of $a_{n,\boldsymbol{m},B}$ in (2.11), different well-known criteria can be considered. Among them one can cite the Akaike criterion (AIC; Akaike (1970)) or the Bayesian Information Criterion (BIC; Schwarz (1978)) which are obtained with $a_{n,\boldsymbol{m},B} = \nu$ and $a_{n,\boldsymbol{m},B} = \nu\log(n)/2$ respectively, where $\nu = (K-1) + KJ(B-1)$ is the model complexity.

To select the number of components, we consider the set of competing models $\mathcal{M}$ defined by all the mixture models with at most $K_{\max}$ components and at least three relevant variables (for identifiability reasons), so that

$$\mathcal{M} = \{\boldsymbol{m} = \{K, \Omega\} : K \leq K_{\max}, \boldsymbol{\Omega} \subseteq \{1,\ldots,J\} \text{ and } |\boldsymbol{\Omega}| \geq 3\}.$$

The estimator $\widehat{\boldsymbol{m}}_{n,B}$ of the number of components maximizes the penalized likelihood as follows

$$\widehat{\boldsymbol{m}}_{n,B} = \arg\max_{\boldsymbol{m}\in\mathcal{M}} W_{n,\boldsymbol{m},B}.$$

The study of the asymptotic properties of the estimator $\widehat{\boldsymbol{m}}_{n,B}$ is covered by the approach of Keribin (2000) if the number of intervals B does not increase with the sample size n. However,

due to the discretization, the approach provides an estimator that converges to a model included into the true model. Indeed, we only obtain a lower bound on the number of components and a subset of the discriminative variables. By increasing the number of intervals with n, we avoid the issues due to the loss of identifiability. However, we need to investigate the behavior of the statistics $T_{n,\boldsymbol{m},B}$ and to study the convergence of $T_{n,\boldsymbol{m},B}/n$ to the minimum Kullback divergence, which requires controlling empirical processes defined on space having increasing dimension. The next section presents statistical guarantees of the proposed approach.

2.3.3 Convergence in probability of the estimator

This section investigates the convergence in probability of $\widehat{\boldsymbol{m}}_{n,B}$. It starts by presenting the assumptions required to obtain this convergence, which is then stated.

Assumptions The consistency of the estimator is established under four sets of assumptions described below. Assumption 2.1 and Assumption 2.2 state the constraints on the model and on the distribution of the components respectively. Assumption 2.3 gives some conditions on the penalty term. Finally, Assumption 2.4 gives some conditions on the discretization.

Assumption 2.1. The number of variables is at least three (*i.e.,* $3 \leq J$) and each proportion $\pi_k > 0$ is not zero. Moreover, there exists $\Upsilon \subseteq \{1, \ldots, J\}$ such that $|\Upsilon| = 3$ and for any $j \in \Upsilon$ the univariate densities η_{kj} are linearly independent.

Assumption 2.2. (i) There exists a function τ in $L_1(g_0\nu)$ such that: $\forall \boldsymbol{m} \in \mathcal{M}$ and $\forall \boldsymbol{\psi} \in \Psi_{\boldsymbol{m}}$, $|\ln g_{\boldsymbol{m},\boldsymbol{\psi}}| < \tau$ ν-a.e.

(ii) There exists a positive constant $L < \infty$ such that $\forall j \in \{1, \ldots, J\}$ and $\forall x_j \in \mathcal{X}_j$, $|\eta'_{kj}(x_j)| \leq L$.

(iii) Each variable j is defined on a compact space $\mathcal{X}_j$ and its density for each component k, denoted by η_{kj}, are strictly positive except on a set of Lebesgue measure zero.

Assumption 2.3. (i) $a_{n,\boldsymbol{m},B}$ is an increasing function of K, $|\Omega|$ and B.

(ii) For any model $\boldsymbol{m}$, $a_{n,\boldsymbol{m},B}/n$ tends to 0 as n tends to infinity.

(iii) For any model $\boldsymbol{m}$, $B/a_{n,\boldsymbol{m},B}$ tends to 0 as n tends to infinity.

(iv) For any models $\boldsymbol{m}$ and $\widetilde{\boldsymbol{m}}$ with $\boldsymbol{m} \subset \widetilde{\boldsymbol{m}}$, $a_{n,\widetilde{\boldsymbol{m}},B}/a_{n,\boldsymbol{m},B}$ tends to infinity as n tends to infinity.

Assumption 2.4. (i) The number of bins B tends to infinity with n in the following way $\lim_{n\to\infty} B = \infty$ and $\lim_{n\to\infty} B(\ln^3 n)/n = 0$.

(ii) The length of the each interval is not zero and satisfies, for all $j \in \{1, \ldots, J\}$ and $b \in \{1, \ldots, B\}$, $l_{Bjb}^{-1} = O(B)$.

(iii) Let $\mathcal{I}_{jB}$ be the set of the upper bounds of the B intervals, then, for any value $x_j \in \mathcal{X}_j$, $d(x_j, \mathcal{I}_{jB})$ tends to zero as B tends to infinity.

Assumption 2.1 is derived from the conditions of identifiability for finite mixtures of non-parametric measure products (see Theorems 8 and 9 in Allman, Matias, and Rhodes (2009)). Because Theorems 8 and 9 in Allman, Matias, and Rhodes (2009) consider all the variables as relevant for clustering, we need to extend their assumptions such that there are at least three relevant variables to obtain the identifiability of the model (2.4).

Assumption 2.2 gives sufficient conditions on the component distributions to ensure that the results of Dacunha-Castelle and Gassiat (1999) can be applied to the mixture model obtained after discretization.

Assumption 2.3 presents standard conditions for penalized likelihood model selection in the case of embedding models. It generalizes the usual conditions for selecting the number of components (Keribin (2000) and Chambaz (2006)) to the case of feature selection for mixture models. Conditions (i) and (iii) permit avoiding the overestimation of the model (*i.e.,* overestimation of the number of components or of the support of the relevant variables), while condition (ii) permits avoiding the underestimation of the model by making the penalty term negligible with respect to the model bias. Note that Assumption 2.3 allows the BIC penalty to be considered.

Even if Assumption 2.1 provides the identifiability of model (2.4), after the discretization, model (2.10) could be not identifiable if the number of intervals B is fixed. As an example, one can consider a bi-component mixture model with equal proportions defined such that the first component follows a product of $J \geq 3$ beta distributions $\mathcal{B}e(\alpha, \alpha)$ and that the second component follows a product $J \geq 3$ of beta distributions $\mathcal{B}e(2\alpha, 2\alpha)$, with $\alpha \geq 1$. This model is identifiable but the model (2.10) defined after the discretization of each variables into two bins of equal size (*e.g.,* for any j, $\sigma_{j1}(u) = 1$ if $0 \leq u \leq 1/2$, $\sigma_{j1}(u) = 0$ if $1/2 < u \leq 1$ and $\sigma_{j2}(u) = 1 - \sigma_{j1}$) is not identifiable (*i.e.,* the two mixture components follows the same distribution for the discretized data). However, if the number of bins is strictly more than two and that each interval has a length that is not zero, then the model (2.10) becomes identifiable.

The model identifiability is obtained by Assumption 2.4 that states conditions on the discretization. In particular, the number of levels has to tend to infinity when the sample size increases such that the size of the largest intervals tends to zero when the sample size tends to infinity, but its growth rate is upper bounded which is a key point to control the convergence of the estimators. Note that Assumption 2.4(iii) uses the same ideas as Lemma 17 in Allman, Matias, and Rhodes (2009) and that this condition is not stringent. For instance the bounds of the intervals can be determined by the quantiles $1/B, \ldots, B/B$. In addition, the sizes l_{jb} can vary from one bin to another. This is for instance the case when we consider the quantiles. However, we cannot allow a bin to be exponentially small with n, in order to keep the asymptotic behavior of our estimator which is stated in the next subsection. Note that Assumption 2.4 allows to consider the rate $B = n^{1/3}$ that is usual for bin-density estimation.

Finally, the assumption on the compactness of $\mathcal{X}_j$ can be relaxed if some densities defined on $\mathcal{R}$ are wanted to be considered. In such case, the estimates of the densities are considered on the compact $[\min_i x_{ij}, \max_i x_{ij}]$ defined from the observed sample, and the estimates of the densities are zero outside this interval.

Convergence in probability of the estimator We state the consistency of the estimator $\widehat{M}_{n,B}$ then we give its (sketch of) proof. Note that the proof of all the numbered equations are given in Du Roy de Chaumaray and Marbac (2021a). Finally, we explain the key points of the proof which are different from the proof of the consistency of information criteria for parametric mixture models stated in Keribin (2000).

Theorem 2.1. *Assume that independent data arise from* (2.9) *with the true model* $\boldsymbol{m}_0 = \{K_0, \boldsymbol{\Omega}_0\}$*, that Assumptions 2.1, 2.2, 2.3 and 2.4 hold true, and that the set of competing models* $\boldsymbol{m}$ *is defined with a known upper bound for the number of clusters* $K_{\max}$*.* $\widehat{\boldsymbol{m}}_{n,B}$ *then converges in probability to* $\boldsymbol{m}_0$*.*

Proof of Theorem 2.1. The proof is divided into three parts: the case where $\boldsymbol{m}_0$ is underestimated (*i.e.,* $K < K_0$ or $\boldsymbol{\Omega}_0 \not\subseteq \boldsymbol{\Omega}$), the case where the subset of the relevant variables is

overestimated with K_0 (*i.e.,* $K = K_0$ and $\mathbf{\Omega}_0 \subsetneq \mathbf{\Omega}$) and the case where the number of components and the subset of relevant variables are overestimated (*i.e.,* $K > K_0$ and $\mathbf{\Omega}_0 \subseteq \mathbf{\Omega}$).
• Part 1: We consider the case where $\boldsymbol{m}_0$ is underestimated. Thus, we consider a model $\boldsymbol{m} \in \mathcal{N}_1$ where

$$\mathcal{N}_1 = \{\boldsymbol{m} = \{K, \mathbf{\Omega}\} : K \leq K_0 \text{ or } \mathbf{\Omega}_0 \not\subseteq \mathbf{\Omega}\}.$$

For any $g_{\boldsymbol{m},\boldsymbol{\psi}}$ given by model (2.9), Assumption 2.2(i) implies that $\mathbb{E}_{g_0}[\ln g_{\boldsymbol{m},\boldsymbol{\psi}}]$ is defined. The Kullback-Leibler divergence from model $\boldsymbol{m}$ to the true distribution g_0 is defined by

$$\mathrm{KL}(g_0, \mathcal{G}_{\boldsymbol{m}}) := \inf_{\boldsymbol{\psi} \in \Psi_{\boldsymbol{m}}} \mathbb{E}_{g_0}\left[\ln \frac{g_0}{g_{\boldsymbol{m},\boldsymbol{\psi}}}\right].$$

Using the definition of $\mathcal{N}_1$ and the identifiability of g_0 (ensured by Assumption 2.1), for each $\boldsymbol{m} \in \mathcal{N}_1$, there exists some $\delta_{\boldsymbol{m}} > 0$ such that $\mathrm{KL}(g_0, \mathcal{G}_{\boldsymbol{m}}) \geq \delta_{\boldsymbol{m}}$. In Du Roy de Chaumaray and Marbac (2021a), we prove the following convergence in probability:

$$\frac{1}{n}\left(T_{n,\boldsymbol{m},B} - \ell_n(g_0)\right) = -\delta_{\boldsymbol{m}} + o_{\mathbb{P}}(1). \tag{2.12}$$

This convergence and the properties of the penalty (see Assumption 2.3) imply that for any $\boldsymbol{m} \in \mathcal{N}_1$

$$\frac{1}{n}\left(W_{n,\boldsymbol{m},B} - W_{n,\boldsymbol{m}_0,B}\right) \leq -\delta_{\boldsymbol{m}} + o_{\mathbb{P}}(1).$$

Therefore, noting that $\delta_{\boldsymbol{m}} > 0$ and that the cardinal of $\mathcal{N}_1$ is fixed and finite, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\boldsymbol{m}}_{n,B} \in \mathcal{N}_1) = 0. \tag{2.13}$$

Thus, the probability of underestimating the model tends to zero as n tends to infinity.
• Part 2: We consider the case where the number of components is correct but the subset of the relevant variables is overestimated. Thus, we consider a model $\boldsymbol{m} \in \mathcal{N}_2$ where

$$\mathcal{N}_2 = \{\boldsymbol{m} = \{K, \mathbf{\Omega}\} : K = K_0 \text{ and } \mathbf{\Omega}_0 \subsetneq \mathbf{\Omega}\}.$$

We have the following upper-bound

$$\mathbb{P}(\widehat{\boldsymbol{m}} \in \mathcal{N}_2) \leq \sum_{\boldsymbol{m}\in\mathcal{N}_2} \mathbb{P}(W_{n,\boldsymbol{m},B} \geq W_{n,\boldsymbol{m}_0,B}) = \sum_{\boldsymbol{m}\in\mathcal{N}_2} \mathbb{P}\left(\frac{T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}}{a_{n,\boldsymbol{m}_0,B}} \geq \frac{a_{n,\boldsymbol{m},B}}{a_{n,\boldsymbol{m}_0,B}} - 1\right).$$

Using usual results on likelihood ratio, for a fixed value of B, $2\left(T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}\right)$ is asymptotically distributed like a $\chi^2(\Delta)$ where $\Delta = (B-1)(K-1)(|\mathbf{\Omega}| - |\mathbf{\Omega}_0|)$. As Δ goes to infinity with B, thus with n, we have the following asymptotic distribution

$$\frac{1}{\sqrt{2\Delta}}\left[2\left(T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}\right) - \Delta\right] \xrightarrow{d} \mathcal{N}(0,1).$$

We rewrite

$$\frac{T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}}{a_{n,\boldsymbol{m}_0,B}} = \frac{1}{a_{n,\boldsymbol{m}_0,B}}\sqrt{\frac{\Delta}{2}}\left(\frac{1}{\sqrt{2\Delta}}\left[2\left(T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}\right) - \Delta\right]\right) + \frac{\Delta}{2a_{n,\boldsymbol{m}_0,B}},$$

and conclude, by making use of Slutsky's lemma and Assumption 2.3 (iii), that

$$\frac{T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}}{a_{n,\boldsymbol{m}_0,B}} = o_{\mathbb{P}}(1).$$

For any $\boldsymbol{m} \in \mathcal{N}_2$, Assumption 2.3 (iv) implies that $a_{n,\boldsymbol{m},B}/a_{n,\boldsymbol{m}_0,B} - 1 > 0$, thus, as the cardinal of $\mathcal{N}_2$ is finite and does not depend on B, we can conclude that

$$\mathbb{P}(\widehat{\boldsymbol{m}} \in \mathcal{N}_2) = 0. \tag{2.14}$$

• Part 3: We consider the case where the number of components and the subset of the relevant variables are overestimated. Thus, we consider a model $\boldsymbol{m} \in \mathcal{N}_3$ where

$$\mathcal{N}_3 = \{\boldsymbol{m} = \{K, \boldsymbol{\Omega}\} : K > K_0 \text{ and } \boldsymbol{\Omega}_0 \subseteq \boldsymbol{\Omega}\}.$$

Note that $\mathcal{N}_3 = \mathcal{M} \setminus \{\mathcal{N}_1 \cup \mathcal{N}_2 \cup M_0\}$. The probability of overestimating the model (*i.e.,* $\widehat{\boldsymbol{m}} \in \mathcal{N}_3$) can be upper-bounded by

$$\mathbb{P}(\widehat{\boldsymbol{m}} \in \mathcal{N}_3) \leq \sum_{\boldsymbol{m} \in \mathcal{N}_3} \mathbb{P}(W_{n,\boldsymbol{m},B} \geq W_{n,\boldsymbol{m}_0,B}).$$

Note that for any $\boldsymbol{m} \in \mathcal{N}_3$, we have $\delta_{\boldsymbol{m}} = 0$ and thus the reasoning used to demonstrate that $\boldsymbol{m}_0$ is not underestimated cannot be used. We have, for any $\boldsymbol{m} \in \mathcal{N}_3$

$$\mathbb{P}\left(W_{n,\boldsymbol{m},B} \geq W_{n,\boldsymbol{m}_0,B}\right) = \mathbb{P}\left(\frac{T_{n,\boldsymbol{m},B} - T_{n,\boldsymbol{m}_0,B}}{a_{n,\boldsymbol{m}_0,B}} \geq \frac{a_{n,\boldsymbol{m},B}}{a_{n,\boldsymbol{m}_0,B}} - 1\right).$$

Applying the locally-conic parametrization proposed by Dacunha-Castelle and Gassiat (1997) and Dacunha-Castelle and Gassiat (1999) on model (2.10), and noting that Assumption 2.2 holds true, we can rewrite the log-likelihood ratio as in the proof of Lemma 3.3 in Keribin (2000)

$$\begin{aligned} &T_{n,\boldsymbol{m},B} - \ell_n f_{0B} \\ &\quad = \sup\left\{ \sup_{d \in D_B} \frac{1}{2}\mathcal{G}_n^2(d)\mathbb{1}_{\mathcal{G}_n(d)\geq 0}; \sup_{d_1 \in D_{1B}, d_2 \in D_{2B}} \frac{1}{2}\left(\mathcal{G}_n^2(d_1) + \mathcal{G}_n^2(d_2)\mathbb{1}_{\mathcal{G}_n(d_2)\geq 0}\right) \right\}(1 + o_{\mathbb{P}}(1)) \end{aligned} \tag{2.15}$$

where, for each function d, $\mathcal{G}_n(d) = n^{-1/2}\sum_{i=1}^n d(X_i)$; the considered spaces of functions as well as the definition of f_{0B} are detailed in Du Roy de Chaumaray and Marbac (2021a). Note that

$$\sup_{d \in D_B} \frac{1}{2}\mathcal{G}_n^2(d)\mathbb{1}_{\mathcal{G}_n(d)\geq 0} = \frac{1}{2}\left(\sup_{d \in D_B} \mathcal{G}_n(d)\right)^2.$$

In addition, as D_{1B} and D_{2B} are subspaces of D_B, we have

$$\sup_{d_1 \in D_{1B}, d_2 \in D_{2B}} \frac{1}{2}\left(\mathcal{G}_n^2(d) + \mathcal{G}_n^2(d)\mathbb{1}_{\mathcal{G}_n(d)\geq 0}\right) \leq \left(\sup_{d \in D_{B,s}} \mathcal{G}_n(d)\right)^2,$$

where $D_{B,s}$ is the symmetrized space $D_B \cup (-D_B)$. Therefore, we deduce that

$$T_{n,M,B} - \ell_n f_{0B} \leq \left(\sup_{d \in D_{B,s}} \mathcal{G}_n(d)\right)^2 (1 + o_{\mathbb{P}}(1)).$$

Thus, using the fact that $D_{B,s}$ is a symmetric space, we obtain that, for any $\varepsilon > 0$, for n sufficiently large,

$$\left\{\frac{T_{n,\boldsymbol{m},B} - \ell_n f_{0,B}}{a_{n,\boldsymbol{m}_0,B}} > 4\varepsilon\right\} \subset \left\{\left|\sup_{d \in D_{B,s}} \mathcal{G}_n(d)\right| > 2\sqrt{\varepsilon a_{n,\boldsymbol{m}_0,B}}\right\}.$$

It implies that,

$$\mathbb{P}\left(\frac{T_{n,\boldsymbol{m},B}-\ell_n f_{0,B}}{a_{n,\boldsymbol{m}_0,B}}>\varepsilon\right)\leq$$
$$\mathbb{P}\left(\sup_{d\in D_{B,s}}\xi_d>\sqrt{\varepsilon a_{n,M_0,B}}\right)+\mathbb{P}\left(\left|\sup_{d\in D_{B,s}}\mathcal{G}_n(d)-\sup_{d\in D_{B,s}}\xi_d\right|>\sqrt{\varepsilon a_{n,\boldsymbol{m}_0,B}}\right),$$

where $(\xi_d)_d$ is a Gaussian process indexed by $D_{B,s}$, with covariance the usual Hilbertian product on L^2. Note that under our Assumptions, for a fixed value $B^\star$ of B, $\sup_{d\in D_{B^\star,s}}\mathcal{G}_n(d)$ converges in distribution to $\sup_{d\in D_{B^\star,s}}\xi_d$ (see Du Roy de Chaumaray and Marbac (2021a)) as n goes to infinity. However, as in our context, B goes to infinity with n, we need to control its influence on the deviations.

Under Assumption 2.2, we will control the first term on the right-hand side by using existing deviation bounds for the supremum of Gaussian processes (see Du Roy de Chaumaray and Marbac (2021a) for details), which will lead to

$$\lim_{n\to\infty}\mathbb{P}\left(\sup_{d\in D_{B,s}}\xi_d>\sqrt{\varepsilon a_{n,\boldsymbol{m}_0,B}}\right)=0. \tag{2.16}$$

Moreover, using the results of the approximation of suprema of general empirical processes by a sequence of suprema of Gaussian processes Chernozhukov, Chetverikov, and Kato (2014), we obtain

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\sup_{d\in D_{B,s}}\mathcal{G}_n(d)-\sup_{d\in D_{B,s}}\xi_d\right|>\sqrt{\varepsilon a_{n,\boldsymbol{m}_0,B}}\right)=0. \tag{2.17}$$

Thus, we have for any $\varepsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{T_{n,\boldsymbol{m},B}-\ell_n f_{0,B}}{a_{n,\boldsymbol{m}_0,B}}>\varepsilon\right)=0.$$

Noting that for any $\boldsymbol{m}\in\mathcal{N}_3$, Assumption 2.3 implies that $a_{n,\boldsymbol{m},B}/a_{n,\boldsymbol{m}_0,B}-1>0$ and noting that the cardinal of $\mathcal{N}_3$ is finite and fixed (it does not depend on B) then we can conclude that

$$\mathbb{P}(\widehat{\boldsymbol{m}}\in\mathcal{N}_3)=0. \tag{2.18}$$

Combining equations (2.13), (2.14) and (2.18) leads to the convergence in probability of $\widehat{\boldsymbol{m}}$ to M_0. □

Some comments: Note that the arguments used in Keribin (2000) to prove that underestimation is avoided cannot be used here. Indeed, the proof of Theorem 2.1 in Keribin (2000) considers parameters that are defined over a fixed dimensional space and thus cannot be used to obtain (2.12). In our context, we require that B tends to infinity with n (see Assumption 2.4) to ensure the identifiability and thus the convergence of $\inf_\theta \mathbb{E}_{g_0}\left[\ln f_{\boldsymbol{m}_0,B,\boldsymbol{\theta}^\star_{\boldsymbol{m}_0,B}}-\ln f_{\boldsymbol{m},B,\boldsymbol{\theta}}\right]$ to a quantity lower-bounded by $\delta_{\boldsymbol{m}}$ where $\theta^\star_{M_0,B}=\arg\max_{\boldsymbol{\theta}\in\Theta_{\boldsymbol{m}_0,B}}\mathbb{E}_{g_0}\left[\ln f_{\boldsymbol{m}_0,B,\boldsymbol{\theta}}\right]$ (this convergence is ensured by Assumption 2.2(ii), as discussed in the proof of (2.12)). Note that the existence of $\boldsymbol{\theta}^\star_{\boldsymbol{m}_0,B}$ is ensured by the fact that $\Theta_{\boldsymbol{m}_0,B}$ is compact and that the Kullback-Leibler divergence is continuous.

Note also that the arguments used in Keribin (2000) to prove that overestimation is avoided cannot be used here either. Indeed, despite the fact that $T_{n,\boldsymbol{m},B^\star}-\ell_n f_{0B^\star}$ converges in distribution for a fixed $B^\star$ and that $1/a_{n,\boldsymbol{m},B^\star}$ tends to 0, we cannot directly conclude that $\lim_{n\to\infty}\mathbb{P}\left(\left[T_{n,\boldsymbol{m},B}-T_{n,\boldsymbol{m}_0,B}\right]/a_{n,\boldsymbol{m},B}>\varepsilon\right)=0$ for any $\varepsilon>0$ and any $\boldsymbol{m}\in\mathcal{N}_3$ because we require that B tends to infinite with n to avoid the underestimation (see Assumption 2.4).

2.3.4 Estimation of the best model

The estimation of $\widehat{\boldsymbol{m}}_{n,B}$ requires an optimization over a discrete space whose cardinal is of order $2^J K_{\max}$. Thus, an exhaustive approach computing $W_{n,\boldsymbol{m},B}$ for each $\boldsymbol{m}$ in $\mathcal{M}$ is not doable in practice. As the combinatorial issue is mainly due to the feature selection, we follow the approach of Marbac, Sedki, and Patin (2020) that consists of simultaneously performing feature selection and parameter estimation, with a fixed number of components, via a specific EM algorithm optimizing the penalized likelihood. Thus, for a fixed value of K, the goal of the algorithm is to estimate

$$\widehat{\boldsymbol{m}}_{n,B,K} = \underset{\{\boldsymbol{m}=\{K,\boldsymbol{\Omega}\} \text{ with } \boldsymbol{\Omega}\subseteq\{1,\dots,J\} \text{ and } |\boldsymbol{\Omega}|\geq 3\}}{\arg\max} W_{n,\boldsymbol{m},B}.$$

The following EM algorithm permits the estimation of the model parameters and the detection of the subset of relevant variables, for a fixed number of components K. Parameter estimation is achieved by maximum likelihood and model selection is done with an information criterion with penalty $a_{n,\boldsymbol{m},B} = \nu_{K,\boldsymbol{m},B} c_n$ where $\nu_{K,\boldsymbol{m},B} = (K-1) + |\boldsymbol{\Omega}|K(B-1) + (J-|\boldsymbol{\Omega}|)(B-1)$ is the number of model parameters. The algorithm considers a fixed number of components K and starts at an initial point $\{\boldsymbol{\Omega}^{[0]}, \boldsymbol{\theta}^{[0]}\}$. Its iteration $[r]$ is composed of two steps:
E-step Computation of the fuzzy partition

$$t_{ik}^{[r]} := \frac{\pi_k^{[r-1]} \prod_{j\in\boldsymbol{\Omega}^{[r-1]}} \prod_{b=1}^{B} \left(\alpha_{Bkjb}^{[r-1]}\right)^{\sigma_{Bjb}(x_{ij})}}{\sum_{\ell=1}^{K} \pi_\ell^{[r-1]} \prod_{j\in\boldsymbol{\Omega}^{[r-1]}} \prod_{b=1}^{B} \left(\alpha_{B\ell jb}^{[r-1]}\right)^{\sigma_{Bjb}(x_{ij})}},$$

M-step Maximization of the expectation of the penalized complete-data log-likelihood over $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$ such

$$\boldsymbol{\Omega}^{[r]} = \{j : \Delta_j^{[r]} > 0\},\ \pi_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ and } \alpha_{Bkjb}^{[r]} = \begin{cases} \tilde{\alpha}_{Bkjb}^{[r]} & \text{if } j \in \Omega^{[r]} \\ \bar{\alpha}_{Bkjb} & \text{otherwise} \end{cases},$$

where

$$\Delta_j^{[r]} = \sum_{i=1}^{n}\sum_{b=1}^{B} \sigma_{Bjb}(x_{ij}) \sum_{k=1}^{K} t_{ik}^{[r]} \ln\left(\frac{\tilde{\alpha}_{Bkjb}^{[r]}}{\bar{\alpha}_{Bkjb}}\right) - (K-1)(B-1)c_n$$

is the difference between the maximum of the expected value of the penalized complete-data log-likelihood obtained when variable j is relevant and when it is irrelevant, with

$$\tilde{\alpha}_{Bkjb}^{[r]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}^{[r]} \sigma_{jb}(x_{ij}), \quad \bar{\alpha}_{Bkjb} = \frac{1}{n} \sum_{i=1}^{n} \sigma_{jb}(x_{ij}) \quad \text{and } n_k^{[r]} = \sum_{i=1}^{n} t_{ik}^{[r]}.$$

Note that, when less than three variables happen to have a positive value for $\Delta_j^{[r]}$, the M-step selects in $\boldsymbol{\Omega}^{[r]}$ the three variables having the largest values of $\Delta_j^{[r]}$. To obtain the pair $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$ maximizing the penalized observed-data log-likelihood, for a fixed number of components, many random initializations of this algorithm should be done. Hence, the model (*i.e.,* K and $\boldsymbol{\Omega}$) and the parameters maximizing the penalized observed-data log-likelihood are obtained by performing this algorithm for every values of K between 1 and $K_{\max}$. By considering $c_n = (\ln n)/2$, this algorithm carries out the model selection according to the BIC.

From previous algorithm, we obtain an estimator of the model and of its parameters. Indeed, $\hat{\alpha}_{kjb}/l_{jb}$ estimates the density $\eta_{kj}(u)$ for any u such that $\sigma_{jb}(u) = 1$. However, the bin-based density estimators are generally outperformed by kernel-based estimators. Thus, we advice to

use the proposed approach only for model estimation. Then, for the selected model, kernel-based density estimates provides by the EM-like algorithm (Benaglia, Chauveau, and Hunter (2009)) or by maximizing the smoothed log-likelihood (Levine, Hunter, and Chauveau (2011)) should be considered. However, note that establishing asymptotic properties of these kernel-based density estimators is still an open question.

2.3.5 Benchmark data

This section illustrates our procedure on three real data sets. The first data set illustrates the advantage of the procedure for selecting the number of components while the second data set sheds light on the importance of variable selection. The third data set shows that the procedure can be easily extended to the case of mixed-type data sets (a data set composed of continuous and categorical data).

Swiss banknotes data We consider the Swiss banknotes data set (Flury and Riedwyl (1988)) containing six measurements (length of bill, width of left edge, width of right edge, bottom margin width, top margin width and length of diagonal) made on 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes. This data set is available in the R package **mclust** (Scrucca et al. (2016)). The status of the banknote (genuine or counterfeit) is also known. We perform the clustering of the bills based on the six morphological measurements and we evaluate the resulting partition with the status of the bills. Considering all the six morphological measurements as relevant for clustering, the proposed method detects two clusters which are strongly similar to the status of the bill (the ARI is 0.98 and only one genuine bill is assigned to the cluster grouping all the counterfeit bills). Clustering with Gaussian mixture models provides more components (**mclust** selects three components and **VarSelLCM** selects four components) and a partition related but different to the status of the bill (the ARI is 0.84 and 0.48 for **mclust** and **VarSelLCM** respectively). When a full model selection (feature selection and estimation of the number of components) is performed, the proposed method still selects two components and detects all the variables as relevant. Thus, a full model selection approach provides the same results as a method used for selecting the clusters by considering all the variables as relevant. Moreover, the Gaussian mixture models obtains less relevant results because **VarSelLCM** considers that all the measurements are relevant and thus obtains the same results as without performing feature selection.

Chemical properties of coffees We consider the data set collected by Streuli (1973) that reports on the chemical composition of coffee samples collected from around the world. A total of 43 samples were collected from 29 countries, with beans from both Arabica and Robusta species, which is often considered as a pertinent partition. This data is available in the R package **pgmm** (McNicholas et al. (2015)). We cluster the different coffees based on twelve chemical constituents. A full Gaussian mixture clustering implemented in **Mclust** estimates three clusters and provided an ARI of 0.38. The same partition is obtained when the clustering is performed by **VarSelLCM** with a full model selection conducted according to the BIC (all the variables are detected as relevant for clustering). Again, similar results are obtained by the semi-parametric mixture if the proposed method is used to select the number of components. However, if we perform a full model selection, only five of the twelve variables are detected as relevant for clustering and only two components are estimated. Moreover, this simpler model provides a perfect recovery of the species (ARI=1.00). This illustrates the importance of variable selection for clustering. Note that McNicholas and Murphy (2008) proposed a parsimonious Gaussian mixture model that also provides a perfect recovery of the partition.

Cleveland data set We consider the Cleveland dataset (available at `https://www.kaggle.com/ronitf/heart-disease-uci/version/1`). This data set is composed of $n = 303$ subjects. Each subject is described by eight categorical variables having between two and six levels and five continuous variables. The "goal" field refers to the presence of heart disease in the patient (no presence vs presence). Model (2.9) can be easily extended to the case of mixed-type data (data set composed of continuous and categorical variables). Indeed, if variable j is categorical then η_{kj} is the probability mass function of a multinomial distribution. Thus, the discretization procedure used for model selection is applied only on the continuous variables while the number of levels for the categorical variables is fixed (*i.e.,* it is not defined from the sample size). When the model is selected, the estimation of the extension of model (2.9) can be easily achieved by maximizing the smoothed log-likelihood via an MM algorithm. The proposed approach detects the true number of clusters (*i.e.,* two) while the approach implemented in VarSelLCM overestimates it since it selects six components. Moreover, by considering ten variables as relevant for clustering, our procedure returns a more relevant partition with respect to the occurrence of heart disease because it obtains an ARI equals to 0.37 while the procedure implemente in VarSelLCM obtain an ARI equals to 0.12.

2.4 Numerical experiments

This section compares approaches for a full model selection (*i.e.,* estimation of the subset of the relevant variables and on the number of components) on simulated data. We compare the proposed approach with a BIC applied on a Gaussian mixture model, with the sparse K-means approach and with the non-parametric approach considering all the variables as relevant. The results of the proposed clustering method are obtained by performing full model selection with B levels defined by the empirical quantiles $1/B$,...,B/B where $B = [n^{1/6}]$ and a BIC like penalty and then by estimating the mixture components for the selected model by maximizing the smoothed log-likelihood with a bandwidth, for variable j, equal to $\hat{\sigma}_j n^{-1/5}$ where $\hat{\sigma}_j$ is the empirical standard deviation of variable j. Thus, when the discretization is performed, the model selection can be achieved via the R package VarSelLCM (Marbac and Sedki (2020)) then, when the best model is selected, the maximization of the smoothed log-likelihood is achieved via the R package mixtools (Benaglia et al. (2009a)). The parametric mixture model considers that all the components are Gaussian (this approach is also implemented in the R package VarSelLCM) and uses the BIC to perform model selection. The sparse K-means approach is implemented in the R package sparcl (Witten and Tibshirani (2010)) and consists in the sparse K-means algorithm initialized with the partition provided by the sparse hierchical ascendant classification with the "average" method. Finally, the non-parametric mixture model is implemented the R package mixtools (Benaglia et al. (2009a)) and considers the estimator maximizing the smoothed log-likelihood with a bandwidth, for variable j, equal to $\hat{\sigma}_j n^{-1/5}$ where $\hat{\sigma}_j$ is the empirical standard deviation of variable j. To compare the different methods of clustering, we generate data from a mixture with three components and equal proportions ($\pi_k = 1/3$). The density of X_i given Z_i is a product of univariate densities such that $X_{ij} = \sum_{k=1}^{K} z_{ik}\delta_{kj} + \xi_{ij}$ where all the ξ_{ij} are independent and where $\delta_{11} = \delta_{12} = \delta_{23} = \delta_{24} = \delta_{35} = \delta_{36} = \tau$, while all remaining $\delta_{kj} = 0$, which implies that only the first six variables are relevant for clustering. Three distributions are considered for the ξ_{ij} (standard Gaussian, Student with three degrees of freedom and Laplace) and the value of τ is defined to obtain a theoretical misclassification rate of 5% (τ is equal to 1.94, 2.60 and 2.52 for the Gaussian, Student and Laplace distributions respectively). In the Section **??** of the Supplementary Materials., all the experiments are also run with theoretical misclassification rates equal to 10% and 15%.

Selection of the discriminative features To investigate the performances of the competing methods for feature selection, we first consider the situation with a known number of components. Thus, the model selection consists in performing the feature selection. We consider the methods which automatically provide an estimator of the relevant variables (*i.e.,* the proposed method, sparse K-means and VarSelLCM). Accuracy of this selection is measured by sensitivity (probability to detect as relevant a true discriminative variable) and specificity (probability to detect as irrelevant a true non discriminative variable). Table 2.2 and 2.3 present the sensibility and the specificity obtained by the proposed approach and the parametric approach. They exhibit an advantage of the parametric method when the distribution is well-specified, but only for small samples ($n = 100$). The reason is that, for such samples, the proposed method only finds a part of the relevant variables. However, both methods perform well for larger samples. Moreover, the proposed method obtains similar results for the two other distributions of the components while the results of the parametric approach are strongly deteriorated for both sensibility and specificity, especially for heavy tailed distributions (Student distribution).

		Proposed method			VarSelLCM-BIC			VarSelLCM-MICL			Sparcl		
		n			n			n			n		
Component	J	100	250	500	100	250	500	100	250	500	100	250	500
Gaussian	20	0.83	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.85	0.91	0.95
	50	0.64	0.99	1.00	0.93	1.00	1.00	0.97	1.00	1.00	0.90	0.95	0.97
	100	0.44	0.76	1.00	0.70	1.00	1.00	0.83	1.00	1.00	0.76	0.97	0.98
Student	20	0.81	1.00	1.00	0.35	0.41	0.49	0.27	0.39	0.44	0.74	0.80	0.81
	50	0.70	1.00	1.00	0.10	0.14	0.21	0.10	0.17	0.26	0.71	0.73	0.80
	100	0.54	0.87	1.00	0.09	0.15	0.16	0.08	0.14	0.19	0.60	0.75	0.78
Laplace	20	0.84	1.00	1.00	0.89	1.00	1.00	0.85	1.00	1.00	0.80	0.83	0.80
	50	0.75	1.00	1.00	0.59	0.99	1.00	0.39	0.87	0.97	0.89	0.90	0.91
	100	0.54	0.90	1.00	0.17	0.89	1.00	0.10	0.21	0.42	0.85	0.94	0.94

Table 2.2: Mean of the sensitivity (Sen.: $\text{card}(\widehat{\Omega} \cap \Omega)/6$) for the feature selection obtained by the proposed method (Proposed method), the parametric method with the BIC (VarSelLCM-BIC), the parametric method with the MICL (VarSelLCM-MICL) and the sparse K-means (Sparcl) on 100 replicates for each scenario with theoretical misclassification rate of 5%, when the number of components is known.

		Proposed method			VarSelLCM-BIC			VarSelLCM-MICL			Sparcl		
		n			n			n			n		
Component	J	100	250	500	100	250	500	100	250	500	100	250	500
Gaussian	20	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.66	0.49
	50	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.44	0.31
	100	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.73	0.42	0.21
Student	20	0.97	1.00	1.00	0.70	0.57	0.46	0.61	0.42	0.27	0.75	0.59	0.60
	50	0.98	1.00	1.00	0.74	0.67	0.58	0.71	0.56	0.41	0.78	0.83	0.63
	100	0.98	1.00	1.00	0.75	0.69	0.63	0.77	0.66	0.55	0.86	0.87	0.79
Laplace	20	0.98	1.00	1.00	0.92	0.97	0.97	0.91	0.97	0.98	0.78	0.85	0.84
	50	0.98	1.00	1.00	0.93	0.97	0.97	0.86	0.90	0.91	0.75	0.79	0.74
	100	0.99	1.00	1.00	0.92	0.97	0.98	0.86	0.83	0.79	0.76	0.62	0.65

Table 2.3: Mean of the specificity (Spe.: $\text{card}(\widehat{\Omega}^c \cap \Omega^c)/(J-6)$) for the feature selection obtained by the proposed method (Proposed method), the parametric method with the BIC (VarSelLCM-BIC), the parametric method with the MICL (VarSelLCM-MICL), and the sparse K-means (Sparcl) on 100 replicates for each scenario with theoretical misclassification rate of 5%, when the number of components is known.

Full model selection We now compare both non-parametric and parametric approaches on their performances for full model selection. Table 2.4 presents the statistics on the number of components selected by both approaches. Again, when the distribution of the components is well-specified, the parametric approach obtains better results on small samples because the proposed approach tends to underestimate the number of components. However, when the sample size increases, both methods perform similarly. When the distribution of the components is not Gaussian, the parametric method performs poorly and asymptotically overestimate the number of components with probability one. The proposed method is consistent for any number of variables, however, it tends to underestimate the number of components for small samples.

Component	J	Proposed method					
		$n = 100$		$n = 250$		$n = 500$	
		Tr.	Ov.	Tr.	Ov.	Tr.	Ov
Gaussian	20	0.43	0.00	1.00	0.00	1.00	0.00
	50	0.31	0.00	0.96	0.01	1.00	0.00
	100	0.02	0.00	0.64	0.02	1.00	0.00
Student	20	0.54	0.00	1.00	0.00	1.00	0.00
	50	0.38	0.00	1.00	0.00	1.00	0.00
	100	0.11	0.00	0.86	0.00	1.00	0.00
Laplace	20	0.56	0.00	1.00	0.00	1.00	0.00
	50	0.38	0.00	1.00	0.00	1.00	0.00
	100	0.18	0.00	0.89	0.01	1.00	0.00

Component	J	VarSelLCM-BIC						VarSelLCM-MICL					
		$n = 100$		$n = 250$		$n = 500$		$n = 100$		$n = 250$		$n = 500$	
		Tr.	Ov.	Tr.	Ov.	Tr.	Ov.	Tr.	Ov.	Tr.	Ov.	Tr.	Ov.
Gaussian	20	0.94	0.00	1.00	0.00	1.00	0.00	0.85	0.00	1.00	0.00	1.00	0.00
	50	0.87	0.00	1.00	0.00	1.00	0.00	0.83	0.02	1.00	0.00	1.00	0.00
	100	0.53	0.00	1.00	0.00	1.00	0.00	0.42	0.18	1.00	0.00	1.00	0.00
Student	20	0.62	0.15	0.19	0.79	0.00	1.00	0.56	0.30	0.25	0.75	0.00	1.00
	50	0.78	0.16	0.40	0.60	0.02	0.98	0.44	0.54	0.07	0.93	0.00	1.00
	100	0.78	0.22	0.19	0.81	0.01	0.99	0.16	0.84	0.00	1.00	0.00	1.00
Laplace	20	0.78	0.09	0.32	0.68	0.00	1.00	0.54	0.06	0.71	0.29	0.45	0.55
	50	0.36	0.01	0.52	0.48	0.00	1.00	0.25	0.10	0.50	0.37	0.48	0.52
	100	0.08	0.00	0.63	0.17	0.03	0.97	0.13	0.13	0.06	0.17	0.15	0.59

Table 2.4: Probability to select the true number of components (Tr.) and to overestimate it (Ov.) obtained by the proposed method (Proposed method), the parametric method with the BIC (VarSelLCM-BIC) and the parametric method with the MICL (VarSelLCM-MICL) on 100 replicates for each scenario with theoretical misclassification rate of 5%, by performing a selection of the variables.

Table 2.5 presents the sensitivity and the specificity for feature selection obtained by both approaches when the number of components is also estimated. Again, results show the benefits of the proposed approach when the parametric assumptions are violated. In such case, the parametric approach overestimates the number of components and, for heavy tail distributions (*e.g.,* Student distribution), this approach tends to overestimate the subset of relevant variables. Moreover, for the small samples, the sensitivity is quite low explaining the tendency of overestimating the number of components.

Accuracy of the partition We are now interested in investigating the accuracy of the estimated partition. Thus, we compute the Adjusted Rand index (Hubert and Arabie (1985)) between the true partition and the estimators of the partition given by the non-parametric and the parametric methods when K is known and then when it is estimated. Moreover, to illustrate the benefit of feature selection, we also estimate the partition by considering the full

Component	J	Proposed method					
		$n = 100$		$n = 250$		$n = 500$	
		Sen.	Spe.	Sen.	Spe.	Sen.	Spe.
Gaussian	20	0.97	0.81	1.00	1.00	1.00	1.00
	50	0.96	0.76	1.00	0.99	1.00	1.00
	100	0.94	0.55	1.00	0.95	1.00	1.00
Student	20	0.96	0.86	1.00	1.00	1.00	1.00
	50	0.96	0.79	1.00	1.00	1.00	1.00
	100	0.95	0.67	1.00	0.98	1.00	1.00
Laplace	20	0.97	0.87	1.00	1.00	1.00	1.00
	50	0.97	0.81	1.00	1.00	1.00	1.00
	100	0.95	0.72	1.00	0.99	1.00	1.00

Component	J	VarSelLCM-BIC						VarSelLCM-MICL					
		$n = 100$		$n = 250$		$n = 500$		$n = 100$		$n = 250$		$n = 500$	
		Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.
Gaussian	20	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
	50	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
	100	0.99	0.86	1.00	1.00	1.00	1.00	0.99	0.88	1.00	1.00	1.00	1.00
Student	20	0.67	0.35	0.66	0.71	0.56	0.97	0.63	0.32	0.55	0.69	0.42	0.96
	50	0.74	0.09	0.68	0.14	0.59	0.24	0.72	0.09	0.55	0.16	0.37	0.26
	100	0.76	0.11	0.72	0.13	0.66	0.14	0.78	0.08	0.65	0.14	0.49	0.17
Laplace	20	0.92	0.90	0.96	1.00	0.90	1.00	0.90	0.81	0.95	1.00	0.92	1.00
	50	0.86	0.46	0.97	1.00	0.96	1.00	0.84	0.35	0.92	0.90	0.92	1.00
	100	0.84	0.17	0.96	0.82	0.98	1.00	0.85	0.10	0.82	0.21	0.85	0.69

Table 2.5: Mean of the sensitivity (Sen.: $\text{card}(\widehat{\Omega} \cap \Omega)/6$) and the specificity (Spe.: $\text{card}(\widehat{\Omega}^c \cap \Omega^c)/(J-6)$) for the feature selection obtained by the proposed method (Proposed method) and the parametric method (VarSelLCM) on 100 replicates for each scenario with theoretical misclassification rate of 5%, when the number of components also is estimated.

variables as relevant and the true number of components. Results are presented in Figure 2.3. Thus, when the parametric assumptions are satisfied, the parametric approach outperforms the proposed approach only on small samples (few observations with respect to the number of variables), whenever the number of components is known or not. However, when the parametric assumptions are violated, the proposed approach strongly outperforms the parametric approach. Note that, when the number of irrelevant variables increases, the approach considering all the variables for clustering performs poorly (see row 100), illustrating the benefit of feature selection for clustering.

2.5 Conclusion

This chapter addresses the issue of variable selection for mixture models under the assumption of conditional independence between variables given the component. This assumption limits the number of estimators to be considered and thus is relevant for analyzing data composed of many features (which is the situation where variable selection is crucial). However, if this assumption is violated, then the resulting estimators are biased. In a parametric framework, we propose an optimization algorithm to circumvent the computational issues of model selection via information criteria. We also propose a optimization algorithm for performing model selection according to a criterion tuned for clustering (MICL). This criterion is not consistent but seems to be more robust to the misspecification of the distribution of the components.

To avoid the need for specifying the family of the components, we address the issue of model selection in a nonparametric framework. The proposed method is relevant even if we only want to estimate the number of components by assuming that all the variables are relevant. Indeed,

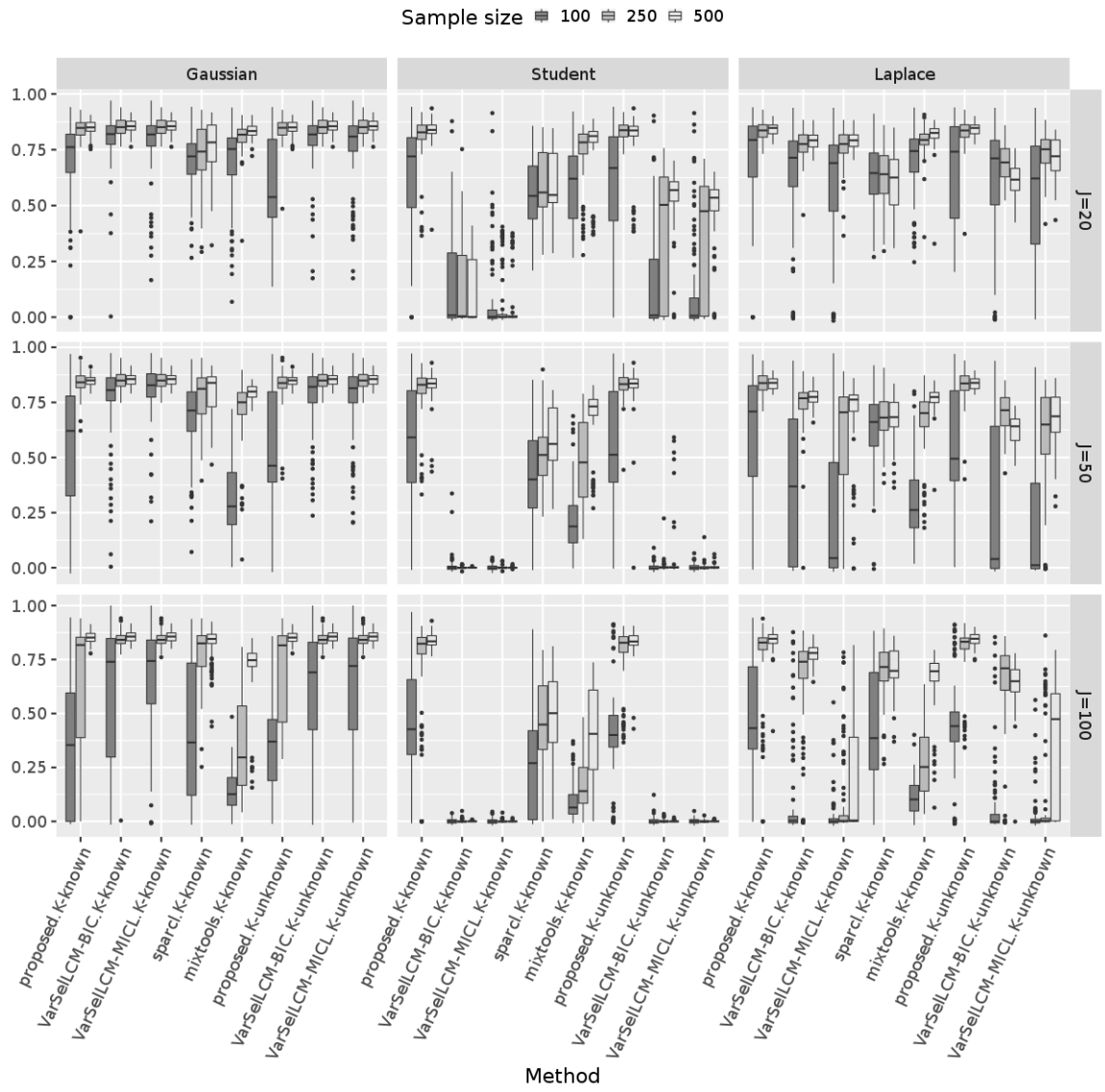


Figure 2.3: Boxplot of the Adjusted Rand Index (ARI) obtained on the resulting partition when feature selection is performed with the true number of components by the proposed method (proposed.K-known) and by the parametric method (VarSelLCM.K-known), by the sparse K-means (Sparcl.K-known) and by the model considering all the variables as relevant components (mixtools.K-known) and when the full model selection (feature selection and estimation of the number of components) is achieved by the proposed approach (proposed.K-unknown) and the parametric approach (VarSelLCM.K-unknown). Data are generated with theoretical misclassification rate of 5%

the approach permits many variables to be considered and thus it is a complementary work to Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020), because these methods suffer

from combinatorial issues if the number of variables is large. In the case where the parametric assumptions are validated, the parametric methods for full model selection slightly outperforms the nonparametric method for small sample size, but these methods obtains similar results on large sample sizes. However, if the parametric assumptions are violated, the nonparametric method outperforms the parametric approaches.

Because the performance of the clustering methods depends on the number of variables (with respect to the sample size), we believe that the choice of the clustering method should consider the dimension of the data to analyze. When few variables are considered, the independence assumption could be too stringent, hence modeling the intra-component dependencies would allow the accuracy of the partition to be improved. It could be done in a parametric Banfield and Raftery (1993) or nonparametric Zhu and Hunter (2019) framework. However, in this case, there is no tool for model selection in a nonparametric framework. Moreover, selecting the variables in a parametric framework could be achieved via stepwise methods for optimizing an information criterion. When the number of variables increases, the assumption of conditional independence has the property of limiting the number of parameters to estimate. Moreover, feature selection allows the accuracy of the estimators to be improved and the clusters to be interpreted. The proposed methods presented in this chapter could be used in this case. If the sample size increases again (*e.g.,* $n << d$), we advise not to use the nonparametric methods or the BIC for model selection in a parametric framework. Exact criterion like the BIC could be considered on a parsimonious model or regularization of the K-means algorithm (Witten and Tibshirani (2010))

The reasoning used in the proof of the Theorem can be considered for investigating the consistency of penalized likelihood in the case of increasing parameter space or increasing model space. This future development is explained in Section 7.1.1.

Chapter 3

Dealing with non-ignorable missingness in clustering

Contents

3.1 Introduction

3.1.1 State of the art

Despite the fact that the data sets often contain missing values, as for example in social surveys, there are only few clustering approaches that consider missingness. Thus, statistical analysis are generally performed on a complete data set where missing values have been either removed or imputed. Removing subjects having missing values leads to severe bias and/or losses of efficiency

(see Molenberghs et al. (2008)). Single imputation of missing values (see Van Buuren (2018)) suffers from a lack of consistency because imputations are generally performed with a model different from the model used to cluster and do not permit the variability of the data to be considered. When the missingness mechanism is *ignorable* (*i.e.,* the mechanism is *Missing at Random* and the property of distinctness is satisfied (see for instance Molenberghs et al. (2014) and Little and Rubin (2002)), then the distribution of the variables can be estimated by modeling the missingness mechanism. Thus, if the parameter of the distribution of the variables is the quantity of interest, likelihood-based methods (Schafer (1997)) or multiple imputations (Van Buuren (2018)) can be used for the estimation. In this chapter, the quantity of interest is not the distribution of the variables but the conditional probability of the cluster memberships given the observed variables.

The case where the mechanism is not ignorable (*e.g.,* the *missing not at random* (MNAR) mechanism, where the missingness is allowed to depend on the missing values even conditionally on the observed covariates) happens frequently in practice (*e.g.,* higher-income respondents may decline to report income data). In such cases, the joint distribution of the variables and the indicators of responses has to be considered. Thus, weighting methods (see Rotnitzky and Robins (1997) and Tsiatis (2007)) can be used if the target is the inference of the distribution of the variables. However, these methods are not really suitable for clustering because they would classify only the subjects with no missingness. Alternatively, multiple imputations could be considered, but because many samples would be generated, it is not easy to consider the aim of the clustering. Thus, in this chapter, we focus on the likelihood-based methods (or smoothed-likelihood-based methods). Note that generally, assumptions should be made (*e.g.,* parametric assumption) on the joint distribution of the variables and the indicators of responses to obtain the identifiability of the model but the distribution of the mechanism cannot be tested on the observed data (see Molenberghs et al. (2008)). Identifiability of the parameter of interest is crucial for consistency of the procedure.

In order to handle missing values in a model-based clustering framework, Hunt and Jorgensen (2003) have implemented the standard EM algorithm based on the observed likelihood. More recently, Serafini, Murphy, and Scrucca (2020) also proposed an EM algorithm to estimate Gaussian mixture models in the presence of missing values by performing multiple imputations (with Monte Carlo methods) in the E-step. However, both works only consider M(C)AR data. Two clustering approaches allow data subject to non-ignorable mechanism to be analyzed. Chi, Chi, and Baraniuk (2016) introduces the K-POD algorithm that extends the K-means to the case of missing data even if the missing mechanism is unknown. K-POD method performs the cluster detection from the observed data via a MM algorithm. Indeed, K-POD algorithm alternates between an imputation of the missing values given the centroids and the cluster assignments and an estimation of the centroids and the cluster assignments based on the imputated data. Thus, the K-POD algorithm is the combination of a formulation that is common to matrix completion problems with a descent algorithm in the MM framework to produce clustering that agrees with the observed data. One argument of the authors is that, by bypassing the completely observed data formulation, K-POD retains all information in the data and avoids committing to distributional assumptions on the missingness patterns. However, this approach suffers from the standard drawbacks of the K-means algorithm (*i.e.,* assumptions of spherical clusters and equal proportions of the clusters). Alternatively, using a *selection model* approach (see Little (1993) and the definition in Section 3.2), Miao, Ding, and Geng (2016) proposed specific univariate Gaussian mixtures and univariate t-mixtures to cluster continuous data under a non-ignorable mechanism. For such an approach, the missingness mechanism must be specified. The authors use probit and logit distributions to model the missingness mechanisms. The authors define gen-

eral conditions for obtaining the identifiability of the resulting distribution. Note that in Miao, Ding, and Geng (2016), there are no details about the computation of the E-step of the EM algorithm used to assess the MLE, despite the fact that this step involves computations of integrals having no closed-form. Moreover, the parametric mixture of Miao, Ding, and Geng (2016) is introduced to cluster univariate data and its extension to the case of multivariate data is not trivial without the assumption of independence within components.

3.1.2 Contributions to clustering under non-ignorable missingness

In the context of missing values, we wrote three book chapters as a companion of the summer school *19èmes Journées d'Étude en Statistique*. These chapters present an introduction of statistical analysis with missing values (Marbac (2022a)), a description of the likelihood-based methods under ignorable mechanisms (Marbac (2022b)) and an introduction to the weighted methods for missing values (Marbac (2022c)). In the framework of model-based clustering, my colleagues and I proposed two approaches to deal with non-ignorable missingness mechanims that we detail in this chapter. The first contribution (Biernacki et al. (2021)) considers a parametric framework and generalizes the approach of Miao, Ding, and Geng (2016). The second approach (Du Roy de Chaumaray and Marbac (2020)) considers a non-parametric framework and can be interpreted as a generalization of (2.3) in the case of non-ignorable missingness. This approach has been implemented in the R package **MNARclust** (Du Roy de Chaumaray and Marbac (2021b)) available on CRAN.

In Biernacki et al. (2021), we presented a relevant inventory of parametric distributions for the MNAR missingness process in the context of unsupervised classification based on parametric mixture models that generalize the approach of Miao, Ding, and Geng (2016) in the multivariate case. We stated the identifiability of the mixture model parameters and of the missingness process parameters, under certain conditions (including the data type and the link functions governing the missingness mechanism distribution). This is a real issue in the context of MNAR data, as models often lead to unidentifiable parameters. When all variables are continuous, all models lead to identifiable parameters. In the categorical case, only the models for which the missingness depends on the class membership have identifiable parameters. For each model or sub-model, an EM or Stochastic EM algorithm is proposed.

In Du Roy de Chaumaray and Marbac (2020), we proposed to perform clustering via a mixture model that uses a *pattern-mixture model* approach (see Little (1993) and the definition in Section 3.2) with non-parametric distributions. Thus, no assumptions were made on the data distribution or on the missingness mechanism except that the variables are independent within components and thus generalizes (2.3) to the case of missingness. Note that this is an implicit assumption made by geometrical clustering (*e.g.,* K-means or K-POD) when a diagonal metric is used to compute the distances between observations. Despite the fact that this assumption is relevant in many situations, especially if the number of variables is large with respect to the number of observations (Hand and Keming (2001), Webb, Boughton, and Wang (2005) and Stephens, Huerta, and Linares (2018)), it can induce a bias when violated. For each mixture component, we estimated, for each variable, its probability to be observed together with its conditional distribution given that the variable is observed. We emphasized that our concern is clustering and not imputation or density estimation. Indeed, the approach presented in Du Roy de Chaumaray and Marbac (2020) permits the conditional probability of the cluster memberships to be estimated given the observed values. Note that, as in any approach developed for a non-ignorable mechanism, the distribution of the variables within a component cannot be estimated by our procedure, without additionnal assumptions. Estimation of the semi-parametric mixture

can be done by maximizing the smoothed likelihood (Levine, Hunter, and Chauveau (2011)). In Du Roy de Chaumaray and Marbac (2020), we extend the concept of smoothed likelihood to mixed-type data. Indeed, the model includes continuous (the covariates) as well as binary (the indicators of the missingness) variables. In our extension, only the distribution of the continuous variables are smoothed. Thus, the smoothed likelihood can be maximized by an MM algorithm implemented in the R package **MNARclust**.

The chapter is organized as follows. Section 3.2 introduces the semi-parametric mixture used for clustering data with non-ignorable missingness and a definition of ignorability for clustering that can be interpreted as an extension of the ignorability for a part of the parameters introduced in Little, Rubin, and Zangeneh (2017). Section 3.3 presents the main elements of the work presented in Biernacki et al. (2021). Section 3.4 presents the main elements of the work presented Du Roy de Chaumaray and Marbac (2020). A conclusion is given in Section 3.5.

3.2 Mixture for non-ignorable missingness

3.2.1 The data

The observed sample is composed of n independent and identically distributed subjects arising from K homogeneous subpopulations. Each subject is described by J continuous variables and some realizations of these variables may be unobserved. The missingness mechanism is allowed to be non-ignorable. Thus, the probability, for a variable, not to be observed is allowed to depend on the values of the variable itself and the subpopulation membership.

Each subject i is described by a vector of three variables $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top, \mathbf{Z}_i^\top)^\top$ where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})^\top \in \mathbb{R}^J$ is a set of continuous variables, $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iJ})^\top \in \{0,1\}^J$ indicates whether X_{ij} is observed ($R_{ij} = 1$) and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})^\top$ indicates the subpopulation of subject i ($Z_{ik} = 1$ if subject i belongs to subpopulation k and otherwise $Z_{ik} = 0$). Each subject belongs to one subpopulation such that $\sum_{k=1}^K Z_{ik} = 1$. The realizations of $\mathbf{Z}_i$ are unobserved and a part of the realizations of $\mathbf{X}_i$ can be unobserved too. Therefore, the observed variables for subject i are $(\boldsymbol{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$ where $\boldsymbol{X}_i^{\text{obs}}$ is composed by the elements of $\mathbf{X}_i$ such that $R_{ij} = 1$ and the unobserved variables for subject i are $(\boldsymbol{X}_i^{\text{miss}\top}, \mathbf{Z}_i^\top)^\top$ where $\boldsymbol{X}_i^{\text{miss}}$ is composed of the elements of $\mathbf{X}_i$ for which $R_{ij} = 0$.

3.2.2 General mixture model

We use mixture models for the purpose of clustering and not for density estimation. Clustering aims to estimate the subpopulation memberships given the observed variables (*i.e.,* the realization of $\mathbf{Z}_i$ given $(\boldsymbol{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$) without any assumption on the missingness mechanism (*i.e.,* no assumption on the conditional distribution of $\boldsymbol{R}_i$ given $(\mathbf{X}_i^\top, \mathbf{Z}_i^\top)^\top$). The probability distribution function (pdf) of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ for subpopulation k (*i.e.,* $Z_{ik} = 1$) is denoted by $f_k(\cdot)$. Thus, the pdf $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ is defined by the pdf of a K-component mixture

$$f(\boldsymbol{x}_i, \boldsymbol{r}_i) = \sum_{k=1}^K \pi_k f_k(\boldsymbol{x}_i, \boldsymbol{r}_i), \tag{3.1}$$

where $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$ and $f_k(\cdot)$ is pdf of component k. From (3.1), the distribution of the observed values $(\mathbf{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$ can be defined by two approaches (see Molenberghs et al. (2014) and Little and Rubin (2002)): the *selection model* and the *pattern-mixture model*.

The approach named *selection model* defines the conditional distribution of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ given $\mathbf{Z}_i$ as the product between the conditional distribution of $\mathbf{X}_i$ given $\mathbf{Z}_i$ and the conditional distribution of $\boldsymbol{R}_i$ given $(\mathbf{Z}_i^\top, \mathbf{X}_i^\top)^\top$ such that

$$f_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = f_k(\boldsymbol{x}_i) f_k(\boldsymbol{r}_i \mid \boldsymbol{x}_i).$$

Thus, the distribution of the observed data is defined for each component k by

$$f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) = \int f_k(\boldsymbol{x}_i) f_k(\boldsymbol{r}_i \mid \boldsymbol{x}_i) d\boldsymbol{x}_i^{\text{miss}}.$$

The *selection model* approach is natural and has been used for many times in different contexts (see Miao, Ding, and Geng (2016) for clustering under a MNAR scenario). When the mechanism is ignorable, it provides an estimation of the marginal distribution of $\mathbf{X}_i$ without considering the distribution of the mechanism. However, when the mechanism is non-ignorable, it requires the missingness mechanism to be modelled, *i.e.* the conditional distribution of $\boldsymbol{R}_i$ given $(\mathbf{Z}_i^\top, \mathbf{X}_i^\top)^\top$. Finally, as it considers the marginal distribution of $\mathbf{X}_i$, the *selection model* should be used when the aim is to fit the marginal distribution of $\mathbf{X}_i$.

Alternatively, the approach named *pattern-mixture model* (Little (1993)) defines the conditional distribution of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ given $\mathbf{Z}_i$ as the product between the conditional distribution of $\boldsymbol{R}_i$ given $\mathbf{Z}_i$ and the conditional distribution of $\mathbf{X}_i$ given $(\mathbf{Z}_i^\top, \boldsymbol{R}_i^\top)^\top$. Thus, using the *pattern-mixture model*, the pdf of component k is given by

$$f_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = f_k(\boldsymbol{r}_i) f_k(\boldsymbol{x}_i \mid \boldsymbol{r}_i). \tag{3.2}$$

The pdf of the observed variables under component k, denoted by $f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)$, is obtained by integrating the pdf of component k over the missing variables $\boldsymbol{X}_i^{\text{miss}}$, which leads to

$$f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) = f_k(\boldsymbol{r}_i) f_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i). \tag{3.3}$$

Note that (3.3) takes into account the missingness mechanism as it involves the whole vector $\boldsymbol{R}_i$. Thus the missing values impact the clustering. To estimate the marginal density of $\mathbf{X}_i$ from (3.3), assumptions should be made on the conditional distribution $\boldsymbol{X}_i^{\text{obs}}$ given $\boldsymbol{R}_i$ (because the realizations under some distributions are never observed, *e.g.,* $\boldsymbol{R}_i = 0$). However, we recall that we focus on the target of clustering that consists of assessing the posterior probabilities of the classification given the observed values using

$$\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)}. \tag{3.4}$$

3.2.3 Weak and strong ignorability for clustering

In a likelihood-based estimation, the missingness mechanism is said to be ignorable for likelihood inference if the missing data are missing at random and if the distinctness property is satisfied by the parameters (see Definition 6.4 in Little and Rubin (2002)). These conditions ensure that it is appropriate to ignore the missingness mechanism, especially for parameter estimation. These conditions have been extended when only a subset of the parameters are of interest (Little, Rubin, and Zangeneh (2017)). Thus, despite the fact that the missingness mechanism is MNAR, these conditions ensure that a subset of the parameters can be consistently estimated by ignoring the missingness mechanism. In clustering, the quantities of interest are the partition and, in some cases, the posterior probabilities of classification. Thus, we introduce the notion of weakly and strongly ignorable mechanisms for clustering that allows the mechanism to be neglected for estimating the partition and the posterior probabilities of classification.

Definition 3.1. Let $g_k(\boldsymbol{x}_i^{\text{obs}})$ be the marginal pdf of the observed variables under component k. The missingness mechanism is said to be strongly ignorable for clustering if

$$\forall \boldsymbol{x}_i^{\text{obs}},\ \frac{\pi_k f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)} = \frac{\pi_k g_k(\boldsymbol{x}_i^{\text{obs}})}{\sum_{\ell=1}^{K} \pi_\ell g_\ell(\boldsymbol{x}_i^{\text{obs}})}.$$

The missingness mechanism is said to be weakly ignorable for clustering if

$$\forall \boldsymbol{x}_i^{\text{obs}},\ \zeta(\boldsymbol{x}_i^{\text{obs}}) = \eta(\boldsymbol{x}_i^{\text{obs}}),$$

where

$$\zeta(\boldsymbol{x}_i^{\text{obs}}) = \underset{k=1,\ldots,K}{\arg\max} \frac{\pi_k f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)} \text{ and } \eta(\boldsymbol{x}_i^{\text{obs}}) = \underset{k=1,\ldots,K}{\arg\max} \frac{\pi_k g_k(\boldsymbol{x}_i^{\text{obs}})}{\sum_{\ell=1}^{K} \pi_\ell g_\ell(\boldsymbol{x}_i^{\text{obs}})}.$$

The strong ignorability for clustering implies the weak ignorability for clustering. When the data are MNAR, the weak ignorability can be interpreted as the condition required for a misspecified model (*e.g.,* the model of the missingness mechanism or the distribution of the mixture components) to provide a consistent estimator of the partition. Note that in clustering, consistency of the estimated partition does not mean a perfect recovery of the partition but that the estimated partition is asymptotically equivalent to the partition obtained by using the rule of the *maximum a posteriori* on the true posterior probabilities of classification. Thus, a misspecified model can provide a consistent estimator of the partition (*e.g.,* if the data arise from a mixture of two univariate Student distributions with the same degrees of freedom, a consistent estimator of the partition can be obtained by considering a mixture of two univariate Gaussian distributions; note that in this case the probabilities of classification are not consistently estimated). As the condition on the missingness mechanism to be weakly ignorable for clustering is quite stringent, we need to introduce an approach based on the joint distribution of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ which allows data to be clustered under a non-ignorable scenario.

3.3 Parametric mixture for non-ignorable missingness

3.3.1 The model

In Biernacki et al. (2021), we consider that the components follow parametric distributions. Thus, it is natural to use the selection model approach to deal with missingness. Thus, we have

$$f(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \int f_k(\boldsymbol{x}_i, \boldsymbol{r}_i; \boldsymbol{\alpha}_k) d\boldsymbol{x}_i^{\text{miss}}, \tag{3.5}$$

with

$$f_k(\boldsymbol{x}_i, \boldsymbol{r}_i; \boldsymbol{\alpha}_k) = f_k(\boldsymbol{x}_i; \boldsymbol{\delta}_k) f_k(\boldsymbol{r}_i | \boldsymbol{x}_i; \boldsymbol{\psi}_k), \tag{3.6}$$

where $\boldsymbol{\theta}$ groups all the parameters, $\boldsymbol{\alpha}_k = (\boldsymbol{\delta}_k^\top, \boldsymbol{\psi}_k^\top)^\top$ groups the parameters of component k, $\boldsymbol{\delta}_k$ groups the parameters related to the distribution of $\mathbf{X}_i$ under component k and $\boldsymbol{\psi}_k$ groups the parameters related to the conditional distribution of $\boldsymbol{R}_i$ given $\mathbf{X}_i$ under component k. Different kinds of distributions can be considered, depending on the types of features at hand. For continuous data, we consider that the conditional distribution of $\mathbf{X}_i$ given the component is the J-variate Gaussian distribution. Similarily, for categorical data, we consider that the conditional distribution of $\mathbf{X}_i$ given the component is a product of J univariate multinomial distributions.

For a combination of continuous and categorical data (*i.e.,* the mixed-type case), the conditional distribution of $\mathbf{X}_i$ given the component is defined as a product of univariate Gaussian and multinomial distributions. In a parsimonious perspective, we assume that the elements of $\boldsymbol{R}_i$'s are independent conditionally on $(\mathbf{X}_i^\top, \mathbf{Z}_i^\top)^\top$, leading to

$$f_k(\boldsymbol{r}_i|\boldsymbol{x}_i;\boldsymbol{\psi}_k) = \prod_{j=1}^{J} f_k(r_{ij}|\boldsymbol{x}_i;\boldsymbol{\psi}_k). \tag{3.7}$$

A general MNAR mechanism for R_{ij} can be written as follows, by giving the probability of missingness for the variable j given the data $\boldsymbol{x}_i$ and the class membership $z_{ik} = 1$,

$$f_k(r_{ij}|\boldsymbol{x}_i;\boldsymbol{\psi}_k) = \rho\left(\upsilon_{kj} + \beta_{kj}x_{ij} + \sum_{j'\neq j}\gamma_{kjj'}x_{ij'}\right)^{r_{ij}} \times \left(1-\rho\left(\upsilon_{kj} + \beta_{kj}x_{ij} + \sum_{j'\neq j}\gamma_{kjj'}x_{ij'}\right)\right)^{1-r_{ij}}, \tag{3.8}$$

where ρ is the cumulative distribution function of any continuous distribution function and $\boldsymbol{\psi}_k$ groups all the υ_{kj}, β_{kj} and $\gamma_{kjj'}$ if the feature j is continuous. Note that constraints between the parameters needs to be added for identifiability reasons when variable j is categorical (see Biernacki et al. (2021) for details). This general MNAR mechanism seems to be over-parameterized. For instance, for a binary dataset, the number of parameters is equal to $2KJ + J(J-1)$ while, for instance, the most parsimonious mixture model on $\boldsymbol{x}_i$, namely the latent class model (*i.e.,* parametric version of (2.3)), has $JK + K - 1$ parameters. Note that the missingness model (3.8) has more parameters than the associated mixture model. Since we are expecting that the individual data $\mathbf{X}$ convey more information on the partition $\mathbf{Z}$ that the pattern $\boldsymbol{R}$ of missing data, it seems to be hazardous to allow the missing data modeling to be more complex than the mixture model itself. Consequently, parsimonious versions of the general MNAR model (3.5)-(3.8) have to be proposed (see Biernacki et al. (2021) for details). Firstly it is reasonable to assume that $\gamma_{jj'} = 0$ (for all $j' \neq j$), which means that a given value is missing mainly because of its own value, much more than the values of the other variables. Therefore, the most complex model that we propose is called the MNAR$\boldsymbol{x}_{[k]}\boldsymbol{z}_{[j]}$ model

$$\text{MNAR}\boldsymbol{x}_{[k]}\boldsymbol{z}_{[j]}\text{:}\quad f_k(r_{ij}|\boldsymbol{x}_i;\boldsymbol{\psi}_k) = \rho(\upsilon_{kj} + \beta_{kj}x_{ij})^{r_{ij}}\left(1-\rho(\upsilon_{kj} + \beta_{kj}x_{ij})\right)^{1-r_{ij}}. \tag{3.9}$$

The parameters $\{\upsilon_{kj}\}$ represent the effect of missingness on the k-th class membership which depends on the variable j (i.e. the effect is not the same for all variables). The parameters $\{\beta_{kj}\}$ represent the direct effect of missingness on the variable j which depends on the class k. In Biernacki et al. (2021), we propose different parsionious models by adding constraints between the υ_{kj} and/or between the β_{kj}.

3.3.2 Identifiability of the parameters

Proving the identifiability of the parameters of a mixture model containing missing values amounts to proving that the joint distribution of $(\boldsymbol{R}_i^\top, \mathbf{X}_i^\top, \mathbf{Z}_i^\top)$ can be uniquely determined from available information. Therefore, we prove the identifiability of the parameters of the observed distribution defined by (3.5)-(3.7) and (3.9). Proposition 3.1 gives sufficient conditions for the identifiability of the parameters for continuous or count data under the followning assumptions.

Assumption 3.1. (i) The parameters of the marginal distribution of $\mathbf{X}_i$ defined by the density $\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\delta}_k)$ are identifiable;

(ii) There exists a total ordering $\preceq$ of $\mathcal{F}_j \times \mathcal{R}$, for $j \in \{1, \ldots, J\}$ fixed, where $\mathcal{F}_j$ is the family of the distributions $\{f_{1j}, \ldots, f_{Kj}\}$ and $\mathcal{R}$ is the family of the mechanism densities $\{f_{1j}(r_{ij}|\boldsymbol{x}_i; \boldsymbol{\psi}_1), \ldots, f_{Kj}(r_{ij}|\boldsymbol{x}_i; \boldsymbol{\psi}_K)\}$. The total ordering is such that $\forall k < \ell$, $F_k \preceq F_\ell$ (denoting $F_k = \rho_k f_{kj}$ and $F_\ell = \rho_\ell f_{\ell j}$) implies

$$\lim_{u \to +\infty} \frac{\rho(\upsilon_{\ell j} + \beta_{\ell j} u) f_{\ell j}(u; \boldsymbol{\delta}_\ell)}{\rho(\upsilon_{kj} + \beta_{kj} u) f_{kj}(u; \boldsymbol{\delta}_k)} = 0;$$

(iii) The missing-data distribution ρ is assumed to be strictly monotone.

Assumption 3.1.(i) means that the identifiability of the parameters $\boldsymbol{\theta}$ of the model defined by (3.5)-(3.7) and (3.9) requires the identifiability of the parameters $(\boldsymbol{\pi}^\top, \boldsymbol{\delta}^\top)$ (*i.e.,* the mixture models has to be identifiable when there is no missingness). Many authors have already studied the identifiability of the mixture models, when $\mathbf{X}_i$ is always fully observed (see Teicher (1963); Teicher (1967) and Yakowitz and Spragins (1968)). Assumption 3.1.(ii) is the core ingredient to prove the identifiability of the parameters and we illustrate it by considering concrete examples in the following. Note that under Assumption 3.1.(iii) the probit and the logistic function may be considered, which are the most widely used for MNAR specifications.

Proposition 3.1. Under Assumptions 3.1, the parameters $\boldsymbol{\theta}$ of the model defined by (3.5)-(3.7) and (3.9) are identifiable up to label swapping (as also the parsimonious models defined in Biernacki et al. (2021)).

The proof of this proposition is detailed in Biernacki et al. (2021) and follows the reasoning used by Theorem 2 in Teicher (1963) which proves the identifiability of a univariate finite mixture using a total ordering of the mixture densities. In the following, we denote by f_{kj}, the marginal density of the variable j under component k. Proposition 3.1 states the identifiability of the Gaussian mixture with a probit missingness mechanism (details are given in Biernacki et al. (2021)). Indeed, finite Gaussian mixtures are identifiable and, for any variable j, there is a total ordering defined by $\sigma^2_{kj} > \sigma^2_{(k+1)j}$ and $\mu_{kj} > \mu_{(k+1)j}$ if $\sigma^2_{kj} = \sigma^2_{(k+1)j}$, where μ_{kj} and σ^2_{kj} are respectively the mean and the variance of variable j under component k. The property stated by Proposition 3.1 has been already etablished, in the case of univariate distributions, by Miao, Ding, and Geng (2016). In particular, the identifiability conditions in Miao, Ding, and Geng (2016) (conditions 1 and 2) imply the existence of the total ordering defined in Proposition 3.1. However, these conditions exclude the case of a Gaussian mixture with a logistic missingness mechanism, which is very much used in practice. In Proposition 3.2, we therefore extend this result to the multivariate case with a logistic missingness mechanism. Note first that with a logistic distribution, a total ordering cannot be defined. Indeed, for variable j, such an ordering cannot be defined if the two univariate variances are equal (*i.e.,* $\sigma^2_{kj} = \sigma^2_{(k+1)j}$) and $\mu_{kj} - \beta_{kj} - \mu_{(k+1)j} + \beta_{(k+1)j} = 0$. However, for the specific case of a Gaussian mixture where all the univariate variances are different between the components, then conditions of Proposition 3.1 hold true with a logistic missing-data distribution and so does its identifiability. In addition, for parsimonious MNAR models for which the effect on the variable j does not depend on the class membership k (i.e. $\beta_{kj} = \beta_{(k+1)j}$), the conditions of Proposition 3.1 hold true with a logistic missingness mechanism. Finally, as stated in Proposition 3.2 (proved in Biernacki et al. (2021)), the condition on the covariance matrices (including the case of a homoscedastic Gaussian mixture) can be relaxed to obtain the generic identifiability of the model (*i.e.,* all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure; Allman, Matias, and Rhodes (2009)).

Proposition 3.2. Assume that $\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\delta}_k)$ is the density of a multivariate Gaussian mixture, that ρ is the logistic function and that the missingness scenario is defined by (3.7) and (3.9), then $\boldsymbol{\theta}$ is generically identifiable (*i.e.,* all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure) up to label swapping.

Proposition 3.2 can also be applied for variables with integer value (*i.e.,* count data), as shown in Biernacki et al. (2021) for the Poisson mixture with probit or logistic missing-data distributions. When the data are categorical, the parameters of the general model defined by (3.5)-(3.7) and (3.9) are no longer identifiable. For such data, only a parsimonious version that allows the missingness process to only depend on the class membership is identifiable.

3.3.3 Parameter estimation

The model defined by (3.5)-(3.7) and (3.9) is not ignorable, thus it requires a specific inference procedure for estimating $\boldsymbol{\theta}$. This section gathers the description of the EM and SEM algorithms for Gaussian, multinomial and mixed data with MNAR models for maximum likelihood estimation. Details of the algorithms are given in Biernacki et al. (2021). The observed log-likelihood is defined as follows

$$\ell(\boldsymbol{\theta}; \mathbf{x}^{\text{obs}}, \mathbf{r}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \int \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\delta}_k) f_k(\boldsymbol{r}_i | \boldsymbol{x}_i; \boldsymbol{\psi}_k) d\boldsymbol{x}_i^{\text{miss}} \right).$$

Thus, the complete log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{r}, \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \log(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\delta}_k) f_k(\boldsymbol{r}_i | \boldsymbol{x}_i; \boldsymbol{\psi}_k)).$$

We first detail the EM algorithm for the different MNAR models at hand with Gaussian, multinomial and mixed mixture models. Initialized at the parameter $\boldsymbol{\theta}^{[0]}$, iteration s of the algorithm is composed of the following two steps

- **E-step:** Computation of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]}) = \mathbb{E}[\ell(\boldsymbol{\theta}; \mathbf{R}, \mathbf{X}, \mathbf{Z}) | \mathbf{X}^{\text{obs}}, \mathbf{R}; \boldsymbol{\theta}^{[s]}]$ which is the expected complete log-likelihood ℓ_{comp} knowing the observed data and a current value of the parameters. This quantity can be decomposed into two parts as follows

  $$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]}) = Q_{\mathbf{x}}(\boldsymbol{\pi}, \boldsymbol{\delta}; \boldsymbol{\theta}^{[s]}) + Q_{\mathbf{r}}(\boldsymbol{\psi}; \boldsymbol{\psi}^{[s]}),$$

  with

  $$Q_{\mathbf{x}}(\boldsymbol{\pi}, \boldsymbol{\delta}; \boldsymbol{\theta}^{[s]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{[s]} \ln \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{[s]} \mathbb{E}\left[\ln f_k(\mathbf{X}_i; \boldsymbol{\delta}_k) \mid \boldsymbol{X}_i^{\text{obs}}, Z_{ik} = 1, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]} \right],$$

  and

  $$Q_{\mathbf{r}}(\boldsymbol{\psi}; \boldsymbol{\psi}^{[s]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{[s]} \mathbb{E}\left[\ln f_k(\boldsymbol{R}_i | \mathbf{X}_i; \boldsymbol{\psi}_k) \mid \boldsymbol{X}_i^{\text{obs}}, Z_{ik} = 1, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]} \right].$$

  where $\tau_{ik}^{[s]} := \mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{X}_i^{\text{obs}}, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]}) \quad \propto \quad \pi_k^{[s]} f_k(\boldsymbol{X}_i^{\text{obs}}; \boldsymbol{\delta}_k^{[s]}) f_k(\boldsymbol{R}_i | \mathbf{X}_i; \boldsymbol{\psi}_k^{[s]})$.

- **M-step:** Maximization over $\boldsymbol{\theta}$ of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]})$, by maximizing $Q_{\mathbf{x}}(\boldsymbol{\pi}, \boldsymbol{\delta}; \boldsymbol{\theta}^{[s]})$ w.r.t. $(\boldsymbol{\pi}, \boldsymbol{\delta})$ and $Q_{\mathbf{r}}(\boldsymbol{\psi}; \boldsymbol{\psi}^{[s]})$ respectively with respect to $\boldsymbol{\psi}$. This step leads to the parameters $\boldsymbol{\theta}^{[s+1]}$.

Computation of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[s]})$ requires the evaluation of integrals defined by the conditional expectations $\mathbb{E}\left[\ln f_k(\mathbf{X}_i; \boldsymbol{\delta}_k) \mid \boldsymbol{X}_i^{\text{obs}}, Z_{ik} = 1, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]}\right]$ and $\mathbb{E}\left[\ln f_k(\boldsymbol{R}_i|\mathbf{X}_i; \boldsymbol{\psi}_k) \mid \boldsymbol{X}_i^{\text{obs}}, Z_{ik} = 1, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]}\right]$ that depend on the MNAR model at hand. Indeed, these integrals can have a closed-form if parsimonious constrains are considered such that the missingness process depends on $\mathbf{Z}_i$ but not on $\boldsymbol{X}_i^{\text{miss}}$ (*i.e.,* all the υ_{kj} are free but all the β_{kj} equals zero in (3.9)). However, in the general case, the integrals involved in these expectations do not have closed forms. Thus, we propose using a stochastic EM (SEM; Celeux, Chauveau, and Diebolt (1996)) algorithm to cirumvent this issue. By imputing missing values using a Gibbs sampler instead of integrating over them. In addition, it has another possible advantage over the EM algorithm since it is not trapped by the first local maximum encountered of the likelihood function (Celeux, Chauveau, and Diebolt (1996)).

The SEM algorithm consists of the following two steps for $s_{\max}$ iterations (details of these steps are given in Biernacki et al. (2021)):

- **SE-step**: Draw the missing data $\mathbf{X}_i^{\text{miss}[s+1]}$ and $\mathbf{Z}_i^{[s+1]}$ according to the conditional distribution of $\boldsymbol{X}_i^{\text{miss}}, \mathbf{Z}_i$ given $\boldsymbol{X}_i^{\text{obs}}, \boldsymbol{R}_i$ and the parameters $\boldsymbol{\theta}^{[s]}$. Since it is not convenient to simulate from this conditional distribution, we simulate instead from the following two easier conditional probabilities using a Gibbs sampling approach:

$$\mathbf{Z}_i^{[s+1]} \sim \mathbf{Z}_i \mid \mathbf{X}_i^{[s]}, \boldsymbol{R}_i, \boldsymbol{\theta}^{[s]} \text{ and } \mathbf{X}_i^{\text{miss}[s+1]} \sim \boldsymbol{X}_i^{\text{miss}} \mid \boldsymbol{X}_i^{\text{obs}}, \mathbf{Z}_i^{[s+1]}, \boldsymbol{R}_i, \boldsymbol{\theta}^{[s]}, \tag{3.10}$$

  where $\mathbf{X}_i^{[s]} = (\mathbf{X}_i^{\text{obs}\top}, \mathbf{X}_i^{\text{miss}[s]\top})^\top$. For the latter distribution, we can draw the membership k of $\mathbf{Z}_i^{[s+1]}$ from the multinomial distribution with probabilities $\mathbb{P}(Z_{ik} = 1 \mid \mathbf{X}_i^{[s]}, \boldsymbol{R}_i; \boldsymbol{\theta}^{[s]})$ for $k = 1, \ldots, K$.

- **M-step**: Maximization of the completed log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{R}, \mathbf{X}^{[s+1]}, \mathbf{Z})$ over $\boldsymbol{\theta}$, which provides $\boldsymbol{\theta}^{[s+1]}$.

3.4 Semi-parametric mixture for non-ignorable missingness

3.4.1 The model

A wide range of literature focuses on models assuming that conditionally on knowing the particular subpopulation the subject i came from, its coordinates $\mathbf{X}_i$ are independent (see model defined by (2.3)). Thus, in Du Roy de Chaumaray and Marbac (2020), we extend this model for non-ignorable missingness. The paris of variables $(X_{ij}, R_{ij})^\top$ are assumed to be conditionally independent given $\mathbf{Z}_i$. Thus, the distribution of $\boldsymbol{R}_i \mid \mathbf{Z}_i$ is a product of Bernoulli distributions and the conditional density of $\mathbf{X}_i \mid \mathbf{Z}_i, \boldsymbol{R}_i$ is defined as the product of univariate densities. Thus, from (3.2), the pdf of component k is also defined by

$$f_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) = \prod_{j=1}^{J} \tau_{kj}^{r_{ij}} (1 - \tau_{kj})^{1 - r_{ij}} \text{ and } f_k(\boldsymbol{x}_i \mid \boldsymbol{r}_i) = \prod_{j=1}^{J} p_{kj}^{r_{ij}}(x_{ij}) q_{kj}^{1-r_{ij}}(x_{ij}), \tag{3.11}$$

where $\boldsymbol{\tau}_k = (\tau_{k1}, \ldots, \tau_{kJ})$, $\tau_{kj} > 0$ is the probability that X_{ij} is observed given that subject i belongs to subpopulation k, $p_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 1$ and $q_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 0$. Thus, clustering is achieved by modeling, for each subpopulation, the marginal probability of missingness and the conditional density given that the variable is observed. Integrating out the unobserved variables

$\boldsymbol{X}_i^{\text{miss}}$ (*i.e.,* the elements of vector $\mathbf{X}_i$ such that $R_{ij} = 0$), we obtain the following pdf for the distribution of the observed variables

$$f(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}), \tag{3.12}$$

where the pdf of component k is a specific version of (2.3) defined by

$$f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = f_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) \prod_{j=1}^{J} p_{kj}^{r_{ij}}(x_{ij}), \tag{3.13}$$

where $\boldsymbol{\theta}$ groups all the finite parameters (π_k and $\boldsymbol{\tau}_k$) and all the infinite parameters $p_{kj}(\cdot)$. Note that cluster analysis does not require estimating $q_{kj}(\cdot)$ because this quantity does not appear in the posterior probabilities of classification given by (3.4). The resulting model allows custering to take into account the missingness mechanism because the whole vector $\boldsymbol{R}_i$ is considered in (3.13) and thus in the computation of the posterio probabilities of classification (see (3.4)). Thus, missing values impact the posterior probabilities of classification through the parameters τ_{kj}'s used for modeling the binary variables R_{ij}. Note that a subject presenting missing values for each variable (*i.e.,* $R_{ij} = 0$ for any j) has a probability $\pi_k f_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) / \sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{r}_i; \boldsymbol{\tau}_\ell)$ to belong to cluster k that is different from the probability obtained under ignorable mechanism (*i.e.,* in this case the probability is π_k). Note that the mechanism is strongly ignorable for clustering is also covered by the approach, because if such a situation occurs, then the conditional distributions of $\boldsymbol{R}_i$ given the cluster membership, are equal for each cluster (*i.e.,* the vector of probability of responses are equal among components: $\boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_K$). Finally, note that the approach allows the mechanism of missingness, for variable j, to depend on the cluster membership and/or on the value of the variable itself. Moreover, model (3.12)-(3.13) allows the missing values to have a wide range of influence on the posterior probabilities of classification as shown by the following example.

Example 3.1 (Impact of the missingness mechanism on clustering). We consider a mixture model of K components such that the distribution of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ under component k is defined by

$$f_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = \prod_{j=1}^{J} g_j(x_{ij} - \mu_{kj}) \left[\Psi(\gamma_k + \delta_j x_{ij})\right]^{r_{ij}} \left[1 - \Psi(\gamma_k + \delta_j x_{ij})\right]^{1-r_{ij}},$$

where $g_1, \ldots, g_J$ are known densities and Ψ is a known function defined on $[0, 1]$ which represents the missingness mechanism. We show that the distribution of $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top)^\top$ under component k can be defined from (3.11) with

$$\tau_{kj} = \int g_j(x_{ij} - \mu_{kj}) \Psi(\gamma_k + \delta_j x_{ij}) dx_{ij},$$

$$p_{kj}(x_{ij}) = \frac{g_j(x_{ij} - \mu_{kj})}{\tau_{kj}} \Psi(\gamma_k + \delta_j x_{ij}) \text{ and } q_{kj}(x_{ij}) = \frac{g_j(x_{ij} - \mu_{kj})}{1 - \tau_{kj}} \left[1 - \Psi(\gamma_k + \delta_j x_{ij})\right].$$

Clustering is achieved by considering the distribution of the observed values (3.13). This framework allows for different situations:

- The missingness mechanism can depend on the component only (*i.e.,* $\delta_j = 0$). If $\gamma_1 \neq \ldots \neq \gamma_K$, then the τ_{kj} are not equals if the μ_{kj} are not. In this case, $\tau_{kj} = \Psi(\gamma_k)$ and $p_{kj} = q_{kj} = g_j$. Usually when the distributions of a *pattern-mixture model* are equal, then

the mechanism is ignorable (see Molenberghs et al. (2014)). However, as the component membership is not observed, the mechanism here is non-ignorable and it can be interpreted as a conditional *MAR* given the component membership.

- If the mechanism only depends on j (*i.e.,* $\gamma_1 = \ldots = \gamma_K$ and $\delta_1 \neq \ldots \neq \delta_J$) then τ_{kj} are different if the μ_{kj} are. Note that the difference of the μ_{kj} is required to have different distributions for the mixture components.
- The clustering is only explained by the mechanism if $\mu_{kj} = 0$ and $\gamma_1 = \ldots = \gamma_K$.
- The mechanism is strongly ignorable for clustering (but not for density estimation) if $\delta_j = 1$ and $\mu_{kj} = -\delta_k$ for any (j, k).

Thus, one can consider the case where the partition is only explained by the missing values: *i.e.,* the distribution of $\mathbf{X}_i$ is the same in each component, but the conditional distribution of $\boldsymbol{R}_i$ given $\mathbf{X}_i$ is not. In such a case, the probabilities τ_{kj} are not the same between the components and the distributions of the observed variables per components p_{kj} are generally not the same either. Alternatively, a strongly ignorable missingness mechanism can be considered and this case can be easily detected because it implies that for any $\boldsymbol{R}_i \in \{0, 1\}^J$, we have for any (k, ℓ), $f_k(\boldsymbol{r}_i; \tau_k) = f_\ell(\boldsymbol{r}_i; \tau_\ell)$.

With model (3.12)-(3.13), we are able to achieve clustering because the posterior probabilities of classification are available, however, as $q_{kj}(\cdot)$ is not estimated, we are not able to estimate the conditional distribution of $\mathbf{X}_i$ given $\mathbf{Z}_i$ or any information on this distribution (*e.g,* the conditional expectation of X_{ij} given $\mathbf{Z}_i$ cannot be computed but only the conditional expectation of X_{ij} given $\mathbf{Z}_i$ and $R_{ij} = 1$). Avoiding the estimation of $q_{kj}(\cdot)$ is the core of the proposed approach. Indeed, estimating $q_{kj}(\cdot)$ requires information about the missingness mechanism (that is generally unknown) as it can only be achieved by estimating the joint distribution of (X_{ij}, R_{ij}). Molenberghs et al. (2008) show that different models used for the distribution of (X_{ij}, R_{ij}) can lead to the same distribution of the observed variables. Thus, supplementary information is needed for consistently estimating q_{kj}. As our approach only considers the marginal probabilities of missingness (for each variable given the component), we avoid the issue of lack of identifiability (see the following lemma) and we are able to estimate the posterior probabilities of classification (but not the pdf of $\mathbf{X}_i$ for each component).

Sufficient conditions for the model identifiability are stated by Lemma 3.1. Its proof uses some results on the identifiability of nonparametric mixtures (Theorem 8 of Allman, Matias, and Rhodes (2009)) and is presented in Du Roy de Chaumaray and Marbac (2020).

Lemma 3.1. If $J \geq 3$, $\pi_k > 0$ and $\tau_{kj} > 0$, and if the densities p_{kj} are linearly independent, then the model defined by (3.11)-(3.13) is identifiable, up to label swapping.

Note that the assumptions of Lemma 3.1 are not stronger than those of Theorem 8 of Allman, Matias, and Rhodes (2009) except that we need one mild condition on the missingness process (*i.e.,* $\tau_{kj} > 0$ means that the probability of observing variable j is not zero for any component k). Indeed, the need to consider at least three variables is explained by the use of the Kruskal's Theorem which is at the core of the results of Allman, Matias, and Rhodes (2009). The assumption of linear independence of the densities is equivalent to the linear independence of the cumulative distribution functions and is not a stringent assumption (see Lemma 17 in Allman, Matias, and Rhodes (2009)). The fact that identifiability holds up to label swapping is standard in clustering, because the labels of the components of mixture models can be permuted without changing the pdf of the model. Finally, note that the assumptions of Lemma 3.1 allow all the τ_{kj} to be equal to one, corresponding to the case where there is no missingness. Note that if the data to cluster are univariate or bivariate, the proposed approach cannot be used because model

identifiability is not proven. In such cases, alternative models (semi-parametric location-scale model or parametric models) should be considered.

3.4.2 Smoothed likelihood

To perform parameter estimation, we maximize the smoothed likelihood by extending the approach of Levine, Hunter, and Chauveau (2011) to the case of mixed-type variables. Indeed, despite the fact that all the elements of $\mathbf{X}_i$ are continuous, the vector of indicator of response $\boldsymbol{R}_i$ is binary, thus the vector of the observed variables $(\mathbf{X}_i^{\text{obs}\top}, \boldsymbol{R}_i^\top)^\top$ is a vector of mixed-type variables. Note that the smoothing is only performed on the densities because these quantities are estimated non-parametrically. Thus, smoothing is performed on the distributions of $\boldsymbol{X}_i^{\text{obs}}$ for each component and there is no need to smooth the distributions of $\boldsymbol{R}_i$ for each component because these distribution are just defined as a product of probabilities (see (3.11)).
Let S be the smoothing operator defined by

$$\mathcal{S}f_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i) = \prod_{j=1}^{J} \left(\mathcal{S}p_{kj}(x_{ij})\right)^{r_{ij}},$$

and

$$\mathcal{S}p_{kj}(x_{ij}) = \int_{\Omega_j} \frac{1}{h_j} K\left(\frac{x_{ij}-u}{h_j}\right) p_{kj}(u)du,$$

where K is a kernel function and $h_j > 0$ its bandwidth. We consider the non-linear smoothing operator defined by

$$\mathcal{N}f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = f_k(\boldsymbol{r}_i; \boldsymbol{\tau}_k) \exp\{\mathcal{S} \ln f_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i)\},$$

where $f_k(\boldsymbol{x}_i^{\text{obs}} \mid \boldsymbol{r}_i) = \prod_{j=1}^{J} p_{kj}^{r_{ij}}(x_{ij})$. The smoothed log-likelihood function is defined by

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}f_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta})\right).$$

Parameter estimation is performed by maximizing the smoothed likelihood over $\boldsymbol{\theta}$. This maximization is achieved by an MM algorithm presented in the next section.

3.4.3 Majorization-Minimization algorithm

The maximization on $\boldsymbol{\theta}$ of the smoothed log-likelihood function is performed via an MM algorithm. This iterative algorithm starts at the initial value of the parameters $\boldsymbol{\theta}^{[0]}$. At iteration $[s]$, it performs the following two steps

- Computing the smoothed probabilities of subpopulation memberships

$$t_{ik}(\boldsymbol{\theta}^{[s]}) = \frac{\pi_k^{[s]} \mathcal{N}g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}^{[s]})}{\sum_{\ell=1}^{K} \pi_\ell^{[s]} \mathcal{N}g_\ell(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}^{[s]})}.$$

- Updating the estimators
  - Updating of the proportions

$$\pi_k^{[s+1]} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[s]}).$$

– Updating of the parameters of the missingness mechanism

$$\tau_{kj}^{[s+1]} = \frac{\sum_{i=1}^{n} r_{ij} t_{ik}(\boldsymbol{\theta}^{[s]})}{\sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{[s]})}.$$

– Updating of the conditional distribution

$$p_{kj}^{[s+1]}(u) = \frac{\sum_{i=1}^{n} r_{ij} t_{ik}(\boldsymbol{\theta}^{[s]}) \frac{1}{h_j} K\left(\frac{x_{ij}-u}{h_j}\right)}{\sum_{i=1}^{n} r_{ij} t_{ik}(\boldsymbol{\theta}^{[s]})}.$$

The monotonicity of the algorithm is stated by Lemma 3.2 whose proof is similar to the proof of Theorem 1 in Levine, Hunter, and Chauveau (2011) and is detailled in Du Roy de Chaumaray and Marbac (2020). This implies that the algorithm converges to a local optimum of the smoothed log-likelihood, hence different random initializations should be performed.

Assumption 3.2. For any $1 \leq j \leq J$, any $1 \leq k \leq K$ and any $x_{ij} \in \mathbb{R}$, we suppose that $p_{kj} \in L_1(\mathbb{R})$ and that $\int_{\mathbb{R}} \frac{1}{h_j} K\left(\frac{x_{ij}-u}{h_j}\right) \ln p_{kj}(u) du < +\infty$.

Lemma 3.2. Let the assumptions of Lemma 3.1, Assumptions 3.2 hold true and that for any j, $\sum_{i=1}^{n} r_{ij} \geq 1$. Let $\boldsymbol{\theta}^{[s]}$ and $\boldsymbol{\theta}^{[s+1]}$ be the estimators obtained at iterations $[s]$ and $[s+1]$ respectively, we have $\ell_n(\boldsymbol{\theta}^{[s]}) \leq \ell_n(\boldsymbol{\theta}^{[s+1]})$.

Note that, due to the evaluation of the integrals required for the computation of the $t_{ik}(\boldsymbol{\theta}^{[r]})$, the MM algorithm has a larger complexity that the EM-like algorithm proposed by Benaglia, Chauveau, and Hunter (2009) for estimating semi-parametric mixture models with no missingness. However, the assumption of independence within components permits to consider integrals that are only univariate. Moreover, extension of the EM-like algorithm for estimating the proposed model is straightforward. However, as explained by Benaglia, Chauveau, and Hunter (2009), the resulting algorithm would not have the ascendant property and the objective function would not be clearly defined. Note that the evaluation of the integrals implied by $t_{ik}(\boldsymbol{\theta}^{[r]})$ as a computational cost that make attractive parametric mixture models and their maximum likelihood estimation via an EM algorithm. However, this argument does not hold under non-ignorable missingness. Indeed, in such case, some integrals having no closed forms appear at the E-step of the EM algorithm used to fit Gaussian mixture model with logit or probit missingness process (see Miao, Ding, and Geng (2016)) and thus needs to be numerically evaluated.

3.4.4 Simulated data

This section compares the different clustering methods that consider the missingness mechanism (including the two approaches we proposed). During all the experiments we use a Gaussian kernel with bandwidths $h_j = C_j n^{-1/5}$ where C_j is the standard deviation of the observed realizations of variable j.In these simulations, different distributions for the components and missingness mechanisms are considered. Moreover, we investigate the influence of four quantities: the rate of missingness, the sample size, the number of variables and the theoretical rate of misclassification. Method comparison is done according to the Adjusted Rand index (ARI; Hubert and Arabie (1985)) computed between the true partition and the estimated partition provided by the competing method. Thus, the closer to one the ARI is, the closer the true and the estimated partitions are.

Competing methods The proposed method, implemented in the R package **MNARclust**, is compared to the following three methods:

- *Ignorable-GMM*: Gaussian mixture assuming that the missingness mechanism is ignorable (implemented in the R package **VarSelLCM** Marbac and Sedki (2018);
- *K-pod*: K-pod approach performed with the function *kpod* of the R package **kpodclustr** Chi and Chi (2014);
- *NPimputed*: non-parametric mixture on the imputed data performed with the functions *np* and *imputePCA* of the R packages **mixtools** Benaglia et al. (2009a) and **missMDA** Josse and Husson (2016).

Simulation setup To compare the different methods of clustering, we generate complete data from mixture models with three components having unequal proportions ($\pi_1 = 1/2$ and $\pi_2 = \pi_3 = 1/4$) and independence between variables within components such that

$$X_{ij} = \delta \sum_{k=1}^{3} \lambda_{kj} Z_{ik} + \varepsilon_{ij},$$

where all the $\lambda_{11} = \lambda_{22} = \lambda_{33} = \lambda_{14} = \lambda_{25} = \lambda_{36} = 1$ and the other $\lambda_{kj} = 0$ and where ε_{ij} are independent from all the variables and define the distribution within-components (Gaussian, Student with three degrees of freedom, Laplace and Skewed Gaussian with shape equals to three). Then, we add missing values from four scenarios:

- MCAR: $\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \mathbf{Z}_i) = (1 + \exp(\gamma))^{-1}$;
- MNAR-logit-Z: $\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \mathbf{Z}_i) = (1 + \exp(\gamma + 2 \sum_{k=1}^{K} z_{ik}))^{-1}$;
- MNAR-logit-X: $\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \mathbf{Z}_i) = (1 + \exp(\gamma + x_{ij}))^{-1}$;
- MNAR-censoring-X: $\mathbb{P}(R_{ij} = 0 \mid X_{ij}, \mathbf{Z}_i) = \mathbb{1}_{\{X_{ij} < \gamma\}}$.

Thus, the parameters δ and γ allow us to set the rates of misclassification and missingness (their values under the different scenarios are given in Du Roy de Chaumaray and Marbac (2020)).

Impact of the rate of missingness To investigate the impact of the rate of missingness, we consider data sets composed of $n = 100$ observations described by $J = 6$ variables with a theoretical misclassification of 10%. For each scenario, we generated 100 data sets. Figure 3.1 presents the boxplots of ARI between the true partition and the estimators of the partition given by the methods. Overall, the proposed method outperforms the competing methods under non-ignorable mechanims. Indeed, its results are robust to the different distributions of the components, the missingness scenarios and the missingness rates. Moreover, when the mechanism is ignorable, all the methods obtain good similar performances. Note that the parametric approach assuming that the missingness mechanism is ignorable yields slightly better results, when the distribution within components is Gaussian or skewed Gaussian. However, this approach yields poor results when the missingness mechanism is not ignorable. Under the non-ignorable scenario, the proposed approach yields the best results. The results of the other methods stay relevant under the logit-X scenario and Gaussian or Skewed-Gaussian components. However, in the other scenarios, they produce poor results. Finally, note that the larger the missingness rate is, the larger the benefit of the proposed method is. Indeed, the results of the proposed method seems not to be impacted by the missingness rate while the results of the other methods are deteriorated when this rate is increasing, under non-ignorable mechanisms.

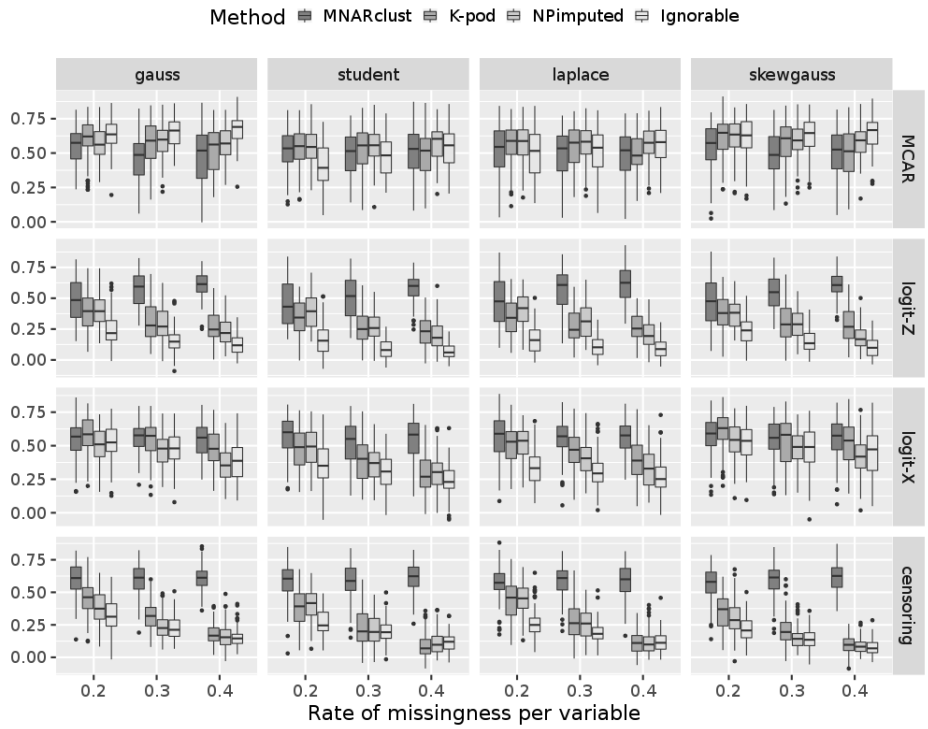


Figure 3.1: Boxplot ARI computed from 100 samples of $n = 100$ observations described by $J = 6$ variables with a misclassification rate of 10%.

Consistency of the estimators To illustrate the consistency of the estimators, we consider data sets composed of observations described by $J = 6$ variables with a theoretical misclassification of 10% and a theoretical missing rate per variable of 30%. For each scenario, we generated 100 data sets. Figure 3.2 presents the boxplot of the ARI between the true partition and the estimators of the partition given by the methods. Again, results show that the method outperforms the competing methods because it is more robust with respect to the distribution of the components and to the missingness scenario. Moreover, despite the fact that the accuracy of the partition is improved when the sample size increases, the results are satisfactory (compared to the results of the parametric methods) even for small samples.

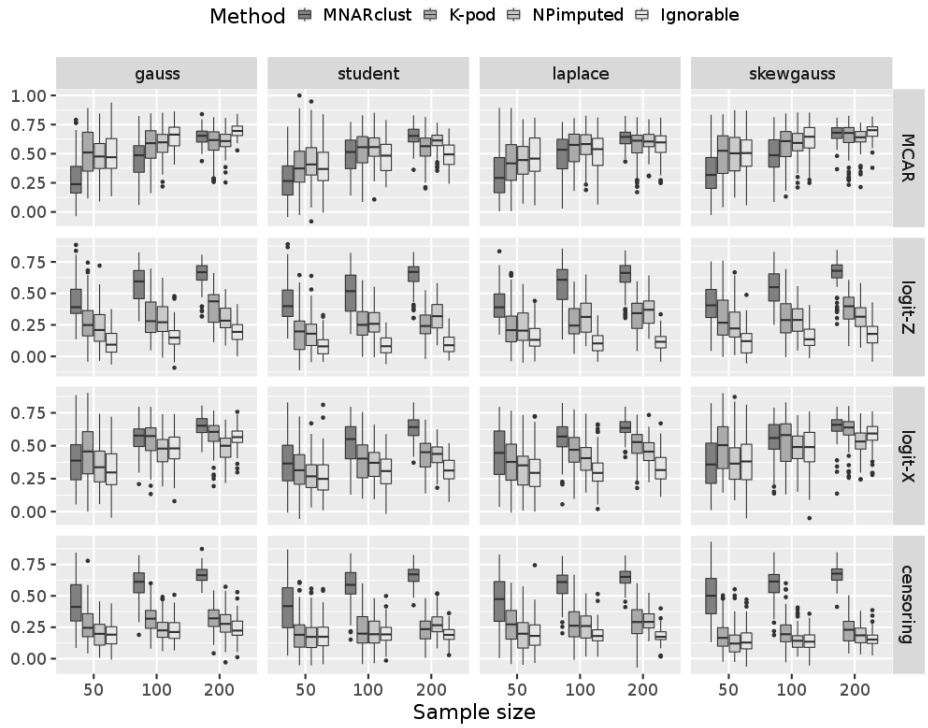


Figure 3.2: Boxplot ARI computed from 100 samples composed of $d = 6$ variables having a missing rate of 30% each with a misclassification rate of 10%.

Impact of the dimension To illustrate the impact of the dimension, we consider data sets composed of $n = 100$ observations generated with a theoretical misclassification of 10% and a theoretical missing rate per variable of 30%. For each scenario, we generated 100 data sets. Figure 3.3 presents the boxplot of the ARI between the true partition and the estimators of the partition given by the methods. Despite the fact that the proposed method is semi-parametric, the results show that it can manage data set with many variables. Indeed, the deterioration of the results of the proposed method when J increases is very weak. This is due to the assumption of conditional independence between the pairs $(X_{ij}, R_{ij})^\top$ given the components membership. Indeed, this assumption permits the impact of the curse of the dimensionality for the nonparametric estimators to be limited.

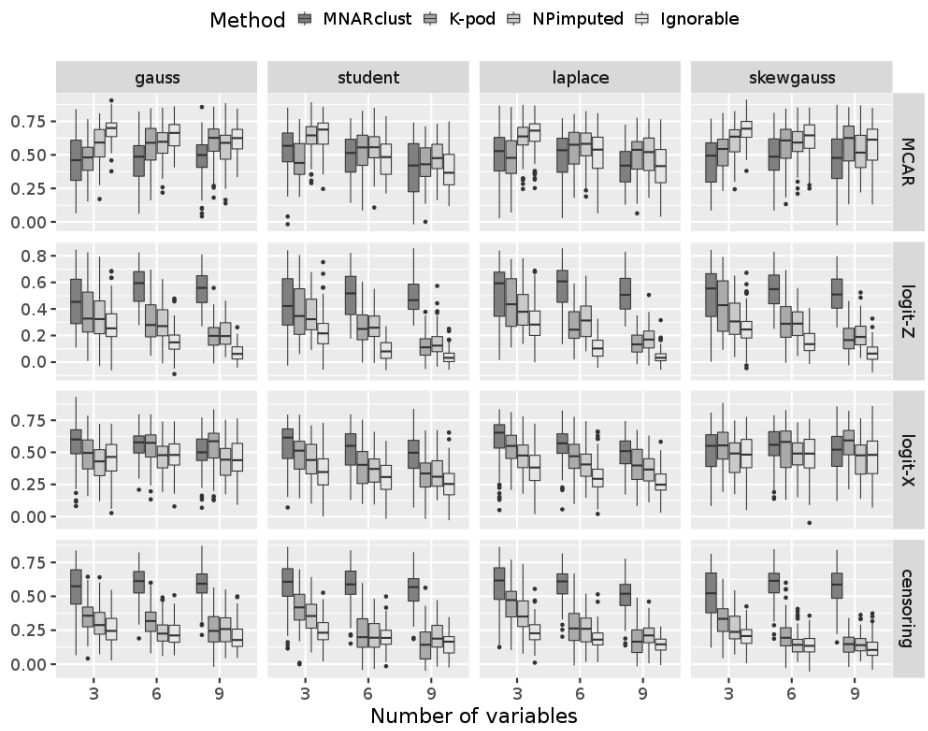


Figure 3.3: Boxplot ARI computed from 100 samples composed of $n = 100$ observations having a missing rate of 30% per variable and a misclassification rate of 10%.

Impact of the theoretical misclassification To illustrate the impact of the overlaps between components, we consider data sets composed of $n = 100$ observations generated with $J = 6$ and a theoretical missing rate per variable of 30%. For each scenario, we generated 100 data sets. Figure 3.4 presents the boxplot of the ARI between the true partition and the estimators of the partition given by the methods. Overall, all the methods perform well under the MCAR mechanism despite the fact that the results of the proposed methods are more deteriorated than those of the other methods when the misclassification rate is high. Howerver, under the non-ignorable scenarios, the proposed method outperforms the competing methods.

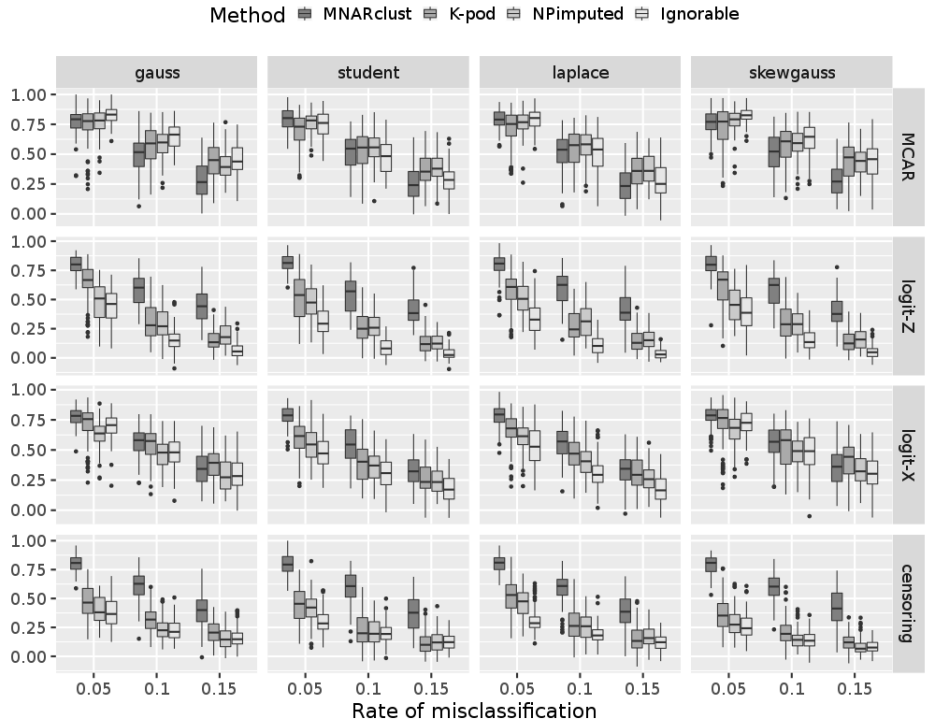


Figure 3.4: Boxplot ARI computed from 100 samples composed of $n = 100$ observations described by $d = 6$ variables having a missing rate of 30%.

3.4.5 Benchmark data

We consider two data sets (*Swiss banknotes* and *Italian wines*) described below, to illustrate the behavior of the proposed method. The Swiss banknotes data set (Flury and Riedwyl (1988)) contains six measurements (length of bill, width of left edge, width of right edge, bottom margin width, top margin width and length of diagonal) made on 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes. This data set is available in the R package **mclust** Scrucca et al. (2016). The status of the banknote (genuine or counterfeit) is also known. We perform the clustering of the bills based on the six morphological measurements and we evaluate the resulting partition with the status of the bills. The Italian wine data set records 27 physical and chemical measurements on 178 Italian wines grown in the same region in Italy but derived from three different cultivars (Barbera, Barolo and Grignolino) and five years of production (1970,

1973, 1974, 1976 and 1979). The data set (Forina et al. (1986)) is available on the R package **MBCbook** (companion R package of Bouveyron et al. (2019)). Clustering of the wines based on the 27 physical was done and chemical measurements and we compare the resulting partition with the three cultivars and the year of production.

The original data does not have missing values. To investigate the behavior of the proposed method, we generate data sets from the original data by adding missing values drawn from three different mecanisms: *MCAR* where the probability to unobserve each value is γ, *MNARZ* where the probability to unobserve a value depends on the true class memberships (*i.e.,* for the Swiss banknote and the Italian wine data sets, the true class memberships are defined by variable bill status and cultivars respectively; so this probability that a variable is unobserved is 0.5γ and γ for the counterfeit and genuine bills respectively and 0.5γ, γ and 1.5γ for the Barbera, Barolo and Grignolino respectively), *MNARcensoring* where a variable is fully observed if its value is more than the empirical quantile of the variable at level $\sqrt{\gamma}$ and observed with probability $\sqrt{\gamma}$ otherwise. For each mecanism, 100 data sets are generated from the original data. Each generated data set is analyzed by the four competing methods: nonparametric proposed method, Kpod algorithm, Gaussian mixture model assuming ignorability of the mecanisms and a two-step approach that first impute data using PCA then fit a non-parametric mixture on the complete data. Figure 3.5 presents the boxplot of the ARI obtained, by the four competing methods, on 100 samples generated under each scenario and for different values of γ. Note that on the original data, the semi-parametric mixture model and the Gaussian mixture model with conditional independence are relevant for detecting the underlying partition. Moreover, the K-means is relevant for detecting the underlying partition on the Swiss banknote data set but it is less relevant for the wine data set. Indeed, on the Swiss banknote data set, the partitions given by semi-parametric mixture model, K-means and Gaussian mixture model have an ARI equal to 0.98, 1.00 and 0.96 respectively. Moreover, on the Italian wine data set, the partitions given by semi-parametric mixture model, K-means and Gaussian mixture model have an ARI equal to 0.98, 0.41 and 0.96 respectively. Results show that increasing the rate of missigness deteriorates the partitions. This phenomenon was expected because less discriminating information is present in the data set. Results show that the proposed method performs well under the non-ignorable scenarios while the results obtained by the alternative methods are strongly deteriorated in such a case (see MNARZ and MNARcensoring).

3.4.6 Echocardiogram data set

We consider the *Echocardiogram Data Set* Salzberg (1988) freely available in the R package **MNARclust**. This data set is composed of $n = 132$ subjects who suffered from heart attacks at some point in the past. The task is generally to determine from the other variables, whether or not the patient will survive at least one year. The data set is composed of 5 continuous variables: *age at heart attack* (missing rate 4.5%), *fractional shortening* (a measure of contracility around the heart, lower numbers are increasingly abnormal, missing rate 6.0%), *epss* (E-point septal separation, another measure of contractility, larger numbers are increasingly abnormal, missing rate 11.4%), *lvdd* (left ventricular end-diastolic dimension; this is a measure of the size of the heart at end-diastole; large hearts tend to be sick hearts, missing rate 8.3%) and *wall-motion-score* (a measure of how the segments of the left ventricle are moving, missing rate 3.0%); one binary variable *pericardial effusion* (pericardial effusion is fluid around the heart, 0=no fluid, 1=fluid, missing rate 0.7%). We also have one binary variable which can be used as a partition among the subjects: *still alive* (0=dead at end of survival period, 1 means still alive). This binary variable is not used for clustering but permits the accuracy of the estimated partition to be evaluated. Among the variables used for clustering, there are 5.7% of missing values and

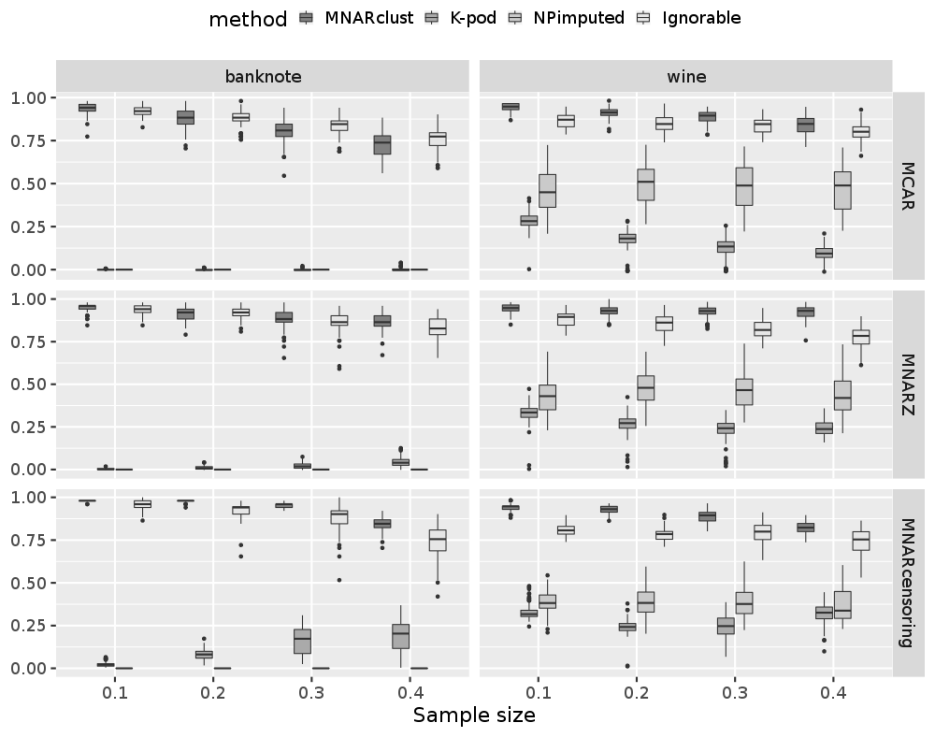


Figure 3.5: Boxplot ARI obtained 25 samples composed genrated from the original data Swiss Banknotes (banknote) and Italian wines (wine).

19.1% of the sujects have at least one missing value. Moreover, the variable, *still alive*, has only one missing value.

Clustering is performed by extending the proposed approach to the case of mixed-type data (data set composed of one binary and five continuous variables). This extension is easy because of the assumption of conditional independence within components. Hence, each categorical variable is modelled by a multinomial distribution given the component and the fact that the variable is observed. Moreover, since non-parametric estimation is only performed for the densities, smoothing is only done for the continuous variables.

Chosing the number of components in a semi-parametric mixture is a difficult problem (even in the complete case). Note that methods have been developed to select this number in the case of a continuous data set (see Kasahara and Shimotsu (2014) and Kwon and Mbakop (2020)). However, they cannot be used directly on mixed-type data. Thus, we used the approach based on discretization described in Section 2.3 with a number of bins $B = [n^{1/6}]$ for selecting the number of components. This approach detects three clusters. This result is confirmed by the evolution of the maximum smoothed log-likelihood with respect to the number of clusters (see Du Roy de Chaumaray and Marbac (2020)). Figures 3.6 and 3.7 show the relation between the missingness rates and the influence on the missingness and on the observed variables on the partition. This quantity is measured by the empirical counterpart of $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid R_{ij})]$ and $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid X_{ij})]$ respectively. Thus, the higher these indexes, the more discriminative the missingness process and the observed variable. Moreover, Figures 3.6 and 3.7 show that both the missingness process and the observed variables influence the partition but that the observed variables are more discriminative (overall the values of $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid R_{ij})]$ are less than those of $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid X_{ij})]$).

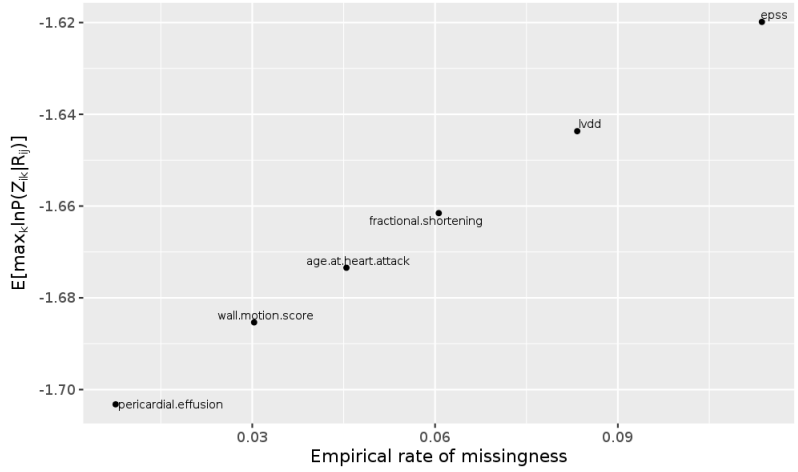


Figure 3.6: Rate of missingness and empirical counterpart of $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid R_{ij})]$ for each variable.

Table 3.1 presents a summary of the conditional distribution of the variables given the clusters.

The three unbalanced classes are mainly explained by two variables: *epss* and *lvdd*, which are highly discriminative for both the missingness mechanism and the conditional densities p_{kj}. Note that if we would consider a full-model selection, the proposed approach based on discretization would select three components and three relevant variables (fractional.shortening, epss and lvdd).

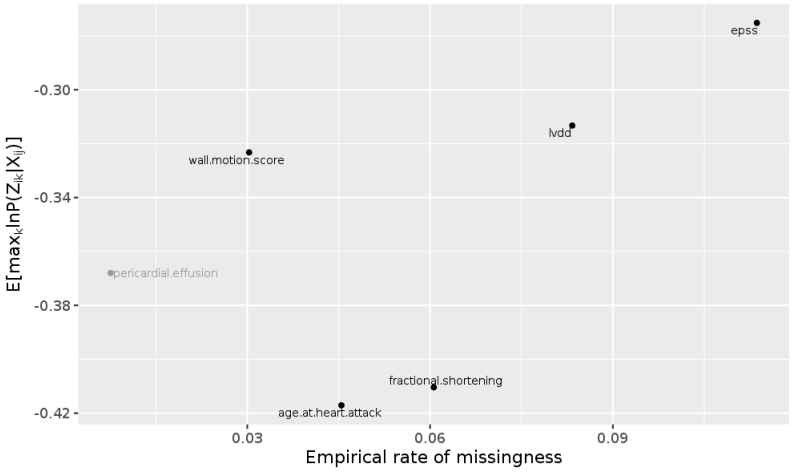


Figure 3.7: Rate of missingness and empirical counterpart of $\mathbb{E}[\max_k \ln \mathbb{P}(Z_{ik} \mid X_{ij})$ for each variable.

		age			effusion		shortening	
	τ_{kj}	mean	sd	τ_{kj}	prob.	τ_{kj}	mean	sd
class-1	0.95	64.61	8.91	1.00	0.07	1.00	0.15	0.07
class-2	0.82	65.10	7.44	0.91	0.03	0.28	0.16	0.08
class-3	0.97	61.83	7.97	1.00	0.09	1.00	0.25	0.11
		epss			lvdd		wall motion	
	τ_{kj}	mean	sd	τ_{kj}	mean	sd	τ_{kj}	mean	sd
class-1	0.97	20.01	6.97	0.97	5.56	0.64	0.98	17.36	6.28
class-2	0.28	9.68	2.07	0.10	5.30	0.08	0.82	11.99	8.05
class-3	0.93	8.88	4.53	1.00	4.43	0.62	0.99	13.50	3.12

Table 3.1: Summary of the conditional distribution of the vairables given the cluster: probability of non missing (τ_{kj}), mean and standard deviation (sd) for the continuous variables and probability of occuring for the binary variable

Note that these three variables are the most discriminative ones for the missingness process and for their conditional distribution within a component, given the fact that the variable is observed (see Figures 3.6 and 3.7). The three estimated classes can be described as follows:

- *class-1* ($\pi_1 = 0.27$) is composed of 33 subjects. These subjects are characterized by high values of the measurements of *epss*, *lvdd* and *wall-motion-score* and small of values of the measurements of *fractional shortening*. This class is characterized by a very low probability of missingness for each variables;

- *class-2* ($\pi_2 = 0.08$) is composed of 11 subjects. These subjects have suffered from heart attack being older than the subjects of the other class and obtain low values for the *wall-motion-score*. They are characterized by high probabilities of missingness for all the variables;

- *class-3* ($\pi_3 = 0.65$) is composed of 88 subjects. These subjects have suffered from heart attacks being young and have low missingness probabilities. They take low values for *epss* and *lvdd* and high values of *shortening*

As shown by the confusion matrix presented in Table 3.2, the estimated partition permits to be partially explained the death of the subject at the end of the survival period.

	Class 1	Class 2	Class 3
dead	12	4	72
still alive	21	6	16

Table 3.2: Confusion matrix between the estimated partition and the *still alive* variable. The Adjusted Rand index is 0.25

Finally, the assumption of independence within components seems to be realistic. Indeed, we investigate this assumption by testing the significance of the correlation coefficients between the conditional distribution of variables X_{ij} given the cluster membership and $R_{ij} = 1$. (the p-values obtained by testing the nullity of the correlation coefficient of the conditional distribution of pair of variables conditionally on component 1 and 3 respectively are available in Du Roy de Chaumaray and Marbac (2020)). The high values of the p-values suggest that the assumption of conditional independence given the component membership is suitable. Note that results related to component 2 are not presented due a lack of subjects assigned to this class.

3.5 Conclusion

We propose two approaches for model-based clustering under a non-ignorable missingness process. One approach uses the *selection model* approach in a parametric framework while the second approach uses the *pattern-mixture model* approach in a nonparametric framework. For clustering, we believe that the *pattern-mixture model* approach should be preferred because it turns out to be more general as it does not require the missingness mechanism to be specified and allows this mechanism to be non-ignorable. Note however that, this approach does not permit the marginal distribution of $\mathbf{X}_i \mid \mathbf{Z}_i$ to be estimated without adding assumptions about the missing mechanism. Thus, the proposed approach can be used for clustering but not for density estimation. If the marginal distribution of $\mathbf{X}_i$ is sought, we advise using the *selection model* approach.

In the context without missingness, a drawback of the MM algorithm is the computation of integrals having no closed form for computing the smoothed probabilities of subpopulation memberships. However, due to the independence within components, those integrals are only univariate. Note that the parametric mixtures (*e.g.,* Gaussian mixtures) do not suffer from this drawback, when the data are complete. However, when missingness occurs, even the estimation of the parametric mixtures via the EM algorithm, leads to compute integrals having no closed form (see Miao, Ding, and Geng (2016)). Thus, when missingness occurs, the estimation of the proposed semiparametric mixture is not more complex than the estimation of the parametric mixture. We believe that this is a supplementary argument in favor of the use of the *pattern-mixture model* approach in a nonparametric framework.

The *pattern-mixture model* approach could be extended to location or location/scale semi-parametric models. However, we believe that these models would be more suitable for modeling the distribution of the variables than rather the conditional distribution of the variables given that their values are not missing. Thus, these models would be more relevant for a *selection model* approach but their estimation would be complex without considering parsiominious contraints on the missingness process (*e.g.,* the probability of missingness only depends on the component membership).

The proposed method allows continuous data sets with non-ignorable missingness to be clustered with no more assumption than the independence within components. In some applications, the assumption of independence within components can be too strong but it can be relaxed. This point is discussed in Section 7.2.

Chapter 4

Simultaneous semi-parametric estimation of clustering and regression

Contents

4.1 Introduction

4.1.1 State of the art

Regression models allow the relationship between some covariates and a target variable to be investigated. These models are defined by an equation on the conditional moment of the transformation of the noise. This transformation is generally the piecewise derivative of the loss function

that defines the type of regression: mean, robust, quantile (see Koenker and Bassett (1978), Horowitz and Lee (2005) and Wei and Carroll (2009)), expectile (see Newey and Powell (1987), Ehm et al. (2016) and Daouia, Girard, and Stupfler (2018)).

The regression model with a fixed group effect is central within this generic paradigm. It considers that the intercept of the regression depends on the group from which the subject belongs (the intercept is common for subjects belonging to the same group but different for subjects belonging to different groups). However, in many applications, the group variable is not observed but other variables related to this variable are observed. For instance, suppose we want to investigate high blood pressure by considering the levels of physical activity among the covariates. In many cohorts, the level of physical activity of a subject is generally not directly available (because such a variable is not easily measurable) but many variables on the mean time spent doing different activities are available. Note that the regression model with a fixed group effect and a latent group variable is a specific mixture of regressions (see Wang et al. (1996), Hunter and Young (2012) and Wu and Yao (2016)) where only the intercepts of the regressions are different among the components and where the mixture weights depend on some other variables. Moreover, the regression model with a fixed group effect and a latent group variable can be interpreted as a regression model with specific quantization of the variables that we use to estimate the group membership (see for instance Charlier, Paindaveine, and Saracco (2015) for the quantization in quantile regression).

The estimation of a regression model with a fixed group effect is generally performed using a *two-step approach* as for instance in epidemiology or in economics (see Auray, Klutchnikoff, and Rouviere (2015), Ando and Bai (2016) and Zhang, Wang, and Zhu (2019)). As a first step, a clustering on the individual based on the group-related variables is performed to obtain an estimator of the group. As a second step, the regression model is fitted by using the estimator of the group variable among the covariates. The second step considers a regression model with measurement errors on the covariates. Indeed, the group variable is estimated in the clustering step with errors. Hence, it is well-known that the resulting estimators of the parameters of regression are biased (see for instance Carroll and Wand (1991), Nakamura (1992) and Bertrand et al. (2017)). The bias depends on the accuracy of the clustering step. Note that, despite the fact that the target variable contains information about the group variable (and so is relevant for clustering), this information is not used in the two-step approach, leading to suboptimal procedures.

Some simultaneous approaches have been considered in the framework of latent variable models, such as latent class and latent profile analysis (see Guo, Wall, and Amemiya (2006) and Kim et al. (2016)). In this framework, the authors introduce latent class and latent factor variables to explain the heterogeneity of observed variables. However, this approach does not focus on the conditional distribution of particular variable given other ones, and it is limited to a parametric framework. Another related reference is the work of Sammel, Ryan, and Legler (1997), where the authors introduce a latent variable mixed effect model, which allows for arbitrary covariate effects, as well as direct modeling of covariates on the latent variable. Some other relevant references can be found in the field of concomitant variables (see Dayton and Macready (1988), Grün and Leisch (2008) and Vaňkátová and Fišerová (2017)), where some additional variables are used to locally adjust the weights of the mixture of regressions. These approaches are rather focused however, on the tasks of the mixture of regressions than on clustering data based on concomitant variables.

4.1.2 Contributions

In Marbac et al. (2022), we propose a new procedure (hereafter referred to as the *simultaneous approach*) that simultaneously estimates the clustering and the regression models in a semi-parametric framework (see Hunter, Richards, and Rosenberger (2011)) thus circumventing the limits of the standard procedure (biased estimators). We demonstrate that this procedure improves both the estimators of the partition and regression parameters. A full parametric setting is also presented, however if one of the clustering or regression models is ill-specified, its bias modeling could contaminate the results of the other. Thus, we focus on a semi-parametric mixture where the component densities are defined as a product of univariate densities (see Chauveau, Hunter, and Levine (2015), Zhu and Hunter (2016a) and Zheng and Wu (2020)). Note that, mixtures of symmetric distributions (see Hunter, Wang, and Hettmansperger (2007) and Butucea and Vandekerkhove (2014)) could also be considered in a similar way. Semi-parametric inference is achieved by a maximization of the smoothed likelihood via a MM algorithm implemented in the R package **ClusPred** (Marbac et al. (2021)).

This chapter is organized as follows. Section 4.2 introduces a general context where a statistical analysis requires both methods of clustering and prediction, and it presents the standard approach that estimates the parameters in two steps. Section 4.3 shows that a procedure that allows for a simultaneous estimation of the clustering and of the regression parameters generally outperforms the two-step approach. This section also briefly presents the simultaneous procedure on a parametric framework, then focuses on the semi-parametric frameworks. Section 4.4 presents numerical experiments on simulated data showing the benefits of the proposed approach. Section 4.5 illustrates our proposition for problems associated with high blood pressure prevention. Section 4.6 provides a conclusion.

4.2 Embedding clustering and prediction models

4.2.1 Data presentation

Let $(\mathbf{V}^\top, \mathbf{X}^\top, Y)^\top$ be the set of the random variables where $\mathbf{V} = (\mathbf{U}^\top, \mathbf{Z}^\top)^\top$ is a $d_V = d_U + K$ dimensional vector used as covariate for the prediction of the univariate variable $Y \in \mathbb{R}$, $\mathbf{X}$ is a d_X-dimensional vector and $\mathbf{Z} = (Z_1, \ldots, Z_K)^\top \in \mathcal{Z}$ is a categorical variable with K levels. The variable $\mathbf{Z}$ indicates the group membership such that $Z_k = 1$ if the subject belongs to cluster k and otherwise $Z_k = 0$. The realizations of $(\mathbf{U}^\top, \mathbf{X}^\top, Y)^\top$ are observed but the realizations of $\mathbf{Z}$ are unobserved. Thus, $\mathbf{X}$ is a set of proxy variables used to estimate the realizations of $\mathbf{Z}$. Considering the high blood pressure example, Y corresponds to the diastolic blood pressure, $\mathbf{U}$ is the set of observed covariates (gender, age, alcohol consumption, obesity and sleep quality), $\mathbf{X}$ is the set of covariates measuring the level of physical activity and $\mathbf{Z}$ indicates the membership of a group of subjects with similar physical activity behaviors. The observed data are n independent copies of $(\mathbf{U}^\top, \mathbf{X}^\top, Y)^\top$ denoted by $\mathbf{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_n^\top)^\top$, $\mathbf{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ respectively. The n unobserved realizations of $\mathbf{Z}$ are denoted by $\mathbf{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$.

4.2.2 Motivating example

We use the following example throughout the chapter, which examines the general objective of high blood pressure prevention. Here, we focus on the detection of indicators related to the diastolic blood pressure (Y) (see Berney, Burnier, and Wuerzner (2018) for the interest of the study). The indicators we wish to consider are the gender, the age, the alcohol consumption, the obesity, the sleep quality and the level of physical activity ($\mathbf{V}$). However, the level of

physical activity ($\mathbf{Z}$) of a patient is not directly measured and we only have a set of variables which describes the physical activity ($\mathbf{X}$), such as practice of that recreational activity, hours spent watching TV, hours spent on the computer, *etc.* More details of the data are provided in Section 4.5. The study of the different indicators is performed using a regression model that explains the diastolic blood pressure with a set of covariates where one variable (the physical activity) was not directly observed. Information about this latter variable is available from other variables that do not appear in the regression.

4.2.3 Introducing the joint predictive clustering model

Regression model Let a loss function be $\mathcal{L}(\cdot)$ and $\rho(\cdot)$ its piecewise derivative. The loss function $\mathcal{L}$ allows the regression model of Y on $\mathbf{V}$ to be specified with a fixed group effect given by

$$Y = \mathbf{V}^\top \boldsymbol{\beta} + \varepsilon \text{ with } \mathbb{E}[\rho(\varepsilon)|\mathbf{V}] = 0, \tag{4.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top \in \mathbb{R}^{d_V}$, $\boldsymbol{\gamma} \in \mathbb{R}^{d_U}$ are the coefficients of $\mathbf{U}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^\top \in \mathbb{R}^K$ are the coefficients of $\mathbf{Z}$ (*i.e.,* the parameters of the group effect), and ε is the noise. Note that for reasons of identifiability, the model does not have an intercept. The choice of $\mathcal{L}$ allows many models to be considered and, among them, one can cite the mean regression (with $\mathcal{L}(t) = t^2$ and $\rho(t) = 2t$), the τ-quantile regression (with $\mathcal{L}(t) = |t| + (2\tau - 1)t$ and $\rho(\varepsilon) = \tau - \mathbb{1}_{\{\varepsilon \leq 0\}}$; Koenker and Bassett (1978)), the τ-expectile regression (with $\mathcal{L}(t) = |\tau - \mathbf{1}\{t \leq 0\}|t^2$ and $\rho(t) = 2t((1-\tau)\mathbf{1}\{t \leq 0\} + \tau \mathbf{1}\{t > 0\})$; Newey and Powell (1987)), *etc.*

The restriction on the conditional moment of $\rho(\varepsilon)$ given $\mathbf{V}$ is sufficient to define a model and allows for parameter estimation. However, obtaining a maximum likelihood estimate (MLE) needs specific assumptions on the noise distribution. For instance, parameters of the mean regression can be consistently estimated with MLE by assuming centered Gaussian noise. Similarly, the parameters of τ-quantile (or τ-expectile) regression can be consistently estimated with MLE by assuming that the noise follows an asymmetric Laplace (or an asymmetric normal) distribution (see Yu and Moyeed (2001) and Xing and Qian (2017)). Hereafter, we denote the density of the noise ε by f_ε.

Clustering model The distribution of $\mathbf{X}$ given $Z_k = 1$ is defined by the density $f_k(\cdot)$. Therefore, the marginal distribution of $\mathbf{X}$ is a mixture model defined by the density

$$f(\boldsymbol{x}_i; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i), \tag{4.2}$$

where $\boldsymbol{\psi} = \boldsymbol{\pi} \cup \{f_1, \ldots, f_K\}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ is the vector of proportions defined on the simplex of dimension K (*i.e.,* $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$) and where f_k is the density of component k. In a parametric approach, f_k is assumed to be parametric so it is denoted by $f_k(\cdot; \boldsymbol{\alpha}_k)$ where $\boldsymbol{\alpha}_k$ are the parameters of component k. In a semi-parametric approach, some assumptions are required to ensure model identifiability (see for instance Chauveau, Hunter, and Levine (2015)). In the following, the semi-parametric approaches are considered with the assumption that each f_k is a product of univariate densities (see Section 4.3.3).

Joint clustering and regression model The joint model assumes that $\mathbf{Z}$ explains the dependency between Y and $\mathbf{X}$ (*i.e.,* Y and $\mathbf{X}$ are conditionally independent given $\mathbf{Z}$) and that $\mathbf{U}$ and $(\mathbf{X}^\top, \mathbf{Z}^\top)^\top$ are independent. Moreover, the conditional distribution of $\mathbf{W} = (\mathbf{X}^\top, Y)^\top$ given

$\mathbf{U}$ is also a mixture model defined by the density

$$f(\boldsymbol{w}_i|\boldsymbol{u}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k), \tag{4.3}$$

where $\boldsymbol{\theta} = \boldsymbol{\pi} \cup \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\} \cup \varsigma_\varepsilon$, $\boldsymbol{\phi}_k$ grouping the parameters specific to component k (*i.e.,* the finite parameter $\boldsymbol{\delta}_k$ and the infinite parameters f_k) and $\boldsymbol{\varsigma}_\varepsilon$ grouping the parameters shared among the components (*i.e.,* the finite parameter $\boldsymbol{\gamma}$ and the infinite parameter f_ε), we have

$$\mathbb{E}[\rho(Y - \mathbf{U}^\top \boldsymbol{\gamma} - \mathbf{Z}^\top \boldsymbol{\delta})|\mathbf{V}] = 0, \tag{4.4}$$

Note that (4.3) is a particular mixture of regressions models where the mixture weights are proportional to $\pi_k f_k(\boldsymbol{x}_i)$ (thus depending on covariates that do not appear in the regressions) and where only the intercepts (*i.e.,* $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$) are different among the regressions. Contrary to Grün and Leisch (2008) who consider the density $f(y_i|\boldsymbol{u}_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ thus focusing on the regression framework, here we propose considering the density $f(\boldsymbol{w}_i|\boldsymbol{u}_i;\boldsymbol{\theta})$ which balances the regression and the clustering frameworks.

Moment condition The following lemma gives the moment equation verified on the joint model and only consider observed variables in conditioning (see Marbac et al. (2022) for the proof). It will be used later to justify the need for a simultaneous approach.

Lemma 4.1. Assume that the model is defined by (4.3), that the condition (4.4) holds true, that the covariance matrix of $\mathbf{U}$ has full rank and finally that f_{kj} and f_ε are strictly positive. Denoting $\boldsymbol{\beta}_0$ as the single parameter satisfying (4.4) and $r_k^{\mathbf{U},\mathbf{X},Y}(\boldsymbol{u}, \boldsymbol{x}, y) = \frac{\pi_k f_k(\boldsymbol{x}) f_\varepsilon(y - \boldsymbol{u}^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}) f_\varepsilon(y - \boldsymbol{u}^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_\ell)}$, we have

$$\forall k = 1, \ldots, K,\ \mathbb{E}[r_k^{\mathbf{U},\mathbf{X},Y}(\mathbf{U}, \mathbf{X}, Y)\rho(Y - \mathbf{U}^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k)|\mathbf{U}, \mathbf{X}] = 0 \Longleftrightarrow \boldsymbol{\beta} = \boldsymbol{\beta}_0. \tag{4.5}$$

4.3 Simultaneous estimation of clustering and prediction models

4.3.1 Limits of the standard two-step approach estimation

The aim is to explain the distribution of Y given $\mathbf{V} = (\mathbf{U}^\top, \mathbf{Z}^\top)^\top$ from an observed sample. A direct estimation of the model (4.1) is not doable because the realizations of $\mathbf{Z}$ are unobserved. The standard approach considers the following two-steps:

1. **Clustering step** Perform a clustering of $\mathbf{x}$ to obtain an estimated hard classification rule $\hat{r}^{\mathbf{X}} : \mathbb{R}^{dx} \to \mathcal{Z}$ or an estimated fuzzy classification rule $\hat{r}^{\mathbf{X}} : \mathbb{R}^{dx} \to \tilde{\mathcal{Z}}_K$ where $\tilde{\mathcal{Z}}_K$ is the simplex of size K.

2. **Regression step** Estimation of the regression parameters given the estimator of the group memberships $\hat{\boldsymbol{\beta}}^{\hat{r}^{\mathbf{X}}} := (\hat{\boldsymbol{\gamma}}^{\hat{r}^{\mathbf{X}}\top}, \hat{\boldsymbol{\delta}}^{\hat{r}^{\mathbf{X}}\top})^\top$ with

$$\hat{\boldsymbol{\beta}}^{\hat{r}^{\mathbf{X}}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i) \mathcal{L}(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k),$$

where $\hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i)$ is the element k of vector $\hat{r}^{\mathbf{X}}(\boldsymbol{x}_i)$. Note that $\hat{r}_k^{\mathbf{X}}(\boldsymbol{x}_i)$ is an estimator of the conditional probability that observation i belongs to cluster k given $\boldsymbol{x}_i$, if the fuzzy classification rule is used.

The following lemma (see Marbac et al. (2022) for the proof) states that the two-step approach is suboptimal . Indeed, even if the optimal classification rule on $\mathbf{X}$ is used, its expected good-classification rate is strictly smaller than that obtained by the best approach (see statement 1) and the estimators of the regression parameters are asymptotically biased (see statement 2).

Lemma 4.2. Let the model be defined by (4.3)-(4.4) where f_k and f_ε are continuous and strictly positive where there exists (k, ℓ) such f_k and f_ℓ have no disjoint support and also $\boldsymbol{\delta}_k \neq \boldsymbol{\delta}_\ell$, and finally where f_ε is not constant. Suppose that f_ε defines a random variable with finite variance and that $\mathbf{U}$ has a full rank covariance matrix. Then,

1. Any hard classification rule $\tilde{r}^X : \mathbb{R}^{d_X} \to \mathcal{Z}$ is suboptimal in the sense that
$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{r}_k^{\mathbf{X}}(\mathbf{X}) Z_k\right] < \mathbb{E}\left[\sum_{k=1}^{K} r_k^{\mathbf{U},\mathbf{X},Y}(\mathbf{U},\mathbf{X},Y) Z_k\right].$$
2. Consider the quadratic loss, the best classification rule $r^{\mathbf{X}}$ computed on $\mathbf{X}$ and its associated estimator of the regression parameters $\hat{\boldsymbol{\beta}}^{r^{\mathbf{X}}}$. The estimator $\hat{\boldsymbol{\gamma}}^{r^{\mathbf{X}}}$ is asymptotically unbiased but the estimator $\hat{\boldsymbol{\delta}}^{r^{\mathbf{X}}}$ is asymptotically biased with an asymptotic bias equals to $\frac{\sum_{\ell=1}^{K} \delta_{k\ell}\boldsymbol{\delta}_\ell}{\sum_{\ell=1}^{K} \delta_{k\ell}} - \boldsymbol{\delta}_k$, where $\boldsymbol{\delta}_{k\ell} = \mathbb{E}[r_k^{\mathbf{X}}(\mathbf{X}) r_\ell^{\mathbf{X}}(\mathbf{X})]$.

Thus the clustering step provides a suboptimal classification rule because the classification neglects the information given by Y. Consequently, the regression step provides estimators that are asymptotically biased and implies fitting the parameters of a regression model with measurement errors in the covariates (for instance, considering the hard assignment, we have no guarantee of obtaining a perfect recovery of the partition, *i.e.*,, $\hat{r}^{\mathbf{X}}(\boldsymbol{x}_i) = \boldsymbol{z}_i$, for $i = 1, \ldots, n$). The measurement errors generally produce biases in the estimation. Finally, the quality of the estimated classification rule directly influences the quality of the estimator of the regression parameters.

4.3.2 Limits of a parametric simultaneous procedure

In this section, we consider a probabilistic approach with a parametric point-of-view. Thus, the family of distributions of each component k is supposed to be known and parameterized by $\boldsymbol{\alpha}_k$ and thus we have $\boldsymbol{\phi}_k = (\boldsymbol{\alpha}_k^\top, \boldsymbol{\delta}_k)^\top$. Moreover, the distribution of the noise f_ε is chosen according to the type of the regression under consideration (see the discussion in Section 4.2.3) and thus the parameters shared among the components are restricted to $\boldsymbol{\varsigma}_\varepsilon = \boldsymbol{\gamma}$. The aim of the simultaneous procedure can be achieved by maximizing the log-likelihood of $\mathbf{x}, \mathbf{y}$ given $\mathbf{u}$ with respect to $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y} \mid \mathbf{u}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k)\right).$$

Indeed, the maximum likelihood inference using $\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y} \mid \mathbf{u})$ simultaneously allows for learning the classification rule based on $(\mathbf{X}^\top, Y)^\top$ and the regression coefficients. This function cannot be directly maximized, so we consider the complete-data log-likelihood with data $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ given $\mathbf{u}$ defined by

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}, \mathbf{z} \mid \mathbf{u}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln\left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k)\right).$$

The MLE $\hat{\theta}$ can be obtained via an EM algorithm (see Marbac et al. (2022) for details). Moreover, if the model defined by (4.3)-(4.4) is identifiable, then

1. If all the parametric distributions are well-specified, then properties of the MLE imply that the classification rule is asymptotically optimal and $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased.

2. If at least one parametric distribution is misspecified, then the classification rule is generally asymptotically suboptimal and $\hat{\boldsymbol{\beta}}$ is generally asymptotically biased.

It should be noticed that the distribution of the noise appears at the E-step and thus influences the classification rule. Hence, the classification rule is deteriorated if the distribution of the noise is misspecified. This is not the case when estimation is performed using the two-step approach, since clustering is performed prior to regression, and regression can still be unbiased if the moment condition (see Lemma 4.1) is well-specified. Thus, in the next section, we propose a semi-parametric approach that circumvents this issue because it does not assume a specific family of distributions for the noise and the components.

4.3.3 Advised simultaneous semi-parametric procedure

Semi-parametric model In this section, we consider the semi-parametric version of the model defined by (4.3) where the densities of the components are assumed to be a product of univariate densities (*i.e.,* $f_k(\boldsymbol{x}_i) = \prod_{j=1}^{d_X} f_{kj}(x_{ij})$). Therefore the parameters specific to component k, denoted by ϕ_k, are $\boldsymbol{\delta}_k$ and $f_{k1}, \ldots, f_{kd_X}$. We have

$$f(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) \text{ with } f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) = \prod_{j=1}^{d_{\mathbf{X}}} f_{kj}(x_{ij}) f_\varepsilon(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k).$$

A sufficient condition implying model identifiability is that the covariance matrix of $\mathbf{U}$ has full rank and that the marginal distribution of $\mathbf{X}$ is identifiable and thus a sufficient condition is to consider linearly independent densities f_{kj}'s and $d_X \geq 3$ Allman, Matias, and Rhodes (2009). Thus, if d_X is less than three, other semi-parametric mixture models should be considered to achieve clustering (*i.e.,* location-scale models; see Hunter, Wang, and Hettmansperger (2007) and Chauveau, Hunter, and Levine (2015)).

Smoothed log-likelihood Let $\mathcal{S}$ be the smoothing operator defined by $\mathcal{S}f_k(\boldsymbol{w} \mid \boldsymbol{u}; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) = \int K_h(\boldsymbol{w} - \tilde{\boldsymbol{w}}) f_k(\tilde{\boldsymbol{w}} \mid \boldsymbol{u}; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) d\tilde{\boldsymbol{w}}$, where $K_h(\boldsymbol{a}) = \prod_{j=1}^{d} K_h(a_j)$ with $\boldsymbol{a} \in \mathbb{R}^d$ and with $K_h(a_j)$ is a rescale kernel function defined by $K_h(a_j) = h^{-1}K(h^{-1}a_j)$ where h is the bandwidth. The estimation is achieved by maximizing the smoothed log-likelihood Levine, Hunter, and Chauveau (2011) defined by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \left(\mathcal{N} f_k\right)(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) \right),$$

where $\left(\mathcal{N} f_k\right)(\boldsymbol{w} \mid \boldsymbol{u}; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) = \exp\left\{ \int K_h(\boldsymbol{w} - \tilde{\boldsymbol{w}}) \ln f_k(\tilde{\boldsymbol{w}} \mid \boldsymbol{u}; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon) d\tilde{\boldsymbol{w}} \right\}$, subject to the empirical counterpart of (4.5):

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{f_k(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_k, \boldsymbol{\varsigma}_\varepsilon)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_\ell, \boldsymbol{\varsigma}_\varepsilon)} \rho(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k) = 0.$$

Majorization-Minimization algorithm Parameter estimation is achieved via a Majorization-Minimization algorithm. Given an initial value $\theta^{[0]}$, this algorithm iterates between a majorization and a minimization step. Thus, an iteration $[r]$ is defined by

- Majorization step:

$$t_{ik}^{[r-1]} = \frac{\pi_k^{[r-1]} \left(\mathcal{N} f_k\right)\left(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \boldsymbol{\phi}_k^{[r-1]}, \boldsymbol{\varsigma}_\varepsilon^{[r-1]}\right)}{\sum_{\ell=1}^{K} \pi_\ell^{[r-1]} \left(\mathcal{N} f_\ell\right)\left(\boldsymbol{w}_i \mid \boldsymbol{u}_i; \rho_\ell^{[r-1]}, \boldsymbol{\varsigma}_\varepsilon^{[r-1]}\right)}.$$

- Minimization step:

$$\pi_k^{[r]} = \frac{1}{n}\sum_i t_{ik}^{[r-1]}, \boldsymbol{\beta}^{[r]} = \arg\min_{\boldsymbol{\beta}} \sum_{i,k} t_{ik}^{[r-1]} \mathcal{L}(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k),$$

$$f_{kj}^{[r]}(a) = \frac{1}{n\pi_k^{[r]}} \sum_i t_{ik}^{[r-1]} K_h(x_{ij} - a) \text{ and } f_\varepsilon^{[r]}(a) = \frac{1}{n} \sum_{i,k} t_{ik}^{[r-1]} K_h(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma}^{[r]} - \boldsymbol{\delta}_k^{[r]} - a),$$

then set $\boldsymbol{\phi}_k^{[r]} = \boldsymbol{\gamma}_k^{[r]} \cup \{f_{k1}^{[r]}, \ldots, f_{kd_X}^{[r]}\}$ and $\boldsymbol{\varsigma}_\varepsilon^{[r]} = \boldsymbol{\delta}^{[r]} \cup f_\varepsilon^{[r]}$.

The Majorization-Minimization algorithm is monotonic for the smoothed log-likelihood. It is a direct consequence of the monotony of the algorithm of Levine, Hunter, and Chauveau (2011) where we use the fact that, in order to satisfy the moment condition defined in (4.5) of Lemma 4.1, we must have $\boldsymbol{\beta}^{[r]} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^{[r-1]} \mathcal{L}(y_i - \boldsymbol{u}_i^\top \boldsymbol{\gamma} - \boldsymbol{\delta}_k)$.

As in Hunter and Young (2012), the majorization step is not explicit. However, because it only implies univariate integrals, it can be efficiently assessed by numerical approximations. Finally, bandwidth selection can be performed as usual for semi-parametric mixtures (see Chauveau, Hunter, and Levine (2015)). However, as in any supervised problem, we can use the cross-validated accuracy of the prediction of Y for bandwidth selection.

4.4 Numerical experiments

4.4.1 Simulation setup

Data are generated such that $\mathbf{U}_i \sim \mathcal{N}_2(0, \mathbf{I}_2)$ and such that $(\mathbf{X}_i, Y_i)^\top$ given $\mathbf{U}_i$ follows a K-component mixture with proportions $\pi_k = 1/2$ if $k = 1$ and $\pi_k = 1/2(K-1)$ otherwise. The density of $\mathbf{X}_i$ given $\mathbf{Z}_i$ is a product of univariate densities such that $X_{ij} = \xi \mathbf{Z}_i^\top \kappa_j + \eta_{ij}$ where $\kappa_j = (\kappa_{j1}, \ldots, \kappa_{jK})^\top$, $\kappa_{jk} = 1$ if $k = (j \bmod K) + 1$ and $\kappa_{jk} = 0$ otherwise. Finally, we have $Y_i = \mathbf{U}_i^\top \boldsymbol{\gamma} + \mathbf{Z}_i^\top \boldsymbol{\delta} + \varepsilon_i$ with $\boldsymbol{\gamma} = (1, 1)^\top$ and $\boldsymbol{\delta}_k = 2\xi k$. η_{ij} and ε_i are independently drawn from a standard Gaussian distribution or a Student distribution with 3 degrees of freedom. The parameter ξ is tuned according to the distributions η_{ij} and ε_i and allows three theoretical misclassification rates (5%, 10% and 15%) to be considered. The approaches are compared with respect to the Mean Square Error (MSE) of the estimator of $\boldsymbol{\beta}$ and the Adjusted Rand Index (ARI) between the true and the estimated partition on 100 replicates. The semi-parametric approach is used with a fixed bandwidth $h = n^{-1/5}$. Note that a tuning of this window could be considered as in Chauveau, Hunter, and Levine (2015).

4.4.2 Method comparison

Considering the quadratic loss, the experiment shows that the simultaneous procedure outperforms the standard two-step procedure, in both parametric and semi-parametric frameworks, where the parametric approaches assume that η_{ij} and ε_i are Gaussian. We consider four scenarios: $\eta_{ij} \sim \mathcal{N}(0, 1)$ for the first two scenarios and $\eta_{ij} \sim \mathcal{T}(3)$ for the last two scenarios, and $\varepsilon_i \sim \mathcal{N}(0, 1)$ for the scenarios 1 and 3 and $\varepsilon_i \sim \mathcal{T}(3)$ for scenarios 2 and 4. Figure 4.1 presents

the results obtained when $K = 3$ and $d = 6$. When the parametric model is well-specified (scenario 1), results are equivalent to those obtained by the semi-parametric model. Moreover, if at least one parametric assumption is violated (scenarios 2, 3 and 4), the results of the parametric approach are deteriorated even if the moment condition of the regression model is well-specified. Thus, we advise using the semi-parametric model if the family of the distributions is unknown to prevent the bias in the estimation.

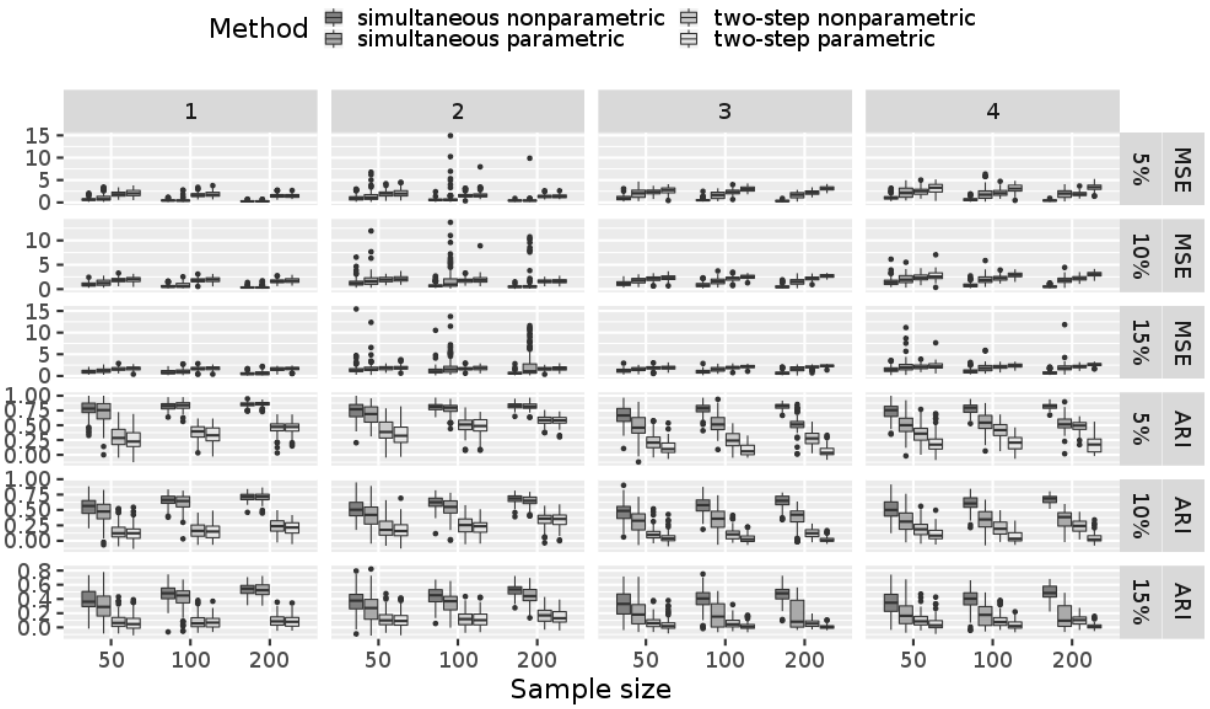


Figure 4.1: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), the scenario (columns) and the sample size obtained when $K = 3$ and $d = 6$.

4.4.3 Robust regression

When the noise of a regression follows an heavy-tail distribution, robust regressions allow the estimators of the regression coefficients to be improved compared to the ordinary least square estimators. Despite this, with a suitable assumption on the noise distribution, the simultaneous parametric approach could consider such regressions. The parametric assumptions made on the noise distribution would be quite unrealistic (*e.g.,* Laplace distribution for the median regression). Thus, we now illustrate that the simultaneous approach can easily consider robust regressions, in a semi-parametric framework, and that the resulting estimators are better than those obtained with the quadratic loss. In this experiment, we consider scenario 4 (*i.e.,* η_{ij} and ε_i both follow independent $\mathcal{T}(3)$) and we consider different robust regressions (median, Huber with parameter 1 and logcosh). Figure 4.2 presents the results obtained when $K = 2$ and $d = 4$. It shows that the simultaneous approach improves the estimators (according to the MSE and the ARI) for any

type of regression and any sample size. Moreover, robust regressions improve the accuracy of the estimator of the regression parameters. However, for this simulation setup, this improvement does not affect the accuracy of the estimated partitions.

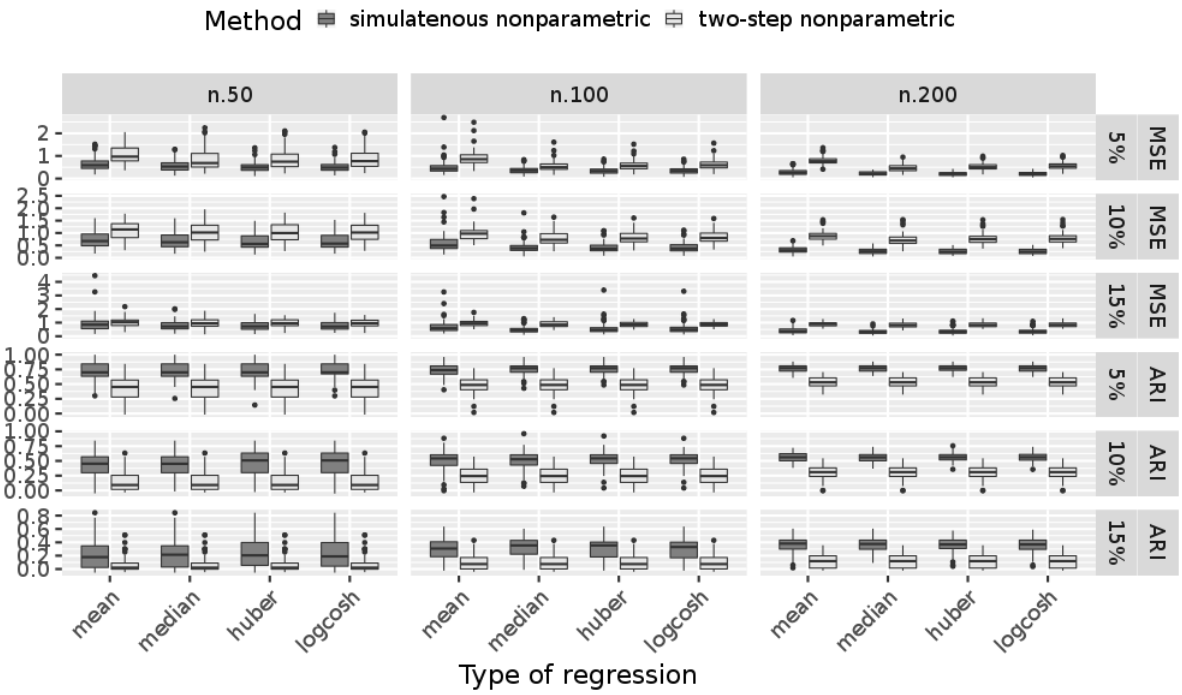


Figure 4.2: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), sample size (columns) and the type of regression obtained when $K = 2$ and $d = 4$ for scenario 4.

4.4.4 Asymmetric losses

Expectile and quantile regressions respectively, generalize the mean and the median regression by focusing on the tails of the distribution of the target variable given the covariates. To illustrate the fact that the semi-parametric simultaneous method allows these regression models to be easily managed, data are generated with $K = 2$ and $d = 4$ such that $\eta_{ij} \sim \mathcal{N}(0, 1)$ and $\varepsilon_i \sim \mathcal{N}(-c_\tau, 1)$. The scalar c_τ is defined according to the regression model. Thus, c_τ is the 0.75-expectile, 0.9-expectile, 0.75-quantile and 0.9-quantile of the standard Gaussian distribution for the 0.75-expectile, 0.9-expectile, 0.75-quantile and 0.9-quantile regression respectively. Figure 4.3 shows that the simultaneous semi-parametric approach improves the estimators compared to those provided by the two-step approach.

4.5 Application on the High blood pressure prevention

Problem summary We consider the problem of high blood pressure prevention where we focus on the detection of indicators related to the diastolic blood pressure. The indicators we

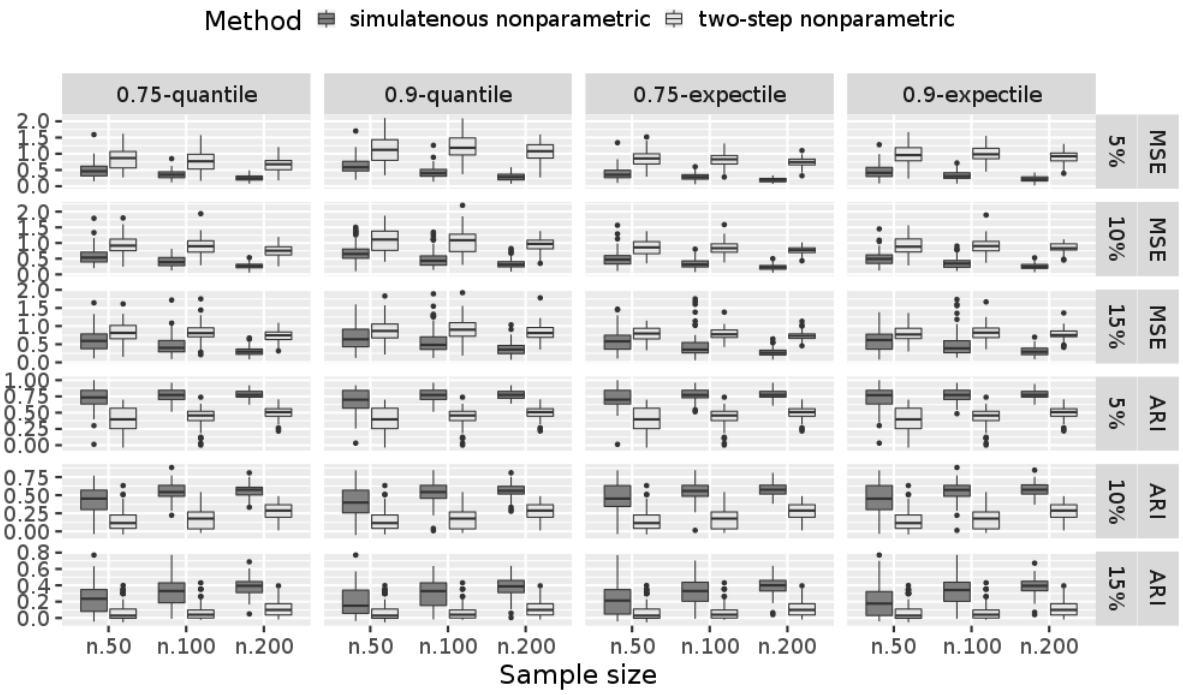


Figure 4.3: Boxplots of the MSE of the estimators of the regression parameters and ARI according to the theoretical misclassification (rows), the type of regression (columns) and regression obtained when $K = 2$ and $d = 4$.

want to consider are gender, age, alcohol consumption, obesity, sleep quality and level of physical activity. However, the level of physical activity of a patient is not directly measured and we only have a set of variables that describe the physical activity. Thus, we want to cluster the subjects based on this set of variables to obtain patterns of similar physical activities and we want to use these patterns in the prediction of the diastolic blood pressure.

Material and methods The data were obtained from National Health and Nutrition Examination Survey of 2011-2012[1]. The target variable is the *diastolic blood pressure* in mmHg (code BPXDI1). The seven covariates in $\mathbf{U}$ are *gender* which was equal to 1 for men et 0 for women (code RIAGENDR), *age* (RIDAGEYR), *alcohol* which indicates whether the subjects consume more than five drinks (for men) and four drinks (for women) of alcoholic beverages almost daily (computed from code ALQ151 and ALQ155), *obesity* which indicates if the body mass index is more than 30 (computed from code BMXBMI), *sleep* which indicates the number of hours of sleeping (computed from code SLD010H), *smoke* which indicates if the subjects used tobacco/nicotine in the last five days (code SMQ680) and *cholesterol* which indicates the total cholesterol in mg/dL (code LBXTC). All the subjects that had missing values for those variables were removed. Seven variables are used in $\mathbf{X}$ to evaluate the level of physical activity. Among these variables, five variables are binary and indicate whether the subject has a vigorous work activity (code PAQ605), whether the subject has a moderate work activity (code

[1]The data are freely downloadable at
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011

PAQ620), whether the subject usually travels on foot or by bike (code PAQ635), whether the subject has vigorous recreational activities (code PAQ650) and whether the subject has moderate recreational activities (code PAQ665). The two remaining variables in $\mathbf{X}$ have 7 levels and indicate the time spent watching TV (code PAQ710) and the time spent using a computer (code PAQ715). Finally, the studied population is composed of 2626 subjects between 18 and 60 years old. To investigate the performances of the different models, 67% of the sample (*i.e.,* 1760 subjects) is used for estimating the model parameters and 33% of the sample (*i.e.,* 866 subjects) is used for investigating the performances of the models. The smoothing is performed on the continuous variables with a Gaussian kernel and a bandwidth $h = \hat{\sigma}_j n^{-1/5}$ where $\hat{\sigma}_j$ is the empirical standard deviation of variable j.

Results We present the main results of the application. Details used for the results interpretation are presented in Appendix 2 of the supplementary materials. We consider a proposed approach in a semi-parametric framework with a quadratic loss. According to the evolution of the smoothed log-likelihood with respect to the number of classes (see Figure 1 in Appendix 2 of the supplementary materials), the model is considered with $K = 3$ classes. To investigate the relevance of the activity level for explaining high blood pressure, we consider three models with a quadratic loss: the proposed approach in a semi-parametric framework (*regquadUZ-K3*), a regression model of Y on $\mathbf{U}$ (*regquadU*) with a selection of variables according to AIC (two variables are removed by the criterion: *alchohol* and *smoke*), a regression model of Y on $(\mathbf{U}^\top, \mathbf{X}^\top)^\top$ (*regquadUX*) with a selection of variables according to AIC (six variables are selected by the criterion: *gender*, *age*, *obesity*, *sleep*, *cholesterol* and the binary variable indicating whether the subject usually travels on foot or by bike). Considering the activity levels seems to be relevant for explaining high blood pressure, since the MSEs of the prediction obtained on the testing samples are 122.34, 122.72 and 122.81 for *regquadUZ-K3*, *regquadUX* and *regquadU* respectively. Thus, the approach allows the information about the physical activity to be summarized and slightly improves the prediction accuracy. Note that a Shapiro-Wilk's normality test performed on the residuals of *regquadUZ-K3* has a pvalue less than 10^{-5} for the learning sample and 0.003 for the testing sample. Thus, the semi-parametric approach avoids the normality assumption which is not relevant for the residuals.

To prevent the variability due to outliers, we fit the proposed approach in a semi-parametric framework with the median loss and the logcosh loss. Again, evolution of the smoothed log-likelihood with respect to the number of classes, leads us to consider $K = 3$ classes for both losses. We now compare the results obtained by the proposed method with $K = 3$ classes in a semi-parametric framework with a quadratic loss, median loss (*regmedUZ-K3*) and logcosh loss (*reglogchUZ-K3*). The three models provided a similar partition since the ARIs between all the couples of partitions is more than 0.83. The regression parameters are presented in Table 1 of Appendix 2 of the supplementary materials. The signs of the coefficients are the same for the three losses. It appears that being a woman lessens the risk of high blood pressure while age, alcohol consumption, overweight, lack of sleeping and cholesterol increase high blood pressure. One can be surprised that the results claim that smoking limits the risk of high blood pressure, but this effect has already been revealed in Omvik (1996) and Li et al. (2017). Note that the robust methods detect a more significant effect of alcohol, smoking and physical activity on high blood pressure. Moreover, they slightly change the prediction accuracy because the MSEs obtained on the testing sample are 122.88 and 123.00 for the median and the logcosh losses respectively.

We now interpret the clustering results provided by the median loss. Class 1 ($\pi_1 = 0.15$ and $\boldsymbol{\delta}_1 = 59.06$) grouping the subjects having high physical activity is the smallest class and contains the subjects having recreational physical activities, traveling by foot or by bike, having

no physical activity at work and spending few hours watching screens. Class 2 ($\pi_2 = 0.44$ and $\boldsymbol{\delta}_2 = 59.29$) groups the subjects having few physical activities but spending little time watching screens. Class 3 ($\pi_3 = 0.37$ and $\boldsymbol{\delta}_3 = 60.34$) groups those having some physical activities but spending a lot of time watching screens. These results show that having moderate physical activities (recreational activities, traveling by bike or foot, not spending many hours watching screens) lessens the risk of high blood pressure.

4.6 Conclusion and perspectives

In Marbac et al. (2022), we propose an alternative to the two-step approach that starts by summarizing some observed variables by clustering and then fits a prediction model using the estimator of the partition as a covariate. Our proposition consists of simultaneously performing the clustering and the estimation of the prediction model to improve the accuracy of the partition and of the regression parameters. This approach can be applied to a wide range of regression models. Our proposition can be applied in a parametric and semi-parametric framework. We advise using the semi-parametric approach to avoid bias in the estimation (due to bias in the distribution modeling).

The quality of the prediction could be used as a tool for selecting the number of components and bandwidth, for semi-parametric mixtures. As in any regression problem, this criterion can also be used for selecting the variables (in the regression part but also in the clustering part). Thus, taking the regression into account is important in model selection for semi-parametric mixtures. Moreover, this could allow for a variable selection in clustering. The semi-parametric approach has been presented by assuming that the components are products of univariate densities. However, the proposed approach can also be used by considering location scale symmetric distributions (Hunter, Wang, and Hettmansperger (2007)) or by incorporating an independent component analysis structure (Zhu and Hunter (2019)). Moreover, we can easily relax the assumption that $(\mathbf{X}^\top, \mathbf{Z}^\top)$ is independent of $\mathbf{U}$. The crucial assumption of the model is the conditional independence of Y and $\mathbf{X}$ given $\mathbf{Z}$.

This approach has been introduced by considering only one latent categorical variable. However, more than one latent categorical variable explained by different sub-groups of variables of $\mathbf{X}$ could be considered. This extension is straightforward if the different sub-groups of variables of $\mathbf{X}$ are known. However, the cases where the sub-groups of variables are also estimated (see the case of multiple partitions in clustering described in Section 2.2.6) could be considered in future work.

Chapter 5

Applications in biostatistics of model-based clustering for functional data

Contents

5.1 Introduction

5.1.1 State of the art

This chapter focuses on functional data clustering (Ferraty and Vieu (2006) and Ramsay and Silverman (2007)). Surveys of clustering techniques for functional data are given in Jacques and

Preda (2014a) and Cheam and Fredette (2020). Model-based clustering methods are available for functional data but require extending the notion of density probability to functional data. Thus, this notion is extended by considering that the curves are defined by a finite number of parameters (Delaigle and Hall (2010)). Moreover, clustering methods for functional data requires to be able to deal with the problem of data dimension. Therefore model-based clustering approaches approximate the observed functions in some functional basis then perform clustering on the coefficients related to the basis (see Chapter 12.1 in Bouveyron et al. (2019)). The choice of the basis function depends on the nature of the data. For instance, James and Sugar (2003) consider the expansion coefficients of the curves into a spline basis of functions while Bouveyron, Côme, and Jacques (2015) proposed approximating the curves into Fourier basis expansion coefficients. Both methods supposed that the coefficients follow a mixture of Gaussian distributions. The use of a spline basis is convenient when the curves are regular, but are not appropriate for peak-like data. Moreover, the use of a Fourier basis is relevant for data having periodicity. Alternatively, feature extraction could be accomplished via an orthogonal wavelet basis (Antoniadis et al. (2013) and Giacofci et al. (2013)). An alternative have been proposed by Samé et al. (2011) who consider a mixture model where each component follows a polynomial regression mixture in which the logistic weights depend on the time. Each observation of a time series arises independently from one of the polynomial regression models specific to the cluster to which it belongs. Therefore, unlike other approaches, the observed data do not require any transformation. The aforementioned model is only applied to a univariate temporal framework.

A further issue raised by the functional data is that of curve alignment. This has been addressed by previous works that do not tackle clustering Kneip and Gasser (1992); Wang and Gasser (1997); Ramsay and Li (1998). Recently, this issue has been considered for clustering with distance-based approaches (Paparrizos and Gravano (2015)) and with model-based approaches (Chudova et al. (2003), Gaffney and Smyth (2005) and Liu and Yang (2009)).

5.1.2 Contributions to model-based clustering of functional data

We have worked on three papers dealing with functional data clustering.

In Cheam, Marbac, and McNicholas (2017), we extend the approach of Samé et al. (2011) to spatio-temporal data clustering. The resulting method is a mixture model where each component is an autoregressive polynomial regression mixture in which the logistic weights depend on the spatial and temporal dimensions. The EM algorithm is carried out to obtain the maximum likelihood estimates of the parameters of interest. A key contribution of our work is the introduction of an autoregressive component to the model proposed by Samé et al. (2011) and the ability to model spatial dependencies for multivariate functional data.

In Du Roy de Chaumaray, Marbac, and Navarro (2020), motivated by the analysis of accelerometer data, we introduce a specific finite mixture of hidden Markov models with particular characteristics that adapt well to the specific nature of this type of data. Our model allows for the computation of statistics that characterize the physical activity of a subject (e.g., the mean time spent at different activity levels and the probability of the transition between two activity levels) without specifying the activity levels in advance but by estimating them from the data. In addition, this approach allows the heterogeneity of the population to be taken into account and subpopulations with homogeneous physical activity behavior to be defined. We prove that, under mild assumptions, this model implies that the probability of misclassifying a subject decreases as an exponentially decay with the length of its measurement sequence. Model identifiability is also investigated. We also report a comprehensive suite of numerical simulations to support our theoretical findings. The method is motivated by and applied to the PAT study. This paper is described in Section 5.2.

In Cheam et al. (2020), we investigate the geographical disparities of the COVID-19 diseaseby focusing on clustering the daily death rates reported in several regions of Europe and the United States over eight months. Several methods have been developed to cluster such functional data. However, these methods are not translation-invariant and thus cannot handle different times of arrivals of the disease, nor can they consider external covariates and so are unable to adjust for the population risk factors of each region. We propose a novel three-step clustering method to circumvent these issues. As a first step, feature extraction is performed by translation-invariant wavelet decomposition which allows dealing with the different onsets. As a second step, single-index regression is used to neutralize disparities caused by population risk factors. As a third step, a nonparametric mixture is fitted on the regression residuals to achieve the region clustering. This paper is described in Section 5.3.

5.2 Mixture of hidden Markov models for accelerometer data

5.2.1 Introduction

Inadequate sleep and physical inactivity affect physical and mental well-being while often exacerbating health problems. They are currently considered major risk factors for several health conditions (Kimm et al. (2005) ,Taheri et al. (2004), Lee et al. (2012), Grandner et al. (2013) and McTiernan (2008)). Therefore, appropriate assessment of activity and sleep periods is essential in disciplines such as medicine and epidemiology. The use of accelerometers to evaluate physical activity—by measuring the acceleration of the part of the body to which they are attached—is a classic method that has become widespread in public health research. Indeed, since the introduction in 2003 of the first objective assessment of physical activity using accelerometers, as part of the National Health and Nutrition Examination Survey (NHANES), the analysis of actigraphy data has been the subject of extensive studies over the past two decades. Recently, the New York City (NYC) Department of Health and Mental Hygiene conducted the 2010-2011 Physical Activity and Transit (PAT) Survey[1], a random survey of adult New Yorkers that tracked levels of sedentary behavior and physical activity at work, at home, and for leisure. A subset of interviewees was also invited to participate in a follow-up study to objectively measure their activity level using an accelerometer. One of the objectives of this study is to describe measured physical activity levels and to compare estimates of adherence to recommended physical activity levels, as assessed by accelerometer, with those from self-reports. In contrast to NHANES accelerometer data, PAT data still seem relatively unexplored in the statistical literature.

In Du Roy de Chaumaray and Marbac (2021a), we were interested in the analysis of the accelerometer data worn by 133 individuals aged at least of 65 who responded to the PAT survey. Our objective is to propose a model adapted to the specificities of these data and study its properties. Indeed, this data set raises various challenges, such as managing the heterogeneity of the population or missing data of different natures. In order to motivate the development of a new model, we present an overview of the literature on accelerometer data analysis.

The pioneering approaches used for analyzing accelerometer data have focused on automatic detection of the sleep and wake-up periods (Cole et al. (1992), Sadeh, Sharkey, and Carskadon (1994), Pollak et al. (2001) and Van Hees et al. (2015)). More recent developments are interested in the classification of different levels of activity (see Yang and Hsu (2010) for a

[1]NYC Department of Health and Mental Hygiene. Physical Activity and Transit Survey 2010-2011; public use datasets accessed on May 10, 2019. The data are freely accessible on this page: https://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey-public-use-data.page

review). These methods provide summary statistics such as the mean time spent at different activity levels. In epidemiological studies, time spent by activity level is often used as a covariate in predictive models (see, for instance, the works of Noel et al. (2010), Palta et al. (2015) and Innerd, Harrison, and Coulson (2018), where the links between physical activity and obesity are investigated). These statistics can be computed using deterministic cutoff levels (Freedson, Melanson, and Sirard (1998)). However, with such an approach, the time dependency is neglected and the cutoff levels are pre-specified and not estimated from the data.

Accelerometer data are characterized by a time dependency between the different measurements. They can be analyzed by methods developed for functional data or by Hidden Markov Models (HMM). Methods for functional data need the observed data to be converted into a function of time (Morris et al. (2006), Xiao et al. (2014), Gruen et al. (2017)). For instance, Morris et al. (2006) use a wavelet basis for analyzing accelerometer profiles. The use of a function basis reduces the dimension of the data, and therefore the computing time. However, these methods do not define levels of activity and thus cannot directly provide the time spent at different activity levels.

When considering a discrete latent variable to model time dependence, HMM are appropriate for adjusting sequence data (Scott, James, and Sugar (2005), Altman (2007) and Gassiat, Cleynen, and Robin (2016)). Titsias, Holmes, and Yau (2016) expand the amount of information which can be obtained from HMM including a procedure for finding a maximum *a posteriori* (MAP) of the latent sequences and for computing posterior probabilities of the latent states. HMM are used on activity data for monitoring circadian rythmicity (Huang et al. (2018b)) or directly for estimating the sequence of activity levels from accelerometer data (Witowski et al. (2014)). For simulated data, Witowski et al. (2014) established the superiority of different HMM models, in terms of classification error, over traditional methods based on *a priori* fixed thresholds. While the simplicity of implementing threshold-based methods is an obvious advantage, they have some significant disadvantages compared to the HMM methods, particularly for real data. Indeed, the variation in counts and the resulting dispersion is large, leading to considerable misclassification of counts recorded in erroneous activity ranges. The approach of Witowski et al. (2014) assumes homogeneity of the population and does not consider missingness within the observations. However, heterogeneity in physical activity behaviors is often present (see, for instance, Geraci (2018)) and the use of more than one HMM allows it to be taken into account (see, *e.g.*, Pol and Langeheine (1990)). Thus, recent methods use clustering of accelerometer data to take into account the heterogeneity of the population. For instance, Wallace et al. (2018) use a specific finite mixture to identify novel sleep phenotypes, Huang et al. (2018a) perform a matrix-variate-based clustering on accelerometer data while Lim, Oh, and Cheung (2019) use a clustering technique designed for functional data. Mixed Hidden Markov Models (MHMM) are a combination of HMM and Generalized Linear Mixed Models (Pol and Langeheine (1990) and Bartolucci, Farcomeni, and Pennoni (2012)). These models consider one (or more) random effect(s) coming from either a continuous distribution (Altman (2007)) or a discrete distribution (Bartolucci, Pennoni, and Vittadini (2011) and Maruotti (2011)). Note that an MHMM with a single discrete random effect distribution, having a finite number of states, is a finite mixture of HMM. Such a model allows us to estimate a partition among the population and to consider the population heterogeneity. The impact of the random effect can be on the measurement model or on the latent model.

In Du Roy de Chaumaray and Marbac (2021a), we focus on the analysis of PAT data with a two-fold objective: obtaining summary statistics about physical activity of the subjects without pre-specifying cutoff levels and obtaining a partition which groups subjects in homogeneous classes. We define a class as homogeneous if its subjects have similar average times spent into the different activity levels and similar transition probabilities between activity levels. To achieve

this goal, we introduce a specific finite mixture of HMM for analyzing accelerometer data. This model considers two latent variables: a categorical variable indicating each subject's class membership and a sequence of categorical variables indicating the subject's level of activity each time its acceleration is measured. At time t, the measurement is independent of the class membership, conditionally on the activity level (*i.e.*, the latent state) and follows a zero-inflated distribution—a distribution that allows for frequent zero-valued observations. The activity level defines the parameter of this distribution. The use of a zero-inflated distribution is quite common for modeling accelerometer data (Ae Lee and Gill (2018) and Bai et al. (2018)), as the acceleration is measured every second and many observations are zero. Note that the definitions of the activity levels are equal among the mixture components. This is an important point for the use of summary statistics (*e.g.*, time spent at different activity levels, probabilities of transition between levels) in a future statistical study. The model we consider is thus a specific MHMM with a finite-states random effect that only impacts the distribution of latent physical activity levels. MHMMs with a finite-states random effect have few developments in the literature (see Bartolucci, Pennoni, and Vittadini (2011) and Maruotti (2011)), especially when the random effects only impact the latent model (and not the measurement model). We propose to theoretically study the model properties by showing that the probability of misclassifying an observation decreases at an exponential rate. In addition, since the distribution given the latent state is itself a bi-component mixture (due to the use of zero-inflated distributions), we investigate sufficient conditions for model identifiability.

In practice, the data collected often include missing intervals due to non-compliance by participants (*e.g.*, if the accelerometer is removed). Thus, Geraci and Farcomeni (2016) propose to identifying different profiles of physical activity behaviors using a principal component analysis that allows for missing values. The PAT data contain three types of missing values corresponding to periods when the accelerometer is removed, making statistical analysis more challenging. First, missingness occurs at the beginning and at the end of the measurement sequences due to the installation and the removal of the accelerometer. Second, subjects are asked to remove the accelerometer when they sleep at night. Third, missing values appear during the day (*e.g.*, due to a shower period, napping, *etc*). We remove missing values which occur at the beginning and at the end of the sequence. For missingness caused by night time sleep, we consider that the different sequences describing different days of observations of a subject, are independent and that the starting point (*e.g.*, first observed measurement of the accelerometer of the day) is drawn from the stationary distribution. For missing values measured during the day, the model and the estimation algorithm can handle these data. Moreover, we propose an approximation to the distribution that avoids the computation of large powers of the transition matrices in the algorithm used for parameter inference and thus reducing computation time. Theoretical guarantees and numerical experiments show the relevance of our proposition.

The R package **MHMM** which implements the method introduced in this paper is available on CRAN (Du Roy de Chaumaray, Marbac, and Navarro (2019)). It permits other accelerometer data to be analyzed and thus it is complementary to existing packages for MHMM. Indeed, it takes into account the specificities of accelerometer data (the class membership only impacts the transition matrices, the emission distributions are zero-inflated gamma (ZIG) distributions). Among the R packages implementing MHMM methods, one can cite **LMest** (Bartolucci, Pandolfi, and Pennoni (2017)) and **seqHMM** (Helske and Helske (2019)) which focus on univariate longitudinal categorical data and **mHMMbayes** A. (2019) which focuses on multivariate longitudinal categorical data.

This chapter is organized as follows. Section 5.2.2 presents the PAT data and the context of the study. Section 5.2.3 introduces our specific mixture of HMM and its justification in the context of accelerometer data analysis. Section 5.2.4 presents the model properties (model

identifiability, exponential decay of the probabilities of misclassification and a result for dealing with the non-wearing periods). Section 5.2.5 discusses the maximum likelihood inference and Section 5.2.6 illustrates the model properties on both simulated and real data. Section 5.2.7 illustrates the approach by analyzing a subset of the PAT accelerometer data. Proofs and technical lemmas are not presented here but are available in Du Roy de Chaumaray and Marbac (2021a).

5.2.2 PAT data description

In Du Roy de Chaumaray and Marbac (2021a), we consider a subset of the data from the PAT survey, the subjects who participated in the follow-up study to objectively measure their activity level using an accelerometer. A detailed methodological description of the study and an analysis of the data is provided in Immerwahr et al. (2012). Note that the protocols for accelerometer data for the PAT survey and NHANES were identical. One of the objectives of the PAT study is to investigate the relationships between self-reported physical activity and physical activity measured by the accelerometer in order to provide best practice recommendations for the use of self-reported data (Wyker et al. (2013)). Indeed, self-reported data may be subject to overreporting. This is particularly the case among less active people, due in particular to a social desirability bias or the cognitive challenge associated with estimating the frequency and duration of daily physical activity (see, *e.g.*, Slootmaker et al. (2009); Dyrstad et al. (2014); Lim et al. (2015)). The results of Wyker et al. (2013) show that males tend to underreport their physical activity, while females and older adults (65 years and older) overreported it (see also Troiano et al. (2008) for a detailed study of the differences between self-reported physical activity and accelerometer measurements in NHANES 2003-2004). Consequently, the study of data measured by accelerometer for these specific populations makes it possible to determine methods for correcting estimates from self-reported data, such as stratification by gender and/or age when comparing groups.

In this work, we are particularly interested in the age category above 65 years old ($n = 133$). We present some characteristics related to PAT data and refer to Immerwahr et al. (2012) for a full description[2]. Accelerometers were worn for one week (beginning on Thursday and ending on Wednesday) and measured the activity minute-by-minute. The trajectory associated with each subject is therefore of length 10080. In addition, a participant's day spans from 3am-3am (and not a calendar day) in order to record late night activities and transit and contains missing data sequences of variable length at the beginning and end of the measurement period (these missing data sequences were excluded from the analysis). This length varies from one subject to another, and the mean and minimum trajectory length for the population under consideration (after excluding those missing at the edges) are 9474 and 5199 respectively (with a total number of observations equal to 1259981). The model of accelerometer used was Actigraph GT3X, it was worn on the hips (which results in the fact that certain activities, such as lifting weights or biking, cannot be measured). In addition, participants were also asked to remove it when sleeping, swimming or bathing, hence the data contains approximately 44% of missing values that appear mainly in sequence, appearing at night but also during the day. Figure 5.1 gives an example of accelerometer data measured on one subject (*i.e.*, patcid:1200255) for one week where the three types of missing data can be seen. The four levels of physical activity based on the classification established by the US Department of Health and Human Services (2008) in the Physical Activity Guidelines for Americans (PAGA) report is also shown in the Figure 5.1. Specifically, the PAT protocol for accelerometer data has established a classification according to PAGA, characterizing

[2]Raw accelerometer data, covariates allowing the subset of the population to be selected, as well as providing a detailed dictionary are freely accessible here: https://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey-public-use-data.page

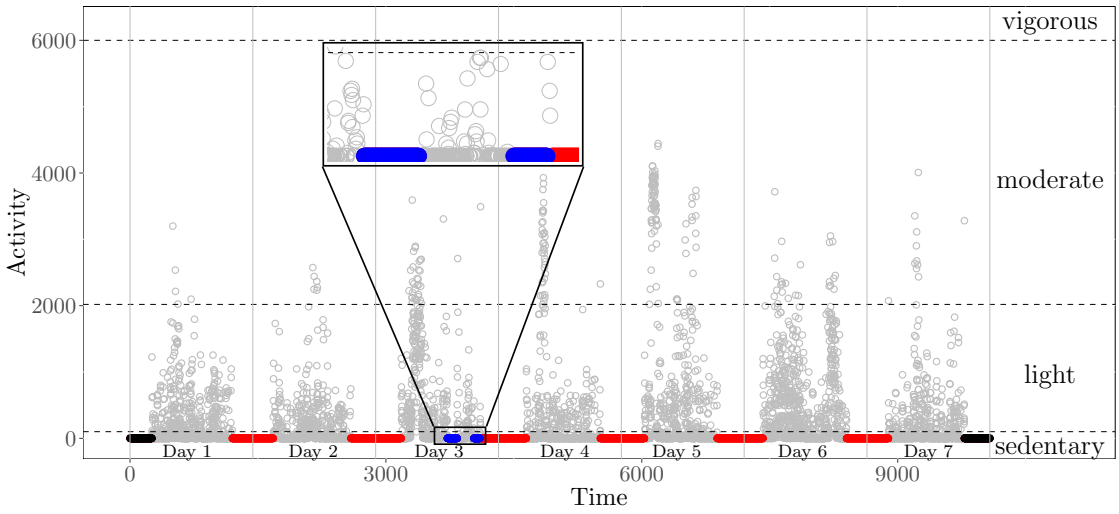


Figure 5.1: Accelerometer data of subject Patcid:1200255 of the PAT study measured for one week (with a zoom on the afternoon of day 3): observed values (in gray), missing values during a daytime period (in blue), missing values during a period of night time sleep(in red) and missing values at the start and end of the measurement period (in black). The dashed horizontal lines represent the four levels of physical activity based on the classification established by the US Department of Health and Human Services (2008).

each minute of activity. Activity minutes with less than 100 activity counts were classified as Sedentary, minutes with 100-2019 counts were classified as Light, the class Moderate corresponds to 2020-5998 counts/minute and Vigorous 5999 counts/minute and more. A comparison between our method and this traditional threshold-based approach is provided in Section 5.2.7.

5.2.3 Mixture of hidden Markov models for accelerometer data

In this section we present the proposed model and the application context for which it has been defined.

The data Observed data $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$ are composed of n independent and identically distributed sequences $\boldsymbol{y}_i$. Each sequence $\boldsymbol{y}_i = (y_{i(0)}, \ldots, y_{i(T)})^\top$ which contains the values measured by the accelerometer at times $t \in \{0, 1, \ldots, T\}$ for subject i, with $y_{i(t)} \in \mathbb{R}^+$. Throughout this section, index i refers to the label of the subject and index (t) refers to the time of measurement.

The model considers M different activity levels (which are unobserved). These levels impact the distribution of the observed sequences of accelerometer data. The sequence of the hidden states $\mathbf{x}_i$ indicates the activity level of subject i at the different times. Thus, $\mathbf{x}_i = (\mathbf{x}_{i(0)}^\top, \ldots, \mathbf{x}_{i(T)}^\top)^\top \in \mathcal{X}$ and the activity level (among the M possible levels) of subject i at time t is defined by the binary vector $\mathbf{x}_{i(t)} = (x_{i(t)1}, \ldots, x_{i(t)M})^\top$ where $x_{i(t)h} = 1$ if subject i is at state h at time t and $x_{i(t)h} = 0$ otherwise.

The heterogeneity (in the sense of different physical activity behaviors) between the n subjects, can be addressed by grouping subjects into K homogeneous classes. This is achieved by clustering that assesses a partition $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$ among the n subjects based on their ac-

celerometer measurements. Thus, the vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})^\top$ indicates the class membership of subject i, as $z_{ik} = 1$ if observation i belongs to class k and $z_{ik} = 0$ otherwise. Throughout the paper, index k refers to the label of a class grouping homogeneous subjects.

Each subject i is described by three random variables: one unobserved categorical variable $\mathbf{z}_i$ (defining the membership of the class of homogeneous physical activity behaviors for subject i), one unobserved categorical longitudinal data $\mathbf{x}_i$ (a univariate categorical discrete-time time series which defines the activity level of subject i at each time) and one observed positive longitudinal data $\boldsymbol{y}_i$ (a univariate positive discrete-time time series which contains the values of the accelerometer measured on subject i at each time).

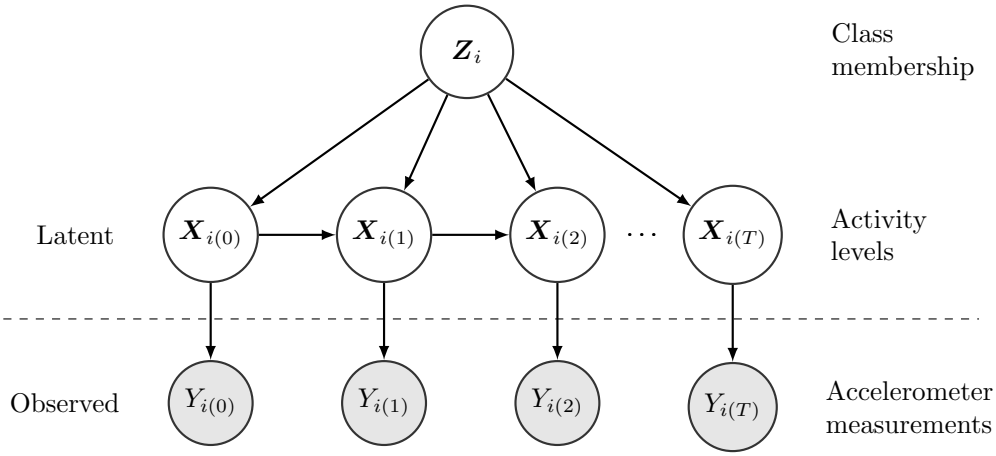


Figure 5.2: Generative model of the specific mixture model of HMM used for the accelerometer data: an arrow between two variables indicates dependency and an absence of arrow indicates conditional independence.

Generative model The model described below considers that the observations are independent between the subjects and identically distributed. It is defined by the following generative model and summarized by Figure 5.2 (note that this figure is similar to Figure 6.2 of Bartolucci, Farcomeni, and Pennoni (2012)):

1. sample class membership $\mathbf{z}_i$ from a multinomial distribution;
2. sample the sequence of activity levels $\mathbf{x}_i$ from a Markov model whose transition matrix depends on class membership;
3. sample the accelerometer measurement sequence given the activity levels (each $\boldsymbol{Y}_{i(t)}$ follows a ZIG distribution whose parameters are defined only by $\mathbf{x}_{i(t)}$).

Finite mixture model for heterogeneity The sequence of accelerometer measurements obtained on each subject is assumed to independently arise from a mixture of K parametric distributions, so that the probability distribution function (pdf) of the sequence $\boldsymbol{y}_i$ is

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \delta_k \, p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}), \tag{5.1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \boldsymbol{\varepsilon}\} \cup \{\delta_k, \boldsymbol{\pi}_k, \mathbf{A}_k; k = 1, \ldots, K\}$ groups the model parameters, δ_k is the proportion of components k with $\delta_k > 0$, $\sum_{k=1}^{K} \delta_k = 1$, and $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is the pdf of component k parametrized by $\{\boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}\}$ defined below. Thus, δ_k is the marginal probability that a subject belongs to class k (*i.e.*, $\delta_k = \mathbb{P}(Z_{ik} = 1)$). Moreover, $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ defines the distribution of a sequence of values measured by the accelerometer on a subject belonging to class k (*i.e.*, $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is the pdf of $\boldsymbol{y}_i$ given $Z_{ik} = 1$).

Hidden Markov model for activity levels The model assumes that the distribution of the hidden state sequence depends on the class membership, and that the distribution of activity measurements depends on the state at time t but not on the component membership given the state (*i.e.*, $\boldsymbol{X}_i \not\perp \boldsymbol{Z}_i$, $Y_{i(t)} \not\perp \boldsymbol{X}_{i(t)}$ and $Y_{i(t)} \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_{i(t)}$). It is crucial that the distribution of $Y_{i(t)}$ given $\boldsymbol{X}_{i(t)}$ is independent of $\boldsymbol{Z}_i$. Indeed, each activity level is defined by the distribution of $Y_{i(t)}$ given the state. Therefore, to extract summary statistics on the whole population (as the average time spent per level of activity) the definition of the activity levels (and the distribution of $y_{i(t)}$ given the state) must be the same among the mixture components.

The pdf of $\boldsymbol{y}_i$ for components k (*i.e.*, given $Z_{ik} = 1$) is

$$p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) = \sum_{\mathbf{x}_i \in \mathcal{X}} p(\mathbf{x}_i; \boldsymbol{\pi}_k, \mathbf{A}_k)\, p(\boldsymbol{y}_i \mid \mathbf{x}_i; \boldsymbol{\lambda}, \boldsymbol{\varepsilon}). \tag{5.2}$$

The Markov assumption implies that

$$p(\mathbf{x}_i; \boldsymbol{\pi}_k, \mathbf{A}_k) = \prod_{h=1}^{\ell} \pi_{kh}^{x_{i(0)h}} \prod_{t=1}^{T} \prod_{h=1}^{M} \prod_{\ell=1}^{M} (\mathbf{A}_k[h, \ell])^{x_{i(t-1)h} x_{i(t)\ell}},$$

where $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kM})^\top$ defines the initial probabilities so that $\pi_{kh} = \mathbb{P}(X_{i(1)h} = 1 \mid Z_{ik} = 1)$, $\mathbf{A}_k$ is the transition matrix so that $\mathbf{A}_k[h, \ell] = \mathbb{P}(X_{i(t)\ell} = 1 \mid X_{i(t-1)h} = 1, Z_{ik} = 1)$. Finally, we have

$$p(\boldsymbol{y}_i \mid \mathbf{x}_i; \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) = \prod_{t=0}^{T} \prod_{h=1}^{M} g(y_{i(t)}; \boldsymbol{\lambda}_h, \varepsilon_h)^{x_{i(t)h}},$$

where $g(\cdot; \boldsymbol{\lambda}_h, \varepsilon_h)$ is the pdf of a zero-inflated distribution defined by

$$g(y_{i(t)}; \boldsymbol{\lambda}_h, \varepsilon_h) = (1 - \varepsilon_h) g_c(y_{i(t)}; \boldsymbol{\lambda}_h) + \varepsilon_h \mathbf{1}_{\{y_{i(t)} = 0\}},$$

where $g_c(\cdot; \boldsymbol{\lambda}_h)$ is the density of a distribution defined on a positive space and parametrized by $\boldsymbol{\lambda}_h$. The choice of considering zero-inflated distributions is motivated by the large number of zeros in the accelerometer data (see Figure 5.1). For the application of Section 5.2.7, we use a gamma distribution of $g_c(\cdot; \boldsymbol{\lambda}_h)$. However, model properties and inference are discussed for a large family of densities $g_c(\cdot; \boldsymbol{\lambda}_h)$.

5.2.4 Model properties

In this section, we present the properties of the mixture of parametric HMMs. We start with a discussion of three assumptions. Model identifiability is then proven. It is shown that the probability of making an error in the partition estimation, exponentially decreases with T, when the model parameters are known. Finally, the analysis of missing data is discussed.

Assumptions

Assumption 5.1. For each component k, the Markov chain is irreducible. Moreover, we assume that the sequence is observed at its stationary distribution (*i.e.*, $\boldsymbol{\pi}_k$ is the stationary distribution so $\boldsymbol{\pi}_k^\top \mathbf{A}_k = \boldsymbol{\pi}_k^\top$). Therefore, there exists $0 \leq \nu < 1$ such that

$$\forall k \in \{1, \ldots, K\},\ \nu_2(\mathbf{A}_k) \leq \nu,$$

where $\nu_2(\mathbf{A}_k)$ is the second-largest eigenvalue of $\mathbf{A}_k$. Denote $\bar{\nu}_2(\mathbf{A}_k) = \max(0, \nu_2(\mathbf{A}_k))$.

Assumption 5.2. The hidden states define different distributions for the observed sequence. Therefore, for $h \in \{1, \ldots, M\}$, $h' \in \{1, \ldots, M\} \setminus \{h\}$, we have $\boldsymbol{\lambda}_h \neq \boldsymbol{\lambda}_{h'}$. Moreover, the parametric family of distributions defining $g_c(\cdot; \boldsymbol{\lambda}_1), \ldots, g_c(\cdot; \boldsymbol{\lambda}_M)$ permits an ordering to be considered such that for a fixed value $\rho \in \mathbb{R}^+ \setminus \{0\}$, we have

$$\forall h \in \{1, \ldots, M-1\},\ \lim_{y_{i(1)} \to \rho} \frac{g_c(y_{i(1)}; \boldsymbol{\lambda}_{h+1})}{g_c(y_{i(1)}; \boldsymbol{\lambda}_h)} = 0.$$

Assumption 5.3. The transition probabilities are different over the mixture components and are not zero. Therefore, for $k \in \{1, \ldots, K\}$, $k' \in \{1, \ldots, K\} \setminus \{k\}$, we have $\forall (h, \ell)$, $\mathbf{A}_k[h, \ell] \neq \mathbf{A}_{k'}[h, \ell]$. Moreover, there exists $\zeta > 0$ such that

$$\forall k \in \{1, \ldots, K\},\ \forall k' \in \{1, \ldots, K\} \setminus \{k\},\ \sum_{h=1}^{M} \sum_{\ell=1}^{M} \pi_{kh} \log \frac{\mathbf{A}_k[h, \ell]}{\mathbf{A}_{k'}[h, \ell]} > \zeta.$$

Finally, without loss of generality, we assume that $A_k[1, 1] > A_{k+1}[1, 1]$.

Assumption 5.1 states that the state at time 1 is drawn from the stationary distribution of the component that the observation belongs to. To obtain model identifiability, we do not need the assumption that the stationary distribution is different over the mixture components. As a result, two components having the same stationary distribution but different transition matrices can be considered. Assumption 5.2 and Assumption 5.3 are required to obtain model identifiability. Assumption 5.3 can be interpreted as the Kullback-Leibler divergence between the distribution of the states under component k and their distribution under component k'. This constraint is required for model identifiability because it is related to the definition of the classes. Consequently, the matrices of the transition probability must be different among components.

Identifiability Model identifiability is crucial for interpreting the estimators of the latent variables and of the parameters. It has been studied for some mixture models (Teicher (1963), Teicher (1967), Allman, Matias, and Rhodes (2009) and Celisse, Daudin, and Pierre (2012)) and HMM (Gassiat, Cleynen, and Robin (2016)), but not for the mixture of HMMs. Generic identifiability (up to switching of the components and of the states) of the model defined in (5.1) implies that

$$\forall \boldsymbol{y}_i,\ p(\boldsymbol{y}_i; \boldsymbol{\theta}) = p(\boldsymbol{y}_i; \tilde{\boldsymbol{\theta}}) \Rightarrow \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}.$$

The following theorem states this property.

Theorem 5.1. *If Assumptions 5.1, 5.2 and 5.3 hold, then the model defined in* (5.1) *is generically identifiable (up to switching of the components and of the states) if* $T > 2K$.

Proof of Theorem 5.1 is given in Du Roy de Chaumaray and Marbac (2021a). The model defined by the marginal distribution of a single $y_{i(t)}$ is not identifiable. Indeed, the marginal distribution of $y_{i(t)}$ is a mixture of zero-inflated distributions and such a mixture is not identifiable

(*i.e.*, different class proportions and inflation proportions can define the same distribution). It is therefore this dependency over time that makes the proposed mixture generically identifiable. Note that such a statement has been made by Gassiat, Cleynen, and Robin (2016) when they discuss the case where the emission distribution for an HMM follows a mixture model.

Probabilities of misclassification In this section, we examine the probability that an observation will be misclassified when the model parameters are known. We consider the ratio between the probability that subject i belongs to class k given $\boldsymbol{y}_i$ and the probability that this subject belongs to its true class, and we quantify the probability of it being greater than some positive constant a. Let θ_0 be the true model parameter and $\mathbb{P}_0 = \mathbb{P}(\cdot \mid Z_{ik_0} = 1, \theta_0)$ denotes the true conditional distribution (the true label of subject i and parameters are known).

Theorem 5.2. *Let $a > 0$, under mild assumptions detailled in Du Roy de Chaumaray and Marbac (2021a), then for every $k \neq k_0$*

$$\mathbb{P}_0\left[\frac{\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i)} > a\right] \leq \mathcal{O}(e^{-cT}),$$

where $c > 0$ is a positive constant

Moreover, the exponential bounds of Theorem 5.2 allows to be used the Borel-Cantelli's lemma to obtain the almost sure convergence.

Corollary 5.1. Assume that Assumptions 5.1 and 5.3 hold. If $\boldsymbol{y}_i$ is generated from component k_0 (*i.e.*, $Z_{ik_0} = 1$), then for every $k \neq k_0$

$$\frac{\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i)} \xrightarrow[T\to+\infty]{a.s.} 0, \quad \mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i) \xrightarrow[T\to+\infty]{a.s.} 1 \text{ and } \mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i) \xrightarrow[T\to+\infty]{a.s.} 0.$$

Therefore, by considering $a = 1$, Theorem 5.2 and Corollary 5.1 show that the probability of misclassifying the subject i based on the observation $\boldsymbol{y}_i$, using the *maximum a posteriori* rule, tends to zero when T increases, if the model parameters are known. Proof of Theorem 5.2 and a sufficient condition that allows us to consider $a = 1$ (value of interest when the partition is given by the MAP rule) are given in Du Roy de Chaumaray and Marbac (2021a). It should be noted that it is not so common to have an exponential rate of convergence for the ratio of the posterior probability of classification. Similar results are obtained for network clustering using the stochastic block model (Celisse, Daudin, and Pierre (2012)) or for co-clustering (Brault and Mariadassou (2015)). For these two models, the marginal distribution of a single variable provides information about the class membership. For the proposed model, this is the dependency between the different observed variables which is the crucial point for recovering the true class membership.

Dealing with missing values Due to the Markovian character of the states, missing values can be handled by iterating the transition matrices. In our particular context, missing values appear when the accelerometer is not worn (see Section 5.2.2 for explanations of the reasons of missingness). We will not observe isolated missing values but rather wide ranges of missing values. Let d be the number of successive missing values, we thus have to compute the matrix A_k^{d+1} to obtain the distribution of the state at time $t+d$ knowing the state at time $t-1$. These powers of transition matrices should be computed many times during the algorithm used for inference (see Section 5.2.5). Moreover, after $d+1$ iterations with d sufficiently large, the transition matrix can be considered sufficiently close to stationarity (*e.g.*, for any (h, ℓ), $A_k^{d+1}[h, \ell] \simeq \pi_{k\ell}$), which has

actually been chosen as the initial distribution. Therefore, for numerical reasons, we will avoid computing the powers of the transition matrices and we will make the following approximation. An observation $\boldsymbol{y}_i$ with S_i observed sequences split with missing value sequences of size at least d are modeled as S_i independent observed sequences with no missing values, all belonging to the same component k. Namely, for each individual i, the pdf $p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ of component k is approximated by the product of the pdf of the S_i observed sequences $\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{iS_i}$:

$$p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) \simeq \prod_{s=1}^{S_i} p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}),$$

where, for each s, $\boldsymbol{y}_{is}$ is an observed sequence of length $T_{is}+1$: $\boldsymbol{y}_{is} = (y_{is(0)}, \ldots, y_{is(T_{is})})$ and $p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is defined as in (5.2). We note that the observation $\boldsymbol{y}_i$ can thus be rewritten as follows

$$\boldsymbol{y}_i = (y_{i1(0)}, \ldots, y_{i1(T_{i1})}, y_{i2(0)}, \ldots, y_{i2(T_{i2})}, \ldots, y_{iS_i(0)}, \ldots, y_{iS_i(T_{iS_i})}),$$

with $y_{i2(0)} = y_{i(T_{i1}+d_{i1}+1)}$ where the d_{i1} values $y_{i(T_{i1}+1)}, \ldots y_{i(T_{i1}+d_{i1})}$ correspond to the first sequence of missing values, and more generally, for each $s = 2, \ldots, S_i$, $y_{is(0)} = y_{i(\sum_{j=1}^{s-1}(T_{ij}+d_{ij}+1))}$, with d_{ij} being the number of missing values between the observed sequences $\boldsymbol{y}_{is_j}$ and $\boldsymbol{y}_{is_{j+1}}$.

Once the estimation of the parameters has been done, we make sure that this assumption was justified by verifying that the width of the smallest range $d_{min} = \min\{d_{i1}, \ldots, d_{i\,S_i-1}\}$ of missing values is sufficiently large to be greater than the mixing time of the obtained transition matrix. To do so, we use an upper bound for the mixing time given by Levin and Peres (2017):Theorem 12.4, p. 155. For each component k, we denote by ν_k^* the second maximal absolute eigenvalue of $\mathbf{A}_k$. For any positive η, if for each k

$$d_{min} \geq \frac{1}{1-\nu_k^*} \log \frac{1}{\eta \min_h \pi_{kh}},$$

then for any integer $D \geq d_{min}$, the maximum distance in total variation satisfies

$$\max_h \|A_k^D[h, \cdot] - \pi_k\|_{TV} \leq \eta.$$

5.2.5 Maximum likelihood inference

This section presents the methodology used to estimate the model parameters.

Inference We propose to estimate the model parameters by maximizing the log-likelihood function where missing values are managed as in Section 5.2.4 and we recall that the log-likelihood is also approximated for numerical reasons, to avoid computing large powers of the transition matrices. We want to find $\hat{\boldsymbol{\theta}}$ which maximizes the following approximated log-likelihood function

$$\ell_K(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \delta_k \prod_{s=1}^{S_i} p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) \right).$$

This maximization is achieved via an EM algorithm which considers the complete-data log-likelihood defined by

$$\ell_K(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \delta_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\sum_{s=1}^{S_i} \log p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) \right).$$

Conditional probabilities Let $\alpha_{ikhs(t)}(\boldsymbol{\theta})$ be the probability of the partial sequence $y_{is(0)}, \ldots, y_{is(t)}$ and ending up in state h at time t under component k. Moreover, let $\beta_{ikhs(t)}(\boldsymbol{\theta})$ be the probability of the ending partial sequence $y_{is(t+1)}, \ldots, y_{is(T_{is})}$ given a start in state h at time t under component k. These probabilities can be easily obtained by the forward/backward algorithm (see Du Roy de Chaumaray and Marbac (2021a)). We deduce that the probability $\gamma_{ikhs(t)}(\boldsymbol{\theta})$ of being in state h at time $t \in \{0, \ldots, T_{is}\}$ for $\boldsymbol{y}_i$ under component k is

$$\gamma_{ikhs(t)}(\boldsymbol{\theta}) = \mathbb{P}(X_{is(t)} = h \mid \boldsymbol{y}_{is}, Z_{ik} = 1; \boldsymbol{\theta}) = \frac{\alpha_{ikhs(t)}(\boldsymbol{\theta})\beta_{ikhs(t)}(\boldsymbol{\theta})}{\sum_{\ell=1}^{M} \alpha_{ik\ell s(t)}(\boldsymbol{\theta})\beta_{ik\ell s(t)}(\boldsymbol{\theta})}.$$

The probability $\xi_{ikh\ell s(t)}(\boldsymbol{\theta})$ of being in state ℓ at time $t \in \Omega_i$ and in state h at time $t-1$ for observation $\boldsymbol{y}_i$ under component k is

$$\begin{aligned}\xi_{ikh\ell s(t)}(\boldsymbol{\theta}) &= \mathbb{P}(X_{is(t)} = \ell, X_{is(t-1)} = h \mid \boldsymbol{y}_{is}, Z_{ik} = 1; \boldsymbol{\theta}) \\ &= \frac{\alpha_{ikhs(t)}(\boldsymbol{\theta})\mathbf{A}_k[h,\ell]g(y_{is(t)}; \boldsymbol{\lambda}_\ell, \varepsilon_\ell)\beta_{ik\ell s(t)}(\boldsymbol{\theta})}{\sum_{h'=1}^{M}\sum_{\ell'=1}^{M} \alpha_{ikh's(t)}(\boldsymbol{\theta})\mathbf{A}_k[h',\ell']g(y_{is(t)}; \boldsymbol{\lambda}_{\ell'}, \varepsilon_{\ell'})\beta_{ik\ell's(t)}(\boldsymbol{\theta})}.\end{aligned}$$

The probability τ_{ik} that one observation arises from component k is

$$\tau_{ik}(\boldsymbol{\theta}) = \mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i, \boldsymbol{\theta}) = \frac{\prod_{s=1}^{S_i}\sum_{h=1}^{M} \alpha_{ikhs(T_{is})}(\boldsymbol{\theta})}{\sum_{k'=1}^{K}\prod_{s=1}^{S_i}\sum_{h=1}^{M} \alpha_{ik'hs(T_{is})}(\boldsymbol{\theta})}.$$

The probability $\eta_{ihs(t)}$ that observation i is in state h at time t of sequence s is

$$\eta_{ihs(t)}(\boldsymbol{\theta}) = \mathbb{P}(X_{is(t)} = h \mid \boldsymbol{y}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_{ik}(\boldsymbol{\theta})\gamma_{ikhs(t)}(\boldsymbol{\theta}).$$

EM algorithm The EM algorithm is an iterative algorithm randomly initialized at the model parameter $\boldsymbol{\theta}^{[0]}$. It alternates between two steps: the Expectation step (E-step) consisting of computing the expectation of the complete-data likelihood under the current parameters, and the maximization step (M-step) consisting of maximizing this expectation over the model parameters. Iteration $[r]$ of the algorithm is defined by

E-step Conditional probability computation, updating of

$$\tau_{ik}(\boldsymbol{\theta}^{[r-1]}),\ \gamma_{ikhs(t)}(\boldsymbol{\theta}^{[r-1]}),\ \eta_{ihs(t)}(\boldsymbol{\theta}^{[r-1]},\ \text{and}\ \xi_{ikh\ell s(t)}(\boldsymbol{\theta}^{[r-1]}).$$

M-step Parameter updating

$$\delta_k^{[r]} = \frac{n_k(\boldsymbol{\theta}^{[r-1]})}{n},\ \pi_{kh}^{[r]} = \frac{n_{kh(0)}(\boldsymbol{\theta}^{[r-1]})}{n_k(\boldsymbol{\theta}^{[r-1]})},\ \mathbf{A}_k[h,\ell]^{[r]} = \frac{n_{kh\ell}(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})},\ \varepsilon_h^{[r]} = \frac{w_h(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})},$$

$$\text{and}\ \ \boldsymbol{\lambda}_h^{[r]} = \arg\max_{\lambda_h} \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}} \eta_{ihs(t)}(\boldsymbol{\theta}^{[r-1]}) g_c(y_{is(t)}; \boldsymbol{\lambda}_h),$$

where

$$n_k(\boldsymbol{\theta}) = \sum_{i=1}^{n} \tau_{ik}(\boldsymbol{\theta}),\ n_{kh}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}} \tau_{ik}(\boldsymbol{\theta})\gamma_{ikhs(t)},\ n_{kh(0)}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i} \tau_{ik}(\boldsymbol{\theta})\gamma_{ikhs(0)}(\boldsymbol{\theta}),$$

$$n_{kh\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=1}^{T_{is}} \tau_{ik}(\boldsymbol{\theta})\xi_{ikh\ell s(t)}(\boldsymbol{\theta})\ \text{ and }\ w_h(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}} \eta_{ihs(t)}(\boldsymbol{\theta})\mathbf{1}_{\{y_{is(t)}=0\}}.$$

5.2.6 Numerical illustrations

This section aims to highlight the main properties of the model used in numerical experiments. First, simulated data are used to illustrate the exponential decay of the probabilities of misclassification (given by Theorem 5.2), the convergence of estimators and the robustness of the approach to missingness. Second, our approach is applied to the data from the PAT study. All the experiments are conducted with the R package **MHMM** available on CRAN.

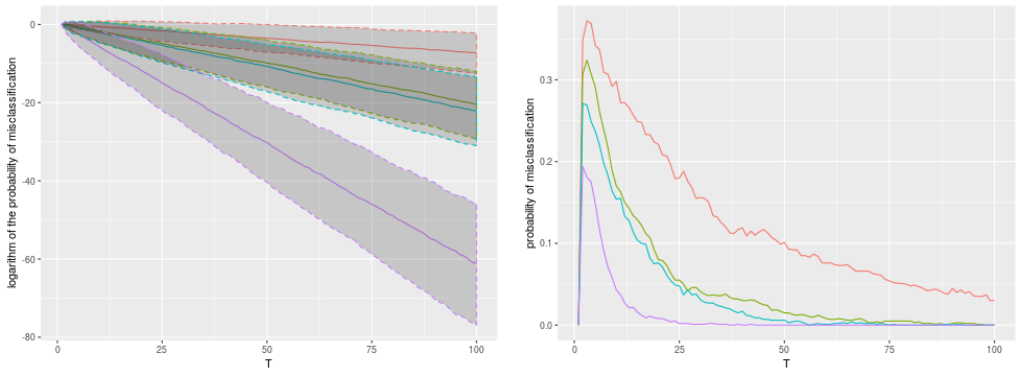
Simulated data All the simulations are performed according to the same model. This model is a bi-components mixture of HMM with two states (*i.e.*, $K = M = 2$) and equal proportions (*i.e.*, $\delta_1 = \delta_2 = 1/2$). The distribution of $Y_{i(t)}$ conditionally on the state h is a ZIG distribution. We have

$$\varepsilon_1 = \varepsilon_2 = 0.1,\ a_1 = 1,\ b_1 = b_2 = 1,\ \mathbf{A}_1 = \begin{bmatrix} e & 1-e \\ 1-e & e \end{bmatrix} \text{ and } \mathbf{A}_2 = \begin{bmatrix} 1-e & e \\ e & 1-e \end{bmatrix}.$$

The parameter $a_2 > 1$ controls the separation of the distribution of $Y_{i(t)}$ given the state. The parameter e controls the separation of the distribution of X given the class (when e increases, the constant c in Theorem 5.2 increases). We consider four cases: hard ($e = 0.75$ and $a_2 = 3$), medium-hard ($e = 0.90$ and $a_2 = 3$), medium-easy ($e = 0.75$ and $a_2 = 5$) and easy ($e = 0.90$ and $a_2 = 5$).

Theorem 5.2 states that the probabilities of misclassification decrease at an exponential rate with T. To illustrate this property, 1000 sequences are generated for $T = 1, \ldots, 100$ and the four cases. For each sequence $\boldsymbol{y}_i$, we compute $\log(\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)/\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i))$ when k_0 is the true class, k the alternative and the true model parameters are used. Figure 3.1(a) shows the behavior of $\log(\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)/\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i))$ (the median of this log ratio is plotted with a plain line and a 90% confidence interval is plotted with a gray area). Note that this log ratio of probabilities linearly decreases with T which illustrates the exponential decay of the probabilities of misclassification. Moreover, Figure 5.3(b) presents the empirical probabilities of misclassification and thus also illustrates Theorem 5.2. As expected, this shows that the decay of the probabilities of misclassification is faster as the overlaps between class decreases.

We illustrate the convergence of the estimators (partition, latent states and parameters) when the model parameters are estimated by maximum likelihood (see Section 5.2.5). We compute the mean square error (MSE) between the model parameters and their estimators. Moreover, we compute the adjusted Rand index (ARI; Hubert and Arabie (1985)) between the true partition and the partition given by the MAP rule, and between the true state sequences and the estimated state sequences given by the MAP rule (obtained with the Viterbi algorithm presented in Viterbi (1967)). Table 5.1 shows the results obtained with two different sample sizes n and two different lengths of sequences T, considering the medium-hard case. It can be seen that the partition and the model parameters are well estimated. Indeed, the MLE converge to the true parameters as T or n increases, except for the proportion of each component δ_k. The convergence of the estimator of the proportions depends mainly on the sample size n. We notice that the partition obtained by our estimation procedure corresponds to the true partition (for n and T sufficiently large) even if we are not under the true parameters but under the MLE, which is not an immediate consequence of Theorem 5.2. On the contrary, we do not find the true state sequences almost surely, as the number of states to be estimated is also growing with n and T. This result was expected because the number of latent states increases with T and n while the number of parameters and the dimension of the partition does not increase with T. Results obtained for the three other cases are similar and are presented in Du Roy de Chaumaray and Marbac (2021a).



(a) Median (in plain) and 90%-confidence region (gray area) of $\log \frac{\mathbb{P}(Z_{ik}=1|\boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1|\boldsymbol{y}_i)}$.

(b) Probability of misclassification.

Figure 5.3: Results obtained on 1000 observations for the hard (orange), medium-hard (green), medium-easy (blue) and easy (purple) cases.

Table 5.1: Convergence of estimators when 1000 replicates are drawn from the medium case: ARI between estimated and true partition, ARI between estimated and true latent states and MSE between the MLE and the true parameters

		ARI (latent variables)		MSE (model parameters)				
n	T	partition	states	$\mathbf{A}_k$	ε_h	a_h	b_h	δ_k
10	100	0.995	0.621	0.021	0.001	0.088	0.024	0.047
10	500	1.000	0.632	0.007	0.000	0.020	0.005	0.048
100	100	0.996	0.630	0.004	0.000	0.011	0.003	0.005
100	500	1.000	0.634	0.003	0.000	0.005	0.002	0.005

We now investigate the robustness of the proposed method with missingness. We compare the accuracy of the estimators (ARI for the latent variables and MSE for the parameters) obtained on samples without missingness to the accuracy of the estimators obtained when missingness is added to the samples. Three situations of missingness are considered: missing completely at random-1 (MCAR-1) (*i.e.*, one sequence of 10 missing values is added to each sequence $\boldsymbol{y}_i$, the location of the sequence follows a uniform distribution), MCAR-2 (*i.e.*, two sequences of 20 missing values are added for each sequence $\boldsymbol{y}_i$, the location of the sequences follows a uniform distribution) and missing not at random (MNAR) (*i.e.*, the probability of observing the value $y_{i(t)}$ being equal to $e^{y_{i(t)}}/(1+e^{y_{i(t)}})$). Note that the last situation adds many missing values when the true value of $y_{i(t)}$ is close to zero, so the occurrence of missing values depends on the latent states. Table 5.2 compares the results obtained with and without missingness, considering the medium-hard case. It shows that estimators are robust to missingness. Results obtained for the other three cases are similar and are reported in Du Roy de Chaumaray and Marbac (2021a).

Using the approach on classical accelerometer data We consider the accelerometer data measured on three subjects available from Huang et al. (2018b). The accelerometer measures the activity every five minutes for one week. Note that the first subject has 2% of missing values. The purpose of this section is to illustrate the differences between the method of Huang

Table 5.2: Convergence of estimators obtained over 1000 replicates with and without missing data when data are sampled from case medium: ARI between estimated and true partition, ARI between estimated and true latent states and MSE between the MLE and the true parameters

			Adjusted Rand index		Mean square error				
n	T	missingness	partition	states	$\mathbf{A}_k$	ε_h	a_h	b_h	δ_k
10	100	no missingness	0.995	0.621	0.021	0.001	0.088	0.024	0.047
		MCAR-1	0.991	0.613	0.024	0.001	0.102	0.028	0.047
		MCAR-2	0.987	0.605	0.028	0.001	0.113	0.032	0.047
		MNAR	0.934	0.497	0.051	0.003	0.398	0.050	0.050
10	500	no missingness	1.000	0.632	0.007	0.000	0.020	0.005	0.048
		MCAR-1	1.000	0.631	0.007	0.000	0.020	0.005	0.048
		MCAR-2	1.000	0.631	0.007	0.000	0.019	0.005	0.048
		MNAR	0.999	0.516	0.021	0.003	0.233	0.028	0.048
100	100	no missingness	0.996	0.630	0.004	0.000	0.011	0.003	0.005
		MCAR-1	0.994	0.624	0.004	0.000	0.013	0.003	0.005
		MCAR-2	0.989	0.618	0.005	0.000	0.014	0.004	0.005
		MNAR	0.951	0.512	0.014	0.002	0.200	0.026	0.005
100	500	no missingness	1.000	0.634	0.003	0.000	0.005	0.002	0.005
		MCAR-1	1.000	0.633	0.002	0.000	0.006	0.002	0.005
		MCAR-2	1.000	0.632	0.002	0.000	0.005	0.002	0.005
		MNAR	1.000	0.520	0.011	0.002	0.198	0.026	0.005

et al. (2018b) and the method proposed in this paper.

Huang et al. (2018b) consider one HMM per subject with three latent states. This model is used for monitoring the circadian rhythm, subject by subject. Because they fit one HMM per sequence measured by the accelerometer of a subject, the definition of the activity level is different for each subject (see, Huang et al. (2018b):Figure 4). This is not an issue for their study because the analysis is done subject by subject. However, the mean time spent per activity levels cannot be compared among the subjects. The method proposed here makes this comparison possible. Figure 5.4 depicts the activity data of the three subjects, the expected value of $Y_{i(t)}$ conditionally on the most likely state and on the most likely component and the probability of each state. Based on the QQ-plots (see, Du Roy de Chaumaray and Marbac (2021a)), we consider $M = 4$ activity levels. These levels can be easily characterized with the model parameters presented in Table 5.3. Moreover, the transition matrices also make sense. For instance, class 1 (subjects 9 and 20) has an almost tri-diagonal transition matrix (by considering an order between the states given through the activity levels per state) and class-2 (subject 2) is composed of a subject with low-overall activity

$$\hat{\mathbf{A}}_1 = \begin{bmatrix} 0.86 & 0.14 & 0.00 & 0.00 \\ 0.12 & 0.81 & 0.06 & 0.01 \\ 0.00 & 0.07 & 0.79 & 0.14 \\ 0.00 & 0.00 & 0.13 & 0.87 \end{bmatrix}.$$

5.2.7 Analysis of PAT data

In this section, we analyze the data presented in Section 5.2.2.

Experimental conditions In order to compare our approach to the cuts defined *a priori* in the PAT study (see Section 5.2.2), the model was fitted with four activity levels. Note that

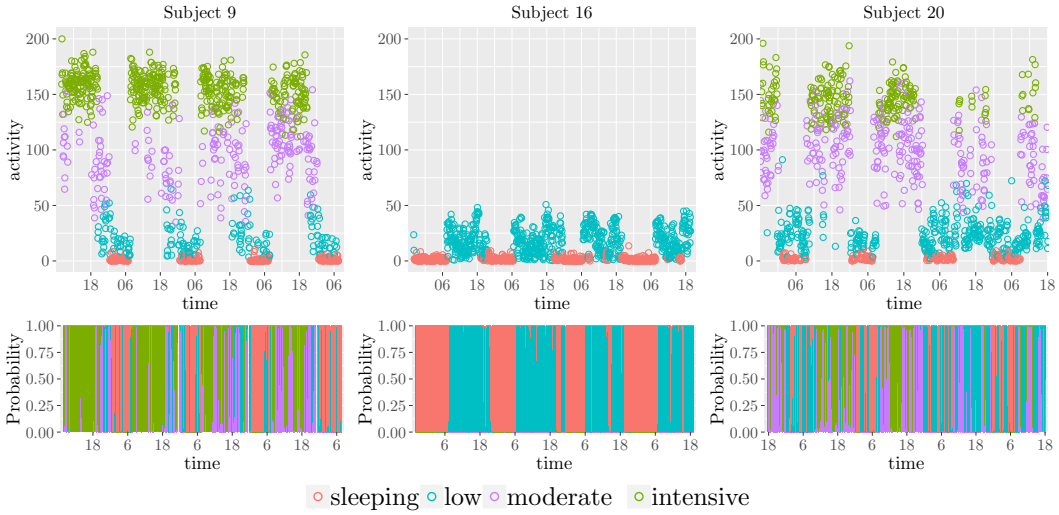


Figure 5.4: State estimation for the three subjects: (top) accelerometer data where color indicates the expected value of $Y_{i(t)}$ conditionally to the most likely state and to the most likely component; (bottom) probability of each state at each time.

Table 5.3: Parameters and mean time per states for the three subjects

State name	ε_h	a_h	b_h	mean	sd
intensive-level	0.00	98.94	0.65	152.76	15.36
moderate-level	0.00	11.09	0.11	99.34	29.84
low-level	0.00	2.32	0.11	20.98	13.79
sleeping	0.22	1.48	0.72	2.06	1.70

selecting the number of states in HMM remains a challenging problem (see the discussion in the conclusion). However, approaches considering four activity levels are standard for accelerometer data. The number of components (*i.e.*, the number of classes) is estimated, using an information criterion unlike the PAT study where it is arbitrarily set at 3 or 4. For each number of components, 5000 random initializations of the EM algorithm are performed. The analysis needs about one day of computation on a 32-Intel(R) Xeon(R) CPU E5-4627 v4 @ 2.60GHz.

Model selection To select the number of components, we use two information criteria which are generally used in clustering:

$$\text{BIC}(K) = \ell_K(\boldsymbol{\theta}; \boldsymbol{y}) - \frac{\nu_K}{2} \log(\sum_{i=1}^{n} \sum_{s=1}^{S_i} T_{is} + 1),$$

and

$$\text{ICL}(K) = \text{BIC}(K) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik}(\hat{\boldsymbol{\theta}}) \log \tau_{ik}(\hat{\boldsymbol{\theta}}),$$

where $\nu_K = (K-1) + K(M + M^2) + 3M$ is the number of parameters for a model with K components and M states and $\hat{z}_{ik}(\hat{\boldsymbol{\theta}})$ defines the partition by the MAP rule associated to the

MLE such that

$$\hat{z}_{ik}(\hat{\boldsymbol{\theta}}) = \begin{cases} 1 & \text{if } \tau_{ik}(\hat{\boldsymbol{\theta}}) = \text{argmax}_{\ell=1,\dots,K}\ \tau_{i\ell}(\hat{\boldsymbol{\theta}}) \\ 0 & \text{otherwise} \end{cases} .$$

The ICL is defined according as the integrated complete-data likelihood computed with the partition given by the MAP rule with the MLE. The values of the information criteria are given in Table 5.4, for different number of classes. Both criteria select five components. The values of ICL(K) are close to those of the BIC(K), implying that the entropy $\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\log\tau_{ik}(\hat{\boldsymbol{\theta}}) \approx 0$. This is a consequence of Theorem 5.2 (see also numerical experiments in Section 5.2.6). In the following, we interpret the results obtained with $M = 4$ activity levels and $K = 5$ classes.

Table 5.4: Information criteria obtained on PAT data with four levels of activity (minima are in bold)

K	1	2	3	4	5	6	7
BIC	-2953933	-2952313	-2951809	-2951705	**-2951308**	-2951364	-2951696
ICL	-2953933	-2952313	-2951810	-2951707	**-2951309**	-2951364	-2951697

Description of the activity levels The parameters of the ZIG distributions are presented in Table 5.5. The four distributions are ordered by the value of their means. The *sleeping state* is characterized by a large probability of observing zero (*i.e.*, ε_h is close to one). However, ε_h is not equal to zero for the other states, but the more active the state is, the smaller ε_h is. We also compute the marginal cutoffs (*i.e.*, the cutoffs by considering the MAP of $\mathbb{P}(X_{i(t)} \mid Y_{i(t)})$). These cutoffs neglect the time dependency due to the Markov structure, but can be compared to the cutoffs proposed by the PAT study. Indeed, according to the PAT study, minutes with < 100 counts are assigned to Sedentary activity, minutes with 100-2019 counts were classified as Light, the Moderate class corresponds to 2020-5998 counts/minute and Vigorous 5999 and above counts/minute. The marginal cutoff associated with the *low-level* state is very close to that of the Sedentary class of the PAT. We find, however, that our marginal cutoffs are more accurate for higher levels of activity. PAT cutoffs do not adequately characterize the activity level of the study population. Finally, contrary to classical thresholds, our modeling approach allows us to capture and characterize the variability associated with the different levels of activity, variability which seems important (see Figure 5.5 and Table 5.5).

Table 5.5: Parameters describing the four activity levels for PAT data and statistics of the distributions

Name of the activity level	Parameters			Statistics	
	ε_h	a_h	b_h	mean	marginal cutoffs
sleeping	0.988	7.470	7.470	0.012	[0, 0]
low-level	0.260	0.974	0.020	36.926	]0, 97.7]
moderate-level	0.025	1.408	0.004	329.249	]97.7, 614.4]
intensive-level	0.007	2.672	0.002	1696.935	]614.4, $+\infty$[

Description of the classes Classes can be described using their proportions and their associated parameters presented in Table 5.5. The data are composed of a majority class ($\delta_1 = 0.518$). Three other classes are composed of more sedentary individuals (*e.g.*, their marginal probabilities of being in states 1 and 2 are higher). Finally, there is a small class ($\delta_5 = 0.045$) which contains

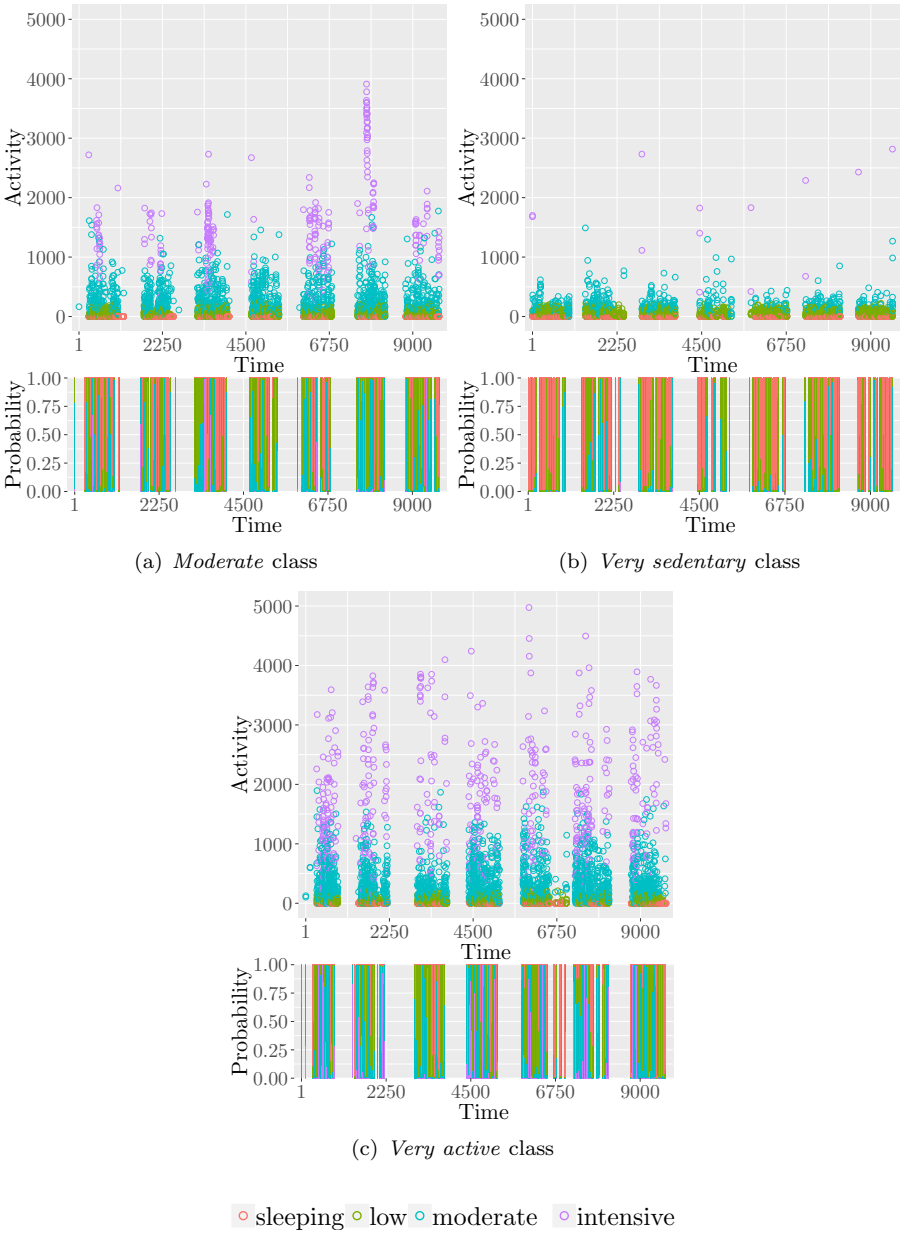
the most active subjects (*i.e.*, $\pi_{k4} = 0.143$). For three of the five classes, Figure 5.5 presents a characteristic subject of each class and the probabilities of the activity levels (the associated graphs for the two remaining classes are given in Du Roy de Chaumaray and Marbac (2021a)). Classes can be interpreted from the mean time spent at different activity levels presented in Table 5.6 and from transition matrices presented in Table 5.7 which are almost tri-diagonal. This could be expected because it seems relevant to obtain a low probability of jumping between the sleeping state and the intensive state. Additionally, the approximation made for efficiently handling the missingness (see Section 5.2.4) turns out to be relevant. The minimal range of missing values is indeed equal to $d_{\min} = 60$ which leads to a distance in total variation between the $d_{\min}$-power of the transition matrices and the stationary distribution being less than 5.10^{-4} for any component.

Table 5.6: Mean time spent at the different activity levels for the five classes

Class	sleeping	low-level	moderate-level	intensive-level
active	0.306	0.284	0.338	0.072
sedentary	0.467	0.209	0.263	0.061
moderate	0.304	0.411	0.225	0.060
very sedentary	0.504	0.366	0.124	0.006
very active	0.189	0.351	0.316	0.143

Table 5.7: Transition matrix for the five classes

			moderate class	
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.76	0.21	0.03	0.00
low-level	0.16	0.73	0.11	0.00
moderate-level	0.03	0.20	0.73	0.04
intensive-level	0.01	0.04	0.16	0.80
			very sedentary class	
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.85	0.08	0.06	0.00
low-level	0.20	0.67	0.13	0.01
moderate-level	0.10	0.11	0.76	0.03
intensive-level	0.01	0.04	0.14	0.82
			very active class	
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.80	0.14	0.05	0.01
low-level	0.08	0.74	0.17	0.01
moderate-level	0.03	0.18	0.69	0.10
intensive-level	0.01	0.05	0.21	0.74



(a) *Moderate* class

(b) *Very sedentary* class

(c) *Very active* class

Figure 5.5: Examples of observations assigned to the five classes with the probabilities of the states.

5.3 Translation-invariant functional clustering to investigate geographical disparities of COVID-19 deaths

5.3.1 Context

In March of 2020, the World Health Organization (WHO) declared pandemic status for the novel coronavirus SARS-Cov-2, denoted COVID-19, indicating that it has reached a critical level of spreading and severity worldwide. The global nature of the COVID-19 pandemic has resulted in much heterogeneity of the data, aggravated further by lack of prior knowledge or coordinated mitigation strategies which have impeded research efforts. For instance, the assumption that the first occurrence emerged concurrently everywhere is improper. Additionally, the number of confirmed cases depends on the number of tests that are being performed in a given region. Hence, a region that has tested very few people can only report very few confirmed cases. Alternatively, the number of COVID-19 deaths is more systematically recorded: countries are asked to follow the 'cause of death' classifications from the WHO's International Classification of Diseases guidelines (World Health Organization (2016)). Though each country is responsible for providing their own guidance on how and when COVID-19 deaths should be recorded, this metric remains more reliable. Undoubtedly, the rapid propagation of this acute infectious respiratory disease has posed governmental challenges. Government responses to contain the virus's spread were multiple (social distancing, travel restrictions, lockdowns, etc.) and their efficiency needs to be investigated. To better understand this virus, it is profoundly useful to cluster regions similarly affected by COVID-19.

In Cheam et al. (2020), we focus on clustering regions of the European Union (EU) and the United States of America (USA) based on the daily COVID-19 deaths recorded over seventeen months. Previous investigations of geographical disparities of COVID-19 (Tang, Wang, and Zhang (2020) and Chen, Yan, and Zhang (2020)) only focus on specific geographical regions (*e.g.*, USA). When considering regions of Europe and North America, a difficulty arises: COVID-19 outbreaks started at different times. The misalignment of the first occurrence between regions should not be neglected, whether between continents or within a country. Another problem to acknowledge is that the mortality occurs at different rates under different population risk factors (Williamson, Walker, and Bhaskaran (2020)). Hence the necessity to adjust these region-specific risk factors is intrinsic to allow regions to be compared fairly. Furthermore, by adjusting the population risk factors, we are able to detect regions more susceptible to COVID-19 and perhaps identify the disparity factors between clusters. For instance, it provides for a retrospective assessment of the effectiveness and the quality of government responses, a concurrent analysis of the economic indicators, and a prospective perception of mental health during this unprecedented period.

In Cheam et al. (2020), we propose a novel three-step approach that circumvents the issues of clustering regions with respect to the COVID-19 dataset: the varying times of arrivals of the virus and the need to incorporate the population risk factors. This approach is named *Clustering Regression residuals of Features given by Translation Invariant Wavelets* (`CRFTIW`). The first step of `CRFTIW` consists of feature extraction using a multiscale approach based on translation-invariant (TI) wavelets (Coifman and Donoho (1995)), which allows the shifted onsets of COVID-19 to be tackled by avoiding any pre-processing step for curve alignment (see Wang and Gasser (1997) and the references cited in Jacques and Preda (2014a):Section 2.3). The objective of constructing clusters that are invariant to time-shifts is somewhat different from conventional clustering in that it allows us to answer slightly different scientific questions about the data. Standard clustering (no time-shifts) will identify regions that peak at the same time, while TI clustering recovers regions that behave in similar patterns that unravel across time. The

features are defined as the logarithm of the norm of the TI wavelet coefficients at each scale. The second step of `CRFTIW` integrates the population characteristics with a single-index regression of the features on the population risk factors. This approach has the benefits of the nonparametric regression but does not suffer from the curse of dimensionality. We show that the residuals of the regression preserve the cluster information. As the third step of `CRFTIW`, clustering of the regions is achieved by fitting a nonparametric mixture on the regression residuals. The only assumption made at this step is to define the density of each component as a product of univariates densities. The proposed approach has differences with the approach of Gaffney and Smyth (2005) despite both methods considering curve translations. First, the scaling that we proposed depends on the covariates (*i.e.,* the risk factors). Second, we use a wavelet approach that permits a greater reduction of the dimension. Finally, we consider a semi-parametric mixture that avoids the bias of the parametric mixtures observed when their parametric assumptions are violated.

For this ongoing COVID-19 dataset, we consider $n = 79$ regions between two continents: the 27 countries within the EU plus the United Kingdom and the 50 states of the USA plus the District of Columbia. There are differences in time of arrival of the peak death rates between regions. This is illustrated by Figure 5.6 which shows that New Hampshire and Pennsylvania have noticeably more delayed peaks than Austria and Italy. Our focus is on the curve $\mathbf{W}_i =$

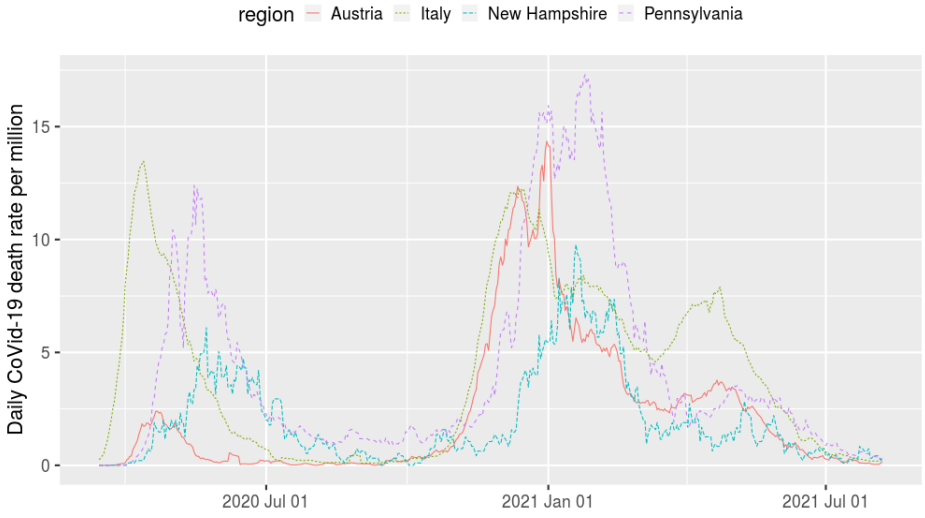


Figure 5.6: Illustration of different arrival times of the COVID-19 among the different regions.

$(W_{i(1)}, \ldots, W_{i(T)})^\top$ recording the daily rate of the number of deaths per million people in each region i for a total of $T = 512$ days (between March 1st, 2020 to July 25, 2021, inclusively), where $W_{i(t)}$ denotes the death rate recorded for region i at time t. Data were extracted from the Center for Systems Science and Engineering at the Johns Hopkins Github repository (file Policy.rds in Badr et al. (2020)) and a 7-day moving average has been performed due to the discrepancy of the data recorded by each region. For instance, this can account for days in the week where data may not be available, such as weekends.

Early findings suggested that differences in COVID-19 disease prevalence and severity may be associated with certain risk factors (Williamson, Walker, and Bhaskaran (2020)). Thus, we consider two groups of risk factors. The first group contains three environmental risk factors (fine particulate matter PM2.5 concentration, nitrogen dioxide NO2 concentration and population

density) and the second group contains eight medical risk factors (age-adjusted percent prevalence of adults with diagnosed diabetes, percent of obese adults, age-adjusted percent prevalence of adults who are current smokers, age-standardized percent prevalence of chronic obstructive pulmonary disease, age-standardized percent prevalence of cardiovascular disease, age-standardized percent prevalence of HIV/AIDS, percent of adults with hypertension and population proportion over 65 years old). All the risk factors were extracted from the Center for Systems Science and Engineering in the Johns Hopkins repository file (file COVID-19_Static.rds in Badr et al. (2020)) and have been scaled. For each group of risk factors, we perform a principal component analysis (PCA) and, in accordance with Kaiser's rule, we consider the first two principal components. For each region i, we store in $\mathbf{X}_i \in \mathbb{R}^4$ the first two principal components of both groups of risk factors.

5.3.2 Method

Outline of the three-step method The daily COVID-19 death curves of the n regions $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are supposed to independently arise from L different clusters. The cluster membership of region i is defined by the latent variable $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iL})^\top$ where $Z_{i\ell} = 1$ if region i belongs to cluster ℓ and $Z_{i\ell} = 0$ otherwise. The model assumes that, conditionally on the cluster ℓ, each $\mathbf{W}_i$ is defined as a product between a noisy version of δ_i-lagged values of an unobserved curve $\boldsymbol{u}_\ell$ and the effect of the population risk characteristics $\mu(\mathbf{X}_i) > 0$, where X_i denotes population risk factors of region i and where we set $\mathbb{E}[\mu(\mathbf{X}_i)] = 1$, for identifiability reasons. The deterministic functions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L$ do not depend on the covariates $\mathbf{X}_i$. Moreover the noises $\boldsymbol{\varepsilon}_{i\ell}$ and the covariates $\mathbf{X}_i$ are independent. Thus, given the cluster membership and the population risk factors of region i, we have

$$\mathbf{W}_i = \sum_{\ell=1}^{L} z_{i\ell}\mu(\boldsymbol{x}_i)(\boldsymbol{u}_\ell^{(\delta_i)} + \boldsymbol{\varepsilon}_{i\ell}^{(\delta_i)}), \tag{5.3}$$

where $\boldsymbol{u}_\ell^{(\delta_i)}$ and $\varepsilon_{i\ell}^{(\delta_i)}$ are δ_i-lagged versions of $\boldsymbol{u}_\ell$ and $\varepsilon_{i\ell}$, and the distribution of each $\varepsilon_{i\ell}$ follows a centered distribution having a finite variance defined by the density f_ℓ (*i.e.*, $\mathbb{E}_{f_\ell}[\varepsilon_{i\ell}] = 0$ and $\mathbb{E}_{f_\ell}[\varepsilon_{i\ell}^2] < \infty$). Thus, the conditional distribution of $\mathbf{W}_i$ given $\mathbf{X}_i = \boldsymbol{x}_i$ is defined by the density

$$f(\boldsymbol{w}_i \mid \boldsymbol{x}_i) = \sum_{\ell=1}^{L} \pi_\ell f_\ell \left(\frac{\boldsymbol{w}_i}{\mu(\boldsymbol{x}_i)} - \boldsymbol{u}_\ell^{(\delta_i)} \right), \tag{5.4}$$

where $\pi_\ell > 0$ is the proportion of cluster ℓ with $\sum_{\ell=1}^{L} \pi_\ell = 1$.

Despite the model defined by (5.4) permitting a clustering of the regions based on the daily COVID-19 death curves with respect to the population risk factors, the estimation the multivariate density f_ℓ is highly complex. Thus, in Cheam et al. (2020), we achieve the clustering with the following three-step approach (see Section 5.3.2):

1. Performing feature extraction of the daily COVID-19 death curves $\mathbf{W}_i$ to obtain $\mathbf{Y}_i \in \mathbb{R}^{J+1}$ using TI wavelets.

2. Fitting single-index regressions of the features $\mathbf{Y}_i$ on the population risk factors $\mathbf{X}_i$ and consider the residuals $\hat{\boldsymbol{\xi}}_i \in \mathbb{R}^{J+1}$.

3. Using the nonparametric mixture to cluster the regions based on the residuals $\hat{\boldsymbol{\xi}}_i$.

This approach is relevant since the specific feature extraction reduces the dimension, allows us to deal with lagged values and keeps the main cluster information. Moreover, the single-index regression keeps the cluster information of the features, allows adjustment to be made on population risk factors, and provides meaningful parameters used for detecting protective or compounding effects of the population characteristics and odd ratios.

Feature extraction and time misalignment A wavelet basis is a set of functions obtained as translations and dilatations of two specific functions: a scaling function denoted by ϕ and a mother wavelet denoted by ψ. For the purpose of this paper, we use Daubechies wavelets and in particular the Symmlet family. Such wavelets are optimal in the sense that they have minimal support for a given number of null moments. We present the essentials below; more details can be found in Daubechies (1992) or Mallat (2008). The decomposition of the observations in a given wavelet basis is defined by

$$W_i(t) = \alpha_{i,0,0}\phi_{0,0}(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \beta_{i,j,k}\psi_{j,k}(t), \quad t \in [1,T],$$

with $J = \log_2(T)$, $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t-k)$, $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t-k)$, $\alpha_{i,0,0} \approx \sqrt{T}\int_1^T W_i(t)\phi_{0,k}(t)dt$ and $\beta_{i,j,k} \approx \sqrt{T}\int_1^T W_i(t)\psi_{j,k}(t)dt$ are the empirical wavelet coefficients of the ith individual. A discrete wavelet transform (DWT) corresponds to the computation of these coefficients. In practice, a fast wavelet decomposition and reconstruction algorithm can be computed using the algorithm proposed by Mallat (1989) (in only $\mathcal{O}(T)$ operations). As mentioned in the introduction, a simple shift in the observed function will potentially result in a significant change in the DWT. Since we use the latter for feature extraction and the observed curves can start at different times, such behavior is not suitable.

In the TI case, we consider the fast translation-invariant discrete wavelet transform (TIDWT) developed by Coifman and Donoho (1995), in a denoising framework. This transformation has been independently discovered, on several occasions, in different communities, and has received several different names, including the "*à trous*" algorithm (Holschneider et al. (1990) and Dutilleux (1990)), the undecimated DWT (Lang et al. (1996)), the shift-invariant DWT (Lang et al. (1995)) or the stationary DWT (Nason and Silverman (1995)), to name just a few (see, *e.g.* Fowler (2005) for a review of some of the various different variants). There are many ways to implement this transformation, and many ways to represent the resulting overcomplete set of wavelet coefficients. We have chosen to focus on the TIDWT of Coifman and Donoho (1995), which provide equivalences and a way to go from one to the other of these representations, for example with the stationary DWT of Nason and Silverman (1995). The main difference with the orthogonal case is that the dictionary is now a tight frame instead of an orthonormal basis (see Mallat (2008):Chapter 5) and the number of coefficients per scale is no longer dyadic but of length T (see Coifman and Donoho (1995) for more details). This wavelet transform is called translation-invariant by Coifman and Donoho (1995) since the whole dictionary is invariant under circular translation. More precisely, for a vector w of size T, let S_h denotes the circulant shift by h defined by $(S_h w)(t) = w(t+h)$ modulo T. As in the traditional case, TIDWT is calculated by a series of decimation and filtering operations, only the additional circulant shift S_h is added and the corresponding wavelet dictionary is obtained by sampling the locations more finely (*i.e.*, one location per sample point). TIDWT consists of calculating the DWT of the shifted data for each shift $h \in \{0,\dots,T-1\}$. Coifman and Donoho (1995) propose an algorithm to perform this transformation in $\mathcal{O}(T\log_2 T)$ operations (we used the R package **rwavelet** which provides an implementation Navarro and Chesneau (2020)). The invariance property of their construction is

formally expressed in terms of the circulant matrix containing the wavelet coefficients (see Coifman and Donoho (1995):eq. (3)). In other words, for a curve W_i translated by h, the wavelet coefficients at each scale will be the same up to some permutation. Thus the norm of the latter is preserved scale-by-scale.

The redundancy of TIDWT makes it possible to detect the presence of hidden information such as stationary or non-stationary patterns as well as their location, making it particularly suitable for clustering purposes. This type of invariant representation has been exploited in many applications (such as denoising, Coifman and Donoho (1995), or texture image classification and segmentation, Unser (1995)). In addition, the use of wavelets allows the information contained in the time series to be compressed into a small number of wavelet coefficients. Following Antoniadis et al. (2013), we characterize each time series by the vector of the energy contribution of their wavelet coefficients at each scale with the difference that the coefficients are calculated by TIDWT instead of DWT. This extension is possible because the expansion being in a *tight frame*, the norm is also conserved (see Mallat (2008):Chapter 11 for more details). More precisely, using Parseval's identity, we have

$$\|\mathbf{W}_i\|_2^2 = 2^{-J}\sum_{k=0}^{T-1}\alpha_{i,0,k}^2 + \sum_{j=1}^{J}2^{-j}\sum_{k=0}^{T-1}\beta_{i,j,k}^2 = 2^{-J}\|\boldsymbol{\theta}_{i0}\|_2^2 + \sum_{j=1}^{J}2^{-j}\|\boldsymbol{\theta}_{ij}\|_2^2, \tag{5.5}$$

where $\boldsymbol{\theta}_{ij} = (\alpha_{i,0,0},\ldots,\alpha_{i,0,T-1},\beta_{i,j,0},\ldots,\beta_{i,j,T-1})^\top$ and the factor 2^{-j} is used to compensate for the redundancy of this representation. Thus, denoting by y_{ij} the log total norm at scale j for the ith individual, we have

$$y_{ij} = \ln\left(\|\boldsymbol{\theta}_{ij}\|_2\right), \quad \forall j = 0,\ldots,J, \quad i = 1,\ldots,n. \tag{5.6}$$

Clustering will therefore be carried out on the basis of the log norm of the TI wavelet coefficients at each scale. Thus, this criterion is not sensitive to the origin of the curves, so it seems relevant given the nature of the data motivating this work.

Adjustment on the population risk factors In this section, we consider the regressions of the features extracted by the wavelet decomposition on the population risk factors. The following lemma shows that the noises of these regressions retain the cluster information given by the daily COVID-19 death curves and permit the information of the population risk factors to be considered in the clustering procedure. Note that the same nonparametric function is used for the regression of each feature (*i.e.*, $j = 0,\ldots,J$). Thus, the lemma shows how the regression function is estimated based on the $n \times (J+1)$ observations.

Lemma 5.1. Let data arise from (5.4) and features are defined by (5.6). Defines the noise of the regression $\boldsymbol{\xi}_i = (\xi_{i1},\ldots,\xi_{iJ})^\top$ for $j = 0,\ldots,J$ as

$$y_{ij}^\star = m(\boldsymbol{x}_i) + \xi_{ij}, \tag{5.7}$$

with

$$\mathbb{E}[m(\mathbf{X}_i)] = 0 \text{ and } \mathbb{E}[\boldsymbol{\xi}_{ij}] = 0,$$

where $\boldsymbol{x}_i$ and $\boldsymbol{\xi}_{ij}$ are independent, $y_{ij}^\star = y_{ij} - \Delta_j$, $m(\boldsymbol{x}_i) = \ln\mu(\boldsymbol{x}_i) - \mathbb{E}[\ln\mu(\mathbf{X}_i)]$, $\Delta_j = \mathbb{E}[\ln\mu(\mathbf{X}_i)] + \sum_{\ell=1}^{L}\pi_\ell\mathbb{E}[\frac{1}{2}\ln\|v_{\ell j} + \varepsilon_{i\ell j}^\star\|_2^2]$, and $v_{\ell j}$ and $\varepsilon_{i\ell j}^\star$ are the features of $\boldsymbol{u}_\ell$ and $\boldsymbol{\varepsilon}_{i\ell}$, respectively (their formal definition is given in the proof of the lemma presented Cheam et al. (2020)). The vector of noises $\boldsymbol{\xi}_i$ then follows a mixture model with latent variable Z_i defined by the density

$$g(\boldsymbol{\xi}_i) = \sum_{\ell=1}^{L}\pi_\ell g_\ell(\boldsymbol{\xi}_i - \boldsymbol{\lambda}_\ell),$$

where $\boldsymbol{\lambda}_\ell = (\lambda_{\ell 0}, \ldots, \lambda_{\ell J})^\top$, $\lambda_{\ell j} = \mathbb{E}[\frac{1}{2}\ln\|v_{\ell j} + \varepsilon^\star_{i\ell j}\|_2^2] - \Delta_j + \mathbb{E}[\ln \mu(\mathbf{X}_i)]$ and $g_1, \ldots, g_\ell$ are densities of centered distributions.

We consider the single-index regression defined by

$$m(\boldsymbol{x}_i) := \nu(\boldsymbol{x}_i^\top \boldsymbol{\gamma}). \tag{5.8}$$

This semiparametric approach is flexible, and avoids the assumptions of the parametric approaches that can be violated and the curse of dimensionality of the full nonparametric approaches. The parameter of the index $\boldsymbol{\gamma}$ permits population characteristics having a protective or compounding effect to be detected. Moreover, considering two sets of covariates X_i and $X_{i'}$, the difference $\nu(\mathbf{X}_i^\top \boldsymbol{\gamma}) - \nu(\mathbf{X}_{i'}^\top \boldsymbol{\gamma})$ can be interpreted as the logarithm of an odd ratio.

The single-index approach requires a methodology for estimating γ and m, with m being in a function space. A common approach, that avoids a simultaneous search involving an infinite-dimensional parameter, is the profiling (Severini and Wong (1992) and Liang et al. (2010)), which defines $\nu(\boldsymbol{x}_i^\top \boldsymbol{\gamma}) := \nu_{\boldsymbol{\gamma}}$ with

$$\nu_{\boldsymbol{\gamma}}(t) = \mathbb{E}[Y_{ij}^\star \mid \mathbf{X}_i^\top \boldsymbol{\gamma} = t], \qquad j \in \{0, \ldots, J\} \text{ and } t \in \mathbb{R}. \tag{5.9}$$

Hence, one expects that, for each x_i, the true value of the parameter, denoted by γ, realizes the minimum of

$$\boldsymbol{\gamma} \mapsto \sum_{j=0}^{J} \mathbb{E}[\{Y_{ij}^\star - \nu_{\boldsymbol{\gamma}}(\boldsymbol{x}_i^\top \boldsymbol{\gamma})\}^2 \mid \mathbf{X}_i = \boldsymbol{x}_i]. \tag{5.10}$$

However, even if $m_{\boldsymbol{\gamma}}$ is well defined for any $\boldsymbol{\gamma} \in \mathbb{R}^d$, the vector $\boldsymbol{\gamma}$ is not identifiable and only its direction could be consistently estimated. Thus, there are two common approaches to restrict $\boldsymbol{\gamma}$ for identification purposes: either fix one component equal to 1 (Ma and Zhu (2013)), or set the norm of $\boldsymbol{\gamma}$ equal to 1 and fix the sign of its first component to be positive (Zhu and Xue (2006)). The estimation of the single-index regressions is performed by considering the empirical counterpart of (5.9) and (5.10)

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \left(\hat{y}_{ij}^\star - \hat{\nu}_{\boldsymbol{\gamma}}(\mathbf{X}_i^\top \boldsymbol{\gamma})\right)^2,$$

$$\hat{y}_{ij}^\star = y_{ij} - \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \quad \text{and} \quad \hat{\nu}_{\boldsymbol{\gamma}}(u) = \frac{\frac{1}{nh}\sum_{i=1}^{n} \hat{y}_{ij}^\star K\left(\frac{\mathbf{X}_i^\top \boldsymbol{\gamma} - u}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i^\top \boldsymbol{\gamma} - u}{h}\right)},$$

where K is a kernel and h a bandwidth. The estimation procedure is implemented in the R package **regpro** Klemela (2016). The clustering of the regions is also performed on the residuals $\hat{\xi}_i = (\hat{\xi}_{i0}, \ldots, \hat{\xi}_{iJ})^\top$ defined by

$$\hat{\xi}_{ij} = \hat{y}_{ij}^\star - \hat{\nu}_{\hat{\boldsymbol{\gamma}}}(\mathbf{X}_i^\top \hat{\boldsymbol{\gamma}}).$$

Nonparametric clustering of the regions A wide range of literature focuses on models assuming that, conditionally on knowing the particular cluster the subject i came from, its features are independent. Thus, we consider that the conditional distribution of the $\hat{\boldsymbol{\xi}}_i$ given cluster membership is defined as a product of univariate densities. Note that this assumption imposes non-explicit constraints on the distribution of the noises $\boldsymbol{\varepsilon}_i$ defined in (5.3). Therefore, the clustering of the region is performed by considering the marginal density defined by

$$g(\hat{\boldsymbol{\xi}}_i; \boldsymbol{\lambda}) = \sum_{\ell=1}^{L} \pi_\ell \prod_{j=0}^{J} g_{\ell j}(\hat{\xi}_{ij}), \tag{5.11}$$

where $\boldsymbol{\lambda}$ groups the mixing proportions $\pi_1, \ldots, \pi_L$ (where $\pi_\ell > 0$ and $\sum_{\ell=1}^{L} \pi_\ell = 1$) and the univariate densities $g_{\ell j}$. The model (5.11) is identifiable, up to a swapping of the cluster labelling, if the densities $g_{\ell j}$ are linearly independent (see Theorem 8 of Allman, Matias, and Rhodes (2009)). Considering a multivariate kernel defined as a product of J univariate kernels K, the maximum smoothed log-likelihood estimator $\hat{\boldsymbol{\lambda}}_L$ (MSLE) is obtained by maximizing the smoothed log-likelihood $\ell(\boldsymbol{\lambda})$, such that

$$\hat{\boldsymbol{\lambda}}_L = \arg\max_{\boldsymbol{\lambda}} \ell(\boldsymbol{\lambda}; L)$$

and

$$\ell(\boldsymbol{\lambda}; L) = \sum_{i=1}^{n} \ln \left\{ \sum_{\ell=1}^{L} \pi_\ell \prod_{j=1}^{J} \mathcal{N} g_{\ell j}(\hat{\xi}_{ij}) \right\},$$

where

$$\mathcal{N} g_{\ell j}(\hat{\xi}_{ij}) = \exp \left\{ \int_{\Omega_j} \frac{1}{h_j} K \left(\frac{\hat{\xi}_{ij} - u}{h_j} \right) \ln g_{\ell j}(u) du \right\},$$

and $h_1, \ldots, h_J$ are the bandwidths (*i.e.*, $h_j > 0$ and $h_j = o(1)$ for $j = 1, \ldots, J$). Considering the MSLE is more convenient than considering the maximum likelihood estimate because the MSLE can be obtained by a Majorization-Minimization algorithm (see Levine, Hunter, and Chauveau (2011) for details on the algorithm and Zhu and Hunter (2016b) for recent developments) implemented in the R package **mixtools** Benaglia et al. (2009b).

Clustering is achieved by computing the MSLE because this estimator permits a soft assignment where the conditional probability that subject i belongs to cluster ℓ, denoted by $t_{i\ell}(\hat{\boldsymbol{\lambda}}_L)$, can be obtained

$$t_{i\ell}(\hat{\boldsymbol{\lambda}}_L) = \frac{\hat{\pi}_\ell \prod_{j=1}^{J} \mathcal{N} \hat{g}_{\ell j}(\hat{\xi}_{ij})}{\sum_{\ell'=1}^{L} \hat{\pi}_{\ell'} \prod_{j=1}^{J} \mathcal{N} \hat{g}_{\ell' j}(\hat{\xi}_{ij})}.$$

Moreover, a hard assignment can be achieved by applying the maximum *a posteriori* rule (leading that $\hat{z}_{i\ell} = 1$ if $\ell = \arg\max_{\ell'} t_{i\ell}(\hat{\boldsymbol{\lambda}}_L)$ and $\hat{z}_{i\ell} = 0$ otherwise).

5.3.3 Numerical experiments

Simulation set-up Data are independently generated from (5.3), (5.4) and (5.8) with $T = 256$, $K = 3$ and unequal proportions such that $Z_i \sim \mathcal{M}(0.5, 0.25, 0.25)$, a function characterizing class k defined by

$$u_{\ell(t)} = r_{\ell(t)} \mathbb{1}_{\{r_{\ell(t)} > 0\}} \text{ with } r_{\ell(t)} = a_{\ell(t)} \sin \left(b_\ell \pi \frac{t}{T} \right),$$

with $a_{1(t)} = 1$, $b_1 = 2.5$, $a_{2(t)} = (1 + \varsigma)$, $b_2 = 2.5$, $a_{3(t)} = (1 + \varsigma \mathbf{1}_{\{t > 128\}})$ and $b_3 = 2.5 - \varsigma$ where $\varsigma = 0.3$. In the attempt to replicate patterns of regions, we devise the three components such that each represents the severity level of COVID-19: moderate throughout, heavy throughout, and moderate during the first wave but drastically more affected in the second wave. Moreover, the lapse between the two waves of the disease is shorter for class 3. We use a heteroscedastic noise defined by $\varepsilon_{ik(1)} \sim \mathcal{N}\left(0, 0.2^2\right)$, where the conditional distribution of $\varepsilon_{i\ell(t)} \mid \mathcal{F}_i$, with $\mathcal{F}_i$ the natural filtration, is equal to the conditional distribution of $\varepsilon_{i\ell(t)} \mid \varepsilon_{i\ell(t-1)}$ where

$$\varepsilon_{i\ell(t)} \mid \varepsilon_{i\ell(t-1)} \sim \mathcal{N}\left(0, \left(0.2 + 0.2\varepsilon_{i\ell(t-1)}^2\right)^2\right).$$

The bivariate vector of covariates $\mathbf{X}_i = (X_{i1}, X_{i2})^\top$ is composed of two independent standard Gaussian random variables and $\mu(\mathbf{X}_i) := \nu(\mathbf{X}_i^\top \boldsymbol{\gamma})$ with $\boldsymbol{\gamma} = (1/\sqrt{2} \;\; 1/\sqrt{2})^\top$. To illustrate the benefits of `CRFTIW`, we consider the following three scenarios, where the first mimics the situation of the COVID-19 daily death curves:

- Scenario 1 (*translations and covariates*): there is a shift and an effect of the covariates since we set $\delta_i = 50$ with probability 0.5 and $\delta_i = 0$ otherwise, and $\nu(a) = (1 + \varsigma(a^2 - 1))$.
- Scenario 2 (*only covariates*): there is no shift but an effect of the covariates since we set $\delta_i = 0$ and $\nu(a) = (1 + \varsigma(a^2 - 1))$.
- Scenario 3 (*only translations*): there is a shift but no effect of the covariates since we set $\delta_i = 50$ with probability 0.5 and $\delta_i = 0$ otherwise, and $\nu(a) = 1$.

For each scenario, we generate 100 replicates by considering three sample sizes: 50, 100 and 250. All nonparametric estimations are done using a Gaussian kernel with a bandwidth $Cn^{-1/5}$ where C represents the empirical standard deviation of the variable considered by the kernel.

Method comparison Results of `CRFTIW` are compared to those of the following methods:

- `depIntra` is similar to `CRFTIW` but clustering is performed by a mixture considering the within component dependencies (Zhu and Hunter (2019)). This model challenges the assumption of a product decomposition of the component densities made in (5.11).
- `noTI` is similar to `CRFTIW` but the feature extraction is performed with orthogonal wavelets commonly used for wavelet-based clustering (Antoniadis et al. (2013)). This model allows the advantages of TI wavelets to be illustrated.
- `noCov` is similar to `CRFTIW` but considers that $\mu(x_i) = 1$. This model illustrates the importance of considering the population risk factors.
- `adjustFirst` fits the estimators $(\bar{\mu}, \bar{\boldsymbol{\gamma}})$ of the regression of $\bar{\mathbf{W}}_i = \frac{1}{T}\sum_{t=1}^{T} W_{i(t)}$ on $\mathbf{X}_i$, then uses (5.11) to cluster the features provided by Step 1 of `CRFTIW` applied on $\mathbf{W}_i / \bar{\mu}(\boldsymbol{x}_i^\top \bar{\boldsymbol{\gamma}})$. This model justifies the relevance of the order between Steps 1 and 2 of `CRFTIW`.
- `funFEM` (Bouveyron, Côme, and Jacques (2015)) is a standard approach to cluster functional data implemented in the R package **funFEM** Bouveyron (2015). We use this approach on the curves adjusted with the covariates $\mathbf{W}_i / \bar{\mu}(\boldsymbol{x}_i^\top \bar{\boldsymbol{\gamma}})$, a decomposition into a Fourier basis with 25 elements and the arguments of the function *funFEM* are *model='AkjBk'*, *init='kmeans'*, *lambda=0* and *disp=TRUE*. Thus, this model investigates the relevance of the order between Steps 1 and 2 of `CRFTIW`, and the choice of a nonparametric clustering.

Clustering accuracy In all scenarios considered, clustering accuracy of the competing methods is measured using the Adjusted Rand Index (ARI, Hubert and Arabie (1985)) between the estimated partition and the true partition shown in Figure 5.7. The results show that by considering a more complex model than (5.11) clustering is not relevant. Indeed, `CRFTIW` and `depIntra` provide similar results for large samples but `CRFTIW` is more accurate for small samples. Recall that for our COVID-19 study, the sample size is 78 regions and we can argue that the model given by (5.11) is more suitable than a complex model. From Scenarios 1 and 3, the method `noTI` reveals that it is imperative to consider a feature extraction that is invariant to time-shifts. Thus, `CRFTIW` can handle remarkably well, situations with different arrival times. When considering the case of population risk factors, such as Scenarios 1 and 2, it is essential to take into account

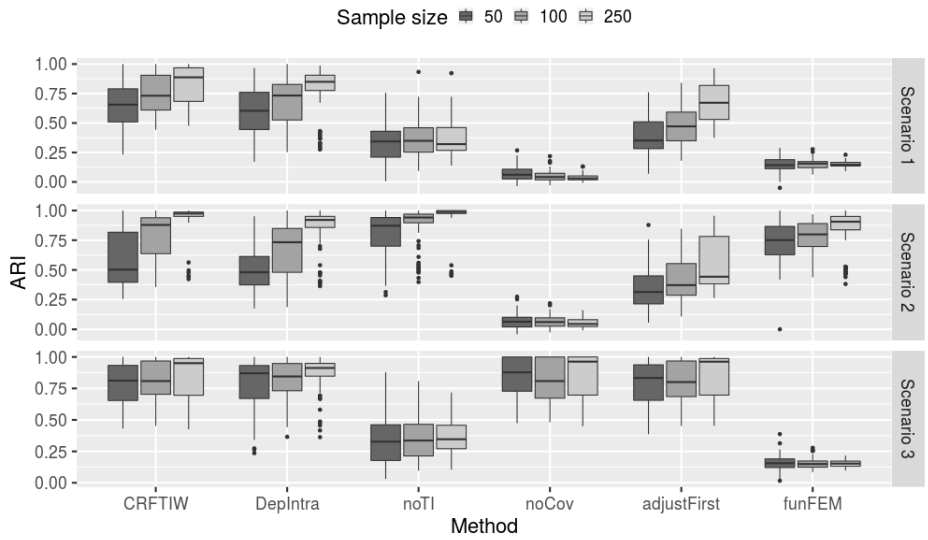


Figure 5.7: Boxplot of the ARI obtained by the competing methods.

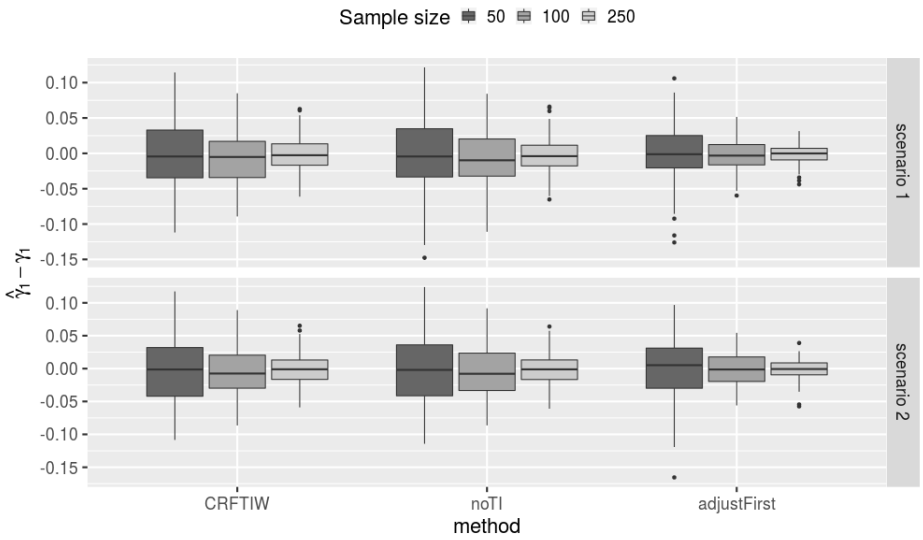
the effects of covariates to fit the partition (see the poor results of `noCov` under these scenarios). Moreover, it seems to be more pertinent to estimate the covariate effects after performing the feature extraction in presence of time-varying arrivals (*i.e.*, `CRFTIW` outperforms `adjustFirst` and `funFEM` under Scenarios 1 and 3).

Covariate effect accuracy The estimation of the covariate effects is a major issue for the COVID-19 application. Indeed, it is essential to properly adjust the covariate effects in order to obtain accurate clusters. Moreover, the parameters and the shape of the function μ make it possible to assert the impact of population risk factors on the death rates. Indeed, the parameters of the index facilitate the interpretation of the population characteristics by distinguishing whether it has a protective or compounding effect. Figure 5.8(a) shows the dispersion of the estimators of γ_1 around the true value of these parameters, for Scenarios 1 and 2. These results illustrate the consistency of the procedure. Note that, as shown by the results of `adjustFirst`, the accuracy of the regression parameters is slightly better when they are obtained directly from the original data and not from the features given by the wavelet decomposition.

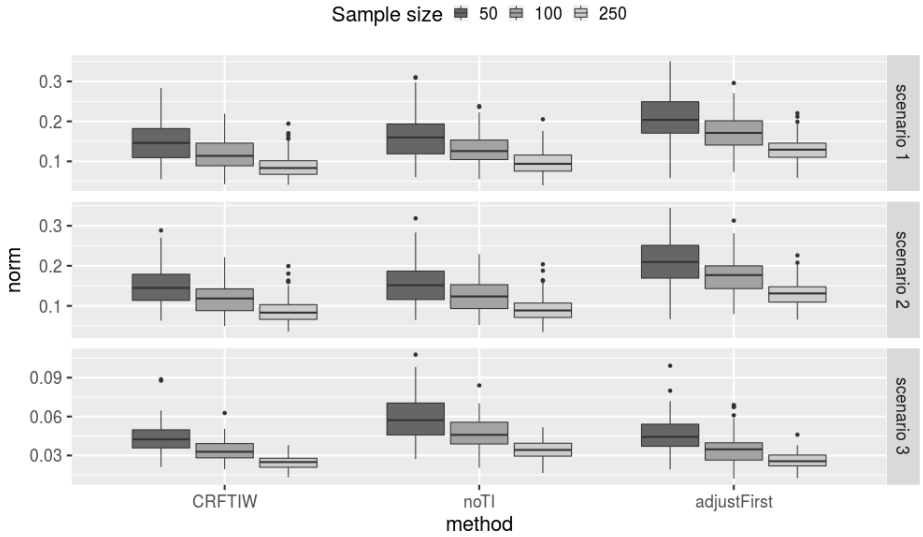
Considering two sets of covariates $\mathbf{X}_i$ and $\mathbf{X}_{i'}$, the difference $\nu(\mathbf{X}_i^\top \boldsymbol{\gamma}) - \nu(\mathbf{X}_{i'}^\top \boldsymbol{\gamma})$ can be interpreted as the logarithm of an odds ratio. To investigate the accuracy of the estimator $\hat{\nu}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}})$, we consider the quantity $e_{\hat{m},\hat{\boldsymbol{\gamma}}}$ defined by

$$e_{\hat{\nu},\hat{\boldsymbol{\gamma}}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \left[\left(\hat{\nu}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}}) - \hat{\nu}(\boldsymbol{x}_{i'}^\top \hat{\boldsymbol{\gamma}}) \right) - \left(\nu(\boldsymbol{x}_i^\top \boldsymbol{\gamma}) - \nu(\boldsymbol{x}_{i'}^\top \boldsymbol{\gamma}) \right) \right]^2 .$$

Figure 5.8(b) depicts the boxplots of $e_{\hat{\nu},\hat{\boldsymbol{\gamma}}}$ obtained from the three scenarios. Thus, the results reflect the consistency of the approach. Moreover, when there is no effect of the covariates, the convergence of the estimators is much faster. However, in the presence of covariate effects, the accuracy of estimators remains the same with or without a translation. Therefore, we conclude that the translation does not increase the difficulty of the estimation of covariates. Finally, note



(a) scenarios with covariate effect



(b) the three scenarios

Figure 5.8: Boxplot of $\hat{\gamma}_1 - \gamma_1$ (a) and $e_{\hat{\nu},\hat{\boldsymbol{\gamma}}}$ (b) obtained by the competing methods.

that despite `adujstFirst` having slightly better parameter estimates, the effects of the covariates (parameters and nonparametric functions) are better estimated with `CRFTIW`.

5.3.4 Investigating geographical disparities for COVID-19

The proposed approach allows risk factors to be taken into account. This is a major issue when comparing regions with respect to COVID-19 daily deaths. Indeed, the adjustment on popula-

tion risk factors enables us to confirm their role (protective or compounding; see Section 5.3.4). Moreover, this adjustment permits similarities and disparities of the disease impacts to be detected, conditional on the population characteristics (see Section 5.3.4). Finally, it allows an *a posteriori* comparison of the clusters to be made with respect to different policy decisions (*e.g.*, lockdown characteristics; see Section 5.3.4).

Population risk factors The estimated coefficients of the single-index regression

$$\hat{\gamma} = (0.433,\ 0.616,\ 0.491,\ 0.439)^\top,$$

and the principal component analysis allow us to compute the contribution of the risk factor to the index. These values are given in Table 5.8 and Table 5.9.

	PM2.5	NO2	population density
Contribution	0.114	0.323	0.565
Axe 1	-0.548	0.747	0.378
Axe 2	0.619	0.057	0.784

Table 5.8: Contribution of the environmental risk factors to the index of the regression and their coordinates on the first two factorial axes (*i.e.,* X_{i1} and X_{i2}).

	Diabetes	Obesity	Smoking	COPD	CVD	HIV	Hypertension	Pop > 65
Contribution	0.303	0.279	0.040	0.076	0.387	0.194	0.441	0.071
Axe 1	0.464	0.431	-0.310	0.329	0.376	0.323	0.301	-0.240
Axe 2	0.056	0.048	0.496	-0.266	0.357	0.002	0.574	0.470

Table 5.9: Contribution of the medical risk factors to the index of the regression and their coordinates on the first two factorial axes (*i.e.,* X_{i3} and X_{i4})

Figure 5.9 shows the estimator of the $\hat{\mu}$ and the density of the index for a range covering 90% of the observed index (this trimming, only performed for this plot, avoids over-interpretation of the curve due to extreme points). This figure confirms that diabetes, overweight, smoking, pulmonary disease (COPD), cardiovascular disease (CVD), HIV, hypertension and age are factors increasing the COVID-19 mortality risk. These results are in agreement with the main mortality risk factors identified in medical publications (see Zhou et al. (2020):Table 1 and Gupta et al. (2020):Figure 2). Moreover, the population density and the concentrations of nitrogen dioxide and fine particulate matter also increase the COVID-19 mortality risk. Additionally, these results align with the findings of other works (Sy, White, and Nichols (2021), Copat et al. (2020) and Pozzer et al. (2020)).

We now illustrate the impact of the adjustment on the population risk factors. Table 5.10 shows the population risk factors and the index of three regions: Portugal, Massachusetts and New Jersey. For example, Portugal and Massachusetts have a population that is less at risk than that of the New Jersey. Taking population risk factors into consideration is vital in order to properly compare the impact of the disease on different regions. Figure 5.10 shows the COVID-19 daily death curves without considering the population risk factors (on the left) and when considering the population risk factors (on the right). Thus, if population characteristics were neglected, one can claim that the first wave of the disease was stronger in New Jersey than in Massachusetts and Portugal. However, after considering the population risk factors, the largest peak of deaths was similar in New Jersey and Massachusetts and it was stronger in Portugal.

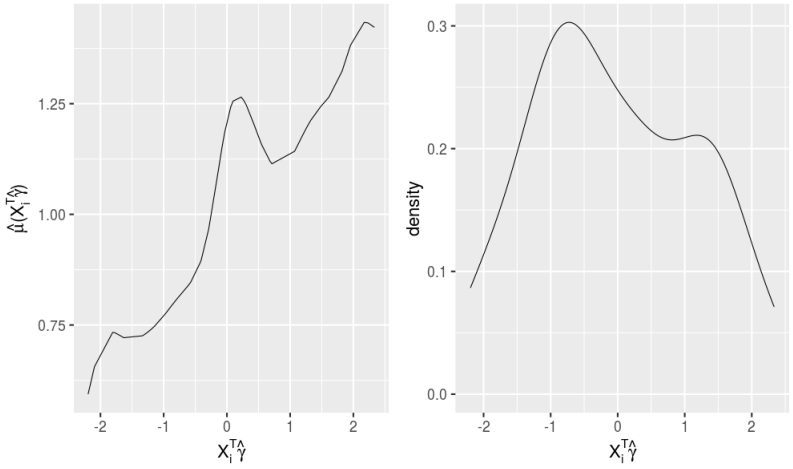


Figure 5.9: Estimator of the impact of the population risk factors $\hat{\mu}$ (on the left) and density of the index $X_i^\top\hat{\gamma}$ (on the right).

Region	$\mathbf{X}_i^\top\hat{\boldsymbol{\gamma}}$	$\hat{\mu}$
Portugal	-1.31	0.73
Massachusetts	-0.28	0.97
New Jersey	1.58	1.26

Table 5.10: Index and effect of the population risk factors of three regions (Portugal, Massachusetts and New Jersey).

Clustering of the regions Clustering is performed by considering a number of clusters between one and ten. Selecting the number of clusters is still a difficult task in nonparametric mixtures. Indeed, despite recent works (Kasahara and Shimotsu (2014); Kwon and Mbakop (2020)) presenting elegant methods for selecting the number of clusters based on the constraint of linear independence between the univariate densities required for model identifiability (Allman, Matias, and Rhodes (2009)), these methods require large samples to perform well (see Section 5 of Kwon and Mbakop (2020)). Because we only have 79 regions, we cannot use these methods and thus select the number of clusters by looking for an elbow in the values of the smoothed log-likelihood. These values are presented in Table 5.11; we focus on the five-clusters partition (note that the difference of $\ell(\hat{\lambda}_L; L)$ obtained with consecutive L is greater than 33 if $L < 5$, otherwise these differences are less than 8 when $L = 5$ and $L = 6$; however seven clusters could also be considered).

Table 5.11: Maximum smoothed log-likelihood $\ell(\hat{\lambda}_L; L)$ with respect to the number of clusters.

L	1	2	3	4	5	6	7	8	9	10
$\ell(\hat{\lambda}_L; L)$	-1161	-1066	-1024	-986	-953	-945	-922	-907	-906	-902

We now describe the clusters based on summary statistics presented in Table 5.12 and the curves adjusted with the population risk factors ($\mathbf{W}_i/\hat{\mu}(\mathbf{X}_i^\top\hat{\boldsymbol{\gamma}})$). Figure 5.11 presents the cluster

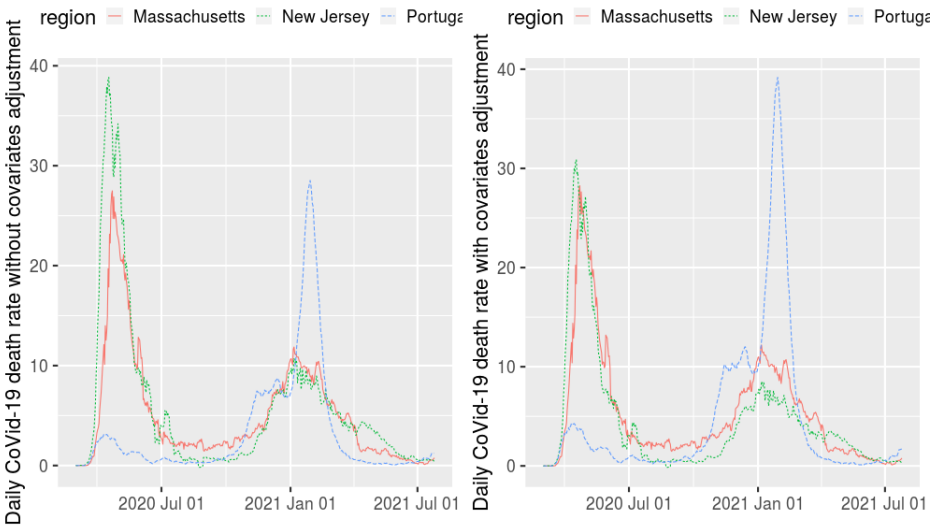


Figure 5.10: COVID-19 death curves without considering the population risk factors (left) and by considering the population risk factors (right) for Portugal, Massachusetts and New Jersey.

memberships of each region and Figure 5.12 presents the curve for one region per cluster to illustrate our cluster interpretation. The labels of clusters are ordered by the impact of the COVID-19 during the studied period. Thus, we notice that cluster 1 has the smallest mean of COVID-19 daily death rates over the studied period and cluster 5 has the highest mean. These

Table 5.12: Statistics per cluster

cluster	proportion	Normalized deaths		Covariate effect	
	π_ℓ	mean	sd.	mean	sd.
1	0.15	777.83	308.08	0.83	0.24
2	0.36	1507.24	274.93	1.08	0.20
3	0.17	1593.68	253.18	1.13	0.25
4	0.13	2230.17	381.75	0.92	0.24
5	0.18	2464.88	459.54	0.91	0.24

results were expected because the distribution of $\hat{\mu}(\mathbf{X}_i^\top \hat{\boldsymbol{\gamma}})$ is supposed to be the same among clusters. Clusters can be interpreted, as follows:

- Cluster 1 contains twelve regions (Cyprus, Denmark, Finland, Alaska, Florida, Hawaii, Maine, North Carolina, Oregon, Utah, Vermont and Washington) that are mainly unaffected by the disease.
- Cluster 2 contains twenty-nine regions (Austria, Germany, Estonia, Latvia, Malta, Netherlands, Romania, California, Delaware, District of Columbia, Georgia, Idaho, Illinois, Indiana, Louisiana, Michigan, Minnesota, Mississippi, Missouri, Nebraska, Nevada, New Hampshire, New Mexico, Pennsylvania, South Carolina, Tennessee, Texas, Virginia and Wisconsin) that suffer from multiple small waves of deaths.

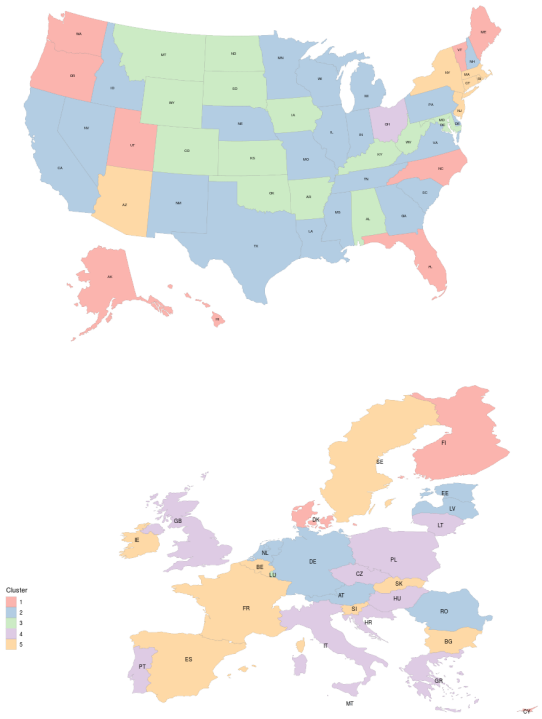


Figure 5.11: Clustering results.

- Cluster 3 contains fourteen regions (Luxembourg, Alabama, Arkansas, Colorado, Iowa, Kansas, Kentucky, Maryland, Montana, North Dakota, Oklahoma, South Dakota, West Virginia and Wyoming) mainly affected by one major wave with steep increasing and decreasing rates.

- Cluster 4 contains ten regions (Croatia, Czechia, Greece, Hungary, Italy, Lithuania, Ohio, Poland, Portugal and United Kingdom) strongly impacted by the disease. These regions suffered from at least three waves of death. Note that there is a wave of deaths which starts before the previous wave ends.

- Cluster 5 contains fourteen regions (Arizona, Belgium, Bulgaria, Connecticut, France, Ireland, Massachusetts, New Jersey, New York, Rhode Island, Slovakia, Slovenia, Spain and Sweden) that begin with a very sharp increase in mortality rate in their first peak, along with a rapid decrease in rate. However, after the second peak, the mortality rate does not constantly decrease and rather experiences three modes in the second wave before dying down.

Clusters analysis example: disparities and policy decisions Amid unforeseen difficulties related to COVID-19, policymakers have resorted to various interventions in attempts to curb the spread of the coronavirus. As seen in Sections 5.3.4 and 5.3.4, by adjusting the risk factors, the proposed approach allows us to detect vulnerable regions and compare them fairly. Additionally, the homogeneity within cluster enables us to analyze the disparities in factors between clusters;

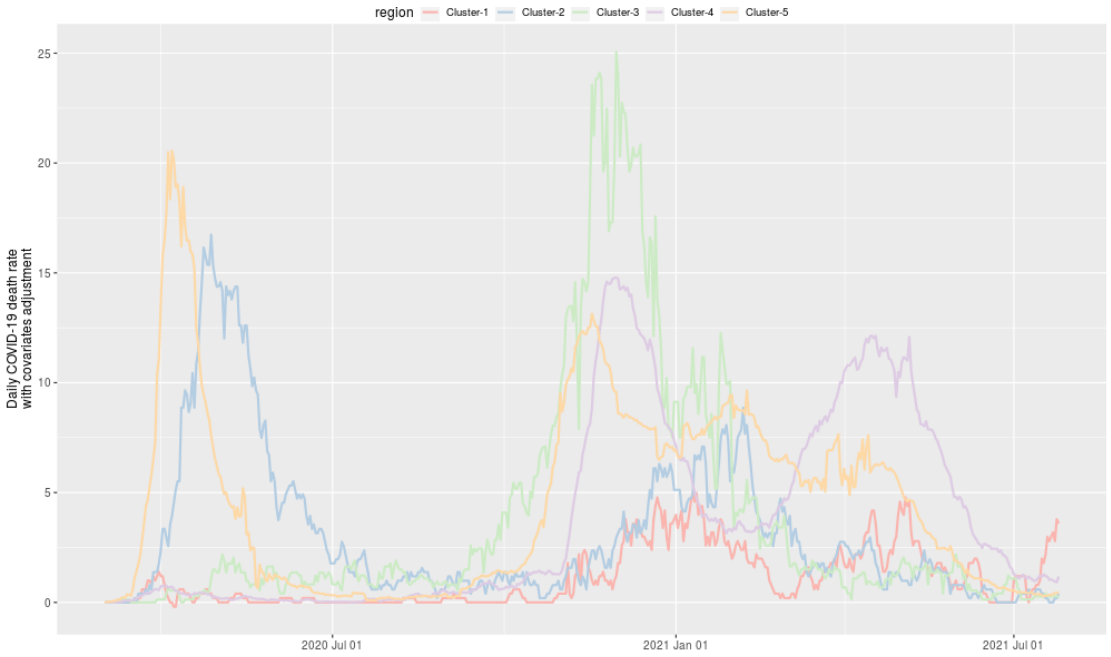
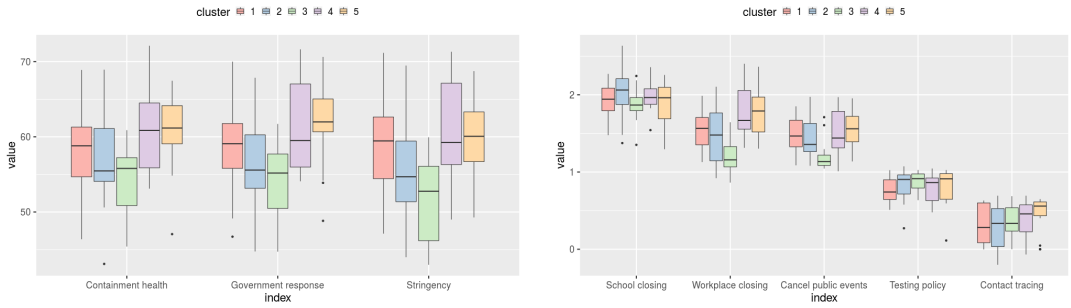


Figure 5.12: Example of COVID-19 daily death curves adjusted with covariates adjustment $W_{ij}/\hat{\mu}(\mathbf{X}_i^\top \hat{\boldsymbol{\gamma}})$. One example per cluster: Cyprus (Cluster-1), District of Columbia (Cluster-2), South Dakota (Cluster-3), Greece (Cluster-4) and France (Cluster-5).

for instance, in this paper, we are looking at the government responses. Note that one may also be interested to study COVID-19's impact on the population mental health or the economic indicators.

Figure 5.13(a) shows the distribution of the value for each indicator, *i.e.*, containment health, government response and stringency, with respect to its clusters. We can observe that the indicators take larger values for regions of clusters 4 and 5 (containing the most impacted regions). The same phenomenon is observed for the specific measures adopted by the governments presented in Figure 5.13(b).

Clusters 4 and 5, which were strongly affected by the virus in the second half of the studied period, seem to take more stringent measures than the other regions, especially cluster 4. One plausible explanation is that the policymakers may abruptly enforce restrictive countermeasures in hopes to lower the increasing mortality rate, thus the observed rapid descent. Hence, the effectiveness of government response may depend on the timing of the measure's implementation, the duration and the stringency (Cheng et al. (2020) and Haug et al. (2020)). To further understand the relationship between the mortality rate and the various policies adopted in reaction to the COVID-19, a more precise analysis can be achieved by fitting a model for multivariate non-stationary time series per clusters (the time series being the daily death rate and the index mentioned above). Thus, a description of the relation between the government responses and the COVID-19 death curve can be done using the model parameters. This can be achieved by modelling the dependencies between the time series (Molenaar, De Gooijer, and Schmitz (1992) and Sanderson, Fryzlewicz, and Jones (2010)) or by a multiple change-point detection (Cho and Fryzlewicz (2015)). Moreover, a comparison of the government interventions with respect to



(a) (b)

Figure 5.13: Boxplot of the overall indicator values for each cluster (a) and for specific measures adopted for each cluster (b): School closing (C1), Workplace closing (C2), Cancelling of public events (C3), Testing policy (H2) and Contact tracing (H3).

COVID-19 can be conducted by comparing the models fitted for each cluster.

5.4 Conclusion and perspectives

The two models presented in this chapter have been developed by considering specificity of the data.

The specific mixture of HMM introduced to analyze accelerometer data, avoids the traditional cutoff point method and provides a better characterization of activity levels for the analysis of these data, while adapting to the population. The proposed model could be applied to a population with different characteristics (*e.g.*, younger) which would lead to different definitions of activity levels. In addition, the use of several HMMs involves taking into account dependency over time and thus improve the traditional method based on cutoff points Witowski et al. (2014). This approach also allows us to take into account the heterogeneity of the population (in the sense of physical activity). An interesting perspective is to consider adjusting for confusing factors (*e.g.*, gender or age). These confusing factors could impact the probabilities of transition between the latent spaces (*e.g.*, using a generalized linear model approach) and/or the definition of the accelerometer measurement given a state (*e.g.*, linear regression on some parameters of the ZIG distribution). In the application, the number of activity levels was not estimated but fixed at a common value for accelerometer data. Estimating the number of states for a mixture of HMMs is an interesting but complex topic: for instance, the use of BIC is criticized (see, Cappé, Moulines, and Rydén (2005):Chapter 15).

To investigate the geographical disparities of the COVID-19 deases, we introduced a new method to cluster functional data when observations are shifted and external covariates are allowed to have a scaling effect. `CRFTIW` is a three-step approach developed for the purpose of analyzing geographical disparities of the COVID-19 impact (measured by the daily number of deaths per million people). As a first step of `CRFTIW`, feature extraction is performed with TI wavelets. While providing an adapted and compact representation of the data, it also allows us to deal with the different times of arrivals of the disease. The main limitation of this approach lies in the dyadic data constraint of the considered sample. This issue could be overcome, for example, by using second generation wavelets and in particular the lifting scheme (Sweldens (1998)), but would lose the property of translation-invariance. However, extending this construction, while preserving both the property of translation-invariance and the property of conservation of the

norm of the coefficients, seems to be an open question that we leave for future work. As a second step of `CRFTIW`, the effect of the population risk factors on the extracted feature is estimated and thus regions can be compared as if they have the same sensitivity of the population to the disease. This step is crucial because we aim to investigate the impact of policy decisions. Obviously, if the purpose is to investigate geographical disparities of the impact of the disease, then no adjustment on the population risk factors should be considered. In such a case, `CRFTIW` can still be used by considering $\mu(\mathbf{X}_i^\top \boldsymbol{\gamma}) = 1$. In the analysis of COVID-19, we consider population adjustment according to the main factors of comorbidity. These factors are well-known to increase the risk of COVID-19 mortality, Zhou et al. (2020); Gupta et al. (2020). However, we advise considering factors already known to be compounded for the disease, and not to use `CRFTIW` to investigate the impact of population risk factors. Indeed, `CRFTIW` does not perform variable selection of the population risk factors and does not permit concluding on causality of the factors. As a third step of `CRFTIW`, a nonparametric mixture is used to achieve the clustering with the assumption that the density of the component is defined as a product of univariate densities. Numerical experiments presented in the paper suggested that considering a more complex model deteriorated the results when the sample size is small (as in the COVID-19 application). However, if the data to be analyzed are composed of several observations, more advanced models could be used Mazo and Averyanov (2019); Zhu and Hunter (2019).

Through the COVID-19 dataset, we had illustrated the importance of adjusting the population risk factors, allowing us to compare regions with a 'standard' comorbidity. Thus, `CRFTIW` found five clusters justified by the mortality rate and curvature. Regions within clusters are varied geographically with different onsets, validating the property of translation-invariance of the proposed method. In addition, as we illustrated, investigations on the effectiveness and agility of government response, the consequences on economic indicators or the impact on human mental health, could be achieved by studying disparities of the indicators between clusters. Despite the model being translation invariant, the time between the arrivals of two waves is discriminative. We argue that this time is important; it determines whether the health facilities have any breaks between waves. Indeed, countries suffering for successive waves of COVID-19 have to postpone non-emergency surgical operations or early cancer detection. Thus, using the proposed clustering, we could investigate the impact of COVID-19 on the global quality of care. Note that an alternative clustering approach could focus only on the death peaks thus neglecting the time between waves. In such a case, the proposed approach is not suitable and we advice using time scaled clustering (Tang and Müller (2009)). Driven by the COVID-19 dataset, we developed this novel approach. However, its application is not limited only to COVID-19. For instance, the problem of time-shifts is also observed in electrocardiogram heartbeats, which Annam, Mittapalli, and Bapi (2011) tackle when clustering heartbeat abnormalities. Our approach could not only handle the time-shift issue, but also allow adjusting plausible factors that may influence the heartbeat. Further, this could extend beyond medical settings, *e.g.,* in motion capture Li and Prakash (2011), there is interest in categorizing types of motion. This could be used for fitness applications to identify whether a person is running or walking, where the motions may begin at different times on separate observed sequences showing a need for the TI property. Since motion can come from different participants, covariate adjustment could also be beneficial for such data.

Chapter 6

Wilks' theorem for semi-parametric regressions with weakly dependent data

Contents

6.1 Introduction

6.1.1 State of the art

We aim modeling and doing inference for one-dimensional time series (Y_i) given a vector-valued time series $(\mathbf{V}_i)$ and the past values of Y_i and $\mathbf{V}_i$, $i \in \mathbb{Z}$. For this purpose, in Du Roy

de Chaumaray, Marbac, and Patilea (2021), we propose flexible semiparametric models for conditional mean and conditional variance of Y_i. Formally, let $(\mathbf{Z}_i)$ be a strictly stationary and strongly mixing sequence of random vectors with $\mathbf{Z}_i = (\mathbf{V}_i^\top, \varepsilon_i)^\top \in \mathbb{R}^{d_X+d_W} \times \mathbb{R}$ where $\mathbf{V}_i = (\mathbf{X}_i^\top, \mathbf{W}_i^\top)^\top \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_W}$. Let $(\mathcal{F}_i)$ be its natural filtration. For any positive integer r, we denote the r lagged values of $\mathbf{Z}_i$ by $\mathbf{Z}_i^{\{r\}} = (\mathbf{V}_{i-1}^\top, Y_{i-1}, \ldots, \mathbf{V}_{i-r}^\top, Y_{i-r})^\top$.

Let us consider the semiparametric model defined by

$$Y_i = \mu(\mathbf{V}_i; \boldsymbol{\gamma}, m) + \varepsilon_i \quad \text{with} \quad \mu(\mathbf{V}_i; \boldsymbol{\gamma}, m) = l(\mathbf{X}_i; \boldsymbol{\gamma}_1) + m(\mathbf{W}_i^\top \boldsymbol{\gamma}_2), \tag{6.1}$$

where

$$\mathbb{E}[\varepsilon_i \mid \mathbf{V}_i, \mathcal{F}_{i-1}] = 0, \tag{6.2}$$

and

$$\mathbb{E}[\varepsilon_i^2 \mid \mathbf{V}_i, \mathcal{F}_{i-1}] = \sigma^2(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta}), \tag{6.3}$$

$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top)^\top$, $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ and $m(\cdot)$ is an infinite dimensional parameter. Thus $\boldsymbol{\theta}$ gathers the finite dimensional parameters, and our interest will focus on this vector, while $m(\cdot)$ is considered as a nuisance parameter. The value of r, as well as the real-valued functions $l(\cdot)$ and $\sigma^2(\cdot)$, are given. Moreover, the functions we consider for $\sigma^2(\cdot)$ do not require to know the infinite dimensional parameter $m(\cdot)$. Let $\boldsymbol{\theta}_0$ and $m_0(\cdot)$ denote the true values of the finite and infinite-dimensional parameters of the model, respectively. The vector V_i may include common random variables and/or lagged values of Y_i, as well as exogenous covariates. We call a model defined by (6.1)-(6.3) a CHPLSIM which stands for *Conditional Heteroscedastic Partially Linear Single-Index Model*. The methodology we will propose in the sequel allows us to replace (6.3) by a higher order moment equation, or to add higher order moments to (6.3). For the sake of simplicity we keep (6.3) and we will only mention such possible extensions in the conclusion section.

CHPLSIM is related to the model proposed by Lian and Liang (2015) in the case of independent observations following the same distribution. Our model covers a wide class of models for weakly dependent and independent data. First, with $l(\mathbf{X}_i; \boldsymbol{\gamma}_1) = \mathbf{X}_i^\top \boldsymbol{\gamma}_1$, CHPLSIM includes the partially linear single-index model (PLSIM) Carroll et al. (1997) in which the errors ε_i are independent and identically distributed (i.i.d.) variables and $\mathbf{V}_i$ are independent covariates. Such semiparametric models were originally used to overcome the curse of dimensionality inherent to nonparametric regression on $\mathbf{W}_i$ by making use of a single-index $\mathbf{W}_i^\top \boldsymbol{\gamma}_2$. The PLSIM includes the partially linear models with a single variable in the nonparametric part. Our non-i.i.d. framework allows for heteroscedasticity in the errors of PLSIM, with the conditional variance of the errors possibly depending of both the covariates and the lagged errors values. For instance, it allows martingale difference errors, as considered by Chen and Cui (2008) and Fan and Liang (2010). Xia, Tong, and Li (1999) considered a model defined by (6.1) for strongly mixing stationary time series, with identity function $l(\cdot)$, $\mathbf{X}_i = \mathbf{W}_i$ and $\mathbf{W}_i$ admitting a density. Their study focuses on the estimation of the parameters in the conditional mean function using kernel smoothing, without investigating the conditional variance, as allows condition (6.3). In the same type of model, using local linear smoothing, Xia and Härdle (2006) allowed for $\mathbf{X}_i$ not necessarily equal to $\mathbf{W}_i$ and, at the price of a trimming, relaxed the condition of a density for $\mathbf{W}_i$ to a density for the index $\mathbf{W}_i^\top \boldsymbol{\gamma}_2$. More recently, using orthogonal series expansions, Dong, Gao, and Tjøstheim (2016) extended the model defined by (6.1) to the case where $\mathbf{X}_i = \mathbf{W}_i$ is a multi-dimensional integrated process.

Model (6.1)-(6.2) is also related to and extends a large class of location-scale type models called conditionnal heteroscedastic autoregressive nonlinear (CHARN) models Härdle, Tsybakov, and Yang (1998); Kanai, Ogata, and Taniguchi (2010). CHARN models include many well-known models widely used with application areas as different as foreign exchange rates Bossaerts, Hafner, and Härdle (1996) or brain and muscular wave analysis Kato, Taniguchi, and Honda (2006). For

general nonlinear autoregressive processes, we refer to the book of Tong (1990) for the basic definitions as well as numerous applications on real data sets. More generally, nonparametric techniques for nonlinear AR processes can be found in the review of Härdle, Lütkepohl, and Chen (1997). CHPLSIM allows for a semiparametric specification of the conditional mean and for exogenous covariates.

We are interested in inference on the finite dimensional parameter $\boldsymbol{\theta}$ constituted of finite-dimensional parameters from both the conditional mean and the conditional variance functions. When the interest focuses on the parameters of the conditional mean, it suffices to consider equations (6.1)-(6.2) with a fully nonparametric conditional variance $\sigma^2(\cdot)$. However, in the time series context, modeling the variance can be important, for instance for forecasting purposes. For our inference purpose, we propose a semiparametric empirical likelihood approach with infinite-dimensional nuisance parameters. Empirical likelihood (EL), introduced by Owen (1988); Owen (2001), is a general inference approach for models specified by moment conditions. Under the assumption of independence between observations, empirical likelihood has been used for inference on finite dimensional parameters into regression models and unconditional moment equations. See Qin and Lawless (1994); see also the review of Chen and Van Keilegom (2009).

Under i.i.d. data assumption, Wang and Jing (1999); Wang and Jing (2003) and Lu (2009) study the conditions implying that the empirical likelihood log-ratio (ELR) still converges to a chi-squared distribution for the partially linear model. Due to the curse of the dimensionality, the performances of the nonparametric estimators decrease dramatically with the number of variables. Xue and Zhu (2006) and Zhu and Xue (2006) show that, if the density of the index is bounded away from zero, the ELR converges to a chi-squared distribution and thus permits parameter testing, for single-index model and PLSIM respectively (see also Zhu et al. (2010)).

6.1.2 Contribution

In Du Roy de Chaumaray, Marbac, and Patilea (2021), we propose a novel general semiparametric regression framework for EL inference which allows for dependent data. Some related cases have been considered in the literature. For instance, the ELR with longitudinal data has been considered by Xue and Zhu (2007), for the partially linear model, and by Li et al. (2010), for PLSIM. In their framework, the convergence of the ELR is guaranteed by the independence between individuals for which a finite bounded number of repeated observations are available. Empirical likelihood has also been used for specific models in times series (see the review of Nordman (2014); see also Chang, Chen, and Chen (2015)). Most of the methods developed in this context are based on a blockwise version of empirical likelihood, first introduced by Kitamura (1997). A large amount of generalizations have been proposed in the literature depending on the type of dependency. We refer to Nordman (2014) for an overview of those techniques of blocking. However, in such an approach, one has to tune additional parameters such as the number, the length or the overlapping of the blocks, which might be a complex task.

Our contribution is the extension of the EL inference approach to the case of CHPLSIM defined by (6.1)-(6.3), for weakly dependent data. This extension is realized without imposing the density of the index bounded away from zero, as it is usually assumed in the literature in the case of i.i.d. data. See, for instance, Zhu and Xue (2006), Zhu et al. (2010) and Lian and Liang (2015). Such a very convenient, though quite stringent, condition implies a bounded support for the index, a restriction which makes practically no sense in a general time series framework. To obtain our results, a preliminary crucial step before using EL consists in building a fixed number of suitable unconditional moment equations equivalent to conditional moment equations defining the regression model. By the definition of these unconditional moment equations, our approach will not require a blocking data technique. Then, we follow the lines of Qin and Lawless (1994),

with the difference of the presence of infinite-dimensional nuisance parameters. We show that the nonparametric estimation of the nuisance parameters does not affect the asymptotics and the ELR still converges to a chi-squared distribution. The negligibility of the nonparametric estimation effect is obtained under mild conditions on the smoothing parameter. Chang, Chen, and Chen (2015) studied the EL inference for unconditional moment equations under strongly mixing conditions, with the number of moment equations allowed to increase with the sample size. Since conditional moment equations models could be approximated by models defined by a large number of unconditional moment equations, in principle, Chang, Chen, and Chen (2015) could also consider semiparametric models. However, the practical effectiveness of their approach remains an uninvestigated issue.

In Section 6.2 we consider the profiling approach for the nuisance parameter $m(\cdot)$ and the identification issue for the finite-dimensional parameters. Next, we establish the equivalence between our model equations and suitable unconditional moment estimating equations for a martingale difference sequence in Section 6.3. The number of unconditional equations is given by the dimension of the vector of identifiable parameters in the (CH)PLSIM. Section 6.4 presents the ELR and the Wilks' Theorem in our context. Section 2.4 illustrates the methodology by numerical experiments and an application using daily pollution data inspired by the study of Lian and Liang (2015). Section 6.6 contains some additional discussion. The proofs and mathematical details are presented in Du Roy de Chaumaray, Marbac, and Patilea (2021).

6.2 Conditional moment equations

6.2.1 The model

Let

$$g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m) = Y_i - \mu(\mathbf{V}_i;\boldsymbol{\gamma},m),$$

with $\mu(\cdot)$ defined in (6.1). The partially linear single index model (PLSIM) is defined by conditional moment equation

$$\mathbb{E}[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m) \mid \mathbf{V}_i,\mathcal{F}_{i-1}] = 0 \Longleftrightarrow \boldsymbol{\gamma} = \boldsymbol{\gamma}_0 \text{ and } m = m_0. \tag{6.4}$$

In such case, the conditional variance of the residuals has to be finite but does not necessarily have a parametric form.

The conditionally heteroscedastic partially linear single index model (CHPLSIM) is defined by two conditional moment equations. For this case, we assume that the second-order conditional moment of the residuals has a semiparametric form. More precisely, the model is defined by the following conditional moment equations

$$\left\{\begin{array}{r} \mathbb{E}[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m) \mid \mathbf{V}_i,\mathcal{F}_{i-1}] = 0 \\ \mathbb{E}[g_\sigma(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\theta},m) \mid \mathbf{V}_i,\mathcal{F}_{i-1}] = 0 \end{array}\right. \Longleftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ and } m = m_0, \tag{6.5}$$

where

$$g_\sigma(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\theta},m) = g_\mu^2(\mathbf{Z}_i;\boldsymbol{\gamma},m) - \sigma^2(\mathbf{V}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\beta}), \tag{6.6}$$

with $\sigma^2(\cdot)$ defined in (6.3).

6.2.2 Profiling nuisance parameter

The model defined by (6.1)-(6.2) requires a methodology for estimating $\boldsymbol{\theta}$ and m, with m being in a function space. A common approach, that avoids a simultaneous search involving an infinite-dimensional parameter, is the profiling Severini and Wong (1992); Liang et al. (2010), which

defines

$$m_{\boldsymbol{\gamma}}(t) = \mathbb{E}[Y_i - l(\mathbf{X}_i; \boldsymbol{\gamma}_1) \mid \mathbf{W}_i^\top \boldsymbol{\gamma}_2 = t], \qquad t \in \mathbb{R}.$$

As usually with such approach, in the following it will be assumed that

$$m_{\gamma_0}(\mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2}) = m_0(\mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2}). \tag{6.7}$$

Hence, one expects that, for each $\boldsymbol{x}, \boldsymbol{w}$, the value $\boldsymbol{\gamma}_0$ realizes the minimum of

$$\boldsymbol{\gamma} \mapsto \mathbb{E}[\{Y_i - l(\boldsymbol{x}_i; \boldsymbol{\gamma}_1) - m_{\boldsymbol{\gamma}}(\boldsymbol{w}_i^\top \boldsymbol{\gamma}_2)\}^2 \mid \mathbf{X}_i = \boldsymbol{x}_i, \mathbf{W}_i = \boldsymbol{w}_i, \mathcal{F}_{i-1}].$$

However, even if $m_\gamma(\cdot)$ is well defined for any $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top)^\top \in \Gamma \subset \mathbb{R}^{d_1} \times \mathbb{R}^{dw}$, in general the value $\boldsymbol{\gamma}_0$ could not be the unique parameter value with this minimum property. More precisely, in general the true value of the vector $\boldsymbol{\gamma}_2$ is not identifiable and only its direction could be consistently estimated. The standard remedies to this identifiability issue are detailed in the following.

6.2.3 Identifiability of the finite-dimensional parameters

Concerning the identification of $\boldsymbol{\gamma}_1 \in \mathbb{R}^{d_1}$, a minimal requirement is that as soon as $l(\mathbf{X}_i; \boldsymbol{\gamma}_1) = l(\mathbf{X}_i; \boldsymbol{\gamma}_1')$ a.s., then necessarily $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_1'$. For instance, when $l(\mathbf{X}_i; \boldsymbol{\gamma}_1) = \mathbf{X}_i^\top \boldsymbol{\gamma}_1$, and thus $d_1 = d_X$, then necessarily $\mathbb{E}(\mathbf{X}_i \mathbf{X}_i^\top)$ invertible. The nonparametric part $m_{\boldsymbol{\gamma}}(\cdot)$ induces some more constraints. It could absorb any intercept in the model equation. Thus, in particular, when $l(\mathbf{X}_i; \boldsymbol{\gamma}_1) = \mathbf{X}_i^\top \boldsymbol{\gamma}_1$, the vectors $\mathbf{X}_i$ and $\mathbf{W}_i$ should not contain constant components.

There are two common approaches to restrict $\boldsymbol{\gamma}_2$ for identification purposes: either fix one component equal to 1 Ma and Zhu (2013), or set the norm of $\boldsymbol{\gamma}_2$ equal to 1 and the sign of one of its components Zhu and Xue (2006). Without loss of generality, we choose the first component of $\boldsymbol{\gamma}_2$ to impose the constraints of value or sign. When the value of the first component is fixed, the parameter $\boldsymbol{\gamma}_2$ could be redefined as $\boldsymbol{\gamma}_2 = (1, \widetilde{\boldsymbol{\gamma}}_2^\top)^\top$ where $\widetilde{\boldsymbol{\gamma}}_2 \in \mathbb{R}^{dw-1}$. The Jacobian matrix of this reparametrization of $\boldsymbol{\gamma}_2$ is the $d_W \times (d_W - 1)$ matrix

$$\mathbf{J}_2(\boldsymbol{\gamma}_2) = \frac{\partial \boldsymbol{\gamma}_2}{\partial \widetilde{\boldsymbol{\gamma}}_2} = \begin{pmatrix} \mathbf{0}_{1\times(d_W-1)} \\ \mathbf{I}_{d_W-1} \end{pmatrix}, \tag{6.8}$$

where here $\mathbf{0}_{1\times(d_W-1)}$ denotes the null $1\times(d_W-1)-$matrix, while $\mathbf{I}_{d_W-1}$ is the $(d_W-1)\times(d_W-1)$ identity matrix. With the second identification approach mentioned above, the reparametrization is

$$\boldsymbol{\gamma}_2 = \left(\sqrt{1 - \|\widetilde{\boldsymbol{\gamma}}_2\|^2},\ \widetilde{\boldsymbol{\gamma}}_2{}^\top\right)^\top,$$

where now $\widetilde{\boldsymbol{\gamma}}_2 \subset \{\boldsymbol{z} \in \mathbb{R}^{dw-1} : \|\boldsymbol{z}\| \leq 1\}$. The Jacobian matrix of this reparametrization using the normalization of $\boldsymbol{\gamma}_2$ is the $d_W \times (d_W - 1)$ matrix

$$\mathbf{J}_2(\boldsymbol{\gamma}_2) = \frac{\partial \boldsymbol{\gamma}_2}{\partial \widetilde{\boldsymbol{\gamma}}_2} = \begin{pmatrix} -\{1 - \|\widetilde{\boldsymbol{\gamma}}_2\|^2\}^{-1/2}\ \widetilde{\boldsymbol{\gamma}}_2{}^\top \\ \mathbf{I}_{d_W-1} \end{pmatrix}. \tag{6.9}$$

Hereafter, when we refer to the true value of the finite-dimensional parameter, we implicitly assume that one of these two approaches for identifying $\boldsymbol{\gamma}_2$ was chosen.

6.3 Unconditional moment estimating equations

This section presents unconditional moment equations which permit parameter inference by using empirical likelihood. The way these equations are constructed will have two important consequences: blocking data is unnecessary and the nonparametric estimation of the infinite-dimensional parameter does not break the chi-squared limit of the ELR statistics. For ease of explanation, we start by introducing an unconditional moment equation which is equivalent to the conditional moment equation of the PLSIM defined in (6.4). Then, we introduce an unconditional moment equation which is equivalent to the conditional moment equation of the CHPLSIM defined in (6.5).

6.3.1 Partially linear single-index model

For the PLSIM, it is quite standard Zhu and Xue (2006) to consider the following unconditional moment equation

$$\mathbb{E}[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\widetilde{\nabla}_{\boldsymbol{\gamma}}g_\mu(Z_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})]=0, \tag{6.10}$$

where $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1^\top,\boldsymbol{\gamma}_2^\top)^\top\in\mathbb{R}^{d_\gamma}$, $d_\gamma=d_1+d_W$, and

$$\widetilde{\nabla}_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})=\mathbf{J}(\boldsymbol{\gamma})\nabla_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\in\mathbb{R}^{d_\gamma-1},$$

with $\mathbf{J}(\boldsymbol{\gamma})$ the $(d_\gamma-1)\times d_\gamma$ Jacobian matrix of the reparametrization chosen to guarantee the identification of the finite-dimensional parameter and $\nabla_{\boldsymbol{\gamma}}$ (resp. $\nabla_{\boldsymbol{\gamma}_1}$) the column matrix-valued operator of the first order partial derivatives with respect to the components of $\boldsymbol{\gamma}\in\mathbb{R}^{d_\gamma}$ (resp. $\boldsymbol{\gamma}_1\in\mathbb{R}^{d_1}$). In our context,

$$\nabla_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})=-\begin{bmatrix}\nabla_{\boldsymbol{\gamma}_1}l(\mathbf{X}_i;\boldsymbol{\gamma}_1)-\mathbb{E}[\nabla_{\boldsymbol{\gamma}_1}l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2]\\ m'(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\left(\mathbf{W}_i-\mathbb{E}[\mathbf{W}_i\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2]\right)\end{bmatrix}$$

$$\text{and}\quad \mathbf{J}(\boldsymbol{\gamma})=\begin{pmatrix}\mathbf{I}_{d_1} & \mathbf{0}_{d_1\times(d_W-1)}\\ \mathbf{0}_{d_W\times d_1} & \mathbf{J}_2(\boldsymbol{\gamma}_2)\end{pmatrix},$$

with $m'(\cdot)$ the derivative of $m(\cdot)$ and $\mathbf{J}_2(\boldsymbol{\gamma}_2)$ the Jacobian matrix of the parametrization of $\boldsymbol{\gamma}_2$, that is either the matrix defined in (6.8) or the one defined in (6.9).

The following lemma proposes new unconditional moment equation by introducing a positive weight function $\omega(\mathbf{V}_i)$ in (6.10). Showing the equivalence between the conditional moment equation (6.4) and our new unconditional moment equation, we deduce that the latter equation could be used for EL inference.

Lemma 6.1. Let $\omega(\cdot)$ be a positive function of $\mathbf{V}_i=(\mathbf{X}_i^\top,\mathbf{W}_i^\top)^\top$ and $H_\mu(\boldsymbol{\gamma})$ be the Hessian matrix of the map $\boldsymbol{\gamma}\mapsto\mathbb{E}[\mathbb{E}^2[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\mid\mathbf{V}_i,\mathcal{F}_{i-1}]\,\omega(\mathbf{V}_i)]$. Assume that conditions (6.4) and (6.7) hold true and $H_\mu(\boldsymbol{\gamma})$ is definite positive. Then

$$\mathbb{E}[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\widetilde{\nabla}_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\omega(\mathbf{V}_i)]=\mathbf{0}\quad\Leftrightarrow\quad\boldsymbol{\gamma}=\boldsymbol{\gamma}_0. \tag{6.11}$$

For the PLSIM, we consider $\omega(\mathbf{V}_i)=\eta_{\boldsymbol{\gamma},f}^4(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)$ where $\eta_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)$ is the density of the index $\mathbf{W}_i^\top\boldsymbol{\gamma}_2$, which is assumed to exist. This choice of the weights $\omega(\mathbf{V}_i)$ allows to cancel all the terms $\eta_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)$ appearing in the denominators, and thus to keep them away from zero. Thus, for the control of the small values in the denominators, it is no longer needed to assume that the density of the index is bounded away from zero. This assumption, often imposed in the

semiparametric literature, is quite unrealistic for bounded vectors $\mathbf{W}_i$ and could not even hold when the $\mathbf{W}_i$'s are unbounded. Imposing bounded $\mathbf{W}_i$ in a time series framework where $\mathbf{W}_i$ could include lagged values of Y_i would be too restrictive.

Thus, we consider that the parameters are defined by the unconditional moment equations

$$\mathbb{E}[\Psi(\mathbf{Z}_i;\boldsymbol{\gamma},\eta_{\boldsymbol{\gamma}})]=\mathbf{0}, \tag{6.12}$$

where $\Psi(\mathbf{Z}_i;\boldsymbol{\gamma},\eta_{\boldsymbol{\gamma}})=g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\tilde{\nabla}_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\eta^4_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\in\mathbb{R}^{d_{\boldsymbol{\gamma}}-1}$. Thus, we have

$$\begin{aligned}\Psi(\mathbf{Z}_i;\boldsymbol{\gamma},\eta_{\boldsymbol{\gamma}})=&\Big(\{Y_i-l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\}\eta_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)-\eta_{\boldsymbol{\gamma},m}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\Big)\\&\times\mathbf{J}(\boldsymbol{\gamma})\begin{bmatrix}\eta^2_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\Big(\nabla_{\boldsymbol{\gamma}_1}l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\eta_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)-\eta_{\boldsymbol{\gamma},X}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\Big)\\\eta_{\boldsymbol{\gamma},m'}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\Big(\mathbf{W}_i\eta_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)-\eta_{\boldsymbol{\gamma},W}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)\Big)\end{bmatrix},\end{aligned} \tag{6.13}$$

where the vector $\eta_{\boldsymbol{\gamma}}=(\eta_{\boldsymbol{\gamma},m},\eta_{\boldsymbol{\gamma},m'},\eta_{\boldsymbol{\gamma},X},\eta_{\boldsymbol{\gamma},W},\eta_{\boldsymbol{\gamma},f})^\top$ groups all the non-parametric elements and, using the stationarity of the process, is given for any $t\in\mathbb{R}$ by

$$\begin{aligned}\eta_{\boldsymbol{\gamma},m}(t)&=m_{\boldsymbol{\gamma}}(t)\eta_{\boldsymbol{\gamma},f}(t)=\mathbb{E}[Y_i-l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2=t]\eta_{\boldsymbol{\gamma},f}(t),\\\eta_{\boldsymbol{\gamma},m'}(t)&=\eta^2_{\boldsymbol{\gamma},f}(t)\frac{\partial}{\partial t}m_{\boldsymbol{\gamma}}(t)=\eta^2_{\boldsymbol{\gamma},f}(t)\frac{\partial}{\partial t}\mathbb{E}[Y_i-l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2=t],\\\eta_{\boldsymbol{\gamma},X}(t)&=\mathbb{E}[\nabla_{\boldsymbol{\gamma}_1}l(\mathbf{X}_i;\boldsymbol{\gamma}_1)\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2=t]\eta_{\boldsymbol{\gamma},f}(t),\\\eta_{\boldsymbol{\gamma},W}(t)&=\mathbb{E}[\mathbf{W}_i\mid\mathbf{W}_i^\top\boldsymbol{\gamma}_2=t]\eta_{\boldsymbol{\gamma},f}(t).\end{aligned}$$

6.3.2 Conditionally heteroscedastic partially linear single-index model

For the CHPLSIM we have to construct an unconditional moment equation to take into account the conditional variance condition in (6.3). In this case, the finite-dimensional parameters are $\theta=(\gamma^\top,\beta^\top)^\top\in\mathbb{R}^{d_\theta}$ with $d_\theta=d_\gamma+d_\beta$. Given the definition (6.6), we have

$$\nabla_{\boldsymbol{\beta}}g_\sigma(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\theta},m)=-\nabla_{\boldsymbol{\beta}}\sigma^2(\mathbf{V}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\beta})\in\mathbb{R}^{d_\beta}.$$

The following lemma provides the unconditional moment equations for EL inference in CHPLSIM. The proof is similar to the proof of Lemma 6.1 and is thus omitted.

Lemma 6.2. Let $\omega_1(\cdot)$ and $\omega_2(\cdot)$ be positive functions of V_i. Let $H_\mu(\boldsymbol{\gamma})$ and $H_\sigma(\boldsymbol{\beta})$ be the Hessian matrices of the maps

$$\boldsymbol{\gamma}\mapsto\mathbb{E}[\mathbb{E}^2[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\mid\mathbf{V}_i,\mathcal{F}_{i-1}]\omega_1(\mathbf{V}_i)]$$

and

$$\boldsymbol{\beta}\mapsto\mathbb{E}[\mathbb{E}^2[g_\sigma(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}},\boldsymbol{\theta},m)\mid\mathbf{V}_i,\mathcal{F}_{i-1}]\omega_2(\mathbf{V}_i)].$$

Assume that conditions (6.5) and (6.7) hold true and $H_\mu(\boldsymbol{\gamma})$ and $H_\sigma(\boldsymbol{\beta})$ are definite positive. Then

$$\begin{cases}\mathbb{E}[g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\widetilde{\nabla}_{\boldsymbol{\gamma}}g_\mu(\mathbf{Z}_i;\boldsymbol{\gamma},m_{\boldsymbol{\gamma}})\omega_1(\mathbf{V}_i)]=\mathbf{0}\\\mathbb{E}[g_\sigma(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\theta},m_{\boldsymbol{\gamma}})\nabla_{\boldsymbol{\beta}}\sigma^2(\mathbf{V}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\beta})\omega_2(\mathbf{V}_i)]=\mathbf{0}\end{cases}\quad\Leftrightarrow\quad\boldsymbol{\theta}=\boldsymbol{\theta}_0.$$

To cancel all the denominators induced by the non-parametric estimator, we take $\omega_1(\mathbf{V}_i)=\eta^4_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)$ and $\omega_2(\mathbf{V}_i)=\eta^2_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top\boldsymbol{\gamma}_2)$. Thus, we consider that the parameters are defined by the unconditional moment equations

$$\mathbb{E}[\Psi(\mathbf{Z}_i,\mathbf{Z}_i^{\{r\}};\boldsymbol{\theta},\eta_{\boldsymbol{\gamma}})]=0, \tag{6.14}$$

where $\eta_{\boldsymbol{\gamma}}$ is defined as in section 6.3 and $\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) \in \mathbb{R}^{d_\theta - 1}$ with

$$\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = \begin{pmatrix} g_\mu(\mathbf{Z}_i; \boldsymbol{\gamma}, m_{\boldsymbol{\gamma}}) \widetilde{\nabla}_{\boldsymbol{\gamma}} g_\mu(\mathbf{Z}_i; \boldsymbol{\gamma}, m_{\boldsymbol{\gamma}}) \eta_{\boldsymbol{\gamma},f}^4(\mathbf{W}_i^\top \boldsymbol{\gamma}_2) \\ g_\sigma(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, m_{\boldsymbol{\gamma}}) \nabla_{\boldsymbol{\beta}} \sigma^2(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta}) \eta_{\boldsymbol{\gamma},f}^2(\mathbf{W}_i^\top \boldsymbol{\gamma}_2) \end{pmatrix}. \tag{6.15}$$

6.4 Parameter inference with weakly dependent data

6.4.1 General framework of empirical likelihood

In the sequel, for EL inference in the CHPLSIM we use condition (6.14), while for EL inference in the PLSIM we use condition (6.12). With a slight abuse of notation, in the sequel we use the notation $\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}})$, with some given integer $r \geq 0$, for both PLSIM and CHPLSIM conditions. By definition, the case $r = 0$ corresponds to the case where $\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}})$ does not depend on the lagged values of $\mathbf{Z}_i$. This is the case for PLSIM, but this situation could also occur in CHPLSIM.

By construction, we have the following important property in the context of dependent data.

Lemma 6.3. The estimating function $\Psi(\cdot, \cdot; \cdot, \cdot)$ satisfies the following property :

$$\forall i \neq j \quad \mathbb{E}\left[\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}_0, \eta_0) \Psi(\mathbf{Z}_j, \mathbf{Z}_j^{\{r\}}; \boldsymbol{\theta}_0, \eta_0)^\top\right] = \mathbf{0}. \tag{6.16}$$

This result is a direct consequence of the fact that $E\left[\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}_0, \eta_0) \mid \mathbf{V}_i, \mathcal{F}_{i-1}\right] = \mathbf{0}$. This property indicates that one can consistently estimate the so-called long-run covariance matrix of the vector-valued sequence $\Psi(\mathbf{Z}_1, \mathbf{Z}_1^{\{r\}}; \boldsymbol{\theta}_0, \eta_0), \ldots, \Psi(\mathbf{Z}_n, \mathbf{Z}_n^{\{r\}}; \boldsymbol{\theta}_0, \eta_0)$ by the standard sample covariance matrix, using our estimating function. Therefore, blocking data is unnecessary in our framework, which is the one of a martingale difference sequence with respect to the filtration $\sigma(\mathbf{V}_i, \mathcal{F}_{i-1})$. See also Kitamura (1997), page 2092, and Chang, Chen, and Chen (2015) page 287.

If $\eta_{\boldsymbol{\gamma}}$ is given, the empirical likelihood, obtained with the unconditional moment conditions we propose for the (CH)PLSIM, is defined by

$$L(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = \max_{\pi_1, \ldots, \pi_n} \prod_{i=1}^n \pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}),$$

where $\sum_{i=1}^n \pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) \Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = \mathbf{0}$, $\pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) \geq 0$, $\sum_{i=1}^n \pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = 1$. Thus, we have

$$\pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = \frac{1}{n} \frac{1}{1 + \lambda(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}})^\top \Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}})},$$

where $\lambda(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) \in \mathbb{R}^{d_1 + d_W - 1}$ are the Lagrange multipliers which permit to satisfy the empirical counterpart of the restriction (6.14), that is

$$\sum_{i=1}^n \pi_i(\boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) \Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \eta_{\boldsymbol{\gamma}}) = \mathbf{0}.$$

The empirical log-likelihood ratio is then defined by

$$\ell_n(\theta, \eta_\gamma) = \sum_{i=1}^n \ln(1 + \lambda(\theta, \eta_\gamma)^\top \Psi(Z_i, Z_i^{\{r\}}; \theta, \eta_\gamma)).$$

As the infinite-dimensional parameter η_γ is unknown, nonparametric estimation using kernel smoothing is used instead. Thus, we propose to consider

$$\ell_n(\boldsymbol{\theta}, \widehat{\eta}_{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \ln\left(1 + \lambda(\boldsymbol{\theta}, \widehat{\eta}_{\boldsymbol{\gamma}})^\top \Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}, \widehat{\eta}_{\boldsymbol{\gamma}})\right), \tag{6.17}$$

where

$$\widehat{\eta}_{\boldsymbol{\gamma}} = (\widehat{\eta}_{\boldsymbol{\gamma},m}, \widehat{\eta}_{\boldsymbol{\gamma},m'}, \widehat{\eta}_{\boldsymbol{\gamma},X}, \widehat{\eta}_{\boldsymbol{\gamma},W}, \widehat{\eta}_{\boldsymbol{\gamma},f})^\top, \tag{6.18}$$

with, for any $t \in \mathbb{R}$,

$$\begin{aligned}
\widehat{\eta}_{\boldsymbol{\gamma},f}(t) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right), \\
\widehat{\eta}_{\boldsymbol{\gamma},m}(t) &= \frac{1}{nh}\sum_{i=1}^{n} \{Y_i - l(\mathbf{X}_i; \boldsymbol{\gamma}_1)\} K\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right), \\
\widehat{\eta}_{\boldsymbol{\gamma},X}(t) &= \frac{1}{nh}\sum_{i=1}^{n} \nabla_{\boldsymbol{\gamma}_1} l(\mathbf{X}_i; \boldsymbol{\gamma}_1) K\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right), \\
\widehat{\eta}_{\boldsymbol{\gamma},W}(t) &= \frac{1}{nh}\sum_{i=1}^{n} \mathbf{W}_i K\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right),
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\eta}_{\boldsymbol{\gamma},m'}(t) = \frac{1}{nh^2}\Bigg[\widehat{\eta}_{\boldsymbol{\gamma},f}(t) \sum_{i=1}^{n} \{Y_i - l(\mathbf{X}_i; \boldsymbol{\gamma}_1)\} K'\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right) \\
- \widehat{\eta}_{\boldsymbol{\gamma},m}(t) \sum_{i=1}^{n} K'\left(\frac{\mathbf{W}_i^\top \boldsymbol{\gamma}_2 - t}{h}\right)\Bigg],
\end{aligned}$$

$K'(\cdot)$ is the derivative of the univariate kernel $K(\cdot)$ and h is the bandwidth.

6.4.2 Assumptions

We will consider weakly dependent data which satisfy strong mixing conditions. We refer the reader to the book of Rio (2000) and to the survey of Bradley (2005) for the basic properties as well as the asymptotic behavior of weakly dependent processes. We will focus our attention on α-mixing sequences. We use the following measure of dependence between two σ-fields $\mathcal{A}$ and $\mathcal{B}$:

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

We recall that a sequence $(\mathbf{Z}_i)_{i\in\mathbb{Z}}$ is said to be α-mixing if $\alpha_m = \sup_{j\in\mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+m}^{\infty})$ goes to zero as m tends to infinity, where for any $-\infty \le j \le l \le \infty$, $\mathcal{F}_j^l = \sigma(\mathbf{Z}_i, j \le i \le l)$. Let

$$\mathbf{U}_i = (l(\mathbf{X}_i; \boldsymbol{\gamma}_{0,1}), \nabla_{\boldsymbol{\gamma}_1} l(\mathbf{X}_i; \boldsymbol{\gamma}_{0,1})^\top, \mathbf{W}_i^\top, \varepsilon_i)^\top.$$

Assumption 6.1. (i) The process $(\mathbf{Z}_i)_{i\in\mathbb{Z}}$, $\mathbf{Z}_i = (\mathbf{X}_i^\top, \mathbf{W}_i^\top, \varepsilon_i)^\top \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_W} \times \mathbb{R}$, is strictly stationary and strongly mixing with mixing coefficients α_m satisfying

$$\alpha_m = O(m^{-\xi}) \text{ with } \xi > 10\frac{s}{s-3}, \tag{6.19}$$

for some $s > 6$ such that

$$\sup_{\|\boldsymbol{c}\|=1} \mathbb{E}[|\mathbf{U}_i^\top \boldsymbol{c}|^s] < \infty. \tag{6.20}$$

(ii) The marginal density of the index $\eta_{\boldsymbol{\gamma}_0,f}(\cdot)$ of the index $\mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2}$ is such that

$$\sup_{t\in\mathbb{R}} \eta_{\boldsymbol{\gamma}_0,f}(t) < \infty,$$

and

$$\sup_{\|\boldsymbol{c}\|=1} \sup_{t\in\mathbb{R}} \mathbb{E}[|\mathbf{U}_i^\top \boldsymbol{c}|\{|t| + |\mathbf{U}_i^\top \boldsymbol{c}|^{s-1}\} \mid \mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2} = t]\eta_{\boldsymbol{\gamma}_0,f}(t) < \infty. \tag{6.21}$$

Moreover, there is some $j^\star < \infty$ such that, for all $j \geq j^\star$,

$$\sup_{(t,t')\in\mathbb{R}^2} \mathbb{E}[|\mathbf{U}_0^\top \mathbf{U}_j| \mid \mathbf{W}_0^\top \boldsymbol{\gamma}_{0,2} = t, \mathbf{W}_j^\top \boldsymbol{\gamma}_{0,2} = t']f_{\mathbf{W}_0^\top \boldsymbol{\gamma}_{0,2}, \mathbf{W}_j^\top \boldsymbol{\gamma}_{0,2}}(t,t') < \infty,$$

where $f_{\mathbf{W}_0^\top \boldsymbol{\gamma}_{0,2}, \mathbf{W}_j^\top \boldsymbol{\gamma}_{0,2}}(\cdot)$ is the joint density of $\mathbf{W}_0^\top \boldsymbol{\gamma}_{0,2}$ and $\mathbf{W}_j^\top \boldsymbol{\gamma}_{0,2}$.

(iii) The second partial derivatives of $\mathbb{E}[\nabla_{\boldsymbol{\gamma}_1} l(\mathbf{X}_i; \boldsymbol{\gamma}_1) \mid \mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2} = \cdot]$, $\mathbb{E}[\mathbf{W}_i \mid \mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2} = \cdot]\eta_{\boldsymbol{\gamma}_0,f}(\cdot)$ and $\eta_{\boldsymbol{\gamma}_0,f}(\cdot)$, as well as the third derivatives of $m_0(\cdot)$, are uniformly continuous and bounded. Moreover, the first derivative of $m_0(\cdot)$ is bounded, and the vector $\nabla_{\boldsymbol{\beta}} \sigma^2(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta}_0)$ is also bounded.

Assumption 6.2. The matrix

$$\Sigma = \mathbb{E}\left[\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}_0, \eta_0)\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}_0, \eta_0)^\top\right]$$

is positive definite.

Assumption 6.3. The Hessian matrix $H_\mu(\boldsymbol{\gamma})$, defined with the weight $\omega_1(\mathbf{V}_i) = \eta^4_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top \boldsymbol{\gamma}_2)$, is positive definite. Moreover, when the model is defined by (6.1)-(6.3), both the Hessian matrices $H_\mu(\boldsymbol{\gamma})$ and $H_\sigma(\boldsymbol{\beta})$ with their corresponding weights $\omega_1(\mathbf{V}_i) = \eta^4_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top \boldsymbol{\gamma}_2)$ and $\omega_2(\mathbf{V}_i) = \eta^2_{\boldsymbol{\gamma},f}(\mathbf{W}_i^\top \boldsymbol{\gamma}_2)$ are positive definite.

Assumption 6.4. The bandwidth h used for the non-parametric part of the estimation is such that $nh^3/\ln n \to \infty$ and $nh^8 \to 0$. The univariate kernel K is symmetric, bounded, integrable, such that $\int_{\mathbb{R}} t^2\{|K(t)| + |tK'(t)|\}dt < \infty$ and $\int_{\mathbb{R}} t^2 K(t)dt \neq 0$. The Fourier Transform of K, denoted by $\mathcal{F}[K]$, satisfies the condition $\sup_{t\in\mathbb{R}} |t|^{c_K}|\mathcal{F}[K](t)| < \infty$ for some $c_K > 3$. Moreover, $t \mapsto |t|^{s/2}\{K(t) + K'(t)\}$ is bounded on $\mathbb{R}$, where s is defined by Assumption 6.1(i).

Assumption 6.1 guarantees suitable rates of uniform convergence for the kernel estimators of the infinite-dimensional parameters gathered in the vector η_γ. More precisely, they imply the conditions used in Theorem 4 of Hansen (2008), with $q = d = 1$. We also use the condition on ξ to apply Davydov's inequality and show that the effect of the nonparametric estimation is negligible and does not alter the pivotalness of the empirical log-likelihood ratio statistic. Due to this purpose, some conditions in Assumption 6.1 are more restrictive than in Theorem 4 of Hansen (2008). Condition (6.19) reveals a link between the existence of some moments of order s and the strength of the dependency given by the coefficient ξ. The more moments for $\mathbf{U}_i$ exist, the stronger the time dependency can be. In particular, if $\mathbf{U}_i$ has finite moments of any order, then $s = \infty$ and thus ξ could be larger but arbitrarily close to 10. There is a wide literature on the mixing properties for time series. The most popular technique for proving this property relies on rewriting the process as a Markov chain and showing the geometrically decay of the

mixing coefficients α_m. For example, ARMA processes were treated in Mokkadem (1988), while some non-linear time series were investigated by Mokkadem (1990), Tjøstheim (1990), Masry and Tjostheim (1995), and more recently by Lu and Jiang (2001), Liebscher (2005), Meitz and Saikkonen (2010). See also the references therein. Another technique has been developed in Fryzlewicz and Subba Rao (2011). They show mixing properties for time-varying ARCH and ARCH(∞) processes by computing explicit bounds for the mixing coefficients using the density function of the processes. Their method could possibly be applied in our context to obtain the conditions of Assumption 6.1. Assumption 6.2 guarantees a non-degenerate limit distribution in the CLT for the sample mean of the $\Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\gamma}_0, \eta_0)$'s. Assumption 6.3 is used to prove Lemma 6.1 and Lemma 6.2. Concerning the bandwidth conditions, one could of course use different bandwidths for the different nonparametric estimators involved. For readability and practical simplicity, we propose a same bandwidth h. Moreover, Assumption 6.4 allows one to use, for instance the Gaussian kernel.

6.4.3 Wilks' Theorem

When the infinite-dimensional parameters $\eta_{\boldsymbol{\gamma}}$ are given and the observations are independent, Theorem 2 of Qin and Lawless (1994) guarantees that the empirical log-likelihood ratio (ELR) statistic $2\ell_n(\theta_0, \eta_0)$ converges in distribution to a $\mathcal{X}^2_{d_\theta - 1}$ as $n \to \infty$ (where d_θ is the dimension of the model parameters). The following theorem states that, under suitable conditions, the chi-squared limit in law is preserved for the ELR defined with our moment conditions for the (CH)PLSIM, with dependent data and estimated $\eta_{\boldsymbol{\gamma}}$. Let us define the ELR statistic

$$W(\boldsymbol{\theta}_0) = 2\ell_n(\boldsymbol{\theta}_0, \widehat{\eta}_{\boldsymbol{\gamma}_0}),$$

where ℓ_n and $\widehat{\eta}_{\boldsymbol{\gamma}_0}$ are respectively given by (6.17) and (6.18). Let $d_\theta = d_\gamma$ for the PLSIM and $d_\theta = d_\gamma + d_\beta$ for the CHPLSIM. In the following $\xrightarrow{d}$ denotes the convergence in distribution.

Theorem 6.1. *Consider that Assumptions 6.1, 6.2, 6.3 and 6.4 hold true. Moreover, condition* (6.7) *is satisfied, as well as condition* (6.5) *in the case of PLSIM or condition* (6.4) *in the case of CHPLSIM. Then,* $W(\theta_0) \xrightarrow{d} \mathcal{X}^2_{d_\theta - 1}$ *as* n *tends to infinity.*

For the proof of Theorem 6.1, we use a central limit theorem for mixing processes that implies that $n^{-1/2} \sum_{i=1}^n \Psi(\mathbf{Z}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\theta}_0, \eta_0)$ converges in distribution to a multivariate centered normal distribution, to deal with the dependency between observations. Moreover, the behavior of the Lagrange multipliers has to be carefully investigated. However, the major difficulty in the proof is to show $\ell_n(\boldsymbol{\theta}_0, \widehat{\eta}_{\boldsymbol{\gamma}_0}) - \ell_n(\boldsymbol{\theta}_0, \eta_0) = o_{\mathbb{P}}(1)$, that is to show that the nonparametric estimation of the nuisance infinite-dimensional parameters does not break the pivotalness of the ELR statistic. This negligibility requirement is a well-known issue, see Remark 2.3 in Hjort, McKeague, and Van Keilegom (2009). See also Chang, Tang, and Wu (2013); Chang, Tang, and Wu (2016); Chang et al. (2020) for a related discussion in the context of high-dimension empirical likelihood inference. However, this type of negligibility, obtained under mild technical conditions, seems to be a new result in the context of semiparametric regression models with weakly dependent data. It is obtained using arguments based on Inverse Fourier Transform and Davydov's inequality in Theorem A.6 of Hall and Heyde (1980). It is also worthwhile to notice that, in order to preserve the chi-squared limit for $W(\boldsymbol{\theta}_0)$, we do not need to follow the general two-step procedure proposed by Bravo, Escanciano, and Van Keilegom (2020) and replace $\Psi(\cdot, \cdot; \cdot, \cdot)$ by some estimated influence function. The reason is given by the gradient $\widetilde{\nabla}_{\boldsymbol{\gamma}} g_\mu(\mathbf{Z}_i; \boldsymbol{\gamma}, m_{\boldsymbol{\gamma}})$ which has the key property $\mathbb{E}[\widetilde{\nabla}_{\boldsymbol{\gamma}} g_\mu(\mathbf{Z}_i; \boldsymbol{\gamma}_0, m_{\boldsymbol{\gamma}_0}) \mid \mathbf{W}_i^\top \boldsymbol{\gamma}_{0,2}] = 0$ a.s.

6.5 Numerical experiments

6.5.1 Simulations

We generated data from model (6.1)-(6.3) with $\varepsilon_i = \sigma(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta})\zeta_i$ and

$$\sigma^2(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta}) = \beta_1 + \beta_2 Y_{i-1}^2,$$

where the ζ_i are independently drawn from a distribution such that $\mathbb{E}(\zeta_i) = 0$ and $\mathrm{Var}(\zeta_i) = 1$. That means, we allow for conditional heteroscedasticity in the mean regression error term. The covariates $\mathbf{X}_i = (Y_{i-1}, Y_{i-2})^\top$ are two lagged values of the target variable Y_i and the covariates $\mathbf{W}_i = (W_{i1}, W_{i2}, W_{i3})^\top$ are generated from a multivariate Gaussian distribution with mean $W_{i-1}/4$ and covariance matrix S defined by $\mathrm{cov}(W_{ik}, W_{i\ell}) = 0.5^{|k-\ell|}$. Thus, the marginal distribution of the index $\mathbf{W}_i^\top \gamma_2$ is a centered Gaussian distribution with variance $(16/15)\boldsymbol{\gamma}_2^\top S \boldsymbol{\gamma}_2$. We set

$$\ell(\mathbf{X}_i; \boldsymbol{\gamma}_1) = \gamma_{11} Y_{i-1} + \gamma_{12} Y_{i-2} \quad \text{and} \quad m(u) = \frac{3}{4}\sin^2(u\pi), \tag{6.22}$$

with $\boldsymbol{\gamma}_1 = (0.1, 0)^\top$, $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$ and $\boldsymbol{\beta} = (0.9, 0.1)^\top$.

Hypothesis testing is based on Wilks' Theorem in Section 6.4.3 (results related to this method are named *estim*), along with the unfeasible EL approach that uses the true density of the index and that previously learns the nonparametric estimators on a sample of size 10^4 (this case mimics the situation where m, m' and the conditional expectations involved in the definition of η_γ are known; results related to this method are named *ref*). The nonparametric elements are estimated by the Nadaraya-Watson method with Gaussian kernel and bandwidth $h = C^{-1}n^{-1/5}$ where C is the standard deviation of the index. In the experiments, we consider four sample sizes (100, 500, 2000 and 5000) and three distributions for ζ_i: a standard Gaussian distribution (*Gaussian*), an uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ (*uniform*) and a mixture of Gaussian distributions (*mixture*) $pN(m_1, v_1^2) + (1-p)N(m_2, v_2^2)$, with $p = 0.5$, $m_2 = -m_1 = 1/\sqrt{6}$, $v_1^2 = 1/6$, $v_2^2 = 3/2$. For each scenario, we generated 5000 data sets.

First, we want to test the order for the lagged values of Y_i in the parametric function ℓ. For this purpose, we use the PLSIM and we consider the following tests:

- *Test Lag(1)* which corresponds to the true order equal to 1, and which is defined by H_0 : $\boldsymbol{\gamma}_1 = (0.1, 0)^\top$ and $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$;
- *Test Lag(0)* which neglects the lagged values of Y_i in the linear part and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0, 0)^\top$ and $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$;
- *Test Lag(2)* which overestimates the order for the lagged values of Y_i and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0.1, 0.1)^\top$ and $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$.

The empirical probabilities of rejection are presented in Table 6.1 for a nominal level of 0.05. A first, not surprising, conclusion: EL inference in such flexible nonlinear models, with dependent data, requires sufficiently large sample sizes. The results with $n = 100$ are quite poor even when $m(\cdot)$ is given, that is in a purely parametric setup. Next, we notice that for the three distributions of the noise, our EL inference approach allows to identify the correct order for the lagged values when the sample size is sufficiently large. Indeed, only *Test Lag(1)* has an asymptotic empirical probability of rejection converging to the nominal level 0.05 while the other tests have a probability of rejection converging to one. Moreover, the differences between the unfeasible EL approach (*ref.* columns) and our approach (*estim.* columns) become quickly negligible. This result was expected because the statistics of both methods converge to the same chi-squared distribution.

Table 6.1: Empirical probabilities of rejection obtained from 5000 replications using the PLSIM for testing the order for the lagged values of Y_i in the parametric part $\ell(\cdot;\boldsymbol{\gamma}_1)$ in (6.22).

Test	ζ_i	$n=100$		$n=500$		$n=1000$		$n=2000$	
		ref.	estim.	ref.	estim.	ref.	estim.	ref.	estim.
Lag(1)	Gaussian	0.167	0.214	0.066	0.075	0.054	0.054	0.055	0.055
	uniform	0.125	0.185	0.058	0.074	0.058	0.056	0.053	0.050
	mixture	0.196	0.229	0.080	0.094	0.063	0.060	0.053	0.051
Lag(0)	Gaussian	0.208	0.243	0.254	0.231	0.705	0.665	0.983	0.980
	uniform	0.160	0.204	0.215	0.207	0.742	0.718	0.991	0.990
	mixture	0.236	0.263	0.241	0.228	0.647	0.619	0.969	0.965
Lag(2)	Gaussian	0.216	0.270	0.266	0.268	0.783	0.760	0.996	0.995
	uniform	0.164	0.227	0.241	0.243	0.775	0.769	0.996	0.997
	mixture	0.263	0.301	0.308	0.299	0.773	0.725	0.993	0.990

We now investigate the order for the lagged values of Y_i in the conditional mean and variance of the noise. Thus, we use the CHPLSIM and we consider the following tests:

- *Test Lag(1)-CH(1)* which corresponds to the true values of the conditional mean and variance and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0.1, 0)^\top$, $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$ and $\boldsymbol{\beta} = (0.9, 0.1)^\top$;
- *Test Lag(0)-CH(1)* which neglects the lagged values of Y_i in the conditional mean and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0, 0)^\top$, $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$ and $\boldsymbol{\beta} = (0.9, 0.1)^\top$;
- *Test Lag(2)-CH(1)* which overestimates the order of the lagged values of Y_i in the conditional mean and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0.1, 0.1)^\top$, $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$ and $\boldsymbol{\beta} = (0.9, 0.1)^\top$;
- *Test Lag(1)-CH(0)* which corresponds to the true value of the conditional mean but neglects the lagged value of Y_i in the conditional variance and which is defined by $H_0 : \boldsymbol{\gamma}_1 = (0.1, 0)^\top$, $\boldsymbol{\gamma}_2 = (1, 1, 1)^\top$ and $\boldsymbol{\beta} = (0.9, 0)^\top$.

The empirical probabilities of rejection are presented in Table 6.2 for a nominal level of 0.05. Again, the true order of the lagged values is detected by the procedure and the differences between the unfeasible EL approach and our approach become quickly negligible. As expected given that the model is more complex, the rate of convergence to the nominal level is slower than for the tests on the PLSIM. However, our procedure allows the conditional heteroscedasticity of the noise to be detected, and meanwhile it identifies the correct order for the lags of Y_i in the mean equation.

6.5.2 Real data analysis

We analyze the data set containing weather (temperature, dew point temperature, relative humidity) and pollution data (PM10 and ozone) for the city of Chicago in the period 1987-2000 from the National Morbidity, Mortality and Air Pollution Study. The analyzed data is freely available in the R package *dlnm* Gasparrini (2011). Lian and Liang (2015) considered the same data set under the assumption of i.i.d. observations.

We use the (CH)PLSIM with a linear function in the parametric part to predict daily mean ozone level ($\widetilde{o3}_i$). For this purpose we use previous daily values of mean ozone level and four other predictors, that are the daily relative humidity ($\widetilde{rhum}_i$), the daily mean temperature (in Celsius

Table 6.2: Empirical probabilities of rejection obtained from 5000 replications using the CH-PLSIM for testing the order of the lagged values of Y_i in the conditional mean and variance.

Test	ζ_i	$n = 100$		$n = 500$		$n = 1000$		$n = 2000$	
		ref.	estim.	ref.	estim.	ref.	estim.	ref.	estim.
Lag(1)	Gaussian	0.292	0.388	0.105	0.111	0.068	0.074	0.074	0.069
CH(1)	uniform	0.167	0.277	0.070	0.077	0.067	0.072	0.084	0.072
	mixture	0.392	0.461	0.151	0.170	0.090	0.098	0.079	0.078
Lag(0)	Gaussian	0.331	0.406	0.260	0.249	0.669	0.641	0.978	0.972
CH(1)	uniform	0.197	0.291	0.198	0.190	0.684	0.675	0.986	0.983
	mixture	0.446	0.493	0.333	0.327	0.653	0.637	0.963	0.958
Lag(2)	Gaussian	0.337	0.426	0.277	0.287	0.743	0.727	0.993	0.992
CH(1)	uniform	0.205	0.304	0.219	0.227	0.724	0.728	0.993	0.993
	mixture	0.438	0.511	0.359	0.352	0.738	0.704	0.990	0.985
Lag(1)	Gaussian	0.289	0.332	0.533	0.523	0.985	0.986	1.000	1.000
CH(0)	uniform	0.283	0.294	0.777	0.748	1.000	1.000	1.000	1.000
	mixture	0.343	0.392	0.489	0.499	0.970	0.970	1.000	1.000

degrees) $\widetilde{temp}_i$, the daily dew point temperature $\widetilde{dptp}_i$ and the daily PM10-level $\widetilde{pm10}_i$. The first step of our analysis was to remove seasonality for each variable we considered in the models. To remove seasonality, we used the function *seasadj* of the R package *forecast* on the data of from year 1994 to year 1997. Thus, we obtain the series $o\widetilde{3}_i$, $rhum_i$, $temp_i$, $dptp_i$ and $pm10_i$ by removing the seasonnality of the series $\widetilde{o3}_i$, $\widetilde{rhum}_i$, $\widetilde{temp}_i$, $\widetilde{dptp}_i$ and $\widetilde{pm10}_i$. Note that the series $temp_i$, $dptp_i$ and $pm10_i$ have been scaled to facilitate the interpretation γ_{12}. Figures S.1-S.5 provided in the Section B.1 of the Supplementary Material of Du Roy de Chaumaray, Marbac, and Patilea (2021) present the original series and the series obtained by removing the seasonality. Thus, all the variables we refer hereafter in this section are deseasonalized. In this application the observations clearly have a time dependency. We split the sample into a learning sample (composed of the observations of years 1994 and 1995) and a testing sample (composed of the observations of years 1996 and 1997). After removing the seasonality, the autocorrelations of $o3$ for the learning and testing samples are 0.469 ($p-$value 0.000) and 0.450 ($p-$value 0.000), respectively; Note that all the covariates have significant autocorrelations (all the $p-$values are 0.000, see Table S.1 in Section B.1 of the Supplementary Material of Du Roy de Chaumaray, Marbac, and Patilea (2021)).

The covariates included in the linear part are the mean relative humidity ($rhum_i$) and the mean ozone level computed on the three previous days ($o3_{i-1}$, $o3_{i-2}$, $o3_{i-3}$). The covariates included in the nonparametric part of the conditional mean are $temp_i$, $dptp_i$ and $pm10_i$. The eigenvalues of the covariance matrix computed on the three variables used in the nonparametric part are 1.995, 0.901 and 0.168 for the data of learning sample, and 1.989, 0.758 and 0.139 for the data of testing sample.

Thus, the equation of the PLSIM is

$$o3_i = \gamma_{11} rhum_i + \gamma_{12} o3_{i-1} + \gamma_{13} o3_{i-2} + \gamma_{14} o3_{i-3} + m(\gamma_{21} temp_i + \gamma_{22} dptp_i + \gamma_{23} pm10_i) + \varepsilon_i. \quad (6.23)$$

We estimate the parameters of the models, on the testing sample, by minimizing the least squares using kernel smoothing (with Gaussian kernel and bandwidth $n^{-1/5}$). Hypothesis testing is conducted on the testing sample. We begin by investigating the order H for the lagged values of the ozone measures to be included in the linear part of the conditional mean. Using PLSIM, we

Table 6.3: Estimators of the parameters obtained by the PLSIM, on the learning sample, with different orders of lagged values, and $p-$values obtained by testing these values on the testing sample for the 'National morbidity and mortality air pollution study' example.

		Lag(0)	Lag(1)	Lag(2)	Lag(3)
$\hat{\gamma}_1$	$rhum_i$	-0.122	-0.157	-0.154	-0.154
	$o3_{(i-1)}$	0.000	0.412	0.459	0.461
	$o3_{(i-2)}$	0.000	0.000	-0.102	-0.116
	$o3_{(i-3)}$	0.000	0.000	0.000	0.025
$\hat{\gamma}_2$	$temp_i$	0.976	0.937	0.941	0.939
	$dptp_i$	-0.215	0.343	0.332	0.336
	$pm10_i$	0.043	0.062	0.066	0.073
	$p-$value	0.000	0.001	0.107	0.044

define different models, called $Lag(H)$ (with $H = 0, 1, 2$ or 3), where only H lagged values of the mean ozone levels are included in the linear part (meaning the coefficients related to the other previous days is zero). The results for different orders H presented in Table 6.3 show that the time dependency cannot be neglected for analyzing these data. It is relevant to include lagged values of the mean ozone level variable to build its daily prediction.

The autocorrelation of the residuals, obtained with the $Lag(2)$ setup, on the testing sample, has a value of 0.035 ($p-$value 0.346). This suggests that $H = 2$ is a reasonable choice. Figure S.6 and Figure S.7, given in Section B.1 of the Supplementary Material of Du Roy de Chaumaray, Marbac, and Patilea (2021), present the estimated density of the index and the estimated function $\hat{m}(\cdot)$, obtained with the $Lag(2)$ setup.

We also calculated the autocorrelation of the squared of the residuals, obtained with the $Lag(2)$ setup, and we obtain the value 0.095 ($p-$value 0.010). This suggests to also investigate the conditional heteroscedasticity of the noise using the CHPLSIM with the $Lag(2)$ setup. For the conditional variance equation we consider

$$\mathbb{E}(\varepsilon_i^2 \mid rhum_i, temp_i, dptp_i, pm10_i, \mathcal{F}_{i-1}) = \beta_1 + \beta_2 \ln\big(\max(o3_{i-1}^2, 1)\big). \tag{6.24}$$

To estimate the parameters of the conditional variance, we use again the learning sample. The estimators for the CHPLSIM with conditional variance as in (6.24) are $\hat{\beta}_1 = 1.553$ and $\hat{\beta}_2 = 3.786$. If we consider constant conditional, we obtain $\tilde{\beta}_1 = 23.816$. The $p-$value obtained by testing the values $\beta_1 = \hat{\beta}_1$ and $\beta_2 = \hat{\beta}_2$ in (6.24) on the testing sample is 0.100. Meanwhile, the $p-$value obtained by testing the values $\beta_1 = \tilde{\beta}_1$ and $\beta_2 = 0$ is 0.020. Thus, we conclude to a non constant conditional variance for the error term in (6.23). This effect should be considered to build forecast confidence intervals.

6.6 Discussion and conclusion

We propose EL inference in a semiparametric mean regression model with strongly mixing data. Our model could include an additional condition on the second order conditional moment of the error term. The regression function has a partially linear single-index form, while for the conditional variance we consider a parametric function. This function could depend on the past values of the observed variables, but it cannot depend directly on the regression error term. A parametric function of the past error terms would break the asymptotic pivotal distribution of the empirical log-likelihood ratio. See Hjort, McKeague, and Van Keilegom (2009) for a description of this common phenomenon in semiparametric models.

We prove Wilks' Theorem under mild technical conditions, in particular without using any trimming and allowing for unbounded series. To obtain this result, first we rewrite the regression model under the form of a fixed number of suitable unconditional moment conditions. These moment conditions include infinite dimensional nuisance parameters estimated by kernel smoothing. Then, we show that estimating the nuisance parameters does not break the asymptotic pivotality of the empirical log-likelihood ratio which behaves asymptotically as if the nuisance parameters were given. Our theoretical result opens the door of the EL inference approach to new applications in nonlinear time series models. We illustrate our result by several simulation experiments and an application to air pollution where assuming time dependency seems reasonable, a fact confirmed by the data.

The models proposed in this paper have several straightforward extensions. First, the variable Y_i could be allowed to be measured with some error. For instance, Y_i could be a function of the error term in a parametric model for some time series (R_i), such as an $AR(1)$ model $R_i = \rho R_{i-1} + u_i$. Taking $Y_i = u_i^2$, (6.1) could be used for inference on the conditional variance of (u_i), while (6.3) could serve to test the value of the kurtosis. This example that could be of interest for financial series is detailed in Section B.3 of the Supplement Du Roy de Chaumaray, Marbac, and Patilea (2021).

Another easy extension is to consider more general conditions than (6.3). Our theoretical arguments apply with practically no change if (6.3) is replaced by one or several conditions like $\mathbb{E}[T(\varepsilon_i) \mid \mathbf{V}_i, \mathcal{F}_{i-1}] = \nu(\mathbf{V}_i, \mathbf{Z}_i^{\{r\}}; \boldsymbol{\beta})$, where the $T(\cdot)$'s are some given twice continuously differentiable functions such that $\mathbb{E}[T'(\varepsilon_i) \mid \mathbf{V}_i, \mathcal{F}_{i-1}] = 0$ a.s., and $\nu(\cdot, \cdot; \cdot)$ is given parametric function. For instance, taking $T(y) = y^4$, we could include a fourth order conditional moment equation in the model, provided $\mathbb{E}[\varepsilon_i^3 \mid \mathbf{V}_i, \mathcal{F}_{i-1}] = 0$ a.s. Such higher-order moment condition could replace or could be added to (6.3).

Finally, one might want to consider some partially linear function, with possibly different index, on the right-had side of (6.3). Lian and Liang (2015) followed a similar idea in the i.i.d. case. While considering several series (Y_i) and equations like (6.1) is a straightforward matter, a semiparametric model for the square of the error term requires some additional effort. We argue that our methodology could be extended to such cases, however the investigation of this extension is left for future work.

Chapter 7

Prospects

Contents

7.1 Estimation for mixture models

7.1.1 Model selection with increasing parameter space and model collection

Context: Considering a family of parametric mixture models, model selection can be performed, under mild assumptions, by information criteria that penalize the log-likelihood and thus provide a consistent estimator of the model (see Keribin (2000)). The main difficulty to state this result is to deal with the model overestimation, due to the lack of identifiability of the model parameters, in such case. This issue is circumvented by the locally conic parametrization (Dacunha-Castelle and Gassiat (1999)) that permits to show the convergence in distribution of the log-likelihood ratio, under mild assumptions. The consistency of the information criteria uses this convergence in distribution and the fact the penalty is an increasing function of the sample size.

Limits: The proof of the consistency of information criteria, provided in Keribin (2000), requires that the parameters are defined on a compact space that is fixed according to the sample size. This assumption can be strong. For instance, considering Gaussian mixture models, it is reasonable to assume that the means of the components are defined on the convex hull of

the observed sample. However, this space increases with the sample size. Moreover the set of competing models is also supposed to be fixed according to the sample size. However, it could be reasonable to allow the upper bound of the number of clusters to increase with the sample size. Moreover, one can consider the case where the number of irrelevant variables increases with the sample size (see Löffler, Wein, and Bandeira (2020)).

Main ideas to explore: In this context, we want to allow the parameter space and the model collection to increase with sample size. To state this result, arguments similar to those used to prove Theorem 2.1 in Section 2.3 will be used. Assumptions made on the rate of increasing of the parameter space would allow us to control the underestimation with bracketing. Moreover, assumptions made on the penalty would allow an upper bound of the probability of overestimating the model to be obtained for a fixed sample size. This task will be achieved by controlling the concentration results of the Gaussian process (Dudley (2014)) and the Gaussian approximation of suprema of empirical processes (Chernozhukov, Chetverikov, and Kato (2014)). The case with an increasing model collection will be treated with the same arguments, by also considering concentration results for the case of underestimation.

7.1.2 Model selection for location scale mixture model

Context: To avoid the bias of the parametric assumptions, semi-parametric mixture models can be considered. In this context, an important family is the family of location-scale mixture models because it generalizes some well-known mixture models, including the Gaussian and the Student mixtures. The mixture components are assumed to be symmetric and to come from the same location-scale family. We refer to these mixtures as semi-parametric because no additional assumptions other than symmetry are made regarding the parametric form of the component distributions (see Hunter, Wang, and Hettmansperger (2007)).

Limits: The estimation of a location-scale mixture model can be achieved by maximizing the smoothed-log-likelihood via an MM algorithm. However, there is a lack of theoretical results on the resulting estimator (consistency, guarantee of convergence to the global optimum). Moreover, there is no procedure for model selection (nor for selecting the number of components or the subset of discriminative variables).

Main ideas to explore: First, we would like to obtain some guarantees of the MM algorithm for location-scale mixture model: *i.e.,* provide conditions that allow for a characterization of the region of convergence of MM algorithm iterates and to define how quickly MM algorithm iterates converge to a small neighborhood of a given global optimum. These results have been recently stated, by Balakrishnan, Wainwright, and Yu (2017), in the case of Gaussian mixtures by following the iterations of the EM algorithm. We would like to extend this work to the semi-parametric mixture estimated by an MM algorithm. This would help to control the accuracy of the kernel density estimator. Theoretical developments for obtaining such a control could be made by obtaining an upper bound of the misclassified observations during the algorithm iterations (similarly to Lu and Zhou (2016)). If a control of the accuracy of the density estimators is stated, then a full model selection (selecting the number of components and the subset of discriminative variables) could be achieved via information criteria by including the uncertainty of the density estimators in a proof similar to Keribin (2000) or those of Theorem 2.1 in Section 2.3.

7.1.3 Non-asymptotic procedure for local false discovery rate estimation

Context: In this work, we will consider the multiple testing problem, where the asymptotic distribution of the test statistic is known while the distribution under the alternative is unknown. We consider n independent test statistics $X_{1,m}, \ldots, X_{n,m}$ where $X_{i,m}$ is the statistic of test i computed over a sample of size m. Thus, under the null hypothesis, $X_{i,m}$ converges in distribution to a known distribution when m tends to infinite. The multiple testing problem can be addressed by considering that, under the null hypothesis, the test statistic follows its asymptotic distribution (Robin et al. (2007) and Patra and Sen (2016)). Thus, the distribution of the test statistic follows a mixture model defined by the density

$$g(x) = (1 - \pi) f_0(x) + \pi \phi(x), \tag{7.1}$$

where the density f_0 is known (*e.g.,* chi-squared distribution for the likelihood ratio test), $0 < \pi < 1$ is the proportion of test statistics drawn under the alternative hypothesis and ϕ is unknown.

Limits: In this work, we will consider the case where the distribution of the test statistic is known only asymptotically. Thus, for a fixed sample size, the distribution of the test statistic is not exactly known meaning that (7.1) is not suitable for assessing the local false discovery rate.

Main ideas to explore: To circumvent this issue, we would try to extend (7.1) by considering that the distribution under the null hypothesis is *close* to the asymptotic distribution. This proximity is measured by constraining the first moments to be close to those of the asymptotic distribution. The distribution of the test statistics could be defined by the density

$$g(x) = (1 - \pi) \psi_m(x) + \pi \phi(x),$$

where the first r_m moments between the distributions defined by ψ_m and f_0 are such that

$$\forall q = 1, \ldots, r_m, \ |\mathbb{E}_{\psi_m}[X_{1,m}^q] - \mathbb{E}_{f_0}[X_{1,m}^q]| < \varepsilon_m,$$

where ε_m tends to 0 and r_m tends to infinity when m tends to infinity ensuring the convergence in distribution of $X_{i,m}$, under the null hypothesis, to the distribution defined by f_0. As a first step, we will focus on non-parametric tests (*e.g.,* Mann-Whitney) where the first moments have a closed form under the null hypothesis for a fixed sample size m. This setup allows us to fix $\varepsilon_m = 0$, for any m such that the moment m has a closed form under the null hypothesis. The identifiability of the resulting mixture model will be obtain by adding constraints on ϕ. The estimation of the resulting bi-component mixture model will be achieved via an MM algorithm where the non-parametric estimation of $\psi_m(x)$ and ϕ will be made via kernel density estimation with constraints on the moments (Hall and Presnell (1999) and Racine, Parmeter, and Du (2009)).

7.2 Developments for biostatistics and epidemiology

7.2.1 Model-based clustering of longitudinal data with non-ignorable missingness

Context: In epidemiology, many cohorts are composed of n subjects described by a longitudinal continuous variable measured at T time moments. Due to attrition of the subjects during the study, some realizations of the longitudinal variable are unobserved. The missingness mechanism

has a monotone pattern and is allowed to be non-ignorable. Thus, if a variable is not observed on a subject at time t, then this variable is no longer observed for any time $t' \in [\![t, T]\!]$. Moreover, the probability, for a variable, to not be observed is allowed to depend on the values of the variable itself and the subpopulation membership. Each subject i is described by a vector of three variables $(\mathbf{X}_i^\top, \boldsymbol{R}_i^\top, \mathbf{Z}_i^\top)^\top$ where $\mathbf{X}_i = (X_{i(1)}, \ldots, X_{i(T)})^\top \in \mathbb{R}^T$, $X_{i(t)}$ is the variable measured on subject i at time t, $\boldsymbol{R}_i = (R_{i(1)}, \ldots, R_{i(T)})^\top \in \{0,1\}^T$ indicates whether $X_{i(t)}$ is observed ($R_{i(t)} = 1$) and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})^\top$ indicates the subpopulation of subject i. The monotone pattern of the missingness mechanisms implies that if for some $t \in [\![1, T]\!]$, $R_{i(t)} = 0$ then $R_{i(t')} = 0$ for any $t' \in [\![t, T]\!]$.

Limits: The variable $D_i = \sum_{t=1}^T R_{i(t)} \geq 1$ indicating the number of elements that are observed in $\mathbf{X}_i$ contains all the information of $\boldsymbol{R}_i$. The nonparametric model described in Chapter 3 is not suitable for a monotone pattern due to its assumption of conditional independence between the elements of $\boldsymbol{R}_i$.

Main ideas to explore: In this work, we will consider that the pair $(\mathbf{X}_i, \boldsymbol{R}_i)$ arises from a K-component mixture models whose pdf of component k is decomposed as follows

$$g_k(\boldsymbol{x}_i, \boldsymbol{r}_i) = g_{k,1}(x_{i(1)}) \prod_{t=2}^{T} g_{k,t}(x_{i(t)}, r_{i(t)} \mid x_{i(1:t-1)}, r_{i(1:t-1)}), \tag{7.2}$$

where $x_{i(1:t)} = (x_{i(1)}, \ldots, x_{i(t)})^\top$ and $r_{i(1:t)} = (r_{i(1)}, \ldots, r_{i(t)})^\top$. The conditional distribution of the observed values $(X_{i(t)}, R_{i(t)})^\top$ given the past and the cluster membership of subject i can be defined by the pattern-mixture model approach leading that

$$g_{k,t}(x_{i(t)}, r_{i(t)} \mid x_{i(t-1)}, r_{i(t-1)}) = g_{k,t}(r_{i(t)} \mid x_{i(1:t-1)}, r_{i(1:t-1)}) g_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}, r_{i(1:t)}).$$

The monotone pattern of the missingness mechanism implies that $R_{i(t)} = 0$ if $t > D_i$ and

$$g_{k,t}(r_{i(t)} \mid x_{i(1:t-1)}, r_{i(1:t-1)}) = g_{k,t}(r_{i(t)} \mid x_{i(1:t-1)}, r_{i(t-1)}),$$

and

$$g_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}, r_{i(1:t)}) = g_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}, r_{i(t)}, r_{i(t-1)}).$$

Indeed, given the past and the cluster membership of subject i, $R_{i(t)}$ follows a Bernoulli distribution with parameter $\tau_{kt}(x_{i(1:t-1)})$, if $t \leq D_i$, where τ_{kt} is an unknown function, while $R_{i(t)}$ is no longer a random variable if $R_{i(t-1)} = 0$. Thus, the pdf of the conditional distribution of $R_{i(t)}$ given the past and the cluster membership of subject i, is defined by

$$\begin{aligned} g_{k,t}(r_{i(t)} \mid x_{i(1:t-1)}, r_{i(t-1)}) = \\ [\tau_{k,t}(x_{i(1:t-1)})]^{r_{i(t)} r_{i(t-1)}} [1 - \tau_{k,t}(x_{i(1:t-1)})]^{(1-r_{i(t)}) r_{i(t-1)}} [1 - r_{i(t)}]^{1 - r_{i(t-1)}}. \end{aligned}$$

Moreover, the pdf of $X_{i(t)}$ given the past and $R_{i(t)}$ is defined by

$$\begin{aligned} g_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}, r_{i(t)}, r_{i(t-1)}) = p_{k,t}^{r_{i(t)} r_{i(t-1)}}(x_{i(t)} \mid x_{i(1:t-1)}) \times \\ \left[q_{k,t}^{r_{i(t-1)}}(x_{i(t)} \mid x_{i(1:t-1)}) \tilde{q}_{k,t}^{1 - r_{i(t-1)}}(x_{i(t)} \mid x_{i(1:t-1)}) \right]^{1 - r_{i(t)}}, \end{aligned}$$

where $p_{k,t}$ ($q_{k,t}$ and $\tilde{q}_{k,t}$, respectively) are the pdf of the conditional distribution of $X_{i(t)}$ given $\mathbf{X}_{i(1:t-1)}$ and $(R_{i(t)}, R_{i(t-1)}) = (1,1)$ ($(R_{i(t)}, R_{i(t-1)}) = (0,1)$ and $(R_{i(t)}, R_{i(t-1)}) = (0,0)$, respectively). The pdf of the observed variables under component k, denoted by $g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i)$, is obtained by integrating the pdf of component k over the missing variables $\boldsymbol{X}_i^{\text{miss}}$, which leads to

$$g(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i) \text{ with } g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \int g_k(\boldsymbol{x}_i, \boldsymbol{r}_i) d\boldsymbol{x}_i^{\text{miss}}.$$

Using the fact that the missingness mechanism has monotone pattern, we have

$$g_k(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = p_{k,1}(x_{i(1)}) \left[\prod_{t=2}^{D_i} \tau_{k,t}(x_{i(1:t-1)}) p_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}) \right]^{r_{i(2)}} [1 - \tau_{k,D_i+1}(x_{i(1:D_i)})].$$

where $\boldsymbol{\theta}$ groups the proportions $\pi_1, \ldots, \pi_K$ and the functions $\tau_{k,t}$ and $p_{k,t}$. To facilitate the estimation of the functions $\tau_{k,t}$ and $p_{k,t}$, we assume a Markov dependence within the components such that the conditional distribution of $(X_{i(t)}, R_{i(t)})$ given the past and the cluster membership of subject i, is equal to the conditional distribution of $(X_{i(t)}, R_{i(t)})$ given $(X_{i(t-1)}, R_{i(t-1)}, \mathbf{Z}_i)$. Therefore, we have

$$\tau_{k,t}(x_{i(1:t-1)}) := \tau_{k,t}(x_{i(t-1)}) \text{ and } p_{k,t}(x_{i(t)} \mid x_{i(1:t-1)}) := \frac{f_{k,t,2}(x_{i(t)}, x_{i(t-1)})}{f_{k,t-1,1}(x_{i(t-1)})},$$

where $f_{k,t,1}$ is the marginal pdf of $X_{i(t)}$ given $R_{i(t)} = 1$ for component k and $f_{k,t,2}$ is the marginal pdf of $(X_{i(t)}, X_{i(t-1)})^\top$ given $R_{i(t)} = 1$ for component k.

7.2.2 Spatial scan statistics

Context: The CoVid-19 pandemic highlighted the need for reliable and responsive public health tools, designed to identify clusters of cases on a fine scale over large geographical areas. In the field of spatial epidemiology, spatial scan statistics (Costa and Kulldorff (2009)) can meet this need for identifying spatial clusters while adjusting for potential confounding factors at the individual or ecological level. The spatial scan statistics (Kulldorff (1997) and Kulldorff (2006)) can be viewed as an extension of bi-dimensional scan statistics (Naus (1965)) to spatial data. Originally, they permit the estimation of a single cluster (named the most likely cluster) and the test of its significance. A cluster is defined by a group of sites where the distribution of the target variable is different to its distribution outside this group. Thus, investigating the significance of the most likely cluster can be viewed as an extension of the homogeneity test with spatial constraints. The difference between the distributions inside and outside the cluster is often summarized by a difference in the conditional mean of the target variable given the location of the sites and potentially other covariates. Thus, different regression models can be used depending on the nature of the target variables (Huang et al. (2009), Jung (2009), Jung, Kulldorff, and Richard (2010), Huang, Kulldorff, and Gregorio (2007), Bhatt and Tiwari (2014) and Zhang and Lin (2009)). The spatial constraints are defined by the assumptions made on the cluster shapes. Irregular shapes (Assuncao et al. (2006), Duczmal, Kulldorff, and Huang (2006) and Duczmal et al. (2007)) that are considered by considering that a cluster is a subset of connected areas. Alternatively, clusters with parametric shapes can be considered. Among them, one can cite the spherical (Kulldorff (1997)), elliptic (Kulldorff et al. (2006)) or rectangular (Walther (2010)) clusters. Thus, by imposing shape constraints, the detection of the most likely cluster can be performed by an exhaustive approach. Testing the significance of the most likely

cluster is often performed by using the quasi likelihood ratio test (McCullagh (1983) and Chiou and Müller (1999)). However, except for very specific cases (Zhang and Lin (2013) and Sharpnack and Arias-Castro (2016)), this statistic does not have an explicit asymptotic distribution. The distribution of the scan statistics under the null hypothesis is also assessed by Monte-Carlo re-sampling (Dwass (1957)). The procedure is consistent and permits detecting the alternative (*i.e.,* existence of one single cluster with the assumptions made on its shape; see Zhang and Lin (2017)). Moreover, it requires independence between the observations (see Loh and Zhu (2007) for the case of dependent data).

Limits: In many applications, more than one cluster can appear. When the most-likely cluster is significant, one can investigate the secondary clusters whose p-values of their significance tests are also computed with an exhaustive approach for the cluster detection. However, this approach is too conservative (Kulldorff et al. (1997)). Thus, Zhang, Assunção, and Kulldorff (2010) propose running an other cluster detection by keeping the most-likely cluster fixed. As we will show, this approach cannot allow us to detect all the alternatives because it keeps fixed the most likely cluster, during the detection of the secondary clusters. However, this cluster can contain two smaller clusters and some other sites not belonging to any clusters. Note that this phenomenon is often observed since spatial scan statistics are known to provide large clusters (that can be difficult to use for the purpose of medical prevention). An alternative consists of estimating simultaneously K clusters and testing the significance of the K clusters (Li et al. (2011)). However, because cluster detection is performed with an exhaustive search that has a complexity of $O(n^{2K})$, for circular clusters and n sites, this approach in not doable in practice.

Main ideas to explore: In this work, we will develop a new method for multiple spatial cluster detection with scan statistics. We want to consider clusters with parametric shape that includes the standard shapes of clusters (*e.g.,* spherical, elliptic and rectangular clusters). We will develop a non-exhaustive cluster detection that uses the parametric definition of the cluster shape to provide a smoothing of the objective function. This would facilitate parameter estimation and thus cluster detection. We want to show that this approximation is negligible, under mild assumptions. The regression model could be general and includes all the generalized linear models. The scan statistics are defined by the quasi likelihood ratio and the distribution under the null hypothesis is assessed by a Monte-Carlo re-sampling. The approach would permit the detection of K clusters and the investigation of the significance of all these clusters. Contrary to Zhang, Assunção, and Kulldorff (2010), the proposed approach would allow us to detect all the alternative hypotheses because the estimation of the K clusters does not use information from the previously estimated $K-1$ clusters.

Limits: One major concern of spatial scan statistics is investigating the significance of the detected clusters. This aim is generally achieved by the statistical test defined by the null hypothesis assuming spatial homogeneity and the alternative hypothesis claiming that there is at least one spatial cluster. However, the test can be extended to investigate the significance of multiple clusters (Lin, Kung, and Clayton (2016)). The test of significance of the clusters is generally conducted by considering the likelihood-ratio test (LRT) that easily allows for the adjustment on covariates or the analysis of aggregated data (Huang et al. (2009)). The asymptotic distribution of the LRT does not generally have a closed form, under the null hypothesis, except in very specific cases (Walther (2010)). Thus, it is generally estimated by Monte Carlo generations which is a major issue for spatial scans because they dramatically increase the computation time and thus limit the size of the data to be analyzed.

Main ideas to explore: The main objective of this work is the development of methods leading to explicit distribution of LRT for spatial scan statistics, in order to avoid the computational limits due to the use of the Monte-Carlo procedure. To achieve this aim, we rewrite the problem of scan statistics as a specific mixture model whose proportions depend on the spatial locations with a parametric link function. Thus, when the null hypothesis holds true, there is a loss of identifiability of the parameters of the mixture models. This phenomenon is well-known for mixture models when the number of components is overestimated and implies that the LRT does not asymptotically follow a chi-squared distribution under the null hypothesis. Using the locally conic parametrization (Dacunha-Castelle and Gassiat (1999)), under mild assumption, we can obtain the asymptotic distribution of the likelihood ratio test, under mild assumptions. We propose extending the locally conic parametrization of mixture model to the case where the mixture proportions depend on the spatial locations with a parametric link function. Moreover, we propose a control of the type-1 error of the procedure, for a fixed sample size, by controlling the concentration results of the Gaussian process (Dudley (2014)) and the Gaussian approximation of suprema of empirical processes (Chernozhukov, Chetverikov, and Kato (2014)). Thus, the procedure will be efficient for small samples (by controlling, in probability, the difference between the fixed-sample size distribution and the asymptotic distribution) and for large samples (by avoiding the multiple Monte-Carlo generations). The relevance of this procedure would depend on the accuracy of the different upper-bounds.

7.2.3 Empirical likelihood with missing values for investigating the link between physical activity and chronic diseases

Context: We aim to investigate the link between physical activity and chronic diseases (especially being overweight reflected by a large value of body mass index; BMI). Hence, we consider the ELFE cohort composed of 3707 subjects. Each subject is described by different families of variables collected during interviews: including BMI, food habits, environmental variables, sleep quality and physical activity. Because physical activity is not subject to bias in the response, a section of the subjects wear an accelerometer for one week. However, for technical reasons, only 159 subjects worn the accelerometer. The challenge is to investigate the significance of the link between physical activity and being overweight. The whole cohort should be considered and thus the surrogate variable related to the physical activity measured over the full cohort.

Limits: To the best of our knowledge, there is no empirical likelihood method that allows considering surrogate variables in a semi-parametric model and that provides a chi-squared distribution for the asymptotic distribution of the empirical log-likelihood ratio.

Main ideas to explore: We aim at modeling and making inference for the one-dimensional variable Y_i given the $d_V + d_W + 1$-dimensional variable $X_i = (U_i, V_i^\top, W_i^\top)^\top$ where U_i is a scalar variable, V_i is a d_V-dimensional variable and W_i is a d_W-dimensional variable. Moreover, we consider a d_S-dimensional surrogate variable S_i that provides information on V_i. We consider the regression model defined by

$$\left\{ \begin{array}{ll} Y_i & = U_i\alpha + V_i^\top\beta + m(W_i^\top\gamma) + \varepsilon_i \\ U_i & = S_i^\top\delta + \xi_i \end{array} \right. , \tag{7.3}$$

such that ε_i and $(U_i, S_i^\top)^\top$ are conditionally independent given $(V_i^\top, W_i^\top)^\top$, ξ_i and $(V_i^\top, W_i)^\top$ are conditionally independent given S_i, ξ_i and ε_i are conditionally independent given $(S_i^\top, V_i^\top, W_i^\top)^\top$

with the conditional first order moments

$$\mathbb{E}[\varepsilon_i \mid U_i] = 0 \text{ and } \mathbb{E}[\xi_i \mid S_i] = 0, \tag{7.4}$$

and the conditional second order moments

$$\mathbb{E}[\varepsilon_i^2 \mid V_i, W_i] = \sigma^2(V_i, W_i; \rho),\ \mathbb{E}[\xi_i^2 \mid S_i] = \varsigma^2(S_i; \varrho) \text{ and } \mathbb{E}[\xi_i \varepsilon_i \mid V_i, W_i, S_i] = 0, \tag{7.5}$$

where $\theta = (\vartheta^\top, \rho^\top, \varrho^\top, \delta^\top)$ groups all the model parameters, $\vartheta = (\alpha, \beta^\top, \gamma^\top)^\top$ and m is an infinite dimensional parameter. Our interest focuses on θ and m is considered as a nuisance parameter. The model defined by (7.3)-(7.4) is a *Partially Linear Single-Index Regression Model* (PLSIRM). The model defined by (7.3)-(7.5) is a *Conditionnal Heteroscedastic Partially Linear Single-Index Regression Model* (CHPLSIRM) where the values of functions σ^2 and ς^2 are known.

Let $Z_i = (S_i^\top, U_i, V_i^\top, W_i^\top, Y_i)^\top$, the PLSIRM model (7.3)-(7.4) is defined by the conditional moment equations

$$\mathbb{E}[f_{\text{PLSIRM}}(Z_i; \vartheta, m, \delta) \mid X_i, S_i] = 0 \Longleftrightarrow \vartheta = \vartheta_0,\ m = m_0 \text{ and } \delta = \delta_0, \tag{7.6}$$

where $f_{\text{PLSIRM}}(Z_i; \vartheta, m, \delta) = (f_1(Z_i; \vartheta, m)^\top, f_2(Z_i; \delta)^\top)^\top$ such that

$$f_1(Z_i; \vartheta, m) = Y_i - U_i\alpha - V_i^\top \beta - m(W_i^\top \gamma),$$

and

$$f_2(Z_i; \delta) = U_i - S_i^\top \delta.$$

Similarly, the CHPLSIRM model (7.3)-(7.5) is defined by the conditional moment equations

$$\mathbb{E}[f_{\text{CHPLSIRM}}(Z_i; \theta, m) \mid X_i, S_i] = 0 \Longleftrightarrow \theta = \theta_0 \text{ and } m = m_0, \tag{7.7}$$

where $f_{\text{CHPLSIRM}}(Z_i; \theta, m) = (f_{\text{PLSIRM}}(Z_i; \vartheta, m, \delta)^\top, f_3(Z_i; \vartheta, m, \rho)^\top, f_4(Z_i; \delta, \varrho)^\top)^\top$ such that

$$f_3(Z_i; \vartheta, m, \rho) = f_1^2(Z_i; \vartheta, m) - \sigma^2(V_i, W_i; \rho),$$

and

$$f_4(Z_i; \delta, \varrho) = f_2^2(Z_i; \delta) - \varsigma^2(S_i; \varrho).$$

Bibliography

A., Emmeke (2019). *mHMMbayes: Multilevel Hidden Markov Models Using Bayesian Estimation*. R package version 0.1.1.

Ae Lee, J. and J. Gill (2018). "Missing value imputation for physical activity data measured by accelerometer". *Statistical Methods in Medical Research* 27(2), pp. 490–506.

Akaike, H. (1970). "Statistical predictor identification". *Annals of the Institute of Statistical Mathematics* 22(1), pp. 203–217.

Alexander, D. H., J. Novembre, and K. Lange (Sept. 2009). "Fast model-based estimation of ancestry in unrelated individuals". *Genome Research* 19 (9). DOI: 10.1101/gr.094052.109.

Allman, E.S., C. Matias, and J.A. Rhodes (2009). "Identifiability of parameters in latent structure models with many observed variables". *The Annals of Statistics* 37(6A), pp. 3099–3132.

Altman, R. M. (2007). "Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting". *Journal of the American Statistical Association* 102(477), pp. 201–210.

Ando, T. and J. Bai (2016). "Panel data models with grouped factor structure under unknown group membership". *Journal of Applied Econometrics* 31(1), pp. 163–191.

Annam, J. R., S. S. Mittapalli, and R. S. Bapi (2011). "Time series Clustering and Analysis of ECG heart-beats using Dynamic Time Warping". *2011 Annual IEEE India Conference*, pp. 1–3.

Antoniadis, A. et al. (2013). "Clustering functional data using wavelets". *International Journal of Wavelets, Multiresolution and Information Processing* 11, p. 1350003.

Arthur, D. and S. Vassilvitskii (2006). *k-means++: The advantages of careful seeding*. Tech. rep. Stanford.

Assuncao, R. et al. (2006). "Fast detection of arbitrarily shaped disease clusters". *Statistics in medicine* 25(5), pp. 723–742.

Auray, S., N. Klutchnikoff, and L. Rouviere (2015). "On clustering procedures and nonparametric mixture estimation". *Electronic journal of statistics* 9(1), pp. 266–297.

Azizyan, M., A. Singh, and L. Wasserman (2013). "Minimax theory for high-dimensional gaussian mixtures with sparse mean separation". *Neural Information Processing Systems, NIPS*.

Badr, H. S. et al. (2020). *Unified COVID-19 Dataset*.

Bai, J.i et al. (2018). "A two-stage model for wearable device data". *Biometrics* 74(2), pp. 744–752.

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). "Statistical guarantees for the EM algorithm: From population to sample-based analysis". *The Annals of Statistics* 45(1), pp. 77–120.

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

Bartolucci, F., A. Farcomeni, and F. Pennoni (2012). *Latent Markov models for longitudinal data*. CRC Press.

Bartolucci, F., S. Pandolfi, and F. Pennoni (2017). "LMest: an R package for latent Markov models for longitudinal categorical data". *Journal of Statistical Software* 81(4), pp. 1–38.

Bartolucci, F., F. Pennoni, and G. Vittadini (2011). "Assessment of school performance through a multilevel latent Markov Rasch model". *Journal of Educational and Behavioral Statistics* 36(4), pp. 491–522.

Baudry, J.-P. (2015). "Estimation and model selection for model-based clustering with the conditional classification likelihood". *Electronic journal of statistics* 9(1), pp. 1041–1077.

Baudry, Jean-Patrick et al. (2010). "Combining mixture components for clustering". *Journal of computational and graphical statistics* 19(2), pp. 332–353.

Benaglia, T., D. Chauveau, and D. R. Hunter (2009). "An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures". *Journal of Computational and Graphical Statistics* 18, pp. 505–526.

Benaglia, T. et al. (2009b). *mixtools: An R package for analyzing finite mixture models.*

Benaglia, T. et al. (2009a). "mixtools: An R Package for Analyzing Finite Mixture Models". *Journal of Statistical Software* 32(6), pp. 1–29.

Berney, M., M. Burnier, and G. Wuerzner (2018). "Isolated diastolic hypertension: do we still have to care about it?" *Revue medicale suisse* 14(618), pp. 1607–1610.

Bertrand, A. et al. (2017). "Robustness of estimation methods in a survival cure model with mismeasured covariates". *Computational Statistics & Data Analysis* 113, pp. 3–18.

Bhatt, V. and N. Tiwari (2014). "A spatial scan statistic for survival data based on Weibull distribution". *Statistics in medicine* 33(11), pp. 1867–1876.

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

Biernacki, C., M. Marbac, and V. Vandewalle (2019). *ClusVis: Gaussian-Based Visualization of Gaussian and Non-Gaussian Model-Based Clustering.* R package version 1.2.0. URL: `https://CRAN.R-project.org/package=ClusVis`.

Biernacki, C., M. Marbac, and V. Vandewalle (2021). "Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering". *Journal of Classification* 38, pp. 129–157. URL: `https://link.springer.com/article/10.1007/s00357-020-09369-y`.

Biernacki, C. et al. (2021). *Model-based Clustering with Missing Not At Random Data.* URL: `https://arxiv.org/abs/2112.10425`.

Biernacki, C. and Celeux, G. and Govaert, G. (2010). "Exact and Monte Carlo calculations of integrated likelihoods for the latent class model". *Journal of Statistical Planning and Inference* 140(11), pp. 2991–3002.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a). "Estimating multivariate latent-structure models". *The Annals of Statistics* 44(2), pp. 540–563.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016b). "Non-parametric estimation of finite mixtures from repeated measurements". *Journal of Royal Statistical Society: Series B*, pp. 211–229.

Booth, J. G. and J. P. Hobert (1999). "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm". *Journal of the Royal Statistical Society: Series B* 61(1), pp. 265–285.

Bossaerts, P., C. Hafner, and W. Härdle (1996). "A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series". *Finanzmarktanalyse und -prognose mit innovativen quantitativen Verfahren: Ergebnisse des 5. Karlsruher Ökonometrie-Workshops.* Physica-Verlag HD: Heidelberg, pp. 71–83.

Bouveyron, C. (2015). *funFEM: clustering in the discriminative functional subspace.*

Bouveyron, C., E. Côme, and J. Jacques (2015). "The discriminative functional mixture model for a comparative analysis of bike sharing systems". *The Annals of Applied Statistics* 9, pp. 1726–1760.

Bouveyron, C. et al. (2019). *Model-based clustering and classification for data science: with applications in R*. Vol. 50. Cambridge University Press.

Bradley, R. C. (2005). "Basic properties of strong mixing conditions. A survey and some open questions". *Probability Surveys* 2. Update of, and a supplement to, the 1986 original, pp. 107–144.

Brault, V. and M. Mariadassou (2015). "Co-clustering through latent bloc model: A review". *Journal de la Société Française de Statistique* 156(3), pp. 120–139.

Bravo, F., J.-C. Escanciano, and I. Van Keilegom (Feb. 2020). "Two-step semiparametric empirical likelihood inference". *The Annals of Statistics* 48(1), pp. 1–26.

Butucea, C. and P. Vandekerkhove (2014). "Semiparametric mixtures of symmetric distributions". *Scandinavian Journal of Statistics* 41(1), pp. 227–239.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, pp. xviii+652. ISBN: 978-0387-40264-2; 0-387-40264-0.

Carroll, R. J. and M. P. Wand (1991). "Semiparametric estimation in logistic measurement error models". *Journal of the Royal Statistical Society: Series B* 53(3), pp. 573–585.

Carroll, R. J. et al. (1997). "Generalized partially linear single-index models". *Journal of the American Statistical Association* 92(438), pp. 477–489. ISSN: 0162-1459.

Celeux, G., D. Chauveau, and J. Diebolt (1996). "Stochastic versions of the EM algorithm: an experimental study in the mixture case". *Journal of statistical computation and simulation* 55(4), pp. 287–314.

Celeux, G. and G. Govaert (1991). "Clustering criteria for discrete data and latent class models". *Journal of classification* 8(2), pp. 157–176.

Celeux, G. and G. Govaert (1995). "Gaussian parsimonious clustering models". *Pattern recognition* 28(5), pp. 781–793.

Celisse, A., J.-J. Daudin, and L. Pierre (2012). "Consistency of maximum-likelihood and variational estimators in the stochastic block model". *Electronic Journal of Statistics* 6, pp. 1847–1899.

Chambaz, A. (2006). "Testing the order of a model". *Annals of statistics* 34(3), pp. 1166–1203.

Chan, K. S. and J. Ledolter (1995). "Monte Carlo EM estimation for time series models involving counts". *Journal of the American Statistical Association* 90(429), pp. 242–252.

Chang, C. et al. (Dec. 2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4 (1).

Chang, J., S. X. Chen, and X. Chen (2015). "High dimensional generalized empirical likelihood for moment restrictions with dependent data". *Journal of Econometrics* 185(1), pp. 283 –304. ISSN: 0304-4076.

Chang, J., C. Y. Tang, and Y. Wu (Apr. 2016). "Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood". *The Annals of Statistics* 44(2), pp. 515–539.

Chang, J., C. Y. Tang, and Y. Wu (Aug. 2013). "Marginal empirical likelihood and sure independence feature screening". *The Annals of Statistics* 41(4), pp. 2123–2148.

Chang, J. et al. (Oct. 2020). "High-dimensional empirical likelihood inference". *Biometrika* 108(1), pp. 127–147. ISSN: 0006-3444.

Charlier, I., D. Paindaveine, and J. Saracco (2015). "Conditional quantile estimation through optimal quantization". *Journal of Statistical Planning and Inference* 156, pp. 14 –30. ISSN: 0378-3758.

Chauveau, D., D. R. Hunter, and M. Levine (2015). "Semi-parametric estimation for conditional independence multivariate finite mixture models". *Statistics Surveys* 9, pp. 1–31.

Cheam, A.M.S., M. Marbac, and P.D. McNicholas (2020). *SpaTimeClus: Model-Based Clustering of Spatio-Temporal Data*. R package version 1.0.1. URL: `https://CRAN.R-project.org/package=SpaTimeClus`.

Cheam, A.M.S. et al. (2020). "Translation-invariant functional clustering on COVID-19 deaths adjusted on population risk factors". URL: `https://arxiv.org/abs/2012.10629`.

Cheam, A.S.M and M. Fredette (2020). "On the importance of similarity characteristics of curve clustering and its applications". *Pattern Recognition Letters* 135, pp. 360 –367. ISSN: 0167-8655.

Cheam, A.S.M., M. Marbac, and P.D. McNicholas (2017). "Model-based clustering for spatiotemporal data on air quality monitoring". *Environmetrics* 28(3), e2437. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2437`.

Chen, J., J. Yan, and P. Zhang (2020). *Clustering US states by time series of COVID-19 new case counts with non-negative matrix factorization*.

Chen, S. X. and I. Van Keilegom (2009). "A review on empirical likelihood methods for regression". *TEST* 18(3), pp. 415–447. ISSN: 1133-0686.

Chen, X. and H. Cui (2008). "Empirical likelihood inference for partial linear models under martingale difference sequence". *Statistics & Probability Letters* 78(17), pp. 2895–2901.

Cheng, C. et al. (2020). "COVID-19 government response event dataset (CoronaNet v.1.0)". *Nature Human Behaviour* 4, pp. 756–768.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). "Gaussian approximation of suprema of empirical processes". *The Annals of Statistics* 42(4), pp. 1564–1597.

Chi, J. T. and E. C. Chi (2014). *kpodclustr: An R package for clustering partially observed data*. version 1.0.

Chi, J. T., E. C. Chi, and R. G. Baraniuk (2016). "k-pod: A method for k-means clustering of missing data". *The American Statistician* 70(1), pp. 91–99.

Chiou, J.-M. and H.-G. Müller (1999). "Nonparametric quasi-likelihood". *The Annals of Statistics* 27(1), pp. 36–64.

Cho, H. and P. Fryzlewicz (2015). "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation". *Journal of the Royal Statistical Society: Series B*, pp. 475–507.

Chudova, D. et al. (2003). "Translation-invariant mixture models for curve clustering". *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 79–88.

Clogg, C. C. (1995). "Latent class models". *Handbook of statistical modeling for the social and behavioral sciences*. Springer, pp. 311–359.

Coifman, R. R. and D. L. Donoho (1995). "Translation-invariant de-noising". *Wavelets and Statistics*. Springer New York: New York, NY, pp. 125–150. ISBN: 978-1-4612-2544-7.

Cole, R. J. et al. (1992). "Automatic sleep/wake identification from wrist activity". *Sleep* 15(5), pp. 461–469.

Compiani, G. and Y. Kitamura (2016). "Using mixtures in econometric models: a brief review and some new results". *The Econometrics Journal* 19(3), pp. C95–C127.

Copat, C. et al. (2020). "The role of air pollution (PM and NO2) in COVID-19 spread and lethality: a systematic review". *Environmental research*, p. 110129.

Costa, M. A. and M. Kulldorff (2009). "Applications of spatial scan statistics: a review". *Scan statistics*, pp. 129–152.

Cruz-Medina, I. R., T. P. Hettmansperger, and H. Thomas (2004). "Semiparametric mixture models and repeated measures: the multinomial cut point model". *Journal of the Royal Statistical Society: Series C* 53(3), pp. 463–474.

Dacunha-Castelle, D. and E. Gassiat (1997). "Testing in locally conic models, and application to mixture models". *ESAIM: Probability and Statistics* 1, pp. 285–317.

Dacunha-Castelle, D. and E. Gassiat (1999). "Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes". *The Annals of Statistics* 27(4), pp. 1178–1209.

Daouia, A., S. Girard, and G. Stupfler (2018). "Estimation of tail risk based on extreme expectiles". *Journal of the Royal Statistical Society: Series B* 80(2), pp. 263–292.

Daubechies, I. (1992). *Ten lectures on wavelets*. Vol. 61. Siam.

Daudin, J.-J., F. Picard, and S. Robin (2008). "A mixture model for random graphs". *Statistics and computing* 18(2), pp. 173–183.

Dayton, C. M. and G. B. Macready (1988). "Concomitant-Variable Latent-Class Models". en. *Journal of the American Statistical Association* 83(401), pp. 173–178.

Delaigle, A. and P. Hall (2010). "Defining probability density for a distribution of random functions". *The Annals of Statistics* 38(2), pp. 1171–1193.

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

Dong, C., J. Gao, and D. Tjøstheim (2016). "Estimation for single-index and partially linear single-index integrated models". *The Annals of Statistics* 44(1), pp. 425–453.

Du Roy de Chaumaray, M. and M. Marbac (2020). "Clustering Data with nonignorable Missingness using Semi-Parametric Mixture Models." URL: `https://arxiv.org/abs/2009.07662`.

Du Roy de Chaumaray, M. and M. Marbac (2021a). "Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models". URL: `https://arxiv.org/abs/2112.05684`.

Du Roy de Chaumaray, M. and M. Marbac (2021b). *MNARclust: Clustering Data with Non-Ignorable Missingness using Semi-Parametric Mixture Models*. R package version 1.0.0. URL: `https://CRAN.R-project.org/package=MNARclust`.

Du Roy de Chaumaray, M., M. Marbac, and F. Navarro (2019). *MHMM: Mixture of hidden Markov models for accelerometer data*. R package version 1.0. URL: `https://cran.rstudio.com/web/packages/MHMM/index.html`.

Du Roy de Chaumaray, M., M. Marbac, and F. Navarro (2020). "Mixture of hidden Markov models for accelerometer data". *The Annals of Applied Statistics* 14(4), pp. 1834–1855. URL: `https://projecteuclid.org/euclid.aoas/1608346901`.

Du Roy de Chaumaray, M., M. Marbac, and V. Patilea (2021). "Wilks' theorem for semiparametric regressions with weakly dependent data". *The Annals of Statistics* 49(6), pp. 3228–3254. URL: `https://doi.org/10.1214/21-AOS2081`.

Duczmal, L., M. Kulldorff, and L. Huang (2006). "Evaluation of spatial scan statistics for irregularly shaped clusters". *Journal of Computational and Graphical Statistics* 15(2), pp. 428–442.

Duczmal, L. et al. (2007). "A genetic algorithm for irregularly shaped spatial scan statistics". *Computational Statistics & Data Analysis* 52(1), pp. 43–52.

Dudley, R. M. (2014). *Uniform central limit theorems*. Vol. 142. Cambridge university press.

Dumas, O. et al. (2021). "Household cleaning and poor asthma control among elderly women". *The Journal of Allergy and Clinical Immunology: In Practice*. URL: `https://www.sciencedirect.com/science/article/abs/pii/S2213219821002026?via%3Dihub`.

Dutilleux, P. (1990). "An implementation of the "algorithme à trous" to compute the wavelet transform". *Wavelets*. Springer, pp. 298–304.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses". *The Annals of Mathematical Statistics*, pp. 181–187.

Dyrstad, S. M. et al. (2014). "Comparison of self-reported versus accelerometer-measured physical activity". *Medicine and Science in Sports and Exercise* 46(1), pp. 99–106.

Ehm, W. et al. (2016). "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings". *Journal of the Royal Statistical Society: Series B* 78(3), pp. 505–562.

Elmore, R. T., T. P. Hettmansperger, and H. Thomas (2004). "Estimating component cumulative distribution functions in finite mixture models". *Communications in Statistics-Theory and Methods* 33(9), pp. 2075–2086.

Fan, G.-L. and H.-Y. Liang (2010). "Empirical likelihood inference for semiparametric model with linear process errors". *Journal of the Korean Statistical Society* 39(1), pp. 55–65.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis*. Springer Series in Statistics: New York.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A practical approach*. London: Chapman and Hall.

Forina, M. et al. (1986). "Multivariate data analysis as a discriminating method of the origin of wines". *Vitis* 25(3), pp. 189–201.

Fowler, J. E. (2005). "The redundant discrete wavelet transform and additive noise". *IEEE Signal Processing Letters* 12(9), pp. 629–632.

Francois, O. et al. (2010). "Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture". *Molecular Biology and Evolution* 27 (6).

Freedson, P. S., E. Melanson, and J. Sirard (1998). "Calibration of the Computer Science and Applications, Inc. accelerometer." *Medicine and Science in Sports and Exercise* 30(5), pp. 777–781.

Friel, N. and J. Wyse (2012). "Estimating the evidence–a review". *Statistica Neerlandica* 66(3), pp. 288–308.

Fruhwirth-Schnatter, S., G. Celeux, and C. P. Robert (2019). *Handbook of mixture analysis*. CRC press.

Fryzlewicz, P. and S. Subba Rao (Feb. 2011). "Mixing properties of ARCH and time-varying ARCH processes". *Bernoulli* 17(1), pp. 320–346.

Frévent, C. et al. (2021). "Detecting spatial clusters on functional data: a parametric scan statistic approach". *Spatial Statistics* 46, p. 100550. ISSN: 2211-6753. URL: `https://www.sciencedirect.com/science/article/pii/S2211675321000609`.

Gaffney, . J. and P. Smyth (2005). "Joint probabilistic curve clustering and alignment". *Advances in neural information processing systems*, pp. 473–480.

Galimberti, G., A. Manisi, and G. Soffritti (2018). "Modelling the role of variables in model-based cluster analysis". *Statistics and Computing* 28(1), pp. 145–169.

Galimberti, G. and G. Soffritti (2007). "Model-based methods to identify multiple cluster structures in a data set". *Computational Statistics & Data Analysis* 52(1), pp. 520 –536. ISSN: 0167-9473.

Gao, L. L., J. Bien, and D. Witten (2020). "Selective Inference for Hierarchical Clustering". *arXiv preprint arXiv:2012.02936*.

Gasparrini, A. (2011). "Distributed lag linear and non-linear models in R: the package dlnm". *Journal of Statistical Software* 43(8), pp. 1–20.

Gassiat, E., A. Cleynen, and S. Robin (2016). "Inference in finite state space non parametric hidden Markov models and applications". *Statistics and Computing* 26(1-2), pp. 61–71. ISSN: 0960-3174. DOI: `10.1007/s11222-014-9523-8`.

Geraci, M. (2018). "Additive quantile regression for clustered data with an application to children's physical activity". *Journal of Royal Statistical Society: Series C*.

Geraci, M. and Al. Farcomeni (2016). "Probabilistic principal component analysis to identify profiles of physical activity behaviours in the presence of non-ignorable missing data". *Journal of Royal Statistical Society: Series C* 65(1), pp. 51–75.

Giacofci, M. et al. (2013). "Wavelet-based clustering for mixed-effects functional models in high dimension". *Biometrics* 69(1), pp. 31–40.

Gollini, I. and T. B. Murphy (2014). "Mixture of latent trait analyzers for model-based clustering of categorical data". *Statistics and Computing* 24(4), pp. 569–588.

Goodman, L. A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models". *Biometrika* 61(2), pp. 215–231.

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

Grandner, M. A. et al. (2013). "Sleep duration, cardiovascular disease, and proinflammatory biomarkers". *Nature and Science of Sleep* 5, p. 93.

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

Gruen, M. E. et al. (2017). "The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats". *PLoS One* 12(1), e0169576.

Grün, B. and F. Leisch (2008). "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters". en. *Journal of Statistical Software* 28(4). ISSN: 1548-7660.

Guo, J., M. Wall, and Y.o Amemiya (2006). "Latent class regression on latent factors". *Biostatistics* 7(1), pp. 145–163.

Gupta, S. et al. (2020). "Factors associated with death in critically ill patients with coronavirus disease 2019 in the US". *JAMA Internal Medicine* 180, pp. 1436–1446.

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied latent class analysis*. Cambridge University Press.

Hall, P. and C. C. Heyde (1980). *Martingale limit theory and its application*. Probability and Mathematical Statistics. Academic Press, Inc. New York-London, pp. xii+308. ISBN: 0-12-319350-8.

Hall, P. and B. Presnell (1999). "Density estimation under constraints". *Journal of Computational and Graphical Statistics* 8(2), pp. 259–277.

Hall, P. and X.-H. Zhou (2003). "Nonparametric estimation of component distributions in a multivariate mixture". *The Annals of Statistics* 31(1), pp. 201–224.

Hall, P. et al. (2005). "Nonparametric inference in multivariate mixtures". *Biometrika* 92(3), pp. 667–678.

Hand, D.J. and Y. Keming (2001). "Idiot's Bayes, not so stupid after all?" *International statistical review* 69(3), pp. 385–398.

Hansen, B. E. (2008). "Uniform convergence rates for kernel estimation with dependent data". *Econometric Theory* 24(3), pp. 726–748.

Härdle, W., H. Lütkepohl, and R. Chen (1997). "A Review of Nonparametric Time Series Analysis". *International Statistical Review / Revue Internationale de Statistique* 65(1), pp. 49–72.

Härdle, W., A. Tsybakov, and L. Yang (1998). "Nonparametric vector autoregression". *Journal of Statistical Planning and Inference* 68(2), pp. 221–245.

Haug, N. et al. (2020). "Ranking the effectiveness of worldwide COVID-19 government interventions". *Nature Human Behaviour* 4, pp. 1303–1312.

Helske, S. and J. Helske (2019). "Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R". *Journal of Statistical Software* 88(3), pp. 1–32.

Hennig, Christian (2010). "Methods for merging Gaussian mixture components". *Advances in data analysis and classification* 4(1), pp. 3–34.

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

Hettmansperger, T. P. and H. Thomas (2000). "Almost nonparametric inference for repeated measures in mixture models". *Journal of the Royal Statistical Society: Series B* 62(4), pp. 811–825.

Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2009). "Extending the scope of empirical likelihood". *The Annals of Statistics* 37(3), pp. 1079–1111.

Holschneider, M. et al. (1990). "A real-time algorithm for signal analysis with the help of the wavelet transform". *Wavelets*. Springer, pp. 286–297.

Horowitz, J. L. and S. Lee (2005). "Nonparametric estimation of an additive quantile regression model". *Journal of the American Statistical Association* 100(472), pp. 1238–1249.

Hu, Y., D. McAdams, and M. Shum (2013). "Identification of first-price auctions with non-separable unobserved heterogeneity". *Journal of Econometrics* 174(2), pp. 186–193.

Huang, L., M. Kulldorff, and D. Gregorio (2007). "A spatial scan statistic for survival data". *Biometrics* 63(1), pp. 109–118.

Huang, L. et al. (2018a). "Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations". *Journal of the American Statistical Association*, pp. 1–12.

Huang, L. et al. (2009). "Weighted normal spatial scan statistic for heterogeneous population data". *Journal of the American Statistical Association* 104(487), pp. 886–898.

Huang, Q. et al. (2018b). "Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data". *Journal of the Royal Society Interface* 15(139), p. 20170885.

Hubert, L. and P. Arabie (1985). "Comparing partitions". *Journal of classification* 2(1), pp. 193–218.

Hunt, L. and M. Jorgensen (2011). "Clustering mixed data". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(4), pp. 352–361. ISSN: 1942-4795.

Hunt, L. and M. Jorgensen (2003). "Mixture model clustering for mixed data with missing information". *Computational Statistics and Data Analysis* 41, pp. 429–440.

Hunter, D. R. and K. Lange (2004). "A tutorial on MM algorithms". *The American Statistician* 58(1), pp. 30–37.

Hunter, D. R., D. S.-P. Richards, and J. L. Rosenberger (2011). *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger, the Pennsylvania State University, USA, 23-24 May 2008*. World Scientific.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). "Inference for mixtures of symmetric distributions". *The Annals of Statistics*, pp. 224–251.

Hunter, D. R. and D. S. Young (2012). "Semiparametric mixtures of regressions". *Journal of Nonparametric Statistics* 24(1), pp. 19–38.

Immerwahr, S. et al. (2012). "The Physical Activity and Transit Survey Device Follow-Up Study: Methodology Report". *The New York City Department of Health and Mental Hygiene*.

Innerd, P., R. Harrison, and M. Coulson (2018). "Using open source accelerometer analysis to assess physical activity and sedentary behaviour in overweight and obese adults". *BMC public health* 18(1), p. 543.

Jacques, J. and C. Preda (2014a). "Functional data clustering: a survey". *Advances in Data Analysis and Classification* 8, pp. 231–255.

Jacques, J. and C. Preda (2014b). "Model-based clustering for multivariate functional data". *Computational Statistics & Data Analysis* 71, pp. 92–106.
James, G. M and C. A. Sugar (2003). "Clustering for sparsely sampled functional data". *Journal of the American Statistical Association* 98(462), pp. 397–408.
Josse, J. and F. Husson (2016). "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis". *Journal of Statistical Software* 70(1), pp. 1–31.
Jung, I. (2009). "A generalized linear models approach to spatial scan statistics for covariate adjustment". *Statistics in medicine* 28(7), pp. 1131–1143.
Jung, I., M. Kulldorff, and O. J. Richard (2010). "A spatial scan statistic for multinomial data". *Statistics in medicine* 29(18), pp. 1910–1918.
Kanai, H., H. Ogata, and M. Taniguchi (2010). "Estimating function approach for CHARN models". *Metron* 68(1), pp. 1–21.
Kasahara, H. and K. Shimotsu (2014). "Non-parametric identification and estimation of the number of components in multivariate mixtures". *Journal of the Royal Statistical Society: Series B* 76(1), pp. 97–111.
Kato, H., M. Taniguchi, and M. Honda (2006). "Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans". *IEEE Transactions on Signal Processing* 54(9), pp. 3414–3425.
Keribin, C. (2000). "Consistent estimation of the order of mixture models". *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 49–66.
Kim, M. et al. (2016). "Modeling predictors of latent classes in regression mixture models". *Structural Equation Modeling: A Multidisciplinary Journal* 23(4), pp. 601–614.
Kimm, S. Y. S. et al. (2005). "Relation between the changes in physical activity and body-mass index during adolescence: a multicentre longitudinal study". *The Lancet* 366(9482), pp. 301–307.
Kitamura, Y. (1997). "Empirical likelihood methods with weakly dependent processes". *The Annals of Statistics* 25(5), pp. 2084–2102. ISSN: 0090-5364.
Klemela, J. (2016). *regpro: Nonparametric regression.*
Kneip, A. and T. Gasser (1992). "Statistical tools to analyze data representing a sample of curves". *The Annals of Statistics*, pp. 1266–1305.
Koenker, R. and G. Bassett (1978). "Regression quantiles". *Econometrica: journal of the Econometric Society*, pp. 33–50.
Kosmidis, I. and D. Karlis (2016). "Model-based clustering using copulas with applications". *Statistics and computing* 26(5), pp. 1079–1099.
Kulldorff, M. (1997). "A spatial scan statistic". *Communications in Statistics-Theory and methods* 26(6), pp. 1481–1496.
Kulldorff, M. (2006). "Tests of spatial randomness adjusted for an inhomogeneity: a general framework". *Journal of the American Statistical Association* 101(475), pp. 1289–1305.
Kulldorff, M. et al. (2006). "An elliptic spatial scan statistic". *Statistics in medicine* 25(22), pp. 3929–3943.
Kulldorff, M. et al. (1997). "Breast cancer clusters in the northeast United States: a geographic analysis". *American journal of epidemiology* 146(2), pp. 161–170.
Kwon, C. and E. Mbakop (2020). "Estimation of the number of components of non-parametric multivariate finite mixture models". *Annals of Statistics (to appear).*
Lang, M. et al. (1996). "Noise reduction using an undecimated discrete wavelet transform". *IEEE Signal Processing Letters* 3(1), pp. 10–12.
Lang, M. et al. (1995). "Nonlinear processing of a shift-invariant discrete wavelet transform (DWT) for noise reduction". *Wavelet Applications II*. Vol. 2491. International Society for Optics and Photonics, pp. 640–651.

Lange, K. (2016). *MM optimization algorithms*. SIAM.
Law, M. H. C., M. A. T. Figueiredo, and A. K. Jain (2004). "Simultaneous feature selection and clustering using mixture models". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26(9), pp. 1154–1166.
Lee, I.-M. et al. (2012). "Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy". *The lancet* 380(9838), pp. 219–229.
Levin, D. A. and Y. Peres (2017). *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.
Levine, M., D. R. Hunter, and D. Chauveau (2011). "Maximum smoothed likelihood for multivariate mixtures". *Biometrika*, pp. 403–416.
Li, G. et al. (2010). "Empirical likelihood inference in partially linear single-index models for longitudinal data". *Journal of Multivariate Analysis* 101(3), pp. 718–732.
Li, G. et al. (2017). "The association between smoking and blood pressure in men: a cross-sectional study". *BMC Public Health* 17(1), p. 797.
Li, L. and B. A. Prakash (2011). "Time series clustering: Complex is simpler!" *ICML*.
Li, X.-Z. et al. (2011). "A spatial scan statistic for multiple clusters". *Mathematical biosciences* 233(2), pp. 135–142.
Lian, H. and R. J. Liang H.and Carroll (2015). "Variance function partially linear single-index models". *Journal of the Royal Statistical Society: Series B* 77(1), pp. 171–194. ISSN: 1369-7412.
Liang, H. et al. (Dec. 2010). "Estimation and testing for partially linear single-index models". *The Annals of Statistics* 38(6), pp. 3811–3836.
Liebscher, E. (2005). "Towards a Unified Approach for Proving Geometric Ergodicity and Mixing Properties of Nonlinear Autoregressive Processes". *Journal of Time Series Analysis* 26(5), pp. 669–689.
Lim, S. et al. (2015). "Measurement error of self-reported physical activity levels in New York City: assessment and correction". *American Journal of Epidemiology* 181(9), pp. 648–655.
Lim, Y., H.-S. Oh, and Y. K. Cheung (2019). "Functional clustering of accelerometer data via transformed input variables". *Journal of Royal Statistical Society: Series C* 68(3), pp. 495–520.
Lin, P.-S., Y.-H. Kung, and M. Clayton (2016). "Spatial scan statistics for detection of multiple clusters with arbitrary shapes". *Biometrics* 72(4), pp. 1226–1234.
Lindsay, B. G. and M. L. Lesperance (1995). "A review of semiparametric mixture models". *Journal of statistical planning and inference* 47(1-2), pp. 29–39.
Little, R. J., D. B. Rubin, and S. Z. Zangeneh (2017). "Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets". *Journal of the American Statistical Association* 112(517), pp. 314–320.
Little, R. J. A. (1993). "Pattern-mixture models for multivariate incomplete data". *Journal of the American Statistical Association* 88(421), pp. 125–134.
Little, .R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.
Liu, C. and D. B. Rubin (1994). "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence". *Biometrika* 81(4), pp. 633–648.
Liu, Xueli and Mark C.K. Yang (2009). "Simultaneous curve registration and clustering for functional data". *Computational Statistics and Data Analysis* 53, pp. 1361–1376.
Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE transactions on information theory* 28(2), pp. 129–137.

Löffler, M., A. S. Wein, and A. S. Bandeira (2020). "Computationally efficient sparse clustering". *arXiv preprint arXiv:2005.10817*.

Loh, J. M. and Z. Zhu (2007). "Accounting for spatial correlation in the scan statistic". *The Annals of Applied Statistics* 1(2), pp. 560–584.

Lu, X. (2009). "Empirical likelihood for heteroscedastic partially linear models". *Journal of Multivariate Analysis* 100(3), pp. 387–396.

Lu, Y. and H. H. Zhou (2016). "Statistical and computational guarantees of lloyd's algorithm and its variants". *arXiv preprint arXiv:1612.02099*.

Lu, Z. and Z. Jiang (2001). "L1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term". *Statistics & Probability Letters* 51(2), pp. 121–130.

Ma, Y. and L. Zhu (2013). "Doubly robust and efficient estimators for heteroscedastic partially linear single-index model allowing high-dimensional covariates". *Journal of the Royal Statistical Society: Series B* 75, pp. 305–322.

Mallat, S. G. (1989). "A theory for multiresolution signal decomposition: the wavelet representation". *IEEE transactions on pattern analysis and machine intelligence* 11, pp. 674–693.

Mallat, S. G. (2008). *A wavelet tour of signal processing: the sparse way*. Academic press.

Marbac, M. (2022a). *Introduction à une étude statistique avec données manquantes, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Marbac, M. (2022b). *Méthodes basées sur la vraisemblance pour données manquantes ayant un mécanisme ignorable, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Marbac, M. (2022c). *Méthodes de pondération pour données manquantes, sous la direction de F. Bertrand, G. Saporta, C. Thomas-Agnan.*

Marbac, M., C. Biernacki, and V. Vandewalle (2016). "Latent class model with conditional dependency per modes to cluster categorical data". *Advances in Data Analysis and Classification* 10(2), pp. 183–207. URL: https://link.springer.com/article/10.1007/s11634-016-0250-1.

Marbac, M., C. Biernacki, and V. Vandewalle (2015). "Model-based clustering for conditionally correlated categorical data". *Journal of Classification* 32(2), pp. 145–175. URL: https://link.springer.com/article/10.1007/s00357-015-9180-4.

Marbac, M., C. Biernacki, and V. Vandewalle (2017). "Model-based clustering of Gaussian copulas for mixed data". *Communications in Statistics-Theory and Methods* 46(23), pp. 11635–11656. URL: https://www.tandfonline.com/doi/abs/10.1080/03610926.2016.1277753.

Marbac, M. and P.D. McNicholas (2016). "Dimension Reduction in Clustering". *Wiley StatsRef: Statistics Reference Online*, pp. 1–7. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat07846.

Marbac, M. and M. Sedki (2017a). "A family of block-wise one-factor distributions for modeling high-dimensional binary data". *Computational Statistics & Data Analysis* 114, pp. 130–145. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167947317300932.

Marbac, M and M. Sedki (2016a). *MHTrajectoryR: Bayesian Model Selection in Logistic Regression for the Detection of Adverse Drug Reactions*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=MHTrajectoryR.

Marbac, M and M. Sedki (2016b). *MvBinary: Modelling Multivariate Binary Data with Blocks of Specific One-Factor Distribution*. R package version 1.1. URL: https://CRAN.R-project.org/package=MvBinary.

Marbac, M. and M. Sedki (2017b). "Variable selection for model-based clustering using the integrated complete-data likelihood". *Statistics and Computing* 27(4), pp. 1049–1063. URL: https://link.springer.com/article/10.1007/s11222-016-9670-1.

Marbac, M. and M. Sedki (2018). "VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values". *Bioinformatics* 35(7), pp. 1255–

1257. URL: https://academic.oup.com/bioinformatics/article/35/7/1255/5091183?login=true.

Marbac, M and M. Sedki (2020). *VarSelLCM: Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values*. R package version 2.1.3.1. URL: https://cran.r-project.org/web/packages/VarSelLCM/index.html.

Marbac, M., M. Sedki, and E. Patin (2020). "Variable selection for mixed data clustering: Application in human population genomics". *Journal of Classification*, pp. 1–19. URL: https://link.springer.com/article/10.1007%2Fs00357-018-9301-y.

Marbac, M., P. Tubert-Bitter, and M. Sedki (2016). "Bayesian model selection in logistic regression for the detection of adverse drug reactions". *Biometrical Journal* 58(6), pp. 1376–1389. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201500098.

Marbac, M. and V. Vandewalle (2019). "A tractable multi-partitions clustering". *Computational Statistics & Data Analysis* 132, pp. 167–179. URL: https://www.sciencedirect.com/science/article/abs/pii/S0167947318301592.

Marbac, M. et al. (2021). *ClusPred: Simultaneous Semi-Parametric Estimation of Clustering and Regression*. R package version 1.0.0. URL: https://CRAN.R-project.org/package=ClusPred.

Marbac, M. et al. (2018). "Patterns of cleaning product exposures using a novel clustering approach for data with correlated variables". *The Annals of Epidemiology* 28(8), pp. 563–569. URL: https://www.sciencedirect.com/science/article/abs/pii/S104727971630504X.

Marbac, M. et al. (2022). "Simultaneous semi-parametric estimation of clustering and regression." *Journal of Computational and Graphical Statistics* forthcoming, pp. 1–9. URL: https://doi.org/10.1080/10618600.2021.2000872.

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

Maruotti, A. (2011). "Mixed hidden markov models for longitudinal data: An overview". *International Statistical Review* 79(3), pp. 427–454.

Masry, E. and D. Tjostheim (1995). "Nonparametric Estimation and Identification of Nonlinear ARCH Time Series Strong Convergence and Asymptotic Normality: Strong Convergence and Asymptotic Normality". *Econometric Theory* 11(2), pp. 258–289.

Maugis, C. (2009). "SelvarClustIndep:c++ software".

Maugis, C., G. Celeux, and M.-L. Martin-Magniette (2009a). "Variable selection for clustering with Gaussian mixture models". *Biometrics* 65(3), pp. 701–709.

Maugis, C., G. Celeux, and M.-L. Martin-Magniette (2009b). "Variable selection in model-based clustering: a general variable role modeling". *Computational Statistics & Data Analysis* 53(11), pp. 3872–3882. ISSN: 0167-9473.

Mazo, G. and Y. Averyanov (2019). "Constraining kernel estimators in semiparametric copula mixture models". *Computational Statistics & Data Analysis* 138, pp. 170–189.

McCullagh, P. (1983). "Quasi-likelihood functions". *The Annals of Statistics*, pp. 59–67.

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

McLachlan, G. J. and T. Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

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

McNicholas, P. D. and T. B. Murphy (2008). "Parsimonious Gaussian mixture models". *Statistics and Computing* 18(3), pp. 285–296.

McNicholas, P. D. et al. (2015). "Package 'pgmm'".

McTiernan, A. (2008). "Mechanisms linking physical activity with cancer". *Nature Reviews Cancer* 8(3), p. 205.

Meitz, M. and P. Saikkonen (2010). "A note on the geometric ergodicity of a nonlinear AR-ARCH model". *Statistics and Probability Letters* 80(7), pp. 631 –638. ISSN: 0167-7152.
Meng, X.-L. and D. B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". *Biometrika* 80(2), pp. 267–278.
Menozzi, P., A. Piazza, and L. Cavalli-Sforza (Sept. 1978). "Synthetic maps of human gene frequencies in Europeans". *Science* 201 (4358).
Meynet, C. (2012). "Sélection de variables pour la classification non supervisée en grande dimension". *Thèse de doctorat, Université Paris-Sud 11*.
Miao, W., P. Ding, and Z. Geng (2016). "Identifiability of normal and normal mixture models with nonignorable missing data". *Journal of the American Statistical Association* 111(516), pp. 1673–1683.
Mokkadem, A. (1988). "Mixing properties of ARMA processes". *Stochastic Processes and their Applications* 29(2), pp. 309–315.
Mokkadem, A. (1990). "Propriétés de mélange des processus autorégressifs polynomiaux". *Annales de l'I.H.P. Probabilités et statistiques* 26(2), pp. 219–260.
Molenaar, P. C. M., J. G. De Gooijer, and B. Schmitz (1992). "Dynamic factor analysis of nonstationary multivariate time series". *Psychometrika* 57(3), pp. 333–349.
Molenberghs, G. et al. (2008). "Every missingness not at random model has a missingness at random counterpart with equal fit". *Journal of the Royal Statistical Society: Series B* 70(2), pp. 371–388.
Molenberghs, G. et al. (2014). *Handbook of missing data methodology*. CRC Press.
Morris, J. S. et al. (2006). "Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study". *Journal of the American Statistical Association* 101(476), pp. 1352–1364.
Moustaki, I. and I. Papageorgiou (2005). "Latent class models for mixed variables with applications in Archaeometry". *Computational statistics & data analysis* 48(3), pp. 659–675.
Nakamura, T. (1992). "Proportional hazards model with covariates subject to measurement error". *Biometrics*, pp. 829–838.
Nason, G. P. and B. W. Silverman (1995). "The stationary wavelet transform and some statistical applications". *Wavelets and statistics*. Springer, pp. 281–299.
Naus, J. L. (1965). "Clustering of random points in two dimensions". *Biometrika* 52(1-2), pp. 263–266.
Navarro, F. and C. Chesneau (2020). "R package rwavelet: Wavelet Analysis". *Version 0.4.1*.
Newey, W. K. and J. L. Powell (1987). "Asymmetric least squares estimation and testing". *Econometrica: Journal of the Econometric Society*, pp. 819–847.
Noel, S. E. et al. (2010). "Use of accelerometer data in prediction equations for capturing implausible dietary intakes in adolescents". *The American Journal of Clinical Nutrition* 92(6), pp. 1436–1445.
Nordman D. J.and Lahiri, S. N. (2014). "A review of empirical likelihood methods for time series". *Journal of Statistical Planning and Inference* 155, pp. 1–18.
Novembre, J. et al. (2008). "Genes mirror geography within Europe". *Nature* 456(7218), pp. 98–101.
Nowicki, K. and T. A. B. Snijders (2001). "Estimation and prediction for stochastic blockstructures". *Journal of the American Statistical Association* 96(455), pp. 1077–1087. ISSN: 0162-1459.
Omvik, P. (1996). "How smoking affects blood pressure". *Blood pressure* 5(2), pp. 71–77.
Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.
Owen, A. B. (1988). "Empirical likelihood ratio confidence intervals for a single functional". *Biometrika* 75(2), pp. 237–249.

Palta, P. et al. (2015). "Self-reported and accelerometer-measured physical activity by body mass index in US Hispanic/Latino adults: HCHS/SOL". *Preventive medicine reports* 2, pp. 824–828.

Pan, W. and X. Shen (2007). "Penalized model-based clustering with application to variable selection". *The Journal of Machine Learning Research* 8, pp. 1145–1164.

Paparrizos, J. and L. Gravano (2015). "k-shape: Efficient and accurate clustering of time series". *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1855–1870.

Patin, E. et al. (2017). "Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America". *Science* 356(6337), pp. 543–546.

Patra, R. K. and B. Sen (2016). "Estimation of a two-component mixture model with applications to multiple testing". *Journal of the Royal Statistical Society: Series B* 78(4), pp. 869–893.

Patterson, N., A. L. Price, and D. Reich (2006). "Population Structure and Eigenanalysis". *PLoS Genetics* 2 (12).

Phillips, C. (2012). *Ancestry informative markers*. Siegel, Jay A and Saukko, Pekka J: Encyclopedia of forensic sciencesAcademic Press, pp. 323–331.

Pol, F. Van de and R. Langeheine (1990). "Mixed Markov latent class models". *Sociological methodology*, pp. 213–247.

Pollak, C. P. et al. (2001). "How accurately does wrist actigraphy identify the states of sleep and wakefulness?" *Sleep* 24(8), pp. 957–965.

Poon, L. K. M. et al. (2013). "Model-based clustering of high-dimensional data: Variable selection versus facet determination". *International Journal of Approximate Reasoning* 54(1), pp. 196–215.

Pozzer, A. et al. (2020). "Regional and global contributions of air pollution to risk of death from COVID-19". *Cardiovascular Research* 116(14), pp. 2247–2253.

Price, A. L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nature genetics* 38(8), pp. 904–909.

Pritchard, J. K., J. K. Pickrell, and G. Coop (2010). "The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation". *Current Biology* 20 (4). DOI: 10.1016/j.cub.2009.11.055.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). "Inference of population structure using multilocus genotype data". *Genetics* 155(2), pp. 945–959.

Qin, J. and J. Lawless (1994). "Empirical likelihood and general estimating equations". *The Annals of Statistics* 22(1), pp. 300–325. ISSN: 0090-5364.

Racine, J. S., C. F. Parmeter, and P. Du (2009). "Constrained nonparametric kernel regression: Estimation and inference". *Working paper*.

Raftery, A. E. and N. Dean (2006). "Variable Selection for Model-Based Clustering". *Journal of the American Statistical Association* 101(473), pp. 168–178.

Ramsay, J. O. and X. Li (1998). "Curve registration". *Journal of the Royal Statistical Society: Series B* 60(2), pp. 351–363.

Ramsay, James O. and Bernard W. Silverman (2007). *Applied functional data analysis: methods and case studies*. Springer.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*. Vol. 31. Mathématiques & Applications (Berlin) [Mathematics & Applications]. Springer-Verlag, Berlin, pp. x+169. ISBN: 3-540-65979-X.

Robin, S. et al. (2007). "A semi-parametric approach for mixture models: Application to local false discovery rate estimation". *Computational statistics & data analysis* 51(12), pp. 5483–5493.

Rotnitzky, A. and J. Robins (1997). "Analysis of semi-parametric regression models with non-ignorable non-response". *Statistics in medicine* 16(1), pp. 81–102.

Sadeh, A., M. Sharkey, and M. A. Carskadon (1994). "Activity-based sleep-wake identification: an empirical test of methodological issues". *Sleep* 17(3), pp. 201–207.

Saldanha Gomes, C. et al. (2020). "Clusters of diet, physical activity, television exposure and sleep habits and their association with adiposity in preschool children: the EDEN mother-child cohort." *International Journal of Behavioral Nutrition and Physical Activity* 17(1). URL: `https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-020-00927-6#citeas`.

Salzberg, S. L. (1988). *Exemplar-based learning: Theory and implementation*. Harvard University, Center for Research in Computing Technology, Aiken ...

Samé, A. et al. (2011). "Model-based clustering and segmentation of time series with changes in regime". *Advances in Data Analysis Classification* 5, pp. 301–321.

Sammel, M. D., L. M. Ryan, and J. M. Legler (1997). "Latent variable models for mixed discrete and continuous outcomes". *Journal of the Royal Statistical Society: Series B* 59(3), pp. 667–678.

Sanderson, J., P. Fryzlewicz, and M. W. Jones (2010). "Estimating linear dependence between nonstationary time series using the locally stationary wavelet model". *Biometrika* 97(2), pp. 435–446.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics* 6(2), pp. 461–464.

Scott, S. L., G. M. James, and C. A. Sugar (2005). "Hidden Markov models for longitudinal comparisons". *Journal of the American Statistical Association* 100(470), pp. 359–369.

Scrucca, L. and A.E. Raftery (2014). "clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R".

Scrucca, L. et al. (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models". *The R journal* 8(1), p. 289.

Serafini, A., T. B.n Murphy, and L. Scrucca (2020). "Handling missing data in model-based clustering". *arXiv preprint arXiv:2006.02954*.

Severini, T. A. and W. H. Wong (1992). "Profile likelihood and conditionally parametric models". *The Annals of Statistics* 20, pp. 1768–1802.

Sharpnack, J. and E. Arias-Castro (2016). "Exact asymptotics for the scan statistic and fast alternatives". *Electronic Journal of Statistics* 10(2), pp. 2641–2684.

Slootmaker, S. M. et al. (2009). "Disagreement in physical activity assessed by accelerometer and self-report in subgroups of age, gender, education and weight status". *International Journal of Behavioral Nutrition and Physical Activity* 6(1), p. 17.

Stephens, C. R., H. F. Huerta, and A. R. Linares (2018). "When is the Naive Bayes approximation not so naive?" *Machine Learning* 107(2), pp. 397–441.

Streuli, H. (1973). "Der heutige stand der kaffeechemie". *In Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemisrty*, pp. 61–72.

Sweldens, W. (1998). "The lifting scheme: A construction of second generation wavelets". *SIAM Journal on Mathematical Analysis* 29, pp. 511–546.

Sy, K. T. L., L. F. White, and B. E. Nichols (2021). "Population density and basic reproductive number of COVID-19 across United States counties". *PloS one* 16(4), e0249271.

Szekely, G. J. and M. L. Rizzo (2005). "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method". *Journal of classification* 22(2), pp. 151–184.

Tabouy, T., P. Barbillon, and J. Chiquet (2020). "Variational inference for stochastic block models from sampled data". *Journal of the American Statistical Association* 115(529), pp. 455–466.

Tadesse, M. G., N. Sha, and M. Vannucci (2005). "Bayesian variable selection in clustering high-dimensional data". *Journal of the American Statistical Association* 100(470), pp. 602–617. ISSN: 0162-1459.

Taheri, S. et al. (2004). "Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index". *PLoS Med.* 1(3), e62.

Tang, C., T. Wang, and P. Zhang (2020). *Functional data analysis: An application to COVID-19 data in the United States*.

Tang, R. and H.-G. Müller (2009). "Time-synchronized clustering of gene expression trajectories". *Biostatistics* 10(1), pp. 32–45.

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

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

Tibshirani, R., G. Walther, and T. Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". *Journal of the Royal Statistical Society: Series B* 63(2), pp. 411–423.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society: Series B* 58(1), pp. 267–288.

Titsias, M. K., C. C. Holmes, and C. Yau (2016). "Statistical inference in hidden Markov models using k-segment constraints". *IEEE Transactions on Information Theory* 111(513), pp. 200–215.

Tjøstheim, D. (1990). "Non-linear time series and Markov chains". *Advances in Applied Probability* 22(3), 587–611.

Tong, H. (1990). *Nonlinear time series*. Vol. 6. Oxford Statistical Science Series. A dynamical system approach, With an appendix by K. S. Chan, Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, pp. xvi+564. ISBN: 0-19-852224-X.

Troiano, R. P. et al. (2008). "Physical activity in the United States measured by accelerometer". *Medicine and Science in Sports and Exercise* 40(1), pp. 181–188.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Unser, M. (1995). "Texture classification and segmentation using wavelet frames". *IEEE Transactions on Image Processing* 4, pp. 1549–1560.

US Department of Health and Human Services (2008). "2008 physical activity guidelines for Americans: Be active, healthy, and happy!"

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Hees, V. T. et al. (2015). "A novel, open access method to assess sleep duration using a wrist-worn accelerometer". *PLoS One* 10(11), e0142533.

Vaňkátová, K. and E. Fišerová (2017). "The Evaluation of a Concomitant Variable Behaviour in a Mixture of Regression Models". en. *Statistika* 97(4), p. 16.

Viterbi, A. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *Journal of the American Statistical Association* 13(2), pp. 260–269.

Wallace, M. L. et al. (2018). "Variable selection for skewed model-based clustering: application to the identification of novel sleep phenotypes". *Journal of the American Statistical Association* 113(521), pp. 95–110.

Walther, G. (2010). "Optimal and fast detection of spatial clusters with scan statistics". *The Annals of Statistics* 38(2), pp. 1010–1033.

Wang, K. and T. Gasser (1997). "Alignment of curves by dynamic time warping". *The Annals of Statistics* 25, pp. 1251–1276.

Wang, P. et al. (1996). "Mixed Poisson regression models with covariate dependent rates". *Biometrics*, pp. 381–400.

Wang, Q.-H. and B.-Y. Jing (2003). "Empirical likelihood for partial linear models". *Annals of the Institute of Statistical Mathematics* 55(3), pp. 585–595.
Wang, Q.-H. and B.-Y. Jing (1999). "Empirical likelihood for partial linear models with fixed designs". *Statistics & Probability Letters* 41(4), pp. 425–433.
Ward, J. H. (1963). "Hierarchical grouping to optimize an objective function". *Journal of the American statistical association* 58(301), pp. 236–244.
Webb, G. I., J. R. Boughton, and Z. Wang (2005). "Not so naive Bayes: aggregating one-dependence estimators". *Machine learning* 58(1), pp. 5–24.
Wei, G. C. G. and M. A. Tanner (1990). "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms". *Journal of the American statistical Association* 85(411), pp. 699–704.
Wei, Y. and R. J. Carroll (2009). "Quantile regression with measurement error". *Journal of the American Statistical Association* 104(487), pp. 1129–1143.
Williamson, E. J., A. J. Walker, and K. Bhaskaran (2020). "Factors associated with COVID-19-related death using OpenSAFELY". *Nature* 584, pp. 430–436.
Witowski, V. et al. (2014). "Using Hidden Markov Models to Improve Quantifying Physical Activity in Accelerometer Data–A Simulation Study". *PLoS One* 9(12), e114089.
Witten, D.M. and R. Tibshirani (2010). "A Framework for Feature Selection in Clustering". *Journal of the American Statistical Association* 105(490), pp. 713–726.
World Health Organization (2016). *ICD-10 : International statistical classification of diseases and related health problems : tenth revision.*
Wu, C. F. J. (1983). "On the convergence properties of the EM algorithm". *The Annals of statistics*, pp. 95–103.
Wu, Q. and W. Yao (2016). "Mixtures of quantile regressions". *Computational Statistics & Data Analysis* 93, pp. 162–176.
Wyker, B. et al. (2013). "Self-reported and accelerometer-measured physical activity: a comparison in New York City". *New York (NY): New York City Department of Health and Mental Hygiene: Epi Research Report*, pp. 1–12.
Xia, Y. and W. Härdle (2006). "Semi-parametric estimation of partially linear single-index models". *Journal of Multivariate Analysis* 97(5), pp. 1162 –1184. ISSN: 0047-259X.
Xia, Y., H. Tong, and W. K. Li (1999). "On extended partially linear single-index models". *Biometrika* 86(4), pp. 831–842.
Xiang, S., W. Yao, and G. Yang (2019). "An overview of semiparametric extensions of finite mixture models". *Statistical science* 34(3), pp. 391–404.
Xiao, L. et al. (2014). "Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach". *Biostatistics* 16(2), pp. 352–367.
Xie, B., W. Pan, and X. Shen (2008). "Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables". *Electronic Journal of Statistics* 2, pp. 168–212. ISSN: 1935-7524.
Xing, J.-J. and X.-Y. Qian (2017). "Bayesian expectile regression with asymmetric normal distribution". *Communications in Statistics-Theory and Methods* 46(9), pp. 4545–4555.
Xue, L. and L. Zhu (2007). "Empirical likelihood semiparametric regression analysis for longitudinal data". *Biometrika* 94(4), pp. 921–937.
Xue, L.-G. and L. Zhu (2006). "Empirical likelihood for single-index models". *Journal of Multivariate Analysis* 97(6), pp. 1295–1312.
Yakowitz, S. J. and J. D. Spragins (1968). "On the identifiability of finite mixtures". *The Annals of Mathematical Statistics*, pp. 209–214.
Yang, C.-C. and Y.-L. Hsu (2010). "A review of accelerometry-based wearable motion detectors for physical activity monitoring". *Sensors* 10(8), pp. 7772–7788.

Yu, K. and R. A. Moyeed (2001). "Bayesian quantile regression". *Statistics & Probability Letters* 54(4), pp. 437–447.

Zhang, T. and G. Lin (2017). "Asymptotic properties of spatial scan statistics under the alternative hypothesis". *Bernoulli* 23(1), pp. 89–109.

Zhang, T. and G. Lin (2013). "On the limiting distribution of the spatial scan statistic". *Journal of Multivariate Analysis* 122, pp. 215–225.

Zhang, T. and G. Lin (2009). "Spatial scan statistics in loglinear models". *Computational Statistics & Data Analysis* 53(8), pp. 2851–2858.

Zhang, Y., H. J. Wang, and Z. Zhu (2019). "Quantile-regression-based clustering for panel data". *Journal of Econometrics*.

Zhang, Z., R. Assunção, and M. Kulldorff (2010). "Spatial scan statistics adjusted for multiple clusters". *Journal of Probability and Statistics* 2010.

Zheng, C. and Y. Wu (2020). "Nonparametric estimation of multivariate mixtures". *Journal of the American Statistical Association* 115(531), pp. 1456–1471.

Zhou, F. et al. (2020). "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study". *The Lancet* 395, pp. 1054–1062.

Zhou, H., W. Pan, and X. Shen (2009). "Penalized model-based clustering with unconstrained covariance matrices". *Electronic Journal of Statistics* 3, pp. 1473–1496. ISSN: 1935-7524.

Zhu, L. and L. Xue (2006). "Empirical likelihood confidence regions in a partially linear single-index model". *Journal of the Royal Statistical Society: Series B* 68, pp. 549–570.

Zhu, L. et al. (2010). "Bias-corrected empirical likelihood in a multi-link semiparametric model". *Journal of Multivariate Analysis* 101(4), pp. 850 –868. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2009.08.009. URL: http://www.sciencedirect.com/science/article/pii/S0047259X09001535.

Zhu, X. and D. R. Hunter (2019). "Clustering via finite nonparametric ICA mixture models". *Advances in Data Analysis and Classification* 13(1), pp. 65–87.

Zhu, X. and D. R. Hunter (2016a). "Theoretical grounding for estimation in conditional independence multivariate finite mixture models". *Journal of Nonparametric Statistics* 28(4), pp. 683–701.

Zhu, X. and D. R. Hunter (2016b). "Theoretical grounding for estimation in conditional independence multivariate finite mixture models". *Journal of Nonparametric Statistics* 28, pp. 683–701.