Dynamic mixture models and longitudinal monitoring for mixed-type and spatio-temporal data inference : application in Public Health

Solange Pruilh

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX065

Thèse de doctorat

Dynamic mixture models and longitudinal monitoring for mixed-type and spatio-temporal data inference: application in Public Health

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques Appliquées

Thèse présentée et soutenue à Palaiseau, le 28 septembre 2023, par

SOLANGE PRUILH

Composition du Jury :

Erwan Le Pennec Professeur, École polytechnique	Président
Christophe Biernacki Professeur, Université de Lille	Rapporteur
Paul McNicholas Professor, McMaster University	Rapporteur
Pierre Pudlo Professeur, Aix-Marseille Université	Examinateur
Stéphanie Allassonnière Professeure, Université Paris-Cité	Directrice de thèse
Anne-Sophie Jannot Maitre de Conférence-Praticienne hospitalière, Université Paris-Cité	Co-directrice de thèse


CMAP
HeKA
MATH
INNOV
Région
îledeFrance


Inria
Inserm
Université
Paris Cité

Résumé

Dans cette thèse, nous nous intéressons à des méthodes d'apprentissage statistique sur données spatio-temporelles et données mixtes. En effet, la croissance rapide des systèmes d'informations en santé publique permet aujourd'hui de disposer de données variées en temps réel pour de nombreuses maladies. L'objectif est de développer des méthodes pour utiliser ces données afin de construire des systèmes d'aide à la décision exploitables.

Nous proposons d'abord un pipeline spatio-temporel pour estimer la distribution de la population et mettre en évidence des différences temporelles. Ce pipeline est une première étape vers un système d'aide à la décision et d'alerte pour l'analyse spatio-temporelle de l'évolution d'une population. Ce pipeline est conçu de manière à ce que différentes distributions et donc différents algorithmes puissent être envisagés. Pour une première application, ce pipeline est combiné avec des algorithmes EM robustes permettant l'estimation de modèles de mélange gaussiens. Il est éprouvé sur des données d'hôpitaux parisiens correspondant aux personnes testées positives à l'infection par le SARS-CoV-2 sur onze semaines en 2020.

Dans une deuxième partie nous proposons un ensemble d'algorithmes pour l'estimation de modèles de mélange sur données mixtes. Nous décrivons d'abord des modèles de mélange pour diverses lois continues et discrètes, en supposant une indépendance conditionnelle entre les variables discrètes et continues. Nous proposons ensuite des algorithmes dynamiques de type EM, permettant l'estimation de tous les paramètres du mélange ainsi que l'estimation du nombre de classes. Nous montrons que nos différents algorithmes dynamiques permettent d'atteindre le nombre réel de classes et d'estimer correctement les paramètres des lois discrètes comme continues. Nous soulignons aussi l'intérêt d'introduire des régularisations sur des paramètres particuliers afin d'améliorer les performances dans des situations où la taille de l'échantillon n'est pas suffisante en regard de la complexité du modèle. Ces algorithmes dynamiques sont ensuite validés sur des données réelles issues de la littérature.

Abstract

In this thesis, we focus on statistical learning methods for spatio-temporal and mixed-type data. With the rapid growth of public health information systems, a wide range of real-time data is now available for many diseases. The aim is to develop methods for using this data to build operational decision support systems.

We first propose a spatio-temporal pipeline for estimating the distribution of a population and highlighting temporal differences. This pipeline is a first step towards a decision support and alert system for the spatio-temporal analysis of population trends. This pipeline is designed so that different distributions and therefore different algorithms can be considered. For an initial application, this pipeline is combined with robust EM algorithms for estimating Gaussian mixture models. It is evaluated using data from Paris hospitals corresponding to people who tested positive for SARS-CoV-2 infection over eleven weeks in 2020.

In the second part, we propose a set of algorithms for estimating statistical models on mixed data. We consider that mixed-type data are distributed according to mixtures of laws. We first describe mixture models for various continuous and discrete laws, assuming local independence between discrete and continuous variables. We then propose dynamic algorithms of the EM type, allowing the estimation of all the parameters of the mixture as well as the estimation of the number of classes. We show that our different dynamic algorithms allow us to reach the real number of classes and to correctly estimate the parameters of the discrete and continuous laws. We also highlight the benefits of introducing regularizations to improve performance in situations where the sample size is insufficient for the complexity of the model. These dynamic algorithms are then validated on real data from the literature.

Pour ma mère, Marie-Laure.

Remerciements

Arrivée au bout de cette thèse, je souhaite ici remercier sincèrement toutes les personnes que j'ai côtoyées ces dernières années et qui m'ont soutenues (peut-être sans le savoir) chaque jour. Je ne peux que commencer par remercier chaleureusement mes directrices, Anne-Sophie et Stéphanie, présentes depuis plus de quatre ans. Merci à vous deux de m'avoir proposé ce projet, d'avoir cru en moi et d'être restées avec moi sur ce long chemin que fut ma thèse. Merci Anne-Sophie de m'avoir ouverte aux opportunités que présentent les données en Santé publique, et pour des discussions toujours intéressantes sur les maladies rares. Merci Stéphanie pour ton soutien permanent, tes conseils et ton optimisme à toute épreuve. Merci de m'avoir poussée à croire en moi (ce qui était déjà une épreuve), et pour toutes nos discussions, sur les maths, la santé et tant d'autres sujets.

I am grateful to Paul McNicholas and Christophe Biernacki for accepting to review my thesis manuscript. Je remercie également Pierre Pudlo d'avoir accepté d'être examinateur à ma soutenance, et Erwan Le Pennec, examinateur, mais aussi président du jury et d'un grand soutien avant même la soutenance.

Au cours de cette thèse, j'ai eu la chance de rencontrer de nombreuses personnes au CMAP que je remercie, dans mon équipe mais aussi dans tout le laboratoire. Un congrès, des confinements, et d'autres épreuves que nous avons pu vivre pendant nos thèses ne nous ont pas empêchés de nous soutenir entre doctorants. La thèse peut être un travail solitaire certes, mais rien n'empêche d'être bien entourée, et ce fut mon cas. Au-delà de la recherche et de l'enseignement, ce laboratoire ne tiendrait pas sans toutes les personnes que l'on raccourcit vite en "ITA", et qui sont bien plus que ces trois lettres. Je les remercie toutes et tous chaleureusement pour leur aide bien précieuse, mais aussi et surtout pour leur accueil, leur soutien et leur gentillesse. Je n'oublie pas ma deuxième équipe, HeKA, qui m'a accueillie dès mon stage, et dont je remercie tous les membres, anciens et présents, avec qui j'ai vécu des expériences diverses et variées, allant du bureau collé aux toilettes au chercheur ronfleur. Je souhaite aussi remercier les nombreuses professeures et chercheuses avec qui j'ai étudié ou travaillé au lycée puis à l'INSA, qui m'ont donné goût aux sciences, aux mathématiques et enfin à la recherche.

Ces quatre années, et toutes les épreuves traversées me permettent aujourd'hui de réaliser ma chance, et ma richesse, que sont mes ami(e)s, toujours présent(e)s pour me supporter (parfois), me soutenir et surtout croire en moi (à ma place, merci à vous). À toutes et tous, j'adresse donc mes remerciements les plus sincères, et j'espère vous apporter dans le futur autant que vous m'avez donné.

Je ne peux oublier ma famille, proche et lointaine, qui parfois sans bien savoir si je travaillais ou étudiais (question difficile), m'a toujours apporté encouragements ou distractions, dans les moments où j'en avais besoin. Un merci particulier à ma mère, qui me montre au quotidien la force et le courage dont nous sommes toutes les deux capables.

Je finis par celui qui me supporte depuis de nombreuses années, qui partage mes doutes et mes joies depuis bien avant cette thèse. Mercés pour ton amour et ton soutien sans faille.

Contents

Résumé en Français

Cette thèse s'intéresse au développement de modèles de mélange dynamiques, et à la construction de pipelines guidés par des enjeux de surveillance spatio-temporelle en Santé Publique. La croissance rapide des systèmes d'information sur la santé permet de disposer de données spatio-temporelles en temps réel de patients touchés par une maladie donnée.

Les modèles spatiaux reposent sur la caractérisation des individus par leur localisation géographique (lieu de naissance, lieu au moment du diagnostic, lieu de résidence, *etc*). L'ensemble de ces individus forme une population. Par ailleurs, la composante temporelle est essentielle dans le suivi des maladies, ce qui nécessite de prendre en compte l'évolution de la distribution de la population dans le temps. L'association des composantes spatiales et temporelles d'une maladie permet d'obtenir une distribution spatio-temporelle.

Un système d'aide à la décision qui pourrait améliorer la gestion des risques sanitaires en utilisant ces données est la mise en évidence en temps réel de groupes de patients nouveaux ou en évolution. Il s'agit par exemple de détecter un sous-groupe spécifique de patients qui évoluera différemment, alors que le reste de la population restera stable. Cela serait particulièrement utile pour identifier rapidement une nouvelle source de contamination pour une maladie transmissible, dès que les premiers cas affectés sont présents dans les systèmes d'information sanitaire.

Méthodes statistiques pour données spatio-temporelles Les analyses statistiques spatio-temporelles sont déjà nombreuses dans la recherche en épidémiologie, que cela soit par des tests sur la significativité de taux d'incidence ou de risque, ou la modélisation de ces taux, permettant l'intégration de variables additionnelles. De nombreuses méthodes requièrent néanmoins une connaissance de la population à risque sous-jacente ou un ensemble cas/témoins. Ou encore nécessitent un horizon temporel déterminé, ne permettant pas leur utilisation pour un suivi à long terme ou en temps réel.

Un pipeline comme outil d'aide à la décision Nous proposons dans une première contribution (Chapitre 3) un pipeline simple, flexible afin de suivre l'évolution de distributions estimées sur des observations arrivant à différents temps (discrets). Nous construisons donc ce pipeline afin qu'il soit facile d'utilisation, interprétable, mais aussi rapide d'exécution afin de s'adapter à différentes granularités temporelles. Ce processus ne nécessite pas non plus d'avoir les données sur l'ensemble des temps considérés, contrairement à certaines méthodes estimant des distributions ou des clusters temporels, qui ont besoin de l'ensemble des données, du premier au dernier temps considérés.

Modèles de mélange Les modèles de mélange sont des modèles probabilistes où la densité de probabilité d'une observation x est formulée comme la somme pondérée de K densités connues, de paramètres différents et généralement de la même famille. Les pondérations

sont données par un vecteur de proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ et les paramètres de chaque densité k du mélange par θ_k. Afin d'inférer les paramètres du modèle, on veut considérer que chaque observation appartient à une des classes, et ainsi estimer indépendamment chaque distribution. Dans un cadre non supervisé, l'appartenance d'une observation x à une des K classes est inconnue et modélisée par une variable latente z distribuée suivant une loi catégorielle de paramètres $\boldsymbol{\pi}$. Une observation x est alors générée suivant le modèle suivant

$$\begin{cases} z & \sim & \text{Catégorielle}(\boldsymbol{\pi})\,, \\ x|z=k & \sim & F_g(\theta_k)\,, \end{cases} \tag{1}$$

avec F_g une distribution continue ou discrète de dimension g et de paramètres θ_k.

Grâce à leur flexibilité, les modèles de mélange sont vite devenus très populaires dans tous types d'applications et pour de nombreux types de données : images, données spatiales, données fonctionnelles, *etc.*

Nous nous tournons vers ces modèles car ils répondent à des critères de flexibilité dans le choix des lois, mais aussi d'estimation relativement rapide, d'interprétation, et ne reposant pas dans le cas général sur des informations *a priori* (à part le nombre de classes que nous évoquons plus bas), telles que la distribution sous-jacente d'une population globale ou la présence d'observations témoins pour des applications en Santé Publique.

Estimer des modèles de mélange De nombreuses méthodes permettent d'estimer des modèles de mélange, dont le très connu algorithme Espérance-Maximisation (EM) (Dempster et al., 1977) et ses variantes. Il permet plus largement d'estimer des modèles probabilistes sur données incomplètes, intégrant des variables latentes ou inconnues (comme z ci-dessus).

Dans le cadre d'estimation de modèles de mélange, les variantes de l'algorithme EM sont déjà nombreuses, pour estimer des mélanges de diverses lois possibles, continues comme discrètes, mais aussi afin de résoudre de nombreuses limites, telles la sensibilité à l'initialisation, la convergence vers des maxima locaux incorrects et la convergence des paramètres vers les bords de l'espace des paramètres notamment lors d'un surapprentissage. Au-delà de ces difficultés, un enjeu avec les modèles de mélange, qu'ils soient estimés par algorithme EM ou d'autres méthodes (les méthodes de Monte-Carlo par chaînes de Markov par exemple), est le choix du nombre de composantes K dans le mélange. Fixé *a priori*, choisi par sélection de modèles *a posteriori* ou incrémenté/décrémenté au cours du processus d'estimation, de nombreuses propositions sur cet enjeu ont vu le jour et essayant souvent de résoudre les autres limites énoncées ci-dessus.

Un algorithme «dynamique» Nous appuyant sur le travail de Yang et al. (2012), nous proposons un algorithme de type EM «dynamique», permettant d'éviter notamment le surapprentissage, une des faiblesses initiales de l'algorithme EM que le travail de Yang et al. ne résolvait pas. Par dynamique, nous voulons exprimer que le nombre de classes K du mélange est inconnu à l'initialisation de l'algorithme et va être estimé conjointement avec les paramètres du modèle, tout en évitant de parcourir de manière exhaustive l'ensemble des valeurs possibles pour K. Évitant donc aussi l'estimation d'une collection de modèles, cela repose notamment sur l'introduction de l'entropie des proportions dans la fonction objectif à maximiser. Cet algorithme permet ici d'estimer des modèles de mélange gaussiens.

Un processus de mélange spatio-temporel Une autre variante de l'algorithme EM est aussi explicitée, simplement contrainte lors d'une courte estimation à rester dans un

voisinage de valeurs initiales. Ces deux types d'algorithmes sont ainsi des composantes du pipeline développé afin d'assurer une modélisation et un suivi temporel de données $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(T)})$, $\mathbf{X}^{(t)}$ étant un ensemble de données arrivant à l'instant t indépendamment des instants précédents. Ce pipeline composé de modèles de mélange, gaussiens dans un premier temps, est nommé *spatio-temporal mixture process (STMP)*. À chaque instant t, deux modèles sont estimés : l'un dont les paramètres sont proches de l'instant $t-1$, l'autre non contraint. Un ratio impliquant les vraisemblances des modèles $p_{\theta_1}(x)$ et $p_{\theta_2}(x)$ permet ensuite de déterminer quel modèle est gardé comme référence à l'instant t. Si le modèle à t dépendant de celui estimé à $t-1$ et exécuté sur les données $\mathbf{X}^{(t-1)}$ n'est pas gardé, on peut traduire cela comme une rupture temporelle dans l'évolution du comportement de la population.

Nous montrons ensuite sur un jeu de données réelles que ce pipeline estime des modèles cohérents et révèle des fractures dans l'évolution de la répartition des observations. Ces données sont composées de patients testés positifs à la COVID-19 dans les hôpitaux de Paris sur onze semaines en 2020. Modélisées après agrégation par semaine, elles suivent une évolution statistique cohérente avec les mesures de santé publique mises en place au printemps 2020.

Des données mixtes En Santé Publique comme dans de nombreux domaines, les données récoltées et disponibles sont bien plus complètes et variées que celles abordées dans la première partie de cette thèse. En Santé Publique justement, on dispose de nombreuses informations sur les patients : sexe, âge, caractères phénotypiques (taille, poids, *etc*), facteurs exogènes (tabac, pollution, *etc*), *etc*.

Après le développement du *spatio-temporal mixture process*, nous nous penchons donc sur les composantes internes du pipeline, qui ne permettaient jusqu'à présent que d'estimer des modèles de mélange gaussiens. Améliorer ces composantes permettrait l'utilisation de ce pipeline sur des données mixtes, suivant l'évolution des distributions par rapport à des caractéristiques spatiales, mais aussi d'autres caractéristiques.

Modèles de mélange pour données mixtes Au cours de cette thèse, nous proposons alors dans une deuxième contribution (Chapitre 4) des modèles de mélange pour données mixtes (continues, discrètes, binaires, *etc*), en s'appuyant sur l'hypothèse d'indépendance conditionnelle entre variables continues et catégorielles mais aussi entre les variables catégorielles. Connaissant l'affectation de chaque observation à une des K classes du mélange, ses caractéristiques continues suivent une unique loi, qui modélise ainsi leurs dépendances, tandis que chaque variable discrète comme binaire ou nominale suit une loi marginale correspondant à sa nature. Cette hypothèse a ses défauts, ne permettant notamment pas de capturer des dépendances directes entre toutes les variables. Mais cela est compensé par une simplicité des modèles qui permet aussi une simplicité d'estimation, une interprétation aisée et l'identifiabilité (Foss et al., 2019).

Les modèles de mélange sont alors décrits suivant le système suivant:

$$\begin{cases} z & \sim & \text{Catégorielle}(\boldsymbol{\pi})\,, \\ x^c | z = k & \sim & F_g^c(\theta_k^c)\,, \\ x^d | z = k & \sim & F^d(\theta_k^d)\ \forall\, d = 1, \ldots, D\,, \end{cases} \tag{2}$$

avec pour chaque composant k, une distribution continue F_g de dimension g de paramètres θ_k^c, et $F^d(\theta_k^d)$ la distribution discrète de l'attribut discret d, de paramètre θ^d. Les D variables discrètes peuvent suivre des lois de différentes familles.

Algorithmes EM dynamiques pour données mixtes Une fois ces modèles définis, la partie importante de notre proposition est de développer des algorithmes de type EM pour les estimer. Dans le prolongement de nos premiers travaux sur des EM dynamiques pour modèles de mélange gaussiens, nous proposons un algorithme de type EM qui estime conjointement le nombre de classes et les paramètres d'un mélange de données mixtes. Nommé *Dynamic EM for Mixed-type Data (DEM-MD)*, cet algorithme permet l'estimation de mélanges pour données mixtes, et nous en proposons des versions adaptées aux lois continues suivantes : Gaussienne, Student, et Asymétrique Laplace non centrée (SAL). Les deux dernières distributions permettant d'approcher des données plus dispersées ou asymétriques.

Pour chaque loi continue, nous adaptons la structure générale de l'algorithme DEM-MD afin d'estimer correctement les différents paramètres requis. Notamment pour la distribution de Laplace asymétrique non centrée, où nous intégrons un E-step intermédiaire pendant l'estimation des paramètres (M-step) au cours de l'algorithme, suivant l'algorithme du multicycle ECM (Meng and Rubin, 1993). Nous nous appuyons aussi sur le recuit déterministe (deterministic annealing), suivant la proposition initiale de Ueda and Nakano (1998), qui permet d'aplanir le profil de la vraisemblance, et ainsi une meilleure exploration au cours de l'optimisation des paramètres, évitant des puits potentiels trop proches de l'initialisation de l'algorithme. Dans le recuit déterministe considéré ici, c'est la densité conditionnelle $p(z|\theta)$ qui est élevée à la puissance $1/T$ avec T la température, au cours de l'E-step d'un algorithme EM. Cette densité tempérée des variables latentes est plus ambigüe, avec des modes plus faibles à chaque étape. Nous appuyant sur des schémas de température oscillants, on considère ici que $T \to 1$ quand le nombre d'itérations croît.

Grandissant en complexité avec l'utilisation de lois comme celle de Student ou Laplace asymétrique, l'estimation d'un modèle de mélange correspondant nécessite des données de taille suffisante. Avec l'explosion du nombre de features en apprentissage statistique ces dernières années, estimer des modèles sur des jeux de données de haute dimension est un sujet de recherche actif dans la communauté. Une proposition simple et intégrable rapidement dans nos algorithmes dynamiques est d'introduire d'autres régularisations sur les paramètres, avec ici des priors $p(\Sigma)$ sur les matrices de covariances Σ (présentes pour les trois distributions continues considérées), suivant le travail de Fraley and Raftery (2007).

Ces différents algorithmes pour données mixtes permettent ainsi d'estimer les paramètres d'un modèle de mélange de lois intégrant une famille de lois continues et une ou plusieurs familles de lois discrètes. Ils permettent d'estimer conjointement à ces paramètres le nombre de classes dans le mélange, et sont ainsi applicables facilement à de nombreuses problématiques où le nombre de clusters est inconnu et à estimer. Les algorithmes DEM-MD et EM-MD peuvent d'ailleurs être intégrés au pipeline proposé dans notre premier travail, comme composants permettant d'estimer les modèles à chaque temps t.

Chapter 1

Introduction

Contents

1 Spatio-temporal models

This thesis was initially interested in the conception of an actionable decision support system that could improve health management using spatio-temporal data, *i.e.*, the real-time identification of new or evolving groups of patients. This would be particularly useful to rapidly identify a new contamination source for a transmissible disease, as soon as the first affected cases are observed in health information systems.

In this first section, our pipeline proposal is motivated by a review of the literature on disease surveillance models. We present the different existing families, which do not always address the same underlying objectives but all use spatial and/or temporal data for public health challenges.

A spatial model is based on the characterization of individuals by their geographical location (place of birth, place at the time of diagnosis, place of residence, *etc.*). Taken together, these individuals form a population. Concurrently the temporal component is essential in disease monitoring. Therefore, it requires to consider that the population distribution evolves over time. The association of spatial and temporal components for a disease yields a spatio-temporal distribution.

1.1 Statistical tests

Space-time interaction tests Statistical tests on space-time interactions are mainly based on testing the number of disease cases under the null hypothesis of the absence of interaction. Cases are first grouped according to predefined criteria of spatial and temporal proximity. First works suffered from an arbitrary "closeness" definition (Knox and Bartlett, 1964) while latter approaches relied on neighbor relations such as k-Nearest Neighbor statistic (Jacquez, 1996). Space-time interaction tests only report presence or absence of space-time interaction, without information about trends nor taking into account background heterogeneity.

Cusum In prospective methods such as cumulative sum (cusum), data are case counts on fixed sub-areas of a wider study area. At each time period, a cumulative sum alarm statistic is computed, by summing over an all-time number of cases in the sub-area and defining a threshold/expected count over which excess cases are accumulated. Cusum is

originally a temporal framework, around which spatial-temporal statistic tests can be computed (Rogerson, 1997). Attractive for prospective disease surveillance, its transition to the multivariate case made it more difficult to interpret and required wide temporal simulation to specify a threshold (Robertson et al., 2010).

Scan statistic Scan statistic was originally developed for temporal clustering, and later on developed for spatial cases by scanning over a map of cases using circular searching areas of varying radii (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997). The spatio-temporal version was then proposed by Kulldorff et al. (1998). Scan statistic proposes to detect spatial and/or temporal clusters from aggregated data (discrete in space and time) using sliding windows to compare cases to reference populations. Space and time are analyzed exhaustively to search for statistically significant spatio-temporal clusters for a given tolerance and test. The spatial scan statistic selects the most likely cluster, defined by maximizing the probability that cases located in the search zone are part of a cluster compared with those located outside the zone. The significance test for this cluster is then assessed using Monte Carlo randomization. Different statistics were proposed, using a known underlying risk population, or cases/controls. In the absence of a population at risk, a method has been developed to estimate the expected number of cases in order to deduce a reference population and be able to continue applying the scanning statistics algorithms (Kulldorff et al., 2005). In both cases, these methods require to set several parameters on the sliding window under consideration, such as minimum area and minimum temporal size. In addition, case/control studies are subject, among other things, to costly selection and effort to find a suitable control group, and are not feasible in all situations (Elliott and Wartenberg, 2004). Thereby, the resulting comparison between cases and controls is exposed to false differences due to inadequate sampling of the control group (Sackett, 1979). Following this, Nakaya and Yano (2010) suggested combining statistical scan methods with the Space-Time Kernel Density Estimation (STKDE) method. It transforms a set of points into a density surface using specific kernel functions. Unfortunately, the STKDE method also requires the choice of a bandwidth parameter, and therefore the definition of a fine mesh before calculating the density estimate at each location (Silverman, 1986; Brunsdon et al., 2007). These last years, scan statistic methods were extended to multivariate functional data (Frévent et al., 2023), exploring the potential of these continuous-scale data.

1.2 Model-based approaches

Model-based approaches allow the integration of additional variables into the expectation of disease incidence. They can model the influence of variables such as age, gender, *etc.* on risk evaluation through space and time. In addition, they enable estimated parameters to be conserved, data to be explained and, in some cases, traced over time. Globally, as statistical tests presented previously, the main aim of these methods is to represent the risk or impact on predefined - or not - spatial areas. There are also methods based on longitudinal data that belong to the same family of methods (such as GLMM), although their objectives are completely different. We also present a few references, which provide a different perspective on what is known as a spatio-temporal method.

Generalized Linear Mixed Models Generalized linear mixed models (GLMM) offer a framework based on regression and make it easy to include a temporal aspect in the data. Generalized linear mixed models provide flexibility in the expected distribution of the response variable and link within variables. These models can include spatial variation in the underlying population at risk as random effects, relying on small temporal areas fixed in advance (Kleinman, 2004).

Bayesian models There exists a vast literature for risk mapping with Bayesian models (Louzada et al., 2020), which infer unknown area-specific risk. Relying on a variety of spatial priors, most of the methods are a subclass of Gaussian Markov Random Fields (GMRF). With observed data and a prior distribution, risk distribution is inferred and model parameters can be sampled for prediction. Several techniques have been developed for the spatial representation, such as conditional autoregressive model (CAR) for spatial autocorrelation (Abellan et al., 2008), Stochastic Partial Differential equation (SPDE) which represent a Gaussian random field as a solution to a stochastic partial differential equation (Lindgren et al., 2011; Clarotto et al., 2023), or *a priori* information on spatial components (Yan and Clayton, 2006). Non-Gaussian Markov Random fields are also part of this extended literature, with for example a discrete MRF on latent membership variables associated with estimation by variational EM (Forbes et al., 2013).

Hidden Markov Random Fields have been widely explored in the last decades, providing new hierarchical models for spatial data. For example, Green and Richardson (2002) proposed a semi-parametric approach with a Potts Model on spatial contiguity as a graph structure prior on the hidden random field, and combining several Markov Chains Monte Carlo (MCMC) methods for parameter estimation. There exists a wide spectrum of hierarchical models, and a major source of their differences is the way in which the different levels are defined. Temporal components are classically introduced as a temporal effect in the expression of case counts (Abellan et al., 2008), or can be seen as jointly defining a cluster membership with the space part as in the work by Yan and Clayton (2006), who considered a uniform prior for the temporal elements.

Most of these methods aim at clustering according to a risk level, predicting incidence or prevalence over a wide area. In contrast, Park et al. (2022) claim to offer a method that makes it possible to observe the temporal evolution of spatial clusters, being closer to our work. Developing a Dirichlet process Gaussian mixture model on a bounded space-time, their defined density function relies on rolling window parameters for each spatial and temporal cluster, as well as the neighbor impact. This also implies a finite temporal set at the time of parameter estimation by an MCMC method.

All these works underline that appropriate prior specification in Bayesian inference remains a challenge, as it is both attractive and limited by its introduced prior information. In addition, many methods need predefined set of areas and/or number of observations jointly with population size, as for the scan statistics. Additionally, these approaches are based on MCMC methods which often lead to heavy computation costs.

Functional data Briefly mentioned in statistical scanning methods, the study of functional data is of growing interest. By definition, this includes a temporal aspect, since each observation is continuous in time. Taking into account spatial information has been the objective of several recent works (Cheam, Marbac, and McNicholas, 2017; Cheam, Fredette, Marbac, and Navarro, 2023). Combining time series and mixture models, Cheam, Marbac, and McNicholas (2017) propose a mixture model of autoregressive polynomial regression mixture with spatial and temporal dependent logistic weights. Cheam, Fredette, Marbac, and Navarro (2023)'s work takes another look at the clustering of functional data: it studies Covid-19 weekly death rates with one time series per specific area of the world, which may arrive time-translated. In this work, a death curve, through feature extraction, gives new variables which are regressed on the population's risk factors and then finally the residuals are clustered using a non-parametric mixture.

Remarks While we tried to present an overview of this rich literature on spatio-temporal diseases, several limitations were underlined, and we do not necessarily retrieve the same objectives. We refer the reader to four interesting and extensive review papers on spatial

statistics and epidemiology (Elliott and Wartenberg, 2004; Robertson et al., 2010; Kirby et al., 2017; Louzada et al., 2020). In this thesis, we have chosen to focus on a population modeling problem, rather than on a retrospective or prospective assessment of pure risk/incidence. Numerous methods have been proposed to achieve these goals, as detailed above, using statistical tests or Bayesian models based, for instance, on Random Markov fields. Our objectives are to infer models on data coming independently at each time step. It is important to keep the estimated parameters so that we can draw interpretations from them, particularly with the possible presence of covariates, and plot their evolution over time. Our spatio-temporal process developed in Chapter 3 was developed in order to be a decision-support tool.

2 Strategies for mixed-type data

In the second contribution of this thesis, we are leaving aside the time monitoring pipeline to focus on the development of new algorithms for mixture models. These mixture models are designed for mixed-type data. However, we are keeping in mind the possibility of combining the spatio-temporal mixture process developed in Chapter 3 with mixed-type data algorithms to estimate mixture models at each time step. In addition to spatial and temporal information, many datasets incorporate additional variables that can be informative, and in particular, be used to establish clusters. This leads us to consider in Chapter 4 extensions of Chapter 3 models for mixed-type data. This section reviews the issues associated with mixed-type data and the different families of methods for estimating them.

Mixed datasets are ubiquitous in many disciplines. With the era of so-called 'big data', the availability of datasets composed of heterogeneous data sources and types will continue to increase. Statistical analysis of mixed-type data is therefore still a hot topic, whether for clustering, inference or dimension reduction. Several types of strategies can be considered to deal with this type of data (Foss et al., 2019; Ahmad and Khan, 2019). As there exists no reference distribution for mixed-type data, the main difficulty lies in modeling inter-variable interactions. An easy approach is to rely on conditional independence between all variables, although it can lead to biases.

Transformation of variables A first way of dealing with mixed-type data is to reduce the set of variables to a unique discrete or continuous space. Of course, these native transformations do not respect the nature of each variable, or change the domain to which it belongs, which can make interpretation or post-processing complex. These methods fall into two types: discretization of continuous variables, which allows the use of categorical data methods (Goodman, 1974; Huang, 1997b), or numerical coding of discrete variables (McCane and Albert, 2008). In this case, the possibilities go from replacing a level by a median to dummy coding, and to more complex methods like copulas. Copulas are multivariate cumulative functions with uniform marginals possessing interesting flexibility. Copulas can be used on mixed-type variables, under the guise of defining the dependency structure between continuous and discrete variables (Smith and Khaled, 2012; Murray et al., 2013). A source of difficulty with copulas lies in the lack of uniqueness and identifiability. In fact, if at least one marginal is not continuous, the copula is not unique. Moreover, the overall joint multivariate distribution can be difficult to understand as marginals and copula are specified separately.

Hybrid distances A second category, in our view, comprises hybrid methods. A popular hybrid distance is Gower's distance, combining relative absolute difference for continuous

variables and indicators for categorical variables. It is used for example in combination with the partitioning around medoids (PAM) method, or euclidean distance and fixed categorical weights in a k-prototypes algorithm (Huang, 1997a, 1998). A frequent limitation of these hybrid distances is the need to properly select weights dictating the relative contribution of each of the variables. Semi-parametric approaches were also proposed such as KAMILA (Foss et al., 2016), providing greater flexibility with adequate considered laws on each type of variable. But depending on the discrete laws under consideration, their implementation can quickly become complicated, and they rely on hard assignment through partition steps as in k-means method. Another type of hybrid distance method, in the sense that distances obtained on different variables are combined, is spectral clustering (Mbuga and Tortora, 2021). It proposes k-means clustering on an eigenvalue decomposition after combining continuous and discrete dissimilarities. Spectral clustering for mixed-type data requires tuning continuous/nominal weight and kernel parameters, but the main limitation is decomposition of sample size matrices. However, this method brings flexibility to non-graph data clustering and is solved by linear algebra methods.

Mixture models Among the direct approaches, mixture models are efficient in mixed-type data contexts because they can produce generative models, take into account many types of data, manage dependencies between and within variables, and capture a wide range of scenarios. Different approaches exist to provide meaningful mixture models on mixed-type data without data space transformation in the best-case scenario. A first approach is the location mixture model (Krzanowski, 1993), which models all categorical variables into multinomial variable. Then continuous variables follow a Gaussian distribution conditionally on this modality variable. A limitation is the number of combinations increasing exponentially with the number of levels and variables. Moreover, these models can lack of identifiability without constraints on some parameters, as proved for the mixture of location models (Willse and Boik, 1999). An important property often required for identifiability is local independence, and this is a main assumption in many works as it allows the definition of complete models with different types of variables, and generally simple estimation formulas. However, considering local independence between all variables would lead to a model with one-dimensional distributions, no intraclass dependencies, and biases. The idea is rather to consider the continuous and categorical variables independent given the class memberships, and also within the categorical variables, as in the normal-multinomial mixture model (Hunt and Jorgensen, 1996; Fraley and Raftery, 2002). Of course, these assumptions are open to question, as there is a lack of intraclass dependencies. Dependencies between continuous and categorical variables inside mixture models can be done by factor analyzers (McParland et al., 2014, 2017). Nevertheless, this relies on numerical coding of categorical variables with a flexible covariance structure, as one-type methods above. Additionally, this increases a degeneracy risk and can lead to expensive computations. Another approach is to combine mixture models and copulas mentioned above (Kosmidis and Karlis, 2016; Marbac et al., 2017; Sahin and Czado, 2022), leading to full parametrization of each marginal and overall dependencies, but also with an exploding number of parameters, requiring additional constraints on models (homoscedasticity, truncation, ...).

Still in mixture models, the latent variable mixture model models categorical variables as realizations of continuous latent variables according to threshold values. This model for categorical variables, named latent trait model, is associated in the work of Browne and McNicholas (2012) with a factor model for continuous variables. Whereas McParland and Gormley (2016) developed an overall factor analyzer structure for the whole latent model which includes continuous variables. In the end, the assumption of local independence is maintained to separate groups of variables of different types (continuous, ordinal, nominal) within the latent model. Ultimately, the latent variable models presented here are also

based on the transformation of discrete variables into a continuous space.

Linear mixed models, considered for mixed-type data with longitudinal data and random effects, can also be integrated within mixture models when homogeneous regression relationship across subjects is violated (Celeux et al., 2005; Proust-Lima and Jacqmin-Gadda, 2005; Bai et al., 2016; Lee and Chen, 2019).

Mixed data is best modeled using the raw data, without any prior transformation to a different space, without loss of information. But it remains a challenge like we saw above, and so far mixture models, whatever the internal structure, are a good way of obtaining an interpretable, generative and flexible representation of mixed-type data.

In the next section, we will look at the challenges of estimating mixture models, whether using continuous, mixed, spatial or temporal data.

3 Challenges on estimation of mixture models

As we will be going into more detail later on about some algorithms for estimating mixture models, we will briefly mention here the scope of the different estimation methods and the associated challenges. Mixture models are probabilistic models assuming that data points are generated from a mixture of several (classical) semi-parametric or parametric distributions. Finite mixture models, which therefore have a fixed number of distributions, are part of hierarchical models.

The Expectation-Maximization algorithm Dempster et al. (1977) developed the well-known Expectation-Maximization (EM) algorithm to estimate mixture models with unknown latent variables representing the component memberships. This algorithm relies on the optimization of the model likelihood, but it only ensures convergence towards a local minimum (under regularity conditions) (Wu, 1983). This makes it very dependent on the initialization step (Baudry and Celeux, 2015). In addition, the algorithm can reach the boundaries of the parameter space, leading to errors on certain types of estimated parameters.

MCMC methods With Bayesian mixture models, sampling methods such as Markov Chain Monte Carlo (MCMC) can be used. Although EM algorithms can still be used in some cases, changing from maximum likelihood estimate (MLE) to maximum *a posteriori* (MAP) on regularized log-likelihood (Fraley and Raftery, 2007; Ciuperca et al., 2003), MCMC methods present a different perspective on mixture models. The inclusion of prior introduces a smoothing effect on the mixture likelihood function and reduces the risk of obtaining incorrect modes. Moreover, MCMC methods may be more efficient in cases with small datasets, or extremely unbalanced component proportions. Common methods for Bayesian inference on mixture models with a finite number of components are Gibbs sampling (Escobar and West, 1995) and Metropolis-Hastings algorithm (Neal, 2000). However, MCMC methods usually converge at a high computational cost and encounter difficulties with multimodal posterior distributions. We refer the reader to the book of Frühwirth-Schnatter (2006) for detailed explanations of MCMC methods on mixture models.

Associate EM algorithms and MCMC methods In some situations, for complex hierarchical models, the E-step of the EM algorithm may be infeasible in closed form. An idea was to sample the unobserved data (latent variables) from its marginal posterior, and stochastically approximate the objective function, this is the SAEM algorithm (Delyon et al., 1999). But for most nonlinear models, the unknown data (latent) cannot be simulated under its conditional distribution. Thus, an additional consideration was to insert MCMC

steps inside the SAEM algorithm, this is the MCMC-SAEM algorithm (Kuhn and Lavielle, 2004; Allassonnière and Kuhn, 2010).

An unknown number of components For EM algorithms as for MCMC methods, choosing the number of components is a main challenge. When they serve objectives in an unsupervised setting, which is often the case, estimating the number of mixture components is also an important consideration. A wide variety of propositions have been developed for EM algorithms, starting with model selection criteria such as well-known BIC (Schwarz, 1978), ICL (Biernacki et al., 1998) or slope heuristic (Birgé and Massart, 2007), for an *a posteriori* model selection. Another group of propositions for model selection are regularizations of estimated parameters, which require defining priors on some parameters (Figueiredo and Jain, 2002; Wang et al., 2011; Yang et al., 2012). Finally, there exist split-and-merge algorithms, which, as their name suggests, dynamically explore parameter space by forcing components to split or merge (Wang et al., 2004; Zhang et al., 2004). An unknown number of components is also a difficulty for MCMC methods on mixtures, which has to reach modes in spaces of different dimensions. A popular trans-dimensional MCMC method is the reversible jump Metropoplis-Hastings (Green, 1995; Richardson and Green, 1997). An alternative to jump processes is proposed by Stephens (2000) with an application of a Markov birth-and-death process to select the number of components.

The various mixture estimation methods face several problems, in particular the choice of the number of components, which in real applications is a complex issue. Additionally, existing methods for mixed-type data also face this challenge when modeling with mixture models. Existing dynamic algorithms, developed only for Gaussian mixture models, are not completely robust to initialization or overfitting.

4 Outline of this thesis

In this thesis, we will propose dynamic mixture models, in an attempt to avoid the pitfalls of certain EM algorithms. In particular, we first focus on models for Gaussian mixtures, integrated into a temporal pipeline that we are also designing with the aim of moving toward a decision support tool. Given the diversity of real data, whatever the domain, which generally includes discrete and categorical data in addition to continuous data, we then propose dynamic mixture models for mixed data. Without losing sight of the fact that these new models can take their place in the previously developed pipeline. The remainder of this thesis is organized into three chapters, as follows:

- Chapter 2 *Background on mixture models and inference algorithms*

  This chapter is divided into two sections, in which some statistical and algorithmic notions about mixture models are recalled since they form the basis of our proposals. The first section provides definitions of mixture models and existing models on different continuous laws that will be considered in the next chapters. The second section contains tools for inference in mixture models, with a reminder of the EM algorithm, then of variants and other aspects of the estimation such as model selection, deterministic annealing and stopping criteria.

- Chapter 3 *Spatio-temporal mixture process estimation to detect dynamic changes in population*

  This chapter proposes a pipeline to model and monitor population distributions over space and time, in order to build an alert system for spatio-temporal data changes.

It relies on mixture models to represent populations. In addition to the pipeline, we are proposing a correct dynamic algorithm for better joint estimation of the number of clusters and the model parameters, and in particular to avoid overfitting problems. This algorithm is compared to existing methods on several simulated datasets. Then, combined with the temporal statistical pipeline, it allows the detection of changes in population distributions, and we call the result a spatio-temporal mixture process (STMP). We test STMPs on synthetic data, and consider several behaviors of the distributions, to fit this process. Finally, we validate STMPs on a real dataset of positive diagnosed patients with coronavirus disease 2019. We show that our pipeline correctly models evolving real data and detects epidemic changes.

This work was published in the journal Artificial Intelligence in Medicine (Pruilh et al., 2022).

- Chapter 4 *Dynamic Expectation-Maximization Algorithms for Mixed-type Data*

  This chapter proposes methodological developments of new mixture models on mixed-type data. Component distributions of the continuous variables can be either Gaussian, Student or Shifted Asymmetric Laplace and categorical variables can be distributed according to Bernoulli, Multinomial or Poisson distributions. Relying on conditional independence between continuous and discrete variables, the joint estimation of the number of classes and model parameters is carried out by EM-type algorithms that we have adapted to perform on these mixed-type data. We show that our different dynamic algorithms allow us to reach the real number of classes, and to properly estimate all the parameters. We also highlight the benefits of introducing regularization to improve performance in situations where the sample size is insufficient for the complexity of the model. Our different models are then tested on real datasets from the literature, assessing that the objective of jointly estimating the number of components and the model parameters is achieved.

Chapter 2

Background on mixture models and inference algorithms

This chapter presents generalities on mixture models, estimation methods for these models, and it introduces particular developments that will serve in the following works. It also provides reminders about model selection.

Contents

1 Generalities on mixture models and continuous distributions

In a finite mixture model, individuals are independently drawn from the same distribution, and we consider here component densities belonging to the same parametric family.

Definition. *The probability density function of a finite mixture distribution of a g-dimensional random vector $\mathbf{X}$ takes the form*

$$p(\boldsymbol{x};\Theta) = \sum_{k=1}^{K} \pi_k p_g(\boldsymbol{x};\theta_k)\,, \tag{2.1}$$

where the mixing proportions π_k are positive and sum to one, $p_g(\boldsymbol{x};\cdot)$ are g-dimensional component densities, where $\Theta = (\xi^\top, \pi_1, \ldots, \pi_{K-1})^\top$ defines the vector of unknown parameters and ξ consists of the elements of all the θ_k, known a priori to be distinct.

In this simple definition, variables $\mathbf{X}$ can be either continuous or discrete. Different mixture models were developed to adapt to the nature of considered data, continuous or categorical. We will briefly recall here three continuous families, and how they relate to more general continuous families. In Chapter 3 we will focus on Gaussian mixture models while in Chapter 4 we will develop mixture models on mixed-type data, including the continuous laws briefly described here.

An incomplete model We consider n random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ where $\mathbf{X}_i$ is a g-dimensional random vector with probability density function (pdf) given by Eq.(2.1). Observed values on the sample are noted $\boldsymbol{x}_{obs} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$. We are working in an unsupervised context, meaning that component labels are unknown, and the number of components K is also finite and unknown.

We now write mixture model definitions for several continuous distributions, including the Gaussian distribution, which is the most widely used. These continuous distributions are then detailed and used in Chapter 3 or Chapter 4.

Gaussian Mixture Model

The Gaussian mixture model, first explicitly mentioned in Pearson (1894), is the most popular mixture model, thanks to several advantages including tractability, flexibility, and interpretability. Knowing the component assignment $\mathbf{Z}^k$, each random variable $\mathbf{X}_i | Z_i^k = 1$ is a g-dimensional Gaussian variable with center $\boldsymbol{\mu}_k \in \mathbb{R}^g$ and positive definite matrix $\boldsymbol{\Sigma_k} \in \mathbb{R}^{g \times g}$.

Definition. *The pdf of a sample $\boldsymbol{x}_i$ from a Gaussian mixture model is given by*

$$p_g(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k \phi_g(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,,$$

where $\phi_g(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}(2\pi)^{g/2}} \exp(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k))$.

Remark The Gaussian mixture model has $(K-1) + Kg + Kg(g+1)/2$ parameters.

Student Mixture Model

In presence of outliers in the data, Student mixture models are a great alternative to Gaussian mixture models, allowing to control the thickness of the tails through an additional parameter: the degrees of freedom (McLachlan and Peel, 1998; Peel and Mclachlan, 2000).

A random variable $\mathbf{X}_i | Z_i^k = 1$ is a g-dimensional Student variable with center $\boldsymbol{\mu}_k \in \mathbb{R}^g$, positive definite matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{g \times g}$ and degrees of freedom $\nu_k \in (0; \infty]$.

Definition. *The pdf of a sample $\boldsymbol{x}_i$ from a Student mixture model is given by Eq.(2.1) with*

$$p_g(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) = \frac{\Gamma(\frac{\nu_k + g}{2})|\boldsymbol{\Sigma}_k|^{-1/2}}{(\pi\nu_k)^{g/2}\Gamma(\frac{\nu_k}{2})\{1 + \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)/\nu_k\}^{\frac{(\nu_k+g)}{2}}}\,,$$

with $\delta(\boldsymbol{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ the Mahalanobis distance.

Remark The Student mixture model has $2K + Kg + Kg(g+1)/2 - 1$ parameters.

Shifted Asymmetric Laplace Mixture Model

In many practical problems, the data implies highly asymmetric distributions as well as thicker tails than those of normal distributions. Skew distributions are a better way to model these data than normal or even t-distributions. The multivariate asymmetric Laplace distribution, which is a skew distribution, was first proposed by Kotz et al. (2001). However, the asymmetric Laplace distribution does not define a center (or shift) parameter. If considered in a mixture model, all distributions should have the same origin. To overcome

this problem, Franczak et al. (2014) introduced a shift parameter $\boldsymbol{\mu} \in \mathbb{R}^g$, and proposed to use the resulting Shifted Asymmetric Laplace distribution in mixture models. A random variable $\mathbf{X}_i | Z_i^k = 1$ is a g-dimensional Shifted Asymmetric Laplace (SAL) variable with center $\boldsymbol{\mu}_k \in \mathbb{R}^g$, positive definite matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{g \times g}$ and skewness $\boldsymbol{\alpha}_k \in \mathbb{R}^g$.

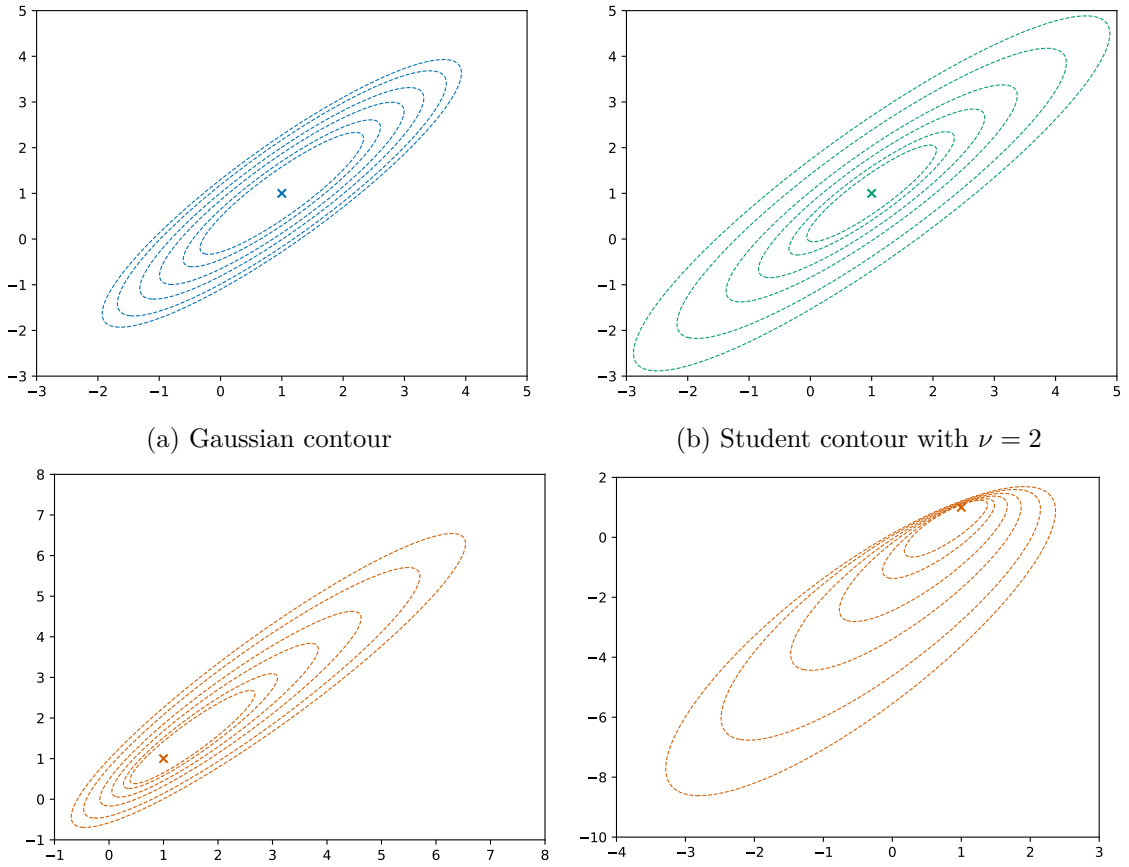
Definition. *The pdf of a sample $\boldsymbol{x}_i$ from a SAL mixture model is given by Eq.(2.1) where*

$$p_g(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k) = \frac{2 \exp\left\{(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} \boldsymbol{\alpha_k}\right\}}{(2\pi)^{g/2} |\boldsymbol{\Sigma}_k|^{1/2}} \times \left(\frac{\delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)}{2 + \boldsymbol{\alpha}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\alpha}_k} \right)^{\nu/2} K_\nu(u)\,,$$

with $u = \sqrt{(2 + \boldsymbol{\alpha}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\alpha}_k)\delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)}$, K_ν modified Bessel function of third kind, with index $\nu = \frac{2-g}{2}$ and δ the Mahalanobis distance.

Remark The Shifted Asymmetric Laplace mixture model has $K + 2Kg + Kg(g+1)/2 - 1$ parameters.

Representation Figure 2.1 depicts contour plots for Gaussian, Student and two SAL distributions, showing the differences introduced by skewness in SAL distributions, and degrees of freedom in Student distribution.



(a) Gaussian contour

(b) Student contour with $\nu = 2$

(c) SAL contour with $\alpha = (1, 1)$.

(d) SAL contour with $\alpha = (-1, -3)$.

Figure 2.1: Bivariate contours of Gaussian, Student or SAL distributions, with $\mu = (1, 1)$, $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1. \end{pmatrix}$ and additional parameters ν or α.

A more general family There exist other skewed distributions such as the skew-normal distribution (Azzalini, 1985; Azzalini and Valle, 1996), the skew t-distribution (Azzalini and Capitanio, 2003; Jones and Faddy, 2003), normal inverse Gaussian distribution (Karlis and Santourian, 2009). Skew distributions but also Gaussian and Student distributions are limiting cases of the Generalized Hyperbolic distribution (McNeil et al., 2015, Chapter 15, page 559), for which the probability density function is the following:

$$\begin{aligned} p_g(x;\theta) = & \left[\frac{\chi + \delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{(\lambda-g/2)/2} \\ & \times \frac{[\psi/\chi]^{\lambda/2} K_{\lambda-g/2}\left(\sqrt{[\psi + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}]\,[\chi + \delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma})]}\right)}{(2\pi)^{g/2}|\boldsymbol{\Sigma}|^{1/2} K_\lambda(\sqrt{\chi\psi})\exp\{(\boldsymbol{\mu}-\boldsymbol{x})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}, \end{aligned} \tag{2.2}$$

with $\delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma}) = (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and K_λ is the modified Bessel function of the third kind with index λ.

Mixtures of Generalized Hyperbolic distributions were proposed later by Browne and McNicholas (2015). We refer the reader to their work and to McNeil et al. (2015)'s one for a detailed description of this distribution. We just include on Figure 2.2 a graphical representation of the Generalized Inverse Hyperbolic distribution as parameterized in Eq.(2.2) (as defined in the paper of Browne and McNicholas (2015)) and its limiting cases, where the distributions of particular interest are highlighted. The Asymmetric Laplace and Laplace distributions are covered by models built for the Shifted Asymmetric Laplace distribution and the Cauchy distribution derives from the Student distribution. However, Cauchy mixture models require additional considerations when estimated (Kalantan and Einbeck, 2019).

2 Estimation of mixture models

In this unsupervised context, a probabilistic clustering of data $\boldsymbol{x}_{obs}$ into K clusters can be obtained by finding the marginal probabilities of component memberships $p(z_i|\theta_k)$. An *a posteriori* hard clustering of observations is done by assigning each x_i to the group k which has the highest posterior probability $k_i = \text{argmax}_j\, p(z_i^j = 1|x_i,\theta_j)$. We now present the Expectation-Maximization algorithm and some variants, that will allow us to estimate in a first work the mixture models of STMP, and to propose in a second work the Dynamic EM for Mixed-type Data algorithms.

2.1 The Expectation-Maximization algorithm and its limitations

The Expectation-Maximization (EM) algorithm was first introduced by Dempster et al. (1977). They proposed a general iterative optimization method to compute maximum likelihood estimates from incomplete data observations. Latent variable models express the dependency of observations x as a function of latent variables z, which are unobserved, to infer unknown information (such as class membership for example). We define $p(\theta)$ prior distribution on the parameters θ, $p(z|\theta)$ distribution of latent variables given the parameters, $p(x|z,\theta)$ distribution of data given the latent variables and the parameters and $p(x,z|\theta)$ likelihood of complete data. In the case of models with latent variables, the incomplete data density writes as an integral over the space of latent variables:

$$p(x|\theta) = \int p(x|z,\theta)p(z|\theta)dz\,.$$

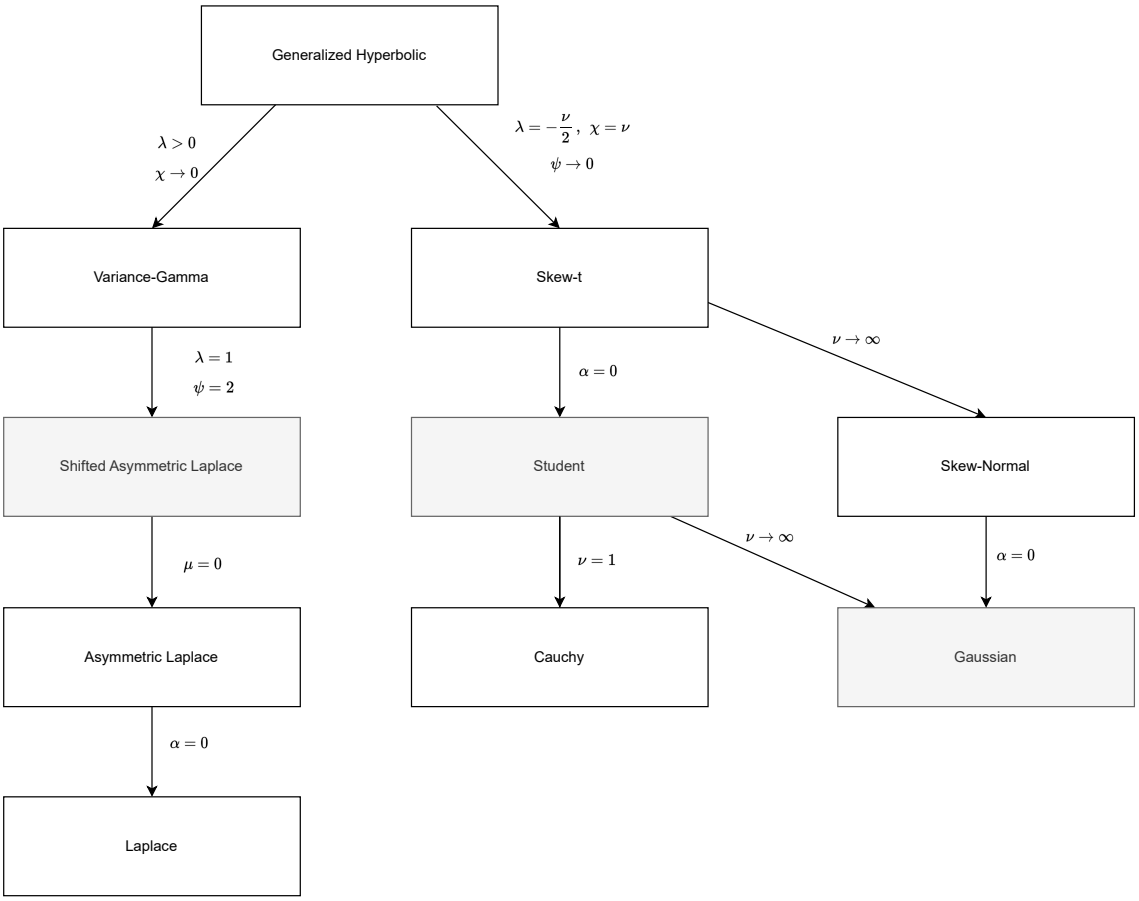


Figure 2.2: Generalized Hyperbolic distribution and some special and limiting cases, with gray boxes for the ones considered in this thesis.

The Expectation-Maximization algorithm of Dempster et al. (1977) is based on the introduction of an auxiliary function Q that relies on current estimate θ' to estimate θ

$$Q(\theta|\theta') = \mathbb{E}_{p(z|x,\theta')} \left[\log\, p(x, z|\theta)\right] .$$

The computation of Q allows summing over logarithms of probabilities, which is more convenient than computing $p(x|\theta)$. A key principle of using this Q function is that increasing Q leads to increasing $p(x|\theta)$, which makes it possible to build a sequence of parameters maximizing Q such as $p(x|\theta)$ is non-decreasing.

The general form of an EM algorithm is given in Algorithm 2.1. We refer the reader to the book by McLachlan and Krishnan (2008) for details on the EM algorithm and some of its variants.

Under fairly general conditions, the likelihood values of an EM algorithm (and variants) converge to stationary values (see McLachlan and Krishnan, 2008, Chapter 3). We refer the reader to the works of Wu (1983) and Delyon et al. (1999), who studied the convergence properties of the EM algorithm.

However, it is frequent that the posterior distribution $p(z|x, \theta)$ does not correspond to a classical distribution, and the computation of Q is intractable. The E-step can be approximated with Monte-Carlo sampling and Q function aggregations, as in the Stochastic Approximation EM (Delyon et al., 1999). Conversely, when the M-step is not achievable in closed form, one can replace the M-step with Newton's method (Lange, 1995).

Algorithm 2.1: Expectation-Maximization algorithm

Initialization : Initial parameters values θ^0
for $1 \leq t \leq T$ **do**
  E-step: with current values of parameters $\theta^{(t-1)}$, compute expectation

$$Q(\theta|\theta^{(t-1)}) = \mathbb{E}_{p(z|x,\theta^{(t-1)})}\left[\ell(x,z|\theta)\right]$$

  M-step: Find $\theta^{(t)} = \text{argmax}_{\theta\in\Theta} Q(\theta|\theta^{(t-1)})$
end

Application to mixture models For the estimation of mixture models, the considered density function is $p(\boldsymbol{x};\Theta)$ given by Eq.(2.1). The complete log-likelihood is given by

$$\ell(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_i^k[\log(\pi_k p_g(x_i|\theta_k))]\,. \tag{2.3}$$

Computing $Q(\cdot|\cdot)$, the expectation of the log-likelihood of the complete data, requires the expression of latent probabilities, which are computed at each E-step of an EM algorithm, with

$$\begin{aligned} p_{\theta^{(t)}}(z_i^k = 1|\boldsymbol{x}_i) &= \frac{\pi_k^{(t)} p_g(\boldsymbol{x}_i|\theta_k^{(t)})}{\sum_{j=1}^{K}\pi_j^{(t)} p_g(\boldsymbol{x}_i|\theta_j^{(t)})} \\ &= \tau_{ik}(\theta^{(t)})\,. \end{aligned} \tag{2.4}$$

Maximum A Posteriori The Maximum A Posteriori (MAP) estimate is the Bayesian equivalent of the MLE and only differs by the presence of a prior term $p(\theta)$, which gives the following estimate

$$\hat{\theta}_{\text{MAP}} = \text{argmax}_\theta\; p(x|\theta)\,p(\theta)\,.$$

The impact on the EM algorithm is minimal: E-step remains unchanged and M-step is modified by addition in Q objective function of the log of prior density.

Initialization A known limit of the EM algorithm is its sensitivity to starting values. Common solutions are: using one random initialization, using multiple random initial positions and choosing the final estimate with the highest likelihood from these runs (Biernacki et al., 2003), using a clustering algorithm like k-means to initialize the clusters, recursive initialization (Baudry and Celeux, 2015) or deterministic annealing (Ueda and Nakano, 1994, 1998) described later on in Subsection 2.4.

Convergence to degenerate local maxima The EM algorithm may converge to boundaries of the parameter space, for example when the proportion of a mixture component approaches zero, and the corresponding covariance matrix becomes close to singular. Overfitting problems lead to this type of behavior when the number of components is larger than the optimal one. Alternatively, the algorithm may also find two classes with equal centers and covariances, which leads to an identifiability problem. Degenerate solutions due to unbounded likelihood can be avoided by adding for example penalty functions, as in the work of Chen and Tan (2009). We will see in Subsection 2.3 other priors and regularizations introduced to try to solve this problem jointly with an estimation of the number of classes K. In Subsection 2.4 we will describe deterministic annealing, which, by flattening the objective function, also tries to avoid the boundaries of the parameter space while allowing to explore the multiple modes of the objective function.

2.2 Select the number of classes in a mixture model

This subsection provides a brief reminder of the model selection criteria, important for mixture models where one main objective is to choose the number of components and intrinsically choose a model. In forthcoming works (Chapters 3 and 4), we will be comparing the performance of our dynamic estimation of components against these selection criteria.

The goal of information criteria is to select the best model m from a collection of models $\mathcal{S} = \{1, \dots, \mathcal{M}\}$. Considering only the log-likelihood $\log p(x|\theta)$ favors complex models, which are prone to overfitting.

The Akaike Information Criterion A first well-known criterion is the Akaike Information Criterion (AIC) (Akaike, 1973). Information criteria in the literature are generally expressed in terms of lack-of-fit and a penalty term that measures the complexity of the model. The AIC selects the model that minimizes

$$\mathrm{AIC}(\theta) = -2 \log p(x|\theta) + 2 \dim(\theta) . \tag{2.5}$$

The AIC assumes that the dataset size is large enough. Many authors (for example Koehler and Murphree (1988)) observed that AIC tends to overfit models, and thus to overestimate the correct number of components in a mixture model.

The Bayesian Information Criterion Some criteria have been derived within a Bayesian framework but can be applied in a non-Bayesian framework. That is the case of the Bayesian Information Criterion (BIC), introduced by Schwarz (1978). Schwarz studied the asymptotic behavior of Bayes estimators and proposed the following criterion

$$\mathrm{BIC}(\theta) = -2 \log p(x|\theta) + \dim(\theta) \log(n) . \tag{2.6}$$

The BIC is interesting because, *if* the true model is in the set $\mathcal{S}$, the BIC will select it consistently for $n \to \infty$. The BIC penalizes heavily complex models compared to the AIC. But it tends to select too low-dimensional models for a small n, even if the correct model is in the collection $\mathcal{S}$. Conversely, if the correct model is not in the family of models, the BIC is not convincing as it is also prone to overfitting, regardless of cluster separation as observed by Biernacki et al. (1998).

Classification-based information criteria Finally, classification-based criteria aim at finding the model with the greatest evidence of clustering. These criteria consider the complete likelihood within the EM framework for the fitting of a mixture model. Considering mixture models, the log-likelihood can be expressed as

$$\log p(x|\theta) = \log p(x, z|\theta) - \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log(\tau_{ik}) ,$$

and $-\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log(\tau_{ik})$ is the estimated entropy of posterior probabilities of the component memberships, $\mathrm{EN}(\tau)$. A first classification criterion is given by

$$-2 \log p(x|\theta) + 2\mathrm{EN}(\tau) .$$

This criterion favors well-separated clusters but does not perform well when the proportions are not constrained to be equal according to Biernacki et al. (1999).

Integrated Complete Likelihood Criterion Biernacki et al. (1998) proposed the Integrated Complete Likelihood (ICL) criterion to overcome the shortcomings of the precedent and BIC criteria.

The ICL criterion to minimize is defined as

$$\log p(x, z) = \log \int_{\theta \in \Theta} p(x, z|\theta) p(\theta) d\theta \,.$$

An approximation of the ICL criterion to minimize has the following expression,

$$\mathrm{ICL}_{\mathrm{BIC}}(\theta) = -2 \log p(x|\theta) + 2\mathrm{EN}(\tau) + \dim(\theta) \log(n) \,. \tag{2.7}$$

This approximation resembles the BIC with the addition of the estimated entropy of posterior probabilities and is satisfying enough compared to the exact ICL according to Biernacki et al. (1998). The ICL criterion favors well-separated clusters contrary to the BIC.

Normalized Entropy Criterion Celeux and Soromenho (1996) proposed a criterion more appropriate for standard cluster analysis in their own words (and improved by Biernacki et al. (1999)). Their criterion uses the estimated entropy $\mathrm{EN}(\tau)$ as a criterion for choosing the number of components. This criterion, named the Normalized Entropy Criterion (NEC), is given by

$$\mathrm{NEC}(\theta, K) = \frac{\mathrm{EN}(\tau, K)}{L(K) - L(1)} \,, \tag{2.8}$$

with $L(K) = \log p(x|\theta)$ for a model with K components and $L(1)$ is the log-likelihood of a model with a single component. However, the NEC favors models with well-separated, non-overlapping clusters. But it remains relevant when comparing several possible partitions.

Remark We do not claim to provide an exhaustive list of model selection criteria, as there are many of them, for a variety of objectives and model types. Rather, we presented some of the most widely used criteria, as we will compare our performances to them in Chapter 3 and 4. We refer the reader to the book of Konishi and Kitagawa (2008) for complete explanations of the concepts on mentioned above criteria and their related criteria.

2.3 Existing parameter priors in EM algorithms

As mentioned above, finding the optimal number of classes K in mixture models is a difficult challenge. The criteria defined above can be characterized as deterministic methods, relying on the estimation of a set of candidate models $\mathcal{S}$ which is often assumed to contain the true/optimal K. Although some of these methods perform well, they still require the estimation of a set of models, and there is still some uncertainty on whether this set contains the optimal model or a correct generalization of the data. In addition, they are more prone to overfitting and parameter space boundary problems and still require careful initialization.

For mixture models, model selection is roughly the same problem as choosing the number of classes. Finding the "best" overall model is a different approach, adopted in several works (Figueiredo and Jain, 2002; Yang et al., 2012). Relying on information theory, these methods associate measures of information with explicit (and automated) rules for moving from a number of components K to a smaller one, and therefore avoiding the limits of the parameter space.

Minimum message length In this field, an important work by Figueiredo and Jain (2002) introduced a criterion based on minimum message length, described below. In addition, a component annihilation part was also introduced, which inspired Yang et al. (2012), because Figueiredo and Jain (2002)'s optimized function is equivalent to using a negative Dirichlet prior for proportions, and as negative Dirichlet prior encourages proportions to be zero or one, annihilation is required.

The minimum message length criterion (and other minimum encoding criteria) is based on the fact that if the data is encoded with the minimum, we obtain a good data generation model (Wallace and Freeman, 1987; Wallace and Dowe, 1999). With a few assumptions, in particular on the parameter priors, Figueiredo and Jain (2002) finally minimizes the following cost function

$$\mathcal{L}(\theta) = \frac{\dim(\theta_k)}{2} \sum_{k=1:\pi_k>0}^{K_{nz}} \log\left(\frac{n\pi_k}{12}\right) + \frac{K_{nz}}{2}\log\left(\frac{n}{12}\right) + \frac{K_{nz}(\dim(\theta_k)+1)}{2} - \log\, p(x|\theta)\,.$$

Their criterion considers full information of the mixture, unlike the works of Yang et al. (2012) and Brand (1999) which restrict to proportions' information.

Figueiredo and Jain (2002) observed the possibility of a failure mode if K is too large, which leads to the situation where no component has enough initial support. Therefore, they decided to update the proportions sequentially (as well as the other component parameters), relying on the Component-Wise EM for mixtures (Celeux et al., 2001). This algorithm avoids the initial support problem, but the choice of update order may affect the final result since the objective function has multiple local minima. In addition, their algorithm, although automatically reducing the number of components, still has to fix a k_{max} and k_{min} number of components. Their algorithm also forces exploration until $\hat{k} = k_{min}$. Their MML criterion penalizes uniform distributions, thus being related to minimum entropy priors (Brand, 1999). As they rely on a "strong" prior, they have difficulties when the components have very different weights.

Penalize with the entropy of proportions Yang et al. (2012) have further developed the dynamic search for the best number of components in a mixture model, with a proposed algorithm named Robust EM. They adjusted the EM mixture objective function by adding a criterion based on the entropy of the mixture proportions $\boldsymbol{\pi}$. Non-informative proportions are given by a high entropy. Consequently, the penalty added to the likelihood is given by the opposite entropy. The objective function to maximize in the M-step with this entropy-based penalty is, therefore:

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} log(\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)) + \beta \sum_{i=1}^{n}\sum_{k=1}^{K} \pi_k log\, \pi_k, \text{with } \beta \geq 0\,. \tag{2.9}$$

With this new criterion to maximize, the updated equation of components proportions $\boldsymbol{\pi}$ inside the EM algorithm becomes:

$$\hat{\pi}_k^{(t)} = \hat{\pi}_{k,EM} + \beta\hat{\pi}_k^{(t-1)}\left(\log\hat{\pi}_k^{(t-1)} - \sum_{s=1}^{K}\hat{\pi}_s^{(t-1)}\log\hat{\pi}_s^{(t-1)}\right), \tag{2.10}$$

with $\hat{\pi}_{k,EM}$ given by original EM update equation, and $\hat{\pi}_k^{(t-1)}$ being the component weight estimate of previous iteration.

A new hyperparameter β comes as a penalty weight in Eq.(2.9). It helps to control the competition between clusters. This parameter is enhanced at each iteration to increase the

entropy weight in Eq.(2.10) if proportions at the previous iteration were too close. Reciprocally β is reduced automatically to undercut the gap between the different proportions. The update equation for β is given by

$$\beta = \frac{\sum_{k=1}^{K} \exp(-\eta n |\hat{\pi}_k^{(t)} - \hat{\pi}_k^{(t-1)}|)}{K}.$$

From the observed behavior of the proportion entropy $\pi \log \pi$, additional constraints are required to avoid all proportions being zero or negative. Boundaries on β are therefore defined to avoid extreme values. An update of β is then given by

$$\beta = \frac{\sum_{k=1}^{K} \exp(-\eta n |\hat{\pi}_k^{(t)} - \hat{\pi}_k^{(t-1)}|)}{K} \wedge \frac{1 - \pi_{(1)}^{\mathrm{EM}}}{-\pi_{(1)}^{(t-1)} E}, \tag{2.11}$$

with $E = \sum_{k=1}^{K} \pi_k \log \pi_k$, and underscript (1) corresponds to the largest value among the K clusters. We refer the reader to (Yang et al., 2012) for detailed computation of β.

Acting on the evolution of proportions with β enables one to check at each iteration that all the components' proportions are above a given threshold, and to delete those of proportion $\pi_k < \frac{1}{n}$. This is the annihilation part of their process. The idea of penalizing proportions in Eq.(2.10) is to reduce the estimated proportions of the component with insufficient information (as given by entropy).

A specific dynamic is imposed to β, defined by Eq.(2.11), throughout iterations. This parameter β is set to zero when the cluster number K is stable, *i.e.*, not decreasing for a period named t_{min}. This is important to avoid oscillating parameters, and thus to reach a maximum. One limitation is that Yang et al. (2012) arbitrarily set this time limit at $t_{min} = 60$ iterations, without any attempt to adapt it to different use cases. This algorithm is however robust to initialization as, to start with, each data point is the center of its own component, which yields the initial number of class K^0 to be n, the sample size. Then, progressive suppression of clusters is done by deleting components of proportion $\pi_k < \frac{1}{n}$.

Starting with a number of mixture components higher than the true value is a solution to the initialization problem, allowing to escape of local maxima in some situations where components are very heterogeneously distributed in space.

However, competition and the instability of component proportions in the Robust EM do not avoid ending up with two (or more) equal classes, which is an overfitting problem leading to an identifiability one. The coefficient β is not high enough to trigger the removal of one of the superimposed clusters.

2.4 Deterministic annealing

Adapted from simulated annealing in combinatorial optimization, and inspired by the theory of thermodynamics and the concept of free energy, deterministic annealing usually helps the algorithm to explore the solution space by flattening the likelihood surface. Deterministic annealing for EM algorithms (Ueda and Nakano, 1994, 1998; Rose, 1998) relies on the introduction of a temperature parameter in the E-step of the algorithm, leading from Eq.(2.4) to the following expression

$$\frac{\left(\pi_k^{(t)} p_g(\boldsymbol{x}_i | \theta_k^{(t)})\right)^{1/T}}{\sum_{j=1}^{K} \left(\pi_j^{(t)} p_g(\boldsymbol{x}_i | \theta_j^{(t)})\right)^{1/T}}. \tag{2.12}$$

Among the drawbacks of the EM algorithm, the bumpy likelihood surface makes parameter estimation complex, and highly dependent on starting values. Zhou and Lange (2010) explored several deterministic annealing strategies to warp a relatively flat surface with handful modes into the bumpy surface of the objective function. Alternative annealing schemes were developed in EM-like algorithms (Lartigue et al., 2022; Allassonnière and Chevallier, 2021; Naim and Gildea, 2012; Pervez and Lee, 2015; Meinicke and Ritter, 2001), pairwise data clustering (Hofmann and Buhmann, Jan./1997) and online learning (Mavridis and Baras, 2022). Deterministic annealing can also be used simply to help initialize the algorithm (Franczak et al., 2014).

2.5 Aitken's criterion

Usual stopping criteria in EM-like algorithms lean on absolute differences of log-likelihoods, which correspond more to a "lack of progress" than to actual convergence. Some algorithms also lean on absolute differences between centers at actual and previous iterations.

Aitken (1927) proposed an acceleration criterion to measure whether the algorithm converged or not. From a sequence of log-likelihoods, assumed to be linearly convergent to some value $l^\star$, an estimate of $l^\star$ is obtained through

$$l_\infty^{t+1} = l^t + \frac{l^{t+1} - l^t}{1 - a^t},$$

where $a^t = \frac{l^{t+1} - l^t}{l^t - l^{t-1}}$.

With this asymptotic estimate, Böhning et al. (1994) proposed the following stopping criterion for an EM algorithm at iteration $t+1$:

$$|l_\infty^{t+1} - l_\infty^t| < \varepsilon\,. \tag{2.13}$$

Later on, variants of this criterion were $l_\infty^{t+1} - l^{t+1} < \varepsilon$ (Lindsay, 1995) or $l_\infty^{t+1} - l^t < \varepsilon$ (McNicholas et al., 2010) which is assessed to be at least as strict as lack of progress. In the end, these different criteria lead to similar results. Aitken's based criteria may lead to algorithms running longer than necessary in cases where the log-likelihood is constantly increasing. They also require additional computations, which leads to their relatively infrequent use, even though they are more coherent.

Chapter 3

Spatio-temporal mixture process estimation to detect dynamic changes in population

We propose in this chapter a method to model and monitor population distributions over space and time, in order to build an alert system for spatio-temporal data changes. Assuming that mixture models can correctly model populations, we propose a new version of the Expectation-Maximization (EM) algorithm to better estimate the number of clusters and their parameters at the same time. This algorithm is compared to existing methods on several simulated datasets. We then combine the algorithm with a temporal statistical model, allowing for the detection of dynamic changes in population distributions, and call the result a spatio-temporal mixture process (STMP). We test STMPs on synthetic data, and consider several behaviors of the distributions, to fit this process. Finally, we validate STMPs on a real data set of positive diagnosed patients to coronavirus disease 2019. We show that our pipeline correctly models evolving real data and detects epidemic changes.

Contents

1 Introduction

The rapid growth of health information systems has led to the availability of a real-time spatio-temporal follow up of patients affected by a given disease. A remaining challenge is to develop methods to use this data to improve public health strategies and to transform this observed data into actionable decision support systems.

Spatial models are based on the characterization of individuals by their geographical location (place of birth, place at the time of diagnosis, place of residence, *etc.*). Taken together, these individuals form a population. Concurrently the temporal component is

essential in disease monitoring, therefore requiring consideration of the population distribution as evolving over time. The association of spatial and temporal components for a disease yields a spatio-temporal distribution. One actionable decision-aid support system that could improve health management using such data is real-time highlighting of new or evolving clusters of patients. To detect for example a specific subgroup of patients which will evolve differently, while the remainder of the population remains stable. This would be particularly useful to rapidly identify a new contamination source for a transmissible disease, as soon as the first affected cases are present in health information systems.

1.1 Related works and motivation

1.1.1 Spatio-temporal statistical analyses in epidemiology

Spatio-temporal statistical analyses are already present in research in epidemiology and are mainly based on statistical tests, coupled, or not, with space-time kernel density estimation, as presented by Kirby et al. (2017). Scan statistic methods proposed by Kulldorff (1997) and Kulldorff et al. (1998) are reference methods for many studies. They propose to detect spatial and/or temporal clusters from aggregated data (discrete in space and time) using sliding windows to compare cases and reference populations. Another scan statistic method is proposed by Kulldorff et al. (2005) in the absence of population-at-risk. In both cases, these methods require to fix several parameters on the considered sliding window (*e.g.*, minimal area and minimal temporal size). Moreover, cases/controls studies are subject among other things to selection and expensive efforts to find a proper control group and are not feasible in all situations (Elliott and Wartenberg, 2004). In addition, these studies are prone to several biases (Sackett, 1979). As it is usually difficult to sample a control group from a reference population distribution, the ensuing comparison between cases and controls is exposed to false differences due to inadequate sampling of the control group (Sackett, 1979). Another important issue is that these methods do not provide a statistical modelling of the population over the whole space and time.

1.1.2 Estimation algorithms for mixture models

Different from looking at data in a sub-window of the space, mixture models are another class of models to spatially model data in statistics. Mixture models come with strong advantages. First, they are flexible as one can set the probability distribution function (pdf) of each cluster depending on the type of observations (scalars, vectors, positive measures, *etc.*). Second, the results are interpretable because subjects can be attributed to estimated classes *a posteriori* which enables one to distinguish homogeneous groups in the whole set. Third, they do not rely on a population reference distribution estimation, unlike scan statistics methods: they only rely on cases distribution. Last, these mixture models are parametric and well understood.

When data are multivariate real valued observations, the customary probability density for each cluster is the multivariate Gaussian distribution. This is particularly relevant when considering geographical data (mapped as lying on the real plane). The use of a multivariate Gaussian distribution inside a mixture model gives a Gaussian Mixture Model (GMM).

To perform the estimation of mixture model parameters, given the number of clusters (*i.e.* classes), a well-known and widely used algorithm is the Expectation-Maximization (EM) algorithm introduced by Dempster et al. (1977). The estimate obtained with the EM algorithm is deterministic and highly dependent on the initialization step. Moreover, the construction of the sequence ensures that the critical points are maxima, but could be either global or local ones. To avoid sensitivity to initial values and selection of a wrong local maximum, several strategies rely on repetitions of a random initialization step or

initialization with K-means algorithm (Baudry and Celeux, 2015). Recently, Lartigue et al. (2022) introduced an annealing E-step to better stride the support and become almost independent of the initialization. However, this method requires to set the temperature profile which may be time-consuming.

Finding an optimal number of components K is not an objective directly included in the original EM algorithm. This objective is often based on a model selection step, which requires a collection of estimated models (Akaike, 1973; Schwarz, 1978; Birgé and Massart, 2007). The well-known criteria for model selection are the Akaike Information Criterion (AIC) (Akaike, 1973), and the Bayesian Information Criterion (BIC) (Schwarz, 1978). They have been proved to be adequate for selecting K, but they are asymptotic criteria, and can select under- (for AIC) or over-adjusted (for BIC) models. Non-asymptotic approaches have been proposed, such as the slope heuristic criterion, introduced by Birgé and Massart (2007) and implemented by Baudry et al. (2012). It provides an optimal penalty of the log-likelihood, and thus an optimal model, but also requires a linear behavior of the log-likelihood. On the other hand, Baudry and Celeux (2015) proposed to introduce a recursive initialization which consists in using the K components solution to initialize the $K + 1$ components mixture. However, their full process requires several GMM estimations, with a varying number of components K, leading to expensive computations.

Subsequently, the last decade has seen the emergence of methods aiming to simultaneously overcome the need for a collection of models, find the optimal number of classes, and avoid bad local maxima (Derman and Pennec, 2017; Figueiredo and Jain, 2002; Wang et al., 2004; Zhang et al., 2004; Yang et al., 2012; Law et al., 2004). Several methods rely on a minimum message length criterion (Wallace and Freeman, 1987; Wallace and Dowe, 1999) which penalizes the cost function (Figueiredo and Jain, 2002; Law et al., 2004). These methods force parameter space exploration to obtains several models. However, these methods have to continue exploration and estimation until they reach a minimal number of clusters fixed in advance. This forced estimation of an internal collection of models is also present in the work of Derman and Pennec (2017), where the authors combine the slope heuristic criterion for model selection (Birgé and Massart, 2007) with a dynamic change of the number of components inside the EM algorithm. Another dynamic algorithm is the stepwise split-and-merge EM algorithm (Wang et al., 2004; Zhang et al., 2004). With split and merge criteria based on Kullback-Leibler divergence or correlation coefficient, these methods explore dynamically the parameters space by forcing clusters to merge together (or split apart). But they may rely on independent split and merge movements or several runs of the EM algorithms, implying computational issues. On the contrary, in the work of Yang et al. (2012), the number of components is estimated in a single-run EM algorithm with a reasonably low computation time. This solution is named Robust EM algorithm. But this algorithm can reach incorrect local maxima as we will see below.

The temporal component to monitor the population distribution is absent of these different procedures using EM algorithms, and the epidemiological models presented previously also cannot meet the criteria for estimating, monitoring and modelling population dynamics over time. As a consequence, these drawbacks prevent us from directly using the presented algorithms to obtain correct approximations of population dynamic and to monitor them.

1.2 Contributions

In this chapter, we propose a complete pipeline named spatio-temporal mixture process (STMP). This pipeline infers population distribution and highlights temporal population distribution differences **as a first step towards a decision support and alert system for spatio-temporal analysis of the evolution of a population.** STMP can be used to initiate a detailed analysis of the environment for example if the pathology may depend on

environmental causes. The STMP can also allow focusing on effects of decisions in specific areas where changes are happening, as we have faced with the COVID-19 pandemic and successive lockdowns for example. Within the proposed STMP, we combine a mixture model with reliable estimation and temporal monitoring of this model. This pipeline will create a temporal process with two mixture models, one time-depending and one totally independent. The adequacy of population dynamic to either of these two models will determine if an alert should be raised or not.

As a module to our STMP, we will introduce an adaptation of the EM algorithm (Dempster et al., 1977) to take into account a temporal dependency during a mixture model evolution. Finally, we will also propose an improvement of the Robust EM algorithm (Yang et al., 2012). We will suggest changes to obtain a more automatic algorithm to avoid overlapping components, observed with the Robust EM algorithm on real data tests. This modified version of Robust EM algorithm is compared to the original one (and other selection model criteria) to show that on synthetic data, there is no loss of performances and on real data we outperform the state-of-the-art algorithm.

To finish designing our STMP, we will perform experiments on synthetic data. And we will study the behavior of our pipeline in different situations to produce a robust monitoring. Applying our process to a dataset of COVID-19 cases from the Paris area, we will demonstrate the adequacy of a mixture model evolving over time and the consistency of the alert response to population epidemic changes.

2 Notations and reminders on mixture models and estimation algorithms

We assume for our future application in Section 4 and 5 that the population is generated from a Gaussian mixture model (GMM). In this section, we first recall the GMM definition. Then we detail the Robust EM algorithm, one of the algorithms used to fit GMMs. These methods are the basic elements on which we build our STMP pipeline described in Section 3.

2.1 The Gaussian Mixture Model

In order to describe a Gaussian mixture model, we consider a set of observations denoted $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$. Let $\mathcal{N}_d(\cdot|\mu_k, \Sigma_k)$ be the probability density function (pdf) of the Gaussian density of dimension d with mean μ_k and covariance matrix Σ_k. To write the GMM in its complete form we introduce latent variables $(z_i)_{i=1,\dots,n}$, such that each z_i is following a categorical distribution of parameter π. This information is then encoded as a K-dimensional binary variable $\mathbf{z}_i$ for each $i \in \{1, \dots, n\}$ with $z_i^k = 1$ if data x_i belongs to cluster k, 0 otherwise.
Then the complete model writes :

$$\begin{cases} z_i & \sim & \text{Categorical}(\pi_1, \dots, \pi_k)\,, \\ x_i | z_i^k = 1 & \sim & \mathcal{N}_d(\mu_k, \Sigma_k)\,, \end{cases}, \tag{3.1}$$

with $\theta = (\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K)$. The problem with estimation of GMM is twofold. The first challenge is to estimate the number of components K in the model. Then, given this estimated K, the second issue is how to estimate the vector of parameters θ, containing the Gaussian distributions parameters and the mixture proportions π. All this has to be performed from the observed data only.

2.2 The Expectation-Maximization algorithm

The EM algorithm (Dempster et al., 1977) was introduced for the purpose of estimating GMMs and has remained the most popular choice. The general principle of this algorithm is to produce a sequence of parameters $(\hat{\theta}^p)_{p\in\mathbb{N}}$ which converges towards an element of the set of critical points of the observed likelihood, which is for a GMM on a set of observations $\mathbf{x}$

$$p(\mathbf{x};\Theta) = \prod_{i=1}^{n}\left[\sum_{k=1}^{K}\pi_k\mathcal{N}_d(x_i|\mu_k\,\Sigma_k)\right]. \tag{3.2}$$

The EM algorithm alternates between an expectation step, and a maximization step which updates the mixture parameters, until a convergence criterion is met. The detailed equations are given in Appendix 3.A.1.

As the EM algorithm presents several drawbacks detailed in Section 1, and that we expect our framework to have a single run to estimate the data distribution at a given time step, we turn to the more "dynamic" algorithms where estimation and selection of the model are performed at the same time (Figueiredo and Jain, 2002; Law et al., 2004; Zhang et al., 2004; Wang et al., 2004; Yang et al., 2012).

In the next part, we will detail a recent dynamic algorithm proposed by Yang et al. (2012), which answers almost all issues and is the base of our proposition.

2.3 The Robust EM algorithm

As mentioned previously, the unknown number of clusters in GMM is a main issue. Yang et al. (2012) go deeper in the dynamic search for the best number of components in the mixture. Their Robust EM adjusts the EM mixture objective function, by adding a penalty criterion based on the entropy of the mixture proportions π_k. Non-informative proportions are given by a high entropy. Consequently, the penalty added to the likelihood is given by the negative entropy. Starting from the complete log-likelihood $\log\mathcal{L}(\theta,\mathbf{x},\mathbf{z})$, the objective function to maximize in the M-step with this entropy-based penalty is therefore

$$\begin{aligned}\tilde{Q}(\Theta;\Theta^{(p)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} p_{\Theta^{(p)}}(z_i^k = 1|\boldsymbol{x}_i)\log(\pi_k\mathcal{N}(x_i|\mu_k,\Sigma_k))\\ &+\beta\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k\log\pi_k, \text{ with } \beta\geq 0\,.\end{aligned} \tag{3.3}$$

With this new criterion to maximize, the update equation of component proportions π inside the EM algorithm becomes

$$\hat{\pi}_k^p = \hat{\pi}_{k,\text{MLE}} + \beta\hat{\pi}_k^{p-1}\left(\ln\hat{\pi}_k^{p-1} - \sum_{s=1}^{K}\hat{\pi}_s^{p-1}\ln\hat{\pi}_s^{p-1}\right), \tag{3.4}$$

with $\hat{\pi}_{k,\text{MLE}}$ obtained by maximization of the original objective function (without penalization) (see Section 3.A.1 Eq.(3.10)), and $\hat{\pi}_k^{p-1}$ being the component weight estimate of previous iteration. The equations to estimate the means $\hat{\mu}_k$ and the covariance matrices $\hat{\Sigma}_k$ in Robust EM remain unchanged. These parameters are estimated at each maximization step by Eq.(3.11) and Eq.(3.12) with the new component weights from Eq.(3.4).

As we can see, a new hyperparameter β comes as a penalty weight in Eq.(3.3). It helps to control the competition between clusters. Acting on the evolution of proportions with β enables one to check at each iteration that all the components proportions are above a given threshold, and therefore to delete those of proportion $\pi_k < \frac{1}{n}$. This is the annihilation part

in their process. A specific dynamic is imposed on β. This parameter is set to zero when the cluster number K is stable, *i.e.* not decreasing for a time period p_{min}. This is important to avoid oscillating parameters, and so to reach a maximum. A limitation is that they fixed this time limit to $p_{min} = 60$ iterations, without any attempt to adapt it to different use cases. This algorithm is however robust to initialization as, to start with, each data point is the center of its own component, which yields the initial number of class K^0 to be n, the sample size.

Although efficient, the entropy-based penalization of Yang et al. (2012) does not prevent from having several components with similar parameters, meaning that two cluster may be superimposed. In their Robust EM algorithm, competition and instability of component proportions do not avoid ending up with a local maximum of this type. The coefficient β is usually not high enough to trigger removal of one of the superimposed clusters. As the competition is not guaranteed at each iteration, we suggest improvements of the Robust EM algorithm in the next section. We also present a temporal process which, combined with estimation algorithms, will provide efficient detection of population dynamic changes.

3 Method: Spatio-temporal mixture model with efficient estimation algorithms for distribution change detection

In this section, we describe our general pipeline for temporal evolution modelling of a population including a distribution change detection, named STMP. Then, we introduce modifications on the Robust EM algorithm to escape local maxima characterized by "overlapping clusters". Finally, we detail another adaptation of the EM algorithm in order to constrain the estimation of GMM parameters. This enables to propose a close estimation of a given distribution, while taking into account new samples. The STMP pipeline and the estimation algorithms are generic enough to apply on different mixture models by using different estimation algorithms.

3.1 A spatio-temporal mixture process (STMP) with dynamic change detection

We consider that the time period is discretized, and the time steps are given by $t = 1, \dots, T$. At each time step, denote the data vector $\mathbf{X}^{(t)} = (\mathbf{X}_1^{(t)}, \dots, \mathbf{X}_{n_t}^{(t)})$ with $\mathbf{X}_i^{(t)} \in \mathbb{R}^d$. We assume that this data is sampled from a statistical time dependent model. We model the data at each time step t by a mixture of probability distributions, parametrized by a vector $\theta^{(t)}$, characterizing the current model $M^{(t)}$.

At each time t, we observe a new vector $\mathbf{X}^{(t)}$, independent of the previous one $\mathbf{X}^{(t-1)}$. Given this new sample, we want to evaluate if the previous model $M^{(t-1)}$, defined as a mixture model estimated on $\mathbf{X}^{(t-1)}$, is likely to fit the new set $\mathbf{X}^{(t)}$. We make the assumption that the distribution of the underlying global population does not change over time. This is in line with the difficulties related to the use of reference populations presented in Section 1 and coherent with our targeted applications. We design our model to monitor population evolution over time in particular for either short time period of time or longer period of time with aggregated data. Thus, the model does not require any datasets other than the vectors $\mathbf{X}^{(t)}$ for each time step t.

However, as $M^{(t-1)}$ depends on the data set at time $t-1$, it suffers from estimation variability, which means that the true model is likely close to but not equal to $M^{(t-1)}$. To deal with this uncertainty, we estimate a constrained model (or candidate model) M' to

fit $\mathbf{X}^{(t)}$ where M' is an adjustment of $M^{(t-1)}$, given by θ' close to $\theta^{(t-1)}$. Through this adaptation of $M^{(t-1)}$, we indirectly keep track of the estimated model at previous time. However, if at time t the data set $\mathbf{X}^{(t)}$ is sampled from a very different distribution, M' should not be able to fit $\mathbf{X}^{(t)}$. In this situation, we would like our process to detect this shift in population dynamic, and propose an alternative model more representative of the new data.

In order to do this, we propose to also estimate an alternative model, M^a only from the dataset $\mathbf{X}^{(t)}$. We do not make any assumption on a previous time step dependence to estimate this model leading to a parameter vector θ^a only driven by $\mathbf{X}^{(t)}$.

With these two estimated models in hands, we are now able to track changes of the population distribution, and determine whether there is a modification in the population geographical spreading. Our proposed warning system is defined as follows. If at time t, the model M', close to $M^{(t-1)}$, is not adapted to describe $\mathbf{X}^{(t)}$, we keep the independent model M^a as the new description of the current population and raise an alert. The aim is now to define the decision rule to select either model and to raise the alert or not as a result.

A simple way to quantify goodness of fit of a statistical model to the data is its likelihood. The likelihoods of estimated mixture models M' and M^a, given by $p_{\theta'}(\mathbf{X}^{(t)})$ and $p_{\theta^a}(\mathbf{X}^{(t)})$ respectively, are used to define a decision rule in our process, called the likelihood ratio or Bayes factor.

As the alternative model is unconstrained, $p_{\theta^a}(\mathbf{X}^{(t)})$ is the maximum value of the likelihood of the data without constraint on parameters' estimation. On the other hand, $p_{\theta'}(\mathbf{X}^{(t)})$ is the maximum value of the likelihood when the parameters θ' are restricted to stay in a neighborhood of $\theta^{(t-1)}$. In the case where the constrained model M', fits well the new data set, the alternative model is likely to be similar and to have a similar likelihood. Therefore, the likelihood ratio will be close to one. On the other hand, if the new data set is sampled from a very different distribution from $M^{(t-1)}$, then the constrained model will have a likelihood that is lower than the alternative model which by design will be able to better fit the new point cloud. Therefore, there should be a notification when this ratio is far above one.

Finally, we define the likelihood ratio as follows:

$$r_t(M', M^a) = \frac{p_{\theta^a}(\mathbf{X}^{(t)})}{p_{\theta'}(\mathbf{X}^{(t)})}\,. \tag{3.5}$$

In order to accept or reject the alternative model at time t, we define a threshold τ such that if $r_t(M', M^a) \geq \tau$, the alternative model is selected and an alert is raised. The detailed behavior of this likelihood ratio depending on the population evolution will be studied in Subsection 4.3. In particular, this empirical study allows us to set the threshold τ and highlight its properties in particular its low dependence w.r.t the sample size.

With all these elements in hand, our space-time complete pipeline, named Spatio-Temporal Mixture Process (STMP), executes at each time t the following steps:

1. Estimate models M' and M^a based on respectively $(M^{(t-1)}, \mathbf{X}^{(t)})$ and $(\mathbf{X}^{(t)})$,
2. Compute likelihood ratio $r_t(M', M^a)$ as in Eq.(3.5),
3. If $r_t(M', M^a) \geq \tau$, raise an alert and set $M^{(t)} = M^a$. Else set $M^{(t)} = M'$.

Note that this pipeline is very versatile with respect to the chosen distributions in the mixture model as well as the estimation algorithms used in first step. Depending on the dataset, the model is able to handle any type of pdfs.

We now describe the two algorithms that we use to perform the candidate and alternative model estimations.

3.2 The Modified Robust EM algorithm: tackling superimposed clusters

In Section 2, we have highlighted two weaknesses of the Robust EM algorithm by Yang et al. (2012). First, the minimal number of iterations (named p_{min}) before setting $\beta = 0$ is too small, which means that the algorithm is untimely stopped in its exploration. Then, the algorithm is stuck in local maxima as soon as the convergence condition ($\|\mu^{(p)} - \mu^{(p-1)}\| < \tau$ where $\tau > 0$ is a threshold) is satisfied, which stops the algorithm too early, revealing aberrant clusters. These aberrant clusters are superimposed clusters, which means that at least two clusters are sharing very similar (or exactly equal) parameters values. This corresponds to local maxima which can be analyzed only by post-processing the results, and it is particularly observable on real and scattered data.

To avoid this local maximum issue inside the estimation algorithm (and avoid post-processing analysis), we propose slight modifications of the Robust EM algorithm, by incorporating an online verification step of superimposed clusters. We consider that two clusters i and j are superimposed if

$$\|\mu_i - \mu_j\|_2 + \|\Sigma_i - \Sigma_j\|_F < \epsilon \tag{3.6}$$

for some small $\epsilon > 0$, where $\|.\|_2$ is the euclidean norm and $\|.\|_F$ is the Frobenius norm. Note that requiring equality in Eq.(3.6) is numerically too strong and would rarely happen. We check Condition (3.6) when the algorithm has reached the convergence condition (Algorithm 3.2, line 1). As long as there are overlapping clusters we force the estimation to continue, as we will see now.

Inside Algorithm 3.2, the "stop-competition" part is the moment in the algorithm where $\beta = 0$ if the component number is stable for at least 100 iterations and if the actual iteration number p is greater than p_{min} (Algorithm 3.2, line 2). At that point in the algorithm, if we set $\beta = 0$ too early, it slows down the competition between clusters, and may prevent components from disappearing. If there are no overlapped clusters and stability conditions are fulfilled then we set $\beta = 0$. Otherwise, we proceed as follows: we first increase p_{min} by increment of 50 iterations (Algorithm 3.2, line 3). By increasing p_{min}, the algorithm has more iterations to try to annihilate some components. Since increasing p_{min} indefinitely can lead to a "stable" configuration where β adopts a cyclical behavior and loops on it, we then check the proximity condition (3.6) again. If Eq.(3.6) is still true for some clusters, we merge these clusters. The weight of the fused clusters is the sum of the weights of the overlapping ones. The means and covariance matrices being almost equal, this fusion of components does not change much the likelihood. This makes the algorithm jump to another configuration with almost the same likelihood and enables it to explore this new region of interest. Other steps of the algorithm stay identical to the original Robust EM, as presented in Subsection 2.3. The full modified Robust EM algorithm is summarized in Algorithm 3.2.

3.3 The Constrained EM algorithm: former parameter based estimation

We name Constrained EM (C-EM) a slight variation of original EM algorithm (Dempster et al., 1977) where the parameters are restricted to a neighborhood of a given vector of parameters denoted θ^0. In particular, we introduce constraints on the estimated components proportions $(\pi_k)_{1 \leq k \leq K}$. Moreover, when the cluster means are involved, restrictions are also put on these means. The initialization of our C-EM algorithm is given by the parameter vector θ^0 as well. The idea behind C-EM algorithm is to obtain estimated parameters highly driven by the initial parameters vector θ^0 but updated on data $\mathbf{X}$. Because the parameters of

our dynamic modeling are estimated empirically, the estimation suffers from the uncertainty given by the sampling. This means that the estimated parameters at time $t-1$ may not be the perfect description of the data set and a new independent sample from the same ground truth distribution will lead to a slightly different estimated parameter vector and a slightly different likelihood. Therefore, we consider that a newly independent estimated mixture and the given estimated one may both come from the same ground truth. For this reason, the C-EM enables us to give a chance to the previously estimated model to explain the data distribution. Otherwise, forcing the comparison of $M^{(t-1)}$ with M^a will always be in favor of M^a. With this parameter dependency, the newly estimated parameters could be incorporated in our temporal process as a time-dependent estimate.

From now on, we propose the details of this algorithm for distributions where the cluster means and covariances are to be estimated. This will be the case in our disease progression use case where the model is a mixture of Gaussian distributions. We now detail constraints we impose on parameter estimations inside an EM-like algorithm to estimate GMM. We name $\hat{\pi}^c$, $\hat{\mu}^c$ and $\hat{\mathbf{\Sigma}}^c$ the constrained proportions, means and covariance matrices obtained through the C-EM algorithm. As in the original EM algorithm, $\hat{\pi}^p$ and $\hat{\mu}^p$ vectors are estimated at iteration p of C-EM following equations (3.10) and (3.11). We then add a third step in the estimation algorithm to obtain $\hat{\pi}^c$ and $\hat{\mu}^c$.

The constraints in the C-EM algorithm continuously require θ^0 over iterations, the initial parameter vector at $p=0$, as we want to restrict the parameters' estimation. The initial parameter vector contains $(\pi_k^0)_{1\leq k\leq K}$, $(\mu_k^0)_{1\leq k\leq K}$ and $(\Sigma_k^0)_{1\leq k\leq K}$ the covariance matrices providing information about the anisotropy we allow for the uncertainty on the means parameters to adapt locally. Component proportions are probability weights and live in $[0,1]$, so we simply constrain component proportion of cluster k, $\hat{\pi}_k^p$ (at iteration p), to vary inside $[\pi_k^0 \pm 0.1]$. This means each proportion varies by at most 10%. We also avoid proportions to become null to avoid the artificial death of a cluster in the mixture. Constrained mean $\hat{\mu}_k^c$ of the component k at iteration p with the C-EM algorithm is a projection of estimated $\hat{\mu}_k^p$ on a rectangular space centered on μ_k^0 and of length and width given by ellipse axis of the covariance matrix Σ_k^0 (square roots of the eigenvalues of Σ_k^0).

These constraints are written here for each iteration p and each cluster k

$$\begin{cases} \hat{\pi}_k^c & = min(max(\pi_k^0 - 0.1, \hat{\pi}_k^p), \pi_k^0 + 0.1)\,, \\ \hat{\mu}_k^c & = \mathcal{P}_{rect(\mu_k^0, \Sigma_k^0)}(\hat{\mu}_k^p)\,. \end{cases} \tag{3.7}$$

Note that the algorithm can converge to final parameters where one covariance matrix is singular, reflecting the aim of the algorithm to delete one component of the mixture model. In the original EM algorithm, implementations usually include a regularization on the covariance matrices, in order to avoid singular ones. As we want to determine when the estimated candidate model does not correspond to the data, we remove this regularization from the C-EM algorithm. Therefore, we raise an alert when one or more covariance matrices become singular. We add this condition as an alert in STMP detailed in Subsection 3.1, before the calculation of the ratio r_t (Eq.(3.5)).

In addition to this, as the covariance matrices are not constrained in the C-EM algorithm, we introduce a condition to check these parameters *a posteriori*. From the C-EM algorithm, covariance matrices are freely estimated, but they can evolve far away from initial covariances matrices Σ_k^0, thus missing the time link. Therefore, we introduce an already existing similarity measure between final estimated $\hat{\Sigma}_k^c$ in C-EM and Σ_k^0 the initial covariance matrices. We use the cosine similarity, also introduced as the correlation matrix distance by Herdin et al. (2005) on correlation matrices. We adopt their formulation and apply it on covariance matrices instead of correlation matrices. Bounded between 0 and 1, this coefficient measures orthogonality between two matrices and is useful to evaluate whether the spatial

structure of the clusters have significantly changed. Low values reflect high similarity while high values reflects orthogonality, and so on dissimilarities. As $\hat{\Sigma}_k^c$ should be similar to Σ_k^0, we only tolerate a value of 0.1 or less, in order to introduce flexibility and sampling error tolerance inside STMP. For higher values, showing dissimilarities between $\hat{\Sigma}_k^c$ and Σ_k^0, we also raise an alert in STMP detailed in Subsection 3.1.

In STMP, θ^0 will be the estimated parameter vector from the previous time step $t-1$ of the pipeline, which corresponds to $\theta^{(t-1)}$. Therefore, we obtain at time t an estimated parameter depending on estimated parameter at time $t-1$, but allowing some adaptation of the model to the newly observed data $\mathbf{X}^{(t)}$. Finally, we should not forget that the C-EM is constrained by initial parameters θ^0, including a fixed number of clusters K^0. It is not possible in C-EM to merge clusters based on their properties, as this would violate the imposed constraints. If the model estimated by C-EM is not correctly fitting data $X^{(t)}$, this will be detected inside STMP.

3.4 Application of the STMP on Gaussian Mixtures Models

To conclude this section, our new process is fully described in Algorithm 3.1 and Figure 3.1, combining the temporal process described in Subsection 3.1 with the C-EM algorithm to estimate M' (Subsection 3.3), and the Modified Robust EM algorithm to estimate M^a (Subsection 3.2) on GMMs.

Algorithm 3.1: The Spatio-Temporal Mixture Process (STMP)

input : For $t = 0, \ldots, T$: data $\mathbf{X}^{(t)}$
$\theta^{(0)}, \hat{K}^{(0)} \leftarrow$ `ModifiedRobustEM(`$\mathbf{X}^{(0)}$`)`
$\theta^{(t)} \leftarrow \theta^{(0)}$; $\hat{K}^{(t)} \leftarrow \hat{K}^{(0)}$
for $t = 1, \ldots, T$ **do**
  $\theta^{(t-1)} \leftarrow \theta^{(t)}$
  $\hat{K}^{(t-1)} \leftarrow \hat{K}^{(t)}$
  $\theta' \leftarrow$ `C-EM(`$\mathbf{X}^{(t)}$, $\theta^{(t-1)}$, $\hat{K}^{(t-1)}$,*maxiter=5*`)`
  $\theta^a, \hat{K}^a \leftarrow$ `ModifiedRobustEM(`$\mathbf{X}^{(t)}$`)`
  $r_t \leftarrow \frac{p_{\theta^a}(\mathbf{X}^{(t)})}{p_{\theta'}(\mathbf{X}^{(t)})}$
  if $\left(\exists\ singular\ \hat{\Sigma}'_k \subset \theta'\right)$ *or* $\left(\exists\ cos\_similarity(\hat{\Sigma}'_k, \hat{\Sigma}_k^{(t-1)}) > 0.1\right)$ **then**
    alert $\leftarrow$ True
    $\theta^{(t)} \leftarrow \theta^a$
    $\hat{K}^{(t)} \leftarrow \hat{K}^a$
  else if $r_t \geq \tau$ **then**
    alert $\leftarrow$ True
    $\theta^{(t)} \leftarrow \theta^a$
    $\hat{K}^{(t)} \leftarrow \hat{K}^a$
  else
    $\theta^{(t)} \leftarrow \theta'$
    $\hat{K}^{(t)} \leftarrow \hat{K}^{(t-1)}$
  end
end

As in the following applications we will only consider geographical data (in $\mathbb{R}^2$), we use Gaussian Mixture Models to represent these data. The GMM parameters are estimated

with the presented algorithms, and the likelihoods are computed with Eq. (3.2). Recall that the Algorithm 3.1 and Figure 3.1 show the STMP with all our propositions, which could be used with different mixture models. Adaptations of the estimation algorithms may also be considered to fit with other distributions. This is the subject of Chapter 4.

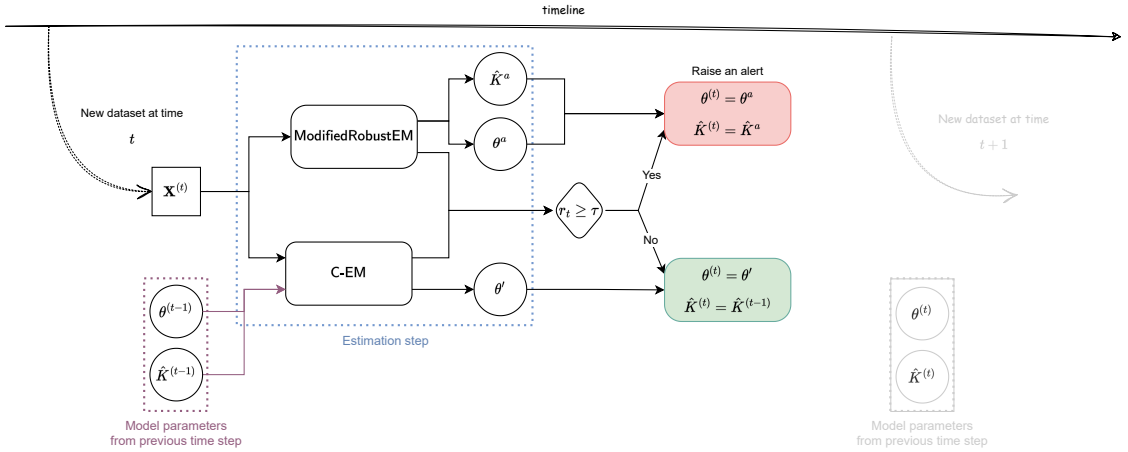


Figure 3.1: Diagram representation of STMP.

4 Experiments on synthetic data

This section is devoted to experimental validations, i.e. based on synthetic data. First, we present comparisons of our Modified Robust EM with other EM-based algorithms and selection criteria. The comparisons are conducted on two mixture distributions from existing benchmarks. Second, we present all the experiments that are tested on our complete pipeline. We study the estimated likelihood ratio for different behaviors of the population distribution (characterized by the experiments) and the resulting performances of the pipeline. By analyzing these performances we can fix a threshold conditioning the raise of an alert in all situations. Then, we focus on the validation of STMP, given by Algorithm 3.1. Finally, we also present experiments on the number of points n in the data sample, and how it affects each step of STMP.

4.1 Comparisons of the Modified Robust EM with other EM-based algorithms and selection criteria

We compare here our Modified Robust EM with several EM-based methods mentioned in Subsection 1.1.2. First, we run the original EM algorithm, which is based on the *a priori* knowledge of the number of clusters. As in practice we do not know the true number of components, we estimate several models and select the best one based on either Bayesian Information Criterion (BIC) (Schwarz, 1978) or Integrated Completed Likelihood (ICL) (Biernacki et al., 1998), two of the most commonly used criteria in model selection. Several mixture estimates are therefore obtained by running an EM algorithm for a range of values of K from K_{min} to K_{max}. The initialization issue is treated by starting from 10 random small K-means runs and then keeping the solution with the highest likelihood as the initialization of the EM algorithm. Second, we also compared our method with the original Robust EM algorithm (REM) (Yang et al., 2012) and Figueiredo and Jain's method (Figueiredo and Jain, 2002) (called FJ method from here). The FJ method requires to fix an initial number of clusters $K_{initial}$. Originally $K_{initial}$ was "far from the true number of components" but

not too high (around $K_{initial} = 30$ in several experiments of Figueiredo and Jain (2002)), but in order to approach the behavior of the Robust EM and Modified Robust EM we use $K_{initial} = n$. The methods are computed on 100 different data sets generated for each of the defined mixture distributions. The methods are then compared based on their capacity to estimate the correct number of components and when this number is correct, to estimate the parameters of the mixture models. They are also compared in terms of the computational cost given by the number of iterations, as iterations times are of the same order.

Results First, we compare the different methods on their ability to estimate the correct number of components. From a first mixture given by Figure 3.2(a) (with $n = 400$ points), REM was 95% successful in identifying the four components, close to the 99% of our method, against 51% for the FJ algorithm and 63% and 61% for EM-BIC and EM-ICL respectively. From a second mixture given by Figure 3.2(b) (with $n = 400$ points), all methods had more difficulty in identifying the four clusters. EM-BIC and EM-ICL were the most performant with 52% and 54% of successful estimation of the number of components, against 46% for our method MREM, and then 37% for the REM and 36% for the FJ method.

Then, we compare the estimated parameter precision over runs with successful component estimation. For each of the two defined mixtures, we computed the relative distance between the true and the estimated parameters. From these mean relative errors (Table 3.1 and Table 3.2), all the values are of the same range. It appears that FJ method, EM-BIC and EM-ICL have slightly lower errors than REM and MREM on the first mixture (Fig. 3.2(a)), but slightly higher errors than REM and MREM on the second mixture. This shows the importance of capturing the correct number of clusters, which is the goal of our algorithm. However, this implies for model selection criteria to have an average guess of the data heterogeneity and to run the estimation algorithm for each of the possible number of components and for several initializations each time.

	REM	FJ	EM-BIC	EM-ICL	MREM
$\hat{\pi}_0$	0.0851 (0.0547)	**0.0670** (0.0460)	0.0683 (0.0465)	0.0683 (0.0465)	0.0870 (0.0615)
$\hat{\pi}_1$	0.0988 (0.0740)	0.0678 (0.0513)	**0.0667**(0.0514)	**0.0667** (0.0514)	0.0990 (0.0737)
$\hat{\pi}_2$	0.0937 (0.0719)	**0.0741** (0.0660)	0.0742 (0.0654)	0.0742 (0.0654)	0.0935 (0.0724)
$\hat{\pi}_3$	0.0888 (0.0749)	**0.0680** (0.0689)	0.0681 (0.0682)	0.0681 (0.0682)	0.0894 (0.0731)
$\hat{\mu}_0$	0.0305 (0.0230)	0.0250 (0.0206)	**0.0248** (0.0205)	**0.0248** (0.0205)	0.0308 (0.0231)
$\hat{\mu}_1$	**0.0167** (0.0111)	0.0186 (0.0120)	0.0185 (0.0119)	0.0185 (0.0119)	0.0187 (0.0129)
$\hat{\mu}_2$	0.0116 (0.0082)	0.0121 (0.0078)	0.0122 (0.0078)	0.0122 (0.0078)	**0.0115** (0.0082)
$\hat{\mu}_3$	0.0187 (0.0122)	0.0207 (0.0123)	0.0210 (0.0123)	0.0210 (0.0123)	**0.0169** (0.0104)
$\hat{\Sigma}_0$	0.1052 (0.0737)	**0.1030** (0.0799)	**0.1030** (0.0792)	**0.1030** (0.0792)	0.1062 (0.0734)
$\hat{\Sigma}_1$	0.7164 (0.5245)	0.4811 (0.5133)	**0.4730** (0.5117)	**0.4730** (0.5117)	0.6087 (0.5353)
$\hat{\Sigma}_2$	0.1121 (0.0777)	0.1075 (0.0725)	**0.1071** (0.0719)	**0.1071** (0.0719)	0.1128 (0.0786)
$\hat{\Sigma}_3$	1.0568 (0.8288)	0.7403 (0.8096)	**0.7275** (0.8070)	**0.7275** (0.8070)	0.9110 (0.8541)

Table 3.1: Mean (standard deviation) relative errors for the estimates parameters of GMM within dataset from Fig.3.2(a). The absolute-value norm is used for proportions, the euclidean norm is used for means, and the Frobenius norm for covariances.

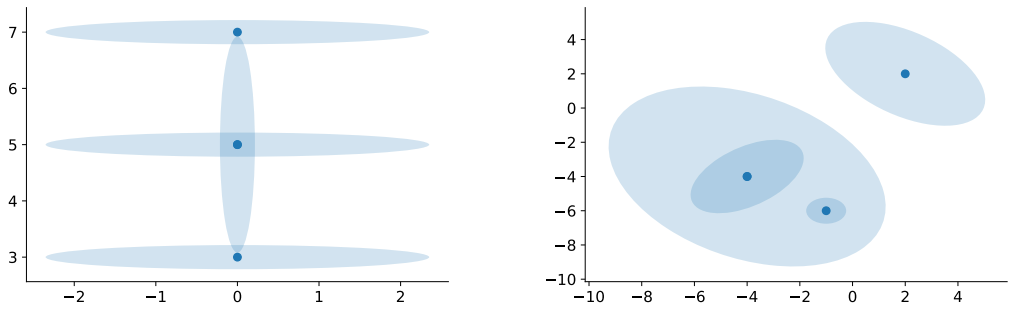
Finally, we compare the mean number of iterations for executions with each mixture distribution. The mean number of iterations on first and second mixture is respectively of 83 and 137 iterations for REM, against 95 and 185 for our method, 170 and 222 for BIC/ICL which used $K_{min} = 2$ and $K_{max} = 6$, and 730 and 711 for FJ method. The number of iterations is slightly higher with our method than the REM one because we put

	REM	FJ	EM-BIC	EM-ICL	MREM
$\hat{\pi}_0$	**0.1427** (0.1393)	0.1665 (0.1795)	0.1645 (0.1775)	0.1645 (0.1775)	0.1430 (0.1605)
$\hat{\pi}_1$	**0.1281** (0.1399)	0.1650 (0.1694)	0.1628 (0.1676)	0.1628 (0.1676)	0.1472 (0.1452)
$\hat{\pi}_2$	**0.0405** (0.0364)	0.0709 (0.0501)	0.0704 (0.0495)	0.0704 (0.0495)	0.0600 (0.0479)
$\hat{\pi}_3$	0.1477 (0.0991)	0.1461 (0.1053)	0.1468 (0.1040)	0.1468 (0.1040)	**0.1268** (0.030)
$\hat{\mu}_0$	0.0761 (0.1174)	0.0669 (0.1720)	0.0661 (0.1697)	0.0661 (0.1697)	**0.0391** (0.0236)
$\hat{\mu}_1$	**0.0368** (0.0265)	0.0455 (0.0356)	0.0458 (0.0351)	0.0458 (0.0351)	0.0650 (0.1150)
$\hat{\mu}_2$	0.0589 (0.0325)	0.0603 (0.0224)	0.0626 (0.0261)	0.0626 (0.0261)	**0.0561** (0.0300)
$\hat{\mu}_3$	**0.0117** (0.0071)	0.0141 (0.0063)	0.0140 (0.0063)	0.0140 (0.0063)	0.0121 (0.0067)
$\hat{\Sigma}_0$	3.7634 (2.1204)	1.9392 (2.4028)	1.8895 (2.3888)	1.8895 (2.3888)	**1.4402** (2.0190)
$\hat{\Sigma}_1$	0.7021 (0.3135)	0.4211 (0.3314)	0.4139 (0.3297)	0.4139 (0.3297)	**0.3611** (0.3156)
$\hat{\Sigma}_2$	0.1213 (0.0611)	0.1253 (0.0620)	0.1277 (0.0627)	0.1277 (0.0627)	**0.1102** (0.0469)
$\hat{\Sigma}_3$	0.3597 (0.1584)	0.3640 (0.1616)	0.3582 (0.1631)	0.3582 (0.1631)	**0.3460** (0.1855)

Table 3.2: Mean (standard deviation) relative errors for the estimates parameters of GMM within dataset from Figure 3.2(b). The absolute-value norm is used for proportions, the euclidean norm is used for means, and the Frobenius norm for covariances.

a "soft" condition on convergence to stop the algorithm. The number of iterations is very high for the FJ method because of the considered initial number of components, which is high here. But it was originally fixed to a lower number by Figueiredo and Jain (2002), and needed to be fixed arbitrarily.

Note that we have provided a narrow range of values including the correct one for model selection criteria with BIC and ICL. The EM algorithm failed with higher number of components as the algorithm tended to remove one cluster by cancelling its proportion and degenerating the covariance matrix. Our Modified Robust EM shows no loss of performances compared to the Robust EM on synthetic data, and solve Robust EM problems on real data as we will see later.



(a) A Gaussian mixture with 4 crossed components, defined in Yang et al. (2012).

(b) A Gaussian mixture with 4 overlapping components, defined initially in Figueiredo and Jain (2002).

Figure 3.2: Two Gaussian mixtures defined in works of Figueiredo and Jain (2002) and Yang et al. (2012).

4.2 Description of the experimental setups to calibrate STMP

All the following experiments are done on a two time steps configuration (only $t = 0$ and $t = 1$). We consider the following situation where we have a Gaussian mixture distribution with three clusters at initial time ($t = 0$). One cluster is isolated on the far right-hand side of the x=axis (first dimension), and the two others are on the left-hand side of the x-axis. This is the basic structure that all initial distributions (at $t = 0$) will follow. Different positions of left-hand side clusters are represented in Figure 3.4, for Setup F. (Far.), Setup M. (Moderate.) and Setup C. (Close.).

From this initial Gaussian mixture, various changes are done at time $t = 1$ considering:

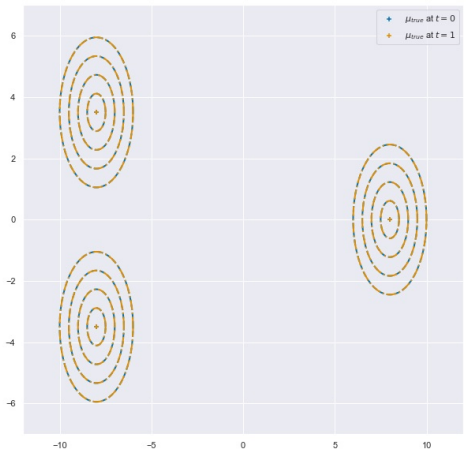
- (Case I.): no evolution at $t = 1$, clusters are properly distinct (corresponds to Setup F. at $t = 0$ and $t = 1$).
- (Case II.): the emergence of one new cluster leading to a distribution with four clusters at time $t = 1$.
- (Case III.): the disappearance of one cluster among the existing three initially present.
- (Case IV.): the movement of one initial cluster, which corresponds to moving centers and changing proportions and covariances.
- (Case V. and Case VI.): no evolution at $t = 1$, as Case I., but here the two left hand side clusters are slightly interfering (Setup M.) for Case V., and finally these two clusters are very closed and barely identifiable without enough samples (Setup C.) for Case VI. .
- (Case VII. to Case IX.): from initial Setup F. or Setup M. at $t = 0$, there is a spatial convergence of the two left-hand side clusters, characterized by Setup M. or C. at $t = 1$.

We denote $K^{(0)}_{true}$ the number of components in the mixture distribution at $t = 0$, $K^{(1)}_{true}$ in the mixture distribution at $t = 1$. A case is finally characterized by its mixture parameters at $t = 0$ and at $t = 1$, and we represent all cases in Table 3.A.3.2. In addition, Figure 3.3 and Figure 3.4 give a simple representation of Cases I. to IV. and of Setup F., M. and C. involved in Cases V. to IX.

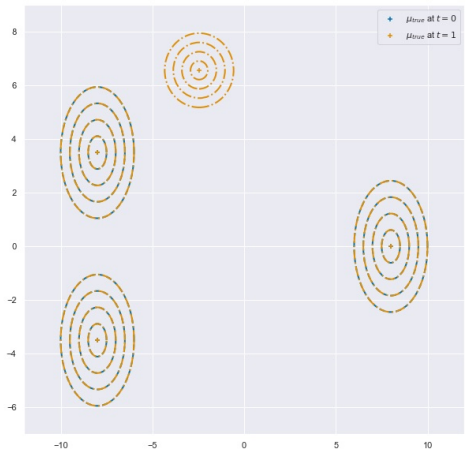
We obtain Gaussian mixture distributions with parameters $\theta^{(0)}$ and $\theta^{(1)}$ from described cases. For each Case I. to IX. , given the two distributions, we can sample $n_0 = n_1 = n$ points, which form our datasets $\mathbf{X}^{(0)}$ and $\mathbf{X}^{(1)}$. The sampling step, for any of the cases presented above, is executed S times and followed by execution of our STMP on each set of sampled data. It produces S different resulting processes, for each case (see Table 3.A.3.2). This enables us to analyze the behavior of STMP and likelihood ratio across runs and evolution cases.

4.3 Estimation of the alert threshold in STMP

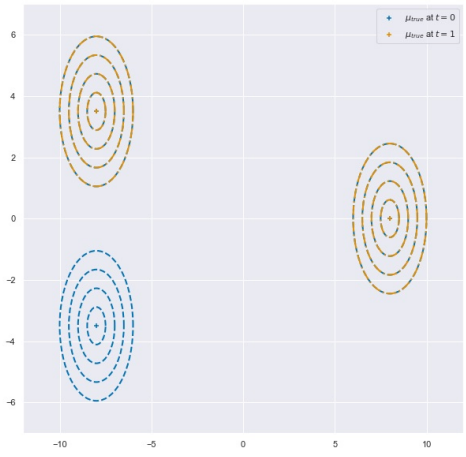
As motivated in Subsection 3.1, the likelihood ratio is a good indicator of how well the alternative model M^a at time t is fitting data $\mathbf{X}^{(t)}$ against the model M'. In case of no evolution of the distribution from $t = 0$ to $t = 1$, both M^a and M' should fit the data correctly, leading to a likelihood ratio around one. Of course, as said previously, due to the sampling of the distribution, it cannot be equal to one exactly. Thus, the goal of the following study is to introduce an empirical threshold of adequacy, over which the alternative model M^a is definitely considered as the best model explaining current data and an alert is raised. With all the experiments above, we study the behavior of our STMP according



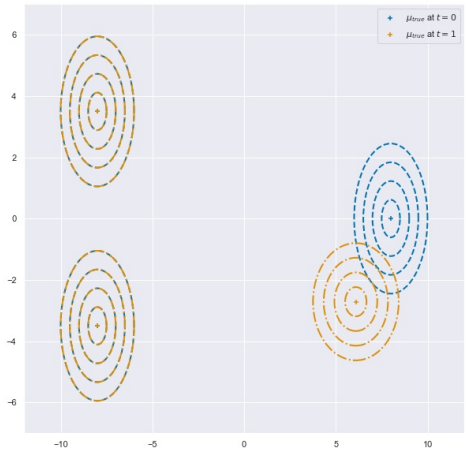
(a) Case I.



(b) Case II.



(c) Case III.



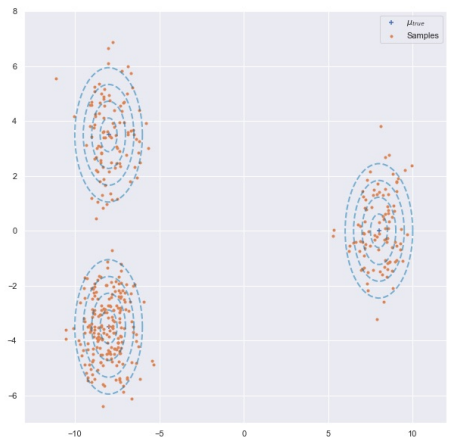
(d) Case IV.

Figure 3.3: Description of Gaussian mixture distributions for Cases I. to IV. (from Table 3.A.3.2). Blue centers and covariance ellipses correspond to Gaussian Mixture parameters at $t = 0$, orange ones to Gaussian Mixture parameters at $t = 1$. Note that when both elements are superimposed, the centers only appear orange and the ellipses have mixed colors dotted lines.

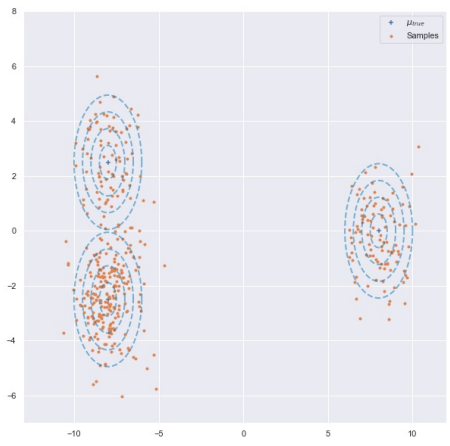
to the alert threshold τ involved in Algorithm 3.1. It is important to fix this threshold in order to raise meaningful alerts and reach a correct performance.

Theoretical likelihood ratio As said previously, we run S sampling steps for each case distributions, here fixed to $S = 100$ runs. We obtain S pairs of datasets $(X^{(0)}, X^{(1)})$. For each pair we compute the theoretical likelihood ratio

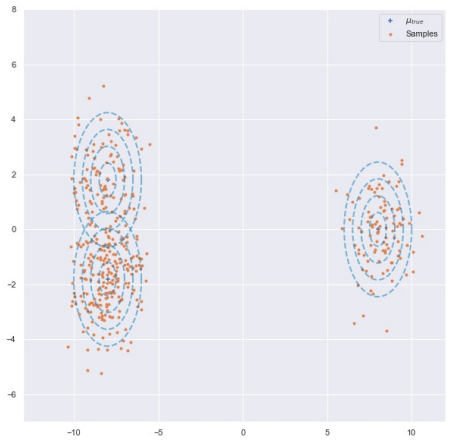
$$r_1^*(M^{(0)}, M^{(1)}) = \frac{p_{\theta^{(1)}}(\mathbf{X}^{(1)})}{p_{\theta^{(0)}}(\mathbf{X}^{(1)})},$$



(a) Setup F. (Far)



(b) Setup M. (Middle)



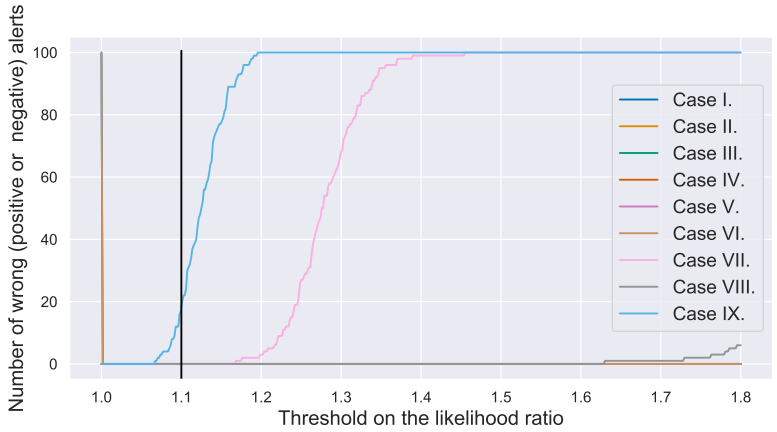
(c) Setup C. (Close)

Figure 3.4: Gaussian mixture distributions for Setups F., M. and C. involved in Cases presented in Table 3.A.3.2 with an example of sampled data sets. Blue crosses correspond to μ_k and ellipses to covariance matrices Σ_k. Orange points are samples.

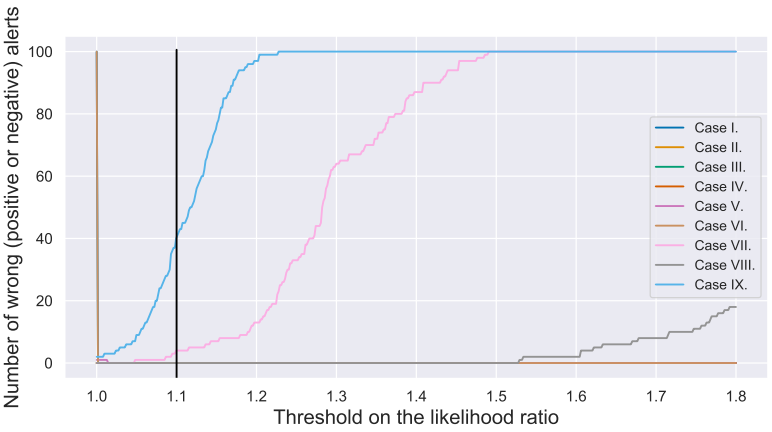
implying the true parameters of models $M^{(0)}$ and $M^{(1)}$. This provides a "theoretical" value of r, depending on the observation sets but not the estimated parameters. We then account for the number of wrong alerts, depending on the value of the likelihood ratio threshold τ.

Validation of the alert threshold Figure 3.5 presents the number of wrong alerts, with one curve by case explained in Subsection 4.2. As the dataset $X^{(1)}$ is sampled from the truth model $M^{(1)}$, the theoretical likelihood ratio should be almost one modulo the variability of the data if $M^{(0)} = M^{(1)}$. In contrary, this theoretical ratio should quickly diverge from one if $X^{(1)}$ is not corresponding to the model $M^{(0)}$. This explains that we obtain 100 % of correct alerts on the majority of the case experiments (Fig. 3.5), as the computed theoretical likelihood ratios are really higher than tested values of the threshold.

The cases which are critical for the choice of the threshold are Case VII. and Case IX



(a) Experiments with datasets of size $n = 400$.



(b) Experiments with datasets of size $n = 100$.

Figure 3.5: For each Case is presented the number of false alerts (positive or negative) on theoretical likelihood ratios, over $S = 100$ runs according to the considered threshold τ. The filled black vertical line is the selected threshold. Note that except for Cases IX. and VII. the other curves are superimposed for a threshold superior to one.

(Fig. 3.5). They imply slight differences of the distributions between $t = 0$ and $t = 1$, so the theoretical likelihood ratio values stay relatively close to one. Therefore, correct alerts are not raised for a threshold over about 1.06 when considering experiments with datasets of size $n = 400$ points (Fig. 3.5(a)) for these two cases. With $n = 100$ points, we clearly see that theoretical likelihood ratios are globally lower. For a same value of the threshold the number of false negative alerts increases (Fig. 3.5(b)). This provides us an intuition on the level of variations that our model can detect.

While the best possible performance would be obtained with a threshold at 1.05 (Fig. 3.5(a)) if we only consider theoretical ratios results, the study of the threshold involving the estimated models M' and M^a is less optimal. The computation of the likelihood ratio in the complete pipeline implies uncertainty on sampled data and on estimated parameters θ' and θ^a. This estimated ratio is defined by Eq.(3.5) between M' and M^a at $t = 1$. We study the performance of the pipeline with these estimated ratios values, by accounting for the number of wrong alerts over $S = 100$ runs as before. The corresponding results, with these estimated likelihood ratios, are given in Figure 3.6. In Table 3.A.3.1 we retrieve

the number of alerts per case for different threshold values and for different dataset sizes.

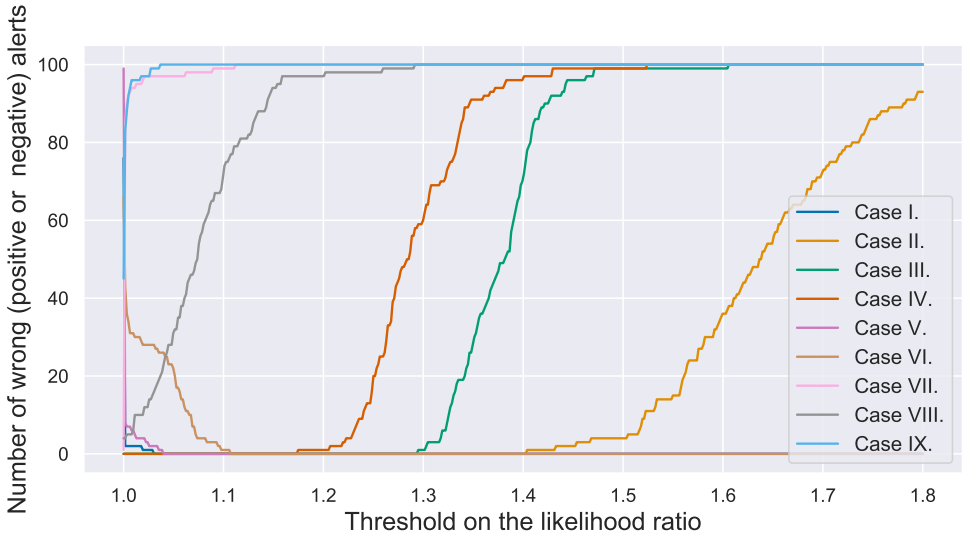


Figure 3.6: For each Case is presented the number of false alerts (positive or negative) depending on estimated likelihood ratios given by Eq.(3.5), over $S = 100$ runs according to the considered threshold τ. Datasets are of size $n = 400$.

For Case I., Case V. and Case VI., the population distribution is the same at $t = 0$ and $t = 1$, but M^a and M' are not estimated by the same algorithm, and the likelihood ratios depend on the sampled data. However, the model M' should still be accepted as the two mixture distributions are very close. From Figure 3.6, we observe that a threshold of 1.0 is not appropriate, as a high number of alerts is raised for these cases, where we should have zero alert. Increasing the threshold allows for model and data variability to be taken into account, and avoid false positive alerts.

On the other hand, if we set a too high threshold τ, there is a risk of not detecting all important changes. We clearly see for Cases II., III., IV., and VIII. that the number of true positive alerts is affected by a too high threshold. If we go above $\tau = 1.2$ we see an important decrease for Case IV., and later for the Cases II. and III.. Case VIII., which corresponds to a move from Setup F. to Setup C. is affected earlier by the likelihood ratio threshold, as the proximity of two clusters (Fig. 3.4(c)) affects the estimation of mixture parameters and so on the likelihood ratio. It leads us to set a threshold relatively closed to one. As on the theoretical likelihood ratio study, we observe here that slight movements corresponding to Case VII. and Case IX. lead to incorrect alerts for a threshold over one. The estimated likelihood ratio values stay relatively close to one because the model M' can adapt to data $\mathbf{X}^{(1)}$. The evolving distributions are not detected.

Therefore, when applied to a specific problem, one has to know that the relocation of one cluster may be detected if it relates to the variance of the estimated clusters. Otherwise, these displacements may be considered as normal variability of the discretization of the distributions. Note that this alert criterion may be adapted given a specific problem with the constraints that are imposed to the candidate model. Finally, we see from analysis of the theoretical ratios and the estimated ratios that we need to make a compromise. The optimistic theoretical likelihood ratios would lead us to take a threshold very close to one. But the obtained values with the estimated models contain more uncertainty that we cannot ignore and require to select a larger threshold. To obtain good performances of our pipeline we fix the threshold to $\tau = 1.1$. We obtain a balance between false negative and

false positive alerts, that we want to maintain as low as possible, considering all possible situations.

4.4 Performances of STMP on synthetic data

4.4.1 Performances of the Modified REM algorithm within STMP

We present here results of the estimation of GMM parameters with the Modified Robust EM algorithm at $t = 0$ and $t = 1$ in STMP experiments on synthetic data. All experimental frameworks described in Subsection 4.2 are tested, with $n_0 = n_1 = n = 400$ points.

Evaluation of $\hat{K}$ For each run of each experiment, we check if the number of estimated clusters at $t = 0$ or $t = 1$ with the Modified REM algorithm is correct. We report the correctly estimated K rates in Table 3.3. We consider that our STMP correctly estimates K over time if and only if $\hat{K}^{(0)} = K^{(0)}_{true}$ and $\hat{K}^{a} = K^{(1)}_{true}$, with $\hat{K}^{(0)}$ and $\hat{K}^{a}$ estimated by Modified Robust EM at $t = 0$ and $t = 1$ respectively. In brief, the correctly estimated K number is given by the intersection of correctly estimated $\hat{K}^{(0)}$ and $\hat{K}^{a}$.

Cases I. to IV. give high rates, explained by the correct separation of the clusters as seen in Figure 3.3. On experiments with configurations bringing closer two clusters (Cases V. to IX.), we obtain high rate (over 90%) for static and well-enough separated clusters (Setup F., Setup M.). This score is also high for displacement from Setup F. to Setup M. (Case VII.).

When we consider moving clusters which are getting too close this score decreases. The global score of STMP executions involving at least one Setup C. distribution is affected by the superposition of two clusters. The correct proportions are not bigger than 54%. By looking at estimated $\hat{K}^{(0)}$ and $\hat{K}^{a}$ in Table 3.3,the Modified REM algorithm estimates at least 30 over 100 times two classes with samples from Setup C. distribution. Although these estimates are incorrect, they lead to understandable results, as samples from the two left-hand side clusters can be confused (see Fig. 3.4(c)). An example of wrong estimated parameters for Setup C. is presented in Figure 3.7, which confirms the interpretability of the results.

Experiment	Proportion of correctly estimated number of components (values for $\hat{K}^{(0)} = 2$, $\hat{K}^{a} = 2$)
Case I.	96%
Case II.	98%
Case III.	100%
Case IV.	100%
Case V.	92%
Case VI.	42% (30%,32%)
Case VII.	99%
Case VIII.	54% (0%,34%)
Case IX.	54% (0%,37%)

Table 3.3: Proportions of correctly estimated number of components among $S = 100$ runs. At each execution, the estimation is correct iff : $\hat{K}^{a} = K^{(1)}_{true}$ and $\hat{K}^{(0)} = K^{(0)}_{true}$. Configurations are described in Table 3.A.3.2.

Estimation of means and covariances parameters Thereafter, we compute estimation errors for means and covariances matrices on experiments with correctly estimated

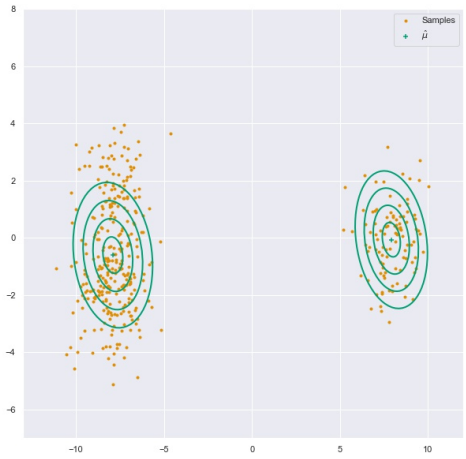


Figure 3.7: An estimated GMM with $\hat{K} = 2 \neq K_{true} = 3$ for a Setup C. distribution. The centers and covariances are represented in green. Orange points are samples.

number of components K (see Table 3.4). It confirms that these estimated Gaussian mixtures are correctly estimated by the Modified Robust EM inside our pipeline STMP. We also notice a poorer average estimate of GMM parameters for datasets from Setup C. As said previously, this parametrization implies that two clusters are mixed up. Estimates of Setup C. models present a slightly higher average Euclidean distance between the true means and the estimated ones. For covariance matrices errors, computed with Frobenius norm, the average errors are less contrasted, but we observe the highest error for M^a estimate in Case VIII. (two clusters are closed to each other at $t = 1$).

	$M^{(0)}$		M^a	
Case	$\hat{\mu}$	$\hat{\Sigma}$	$\hat{\mu}$	$\hat{\Sigma}$
Case I.	1.5 (0.8)	15.4 (7.1)	1.5 (0.9)	14.3 (6.8)
Case II.	1.5 (0.9)	14.4 (7.3)	1.7 (0.9)	16.4 (6.7)
Case III.	1.6 (0.9)	14.3 (6.9)	1.2 (0.7)	11.0 (4.8)
Case IV.	1.5 (0.8)	14.0 (6.7)	1.5 (0.8)	13.6 (5.8)
Case V.	1.7 (1.2)	16.6 (12.2)	1.7 (0.9)	15.6 (7.8)
Case VI.	4.2 (16.6)	22.8 (22.0)	3.2 (3.8)	23.6 (25.5)
Case VII.	1.5 (0.9)	14.8 (6.6)	1.6 (1.0)	15.8 (7.8)
Case VIII.	1.5 (0.9)	14.3 (8.2)	3.7 (4.6)	29.0 (30.3)
Case IX.	1.7 (1.0)	16.2 (8.5)	2.9 (3.7)	24.8 (29.9)

Table 3.4: Mean (standard deviation) relative errors (expressed as a percentage) for the estimated means and covariance matrices within each case, over all runs having correctly estimated $\hat{K}$ inside STMP. The Euclidean norm is used for means, and the Frobenius norm for covariances.

4.4.2 Performances of STMP as an alert system

We have defined in the previous subsection the threshold to alert the user that there may be a population change between two given time points. As explained in Subsection 3.3, there

is also an alert when a component tries to disappear (leading to degenerated covariance matrix) when estimated by the C-EM. And when covariance matrices estimated by the C-EM are too different from the previous step ones. With these warning systems, we defined a whole pipeline, named STMP, to monitor the dynamic of the population and raise alerts when reasonable changes occur. We are now demonstrating the performances of STMP. Using an alert threshold of $\tau = 1.1$, we obtain the following alert rates, that we can retrieve in the Figure 3.6. For Case I., Case V. and Case VI. we obtain 98% true negative alerts respectively. For Cases II. to IV. we obtain a true positive alert rate of 100%, detecting all changes in population distribution with our STMP.

STMP does not raise an alert when the distributions differ barely in time. This is due to our likelihood ratio threshold fixed to $\tau = 1.1$. The true positive number of alerts is of 2% for Case VII. and 1% for Case IX.. In contrary, the bigger movement in Case VIII. leads to a true positive number of 29%. This brings us to the problem that STMP can not raise an alert when GMM are hard to estimate correctly, as here. This experiment involves the Setup C., which is complex to estimate for EM algorithms.

Last but not least, our proposed method is computationally efficient with a very low computational time. All experiments on datasets of size $n = 400$ are performed with an average execution time of 1.33s. We recover average execution time by case type in Table 3.5. Fast execution was also a criterion leading the construction of our method, and satisfying for our future applications.

Experiment	Average computation time over $S = 100$ runs (std)
Case I.	1.24s (0.19)
Case II.	1.14s (0.18)
Case III.	1.55s (0.50)
Case IV.	1.35s (0.41)
Case V.	1.33s (0.14)
Case VI.	1.45s (0.32)
Case VII.	1.23s (0.10)
Case VIII.	1.31s (0.27)
Case IX.	1.35s (0.20)

Table 3.5: Average (and standard deviation) computation time of the different case experiments, with $n_0 = n_1 = 400$ points.

4.4.3 Effects of the dataset size on STMP

In previous explained experiments on synthetic data, we fixed the data set size to $n = 400$. Afterwards, we study the impact of the number of points for Cases I. to IX. described previously, with $n \in \{100, 200, 400\}$. With the same true distributions as in Figures 3.3 and 3.4, we perform $S = 100$ runs of our process with data samples of size $n = 200$ and $n = 100$ at each time step.

As expected, decreasing n decreases the proportion of good estimated $\hat{K}$ over the 100 runs and inherently the quality of estimation of parameters K (Table 3.6). For $n = 200$ points and cases with Gaussian clusters not too close to each other the Modified Robust EM algorithm gives a high rate of correct estimation of K. Concerning Cases I. to V. and Case VII., the rates are between 76% and 92%, allowing us to be confident in the estimates. For cases implying the Setup C. the estimated GMM are worse, because two true Gaussian clusters are almost overlapping. With $n = 100$, it becomes complicated to properly estimate a GMM even with well-defined clusters: the best alert rate is 61% and the worst is 8%.

Therefore, we must be aware that decrease the number of samples decrease the proportion of good estimates $\hat{K}$ and inherently the quality of estimation of parameters θ in our Modified Robust EM.

Experiment	Proportions of correct estimates $\hat{K}$ with $n = 400$	Proportions of correct estimates $\hat{K}$ with $n = 200$	Proportions of correct estimates $\hat{K}$ with $n = 100$
Case I.	**96%**	**90%**	61%
Case II.	**98%**	**87%**	46%
Case III.	**100%**	**92%**	61%
Case IV.	**100%**	**87%**	59%
Case V.	**92%**	76%	42%
Case VI.	42%	20%	8%
Case VII.	**99%**	**81%**	51%
Case VIII.	54%	29%	22%
Case IX.	54%	34%	29%

Table 3.6: Proportions of correctly estimated number of components with Modified REM among $S = 100$ runs. At each execution, the estimation is correct iff: $\hat{K}^a = K^{(1)}_{true}$ and $\hat{K}^{(0)} = K^{(0)}_{true}$. Bold numbers are performances of at least 80%.

But overall, the STMP performance is less affected than the Modified Robust EM by changes of data sets size (Table 3.A.3.1). As we saw in Subsection 4.4.2, we reach 98% true negative alerts for data sets of size $n = 400$. For data sets of size $n = 200$ we have 19 false positive alerts, and for $n = 100$ we have 56 false positive alerts. For Cases II. to IV. the proportion of success is 100% for all n values (Table 3.A.3.1). It even raises more alerts with fewer points on Case VII. to Case IX., due to overlapping Gaussian components which are estimated as one single component. For example if a $t = 1$ we are in Setup C. (Fig. 3.4(c)), as two Gaussian components are hardly separable the pipeline will estimate one cluster for the two components, and raise an alert as it is evolving away from the estimated distribution at $t = 0$ (which could be Setup F. or M.).

Even if the Modified Robust EM becomes less accurate with smaller data sets, our pipeline still produces interpretable and meaningful results. The problem of decreasing performance on small dataset estimation should be solved directly on the Modified Robust EM.

5 Application of STMP on a real life use case

In this section we demonstrate the relevance of STMP with GMM on real epidemiological data from COVID-19 in Paris, France.

5.1 Presentation of the dataset

AP-HP (Assistance Publique-Hopitaux de Paris) is the largest hospital entity in Europe with 39 hospitals (22,474 beds) mainly located in the greater Paris area with 1.5 M hospitalizations per year (10% of all hospitalizations in France). Since 2014, the AP-HP has deployed an analytics platform based on a clinical data repository, aggregating day-to-day clinical data from 8.8 million patients captured by clinical databases. An "EDS-COVID" database stemmed from this initiative. The AP-HP COVID database retrieved electronic health records from all AP-HP facilities and aggregated them into a clinical data warehouse. The clinical data warehouse allows for a large set of data to be retrieved in real time to

deeply characterize hospitalized patients, including their residential address. New patients who tested positive by polymerase chain reaction (PCR) as being infected by SARS-CoV-2 from the 24th of February to the 10th of May 2020 (weeks 9 to 19), in one of the AP-HP hospitals and living in Paris constitute the dataset for this study. During this time period, tests availability outside public hospital facilities were very limited, and therefore we can consider in this study that this sample constitutes a representative sample of patients having been positively tested during this period. To preserve privacy, residential addresses were extracted at the IRIS level, which is a geographical division in France of residential units of 2000 inhabitants on average.

Integration into the spatio-temporal mixture process For each patient we have two pieces of information: the week they were diagnosed positive, and their place of residence at the IRIS level. We therefore use a week as a time step t in our process. Beginning from the first week (week 9), which corresponds to the beginning of the pandemic in France, we apply our STMP, keeping at each time t one of the models M^a or M' according to the criterion defined in Subsection 3.1 with threshold τ given in Section 4. We have 5621 positive diagnosed patients over all weeks and all Paris IRIS areas. Table 3.7 informs us that the number of cases per week is not homogeneous, as in first weeks, few cases living in Paris were detected.

Week	Number of positive diagnosed people per week
9	5
10	18
11	272
12	965
13	1666
14	1297
15	695
16	366
17	209
18	114
19	14

Table 3.7: Distribution of positive diagnosed people to COVID-19 over weeks.

5.2 Comparison of the Robust EM and the Modified Robust EM

As described in Section 3.2, the Robust EM algorithm (Yang et al., 2012) has a convergence problem, revealed on real datasets. Even if this algorithm is dynamic, it can be stuck in an incorrect local maximum involving overlapping clusters. This phenomenon has been detected on the real dataset of COVID-19 positive cases in Paris area. We compare here the estimated GMMs by original Robust EM and by our Modified REM, which is correcting this overlapping effect (Subsection 3.2).

A phenomenon revealed on real data On all weeks except week 13, the Robust EM presents no overlapping clusters. It returns acceptable estimated mixture models. As there are no abnormalities in the estimation process, our Modified Robust EM returns similar results. It is illustrated by Figures 3.A.1(a) and 3.A.1(b) showing estimations on week 12

for both algorithms. On week 13, the Robust EM algorithm presents overlapping clusters. The final number of classes is 11, but the Figure 3.A.1(c) shows us that there are only nine clusters. We can only detect superimposed clusters by doing a post-processing analysis, which consists of checking the estimated parameters. Table 3.A.4.1 gives these estimated parameters for both the Robust EM (Yang et al., 2012) and the Modified REM. From this table we see that there are two pairs of superimposed clusters with mixture estimation by Robust EM. By executing our Modified REM on the same week, independently of the other time steps, we obtain nine clusters (Fig. 3.A.1(d) and Table 3.A.4.1), confirming that if we merge redundant clusters, we obtain a stable solution, accepted by the algorithm. Our Modified REM solves the problem of superimposed clusters. This avoids having to consider post-processing inside STMP, which would require a user action at each time step. It also allows us to solve a specific problem: to correctly model COVID-19 data in space and time.

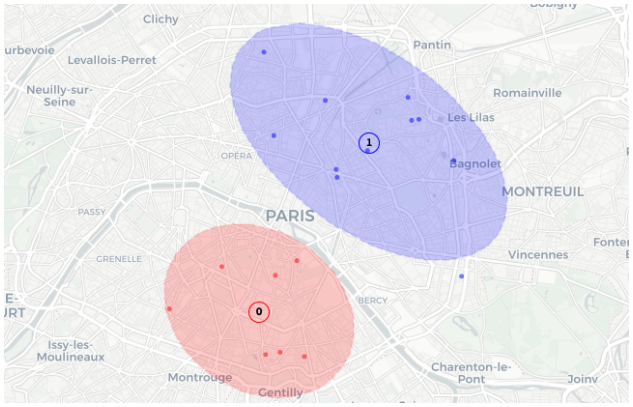
5.3 Results of Spatio-Temporal Mixture Process

The aim here is to underline presence or absence of temporal constancy in COVID-19 data. A temporal constancy would suggest that the population distribution was stable at the peak of the pandemic. This is in line with epidemiological studies that were showing a "peak" around these weeks after the first propagation phase (weeks 9 to 12) (see weekly reports of Public Health Institution (France, 2020, Page 7, Figure 8)).

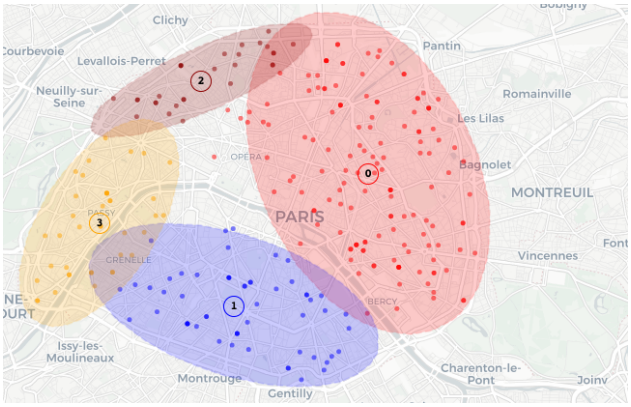
We use the fixed alert threshold $\tau = 1.1$ in our pipeline, as estimated by previous experiments in Section 4. On the first week (week 9), as the number of cases is very low, the initial Modified Robust EM converges towards the removal of all clusters. The final parameters correspond here to the initial ones, so we observe on Figure 3.8(a) initial high variances and that each case is its proper cluster. The week 10 is still presenting a low number of scattered cases, which are modelled by two global clusters, geographically distributed on both sides of the river Seine. As from week 10 to week 11 (and week 11 to week 12) the number of cases is highly increasing, the model accepts new estimated parameters θ^a. Our STMP reveals that the GMM estimated on week 12 with $\hat{K}^{(12)} = 10$ was accepted on 13^{th} and 14^{th} weeks. For reference, week 12 represents the beginning of the lockdown and week 13 represents the peak of the pandemic, in terms of new positive cases. This means that C-EM executed across 13^{th} and 14^{th} weeks fits very well the new dataset each week with a source model estimated on week 12. Even if the number of cases changes over time, STMP is able to detect a constant distribution. This is consistent with the patients' distribution on weeks 12, 13 and 14 as we can see on Figures 3.8 and 3.9. On week 15, STMP rejects the hypothesis that the patients' dataset is approximated by the mixture law estimated on previous weeks. The alternative model M^a is accepted. Parameters $\theta^{(15)}$ on week 15 are newly estimated, evolving too far from $\theta^{(14)}$, estimated parameters of week 14. It can be interpreted as the strong decrease of new positive cases such as the disappearance of large clusters from previous weeks and the detection of large and global clusters, corroborated by the Figure 3.9(a). On week 16, these three clusters from week 15, large and non-informative, are accepted by STMP. On the following weeks (weeks 17, 18 and 19), the number of cases is still decreasing. As on first weeks, the small number of cases leads to accept totally new estimated parameters θ^a each week, without link with previous weeks.

From Table 3.8, the likelihood ratio values are globally distant from our defined threshold $\tau = 1.1$, leaving no doubt about the choice of the best parameters $\theta^{(t)}$ at each time step t. Only on week 15 the likelihood ratio value is smaller that our defined threshold while the temporal-dependent model M' is rejected. This rejection is due to large variations in the covariance matrices during the C-EM stage. The model M' fits the new dataset by excessively moving the covariance parameters inherited from $M^{(14)}$. There is no likelihood

(a) Week 9 (There are only 5 cases, each case is center of its own cluster.)

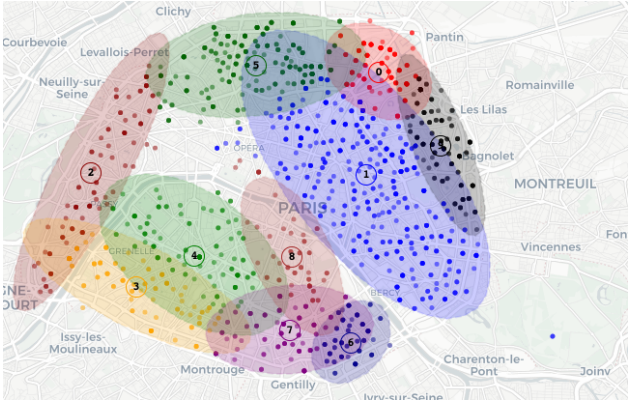


(b) Week 10

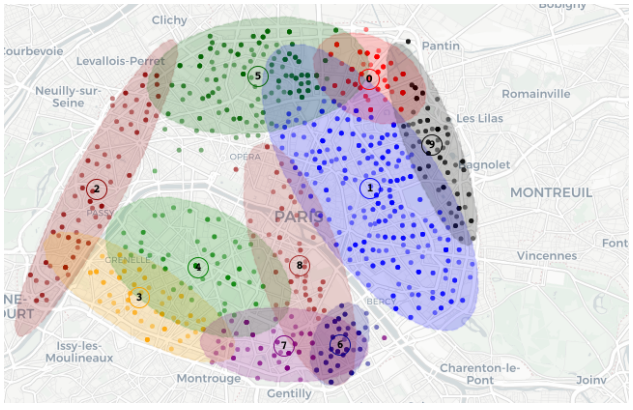


(c) Week 11

(d) Week 12

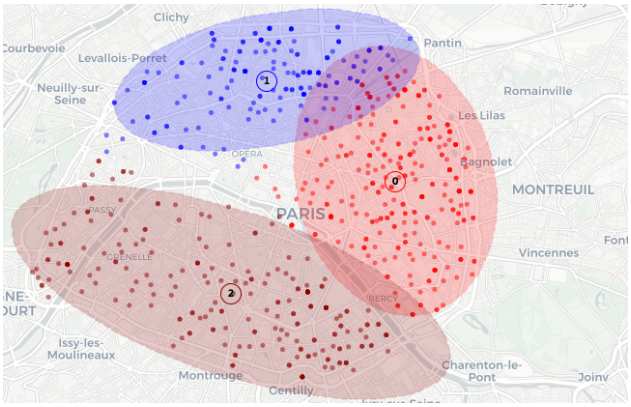


(e) Week 13

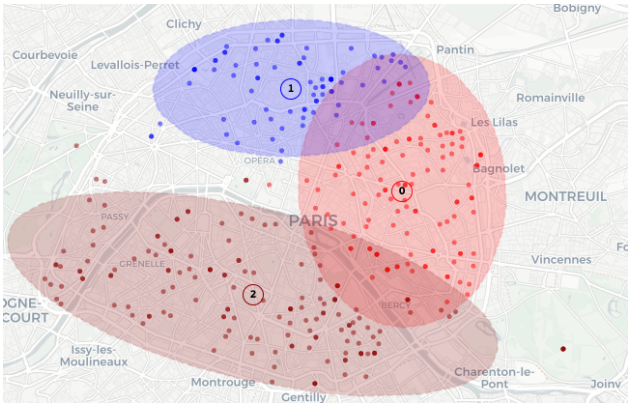


(f) Week 14

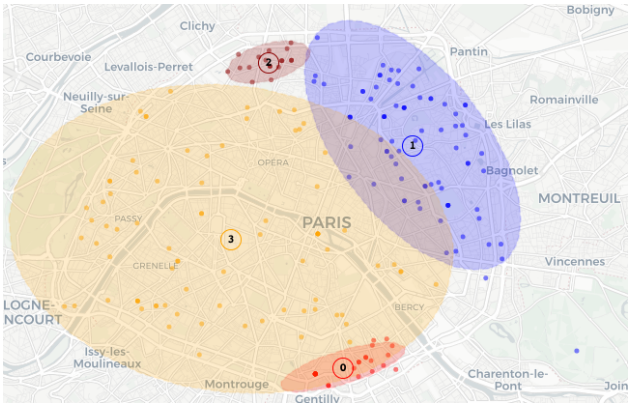
Figure 3.8: Estimated Gaussian Mixture Models on COVID-19 dataset per week (weeks 9 to 14).



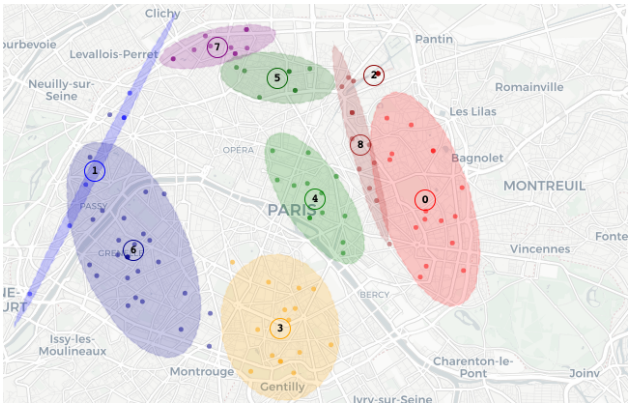
(a) Week 15



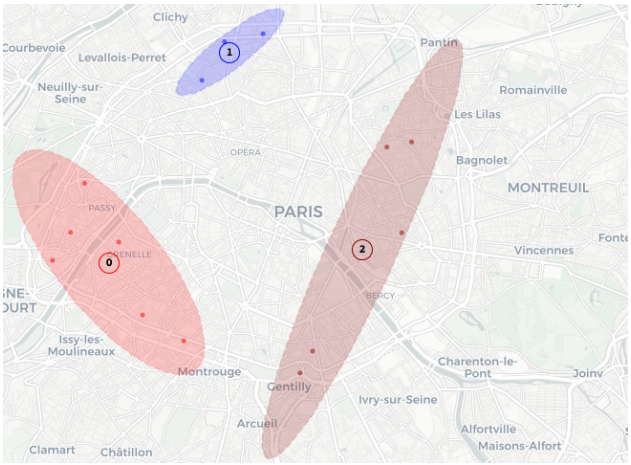
(b) Week 16



(c) Week 17



(d) Week 18



(e) Week 19

Figure 3.9: Estimated Gaussian Mixture Models on COVID-19 dataset per week (weeks 15 to 19).

ratio value in the last week. This ratio is incalculable due to the "empty class phenomenon". The model M' tries to remove a component which leads to an early stop of the estimation process of this model. This triggers the inevitable choice of the alternative model and raises an alert.

Week	Estimated number of classes $\hat{K}$ by M^a	Estimated number of classes $\hat{K}$ by M'	Likelihood Ratio r	Accepted model
9	**5**	None	None	$M^{(0)}$
10	**2**	5	1.383	M^a
11	**4**	2	1.376	M^a
12	**10**	4	1.171	M^a
13	9	**10**	1.088	M'
14	5	**10**	0.950	M'
15	**3**	10	0.87	M^a
16	5	**3**	1.080	M'
17	**4**	3	1.307	M^a
18	**9**	4	2.215	M^a
19	**3**	9	computationally invalid	M^a

Table 3.8: Results of our process on positive diagnosed people in AP-HP hospitals with a time step being a week. Bold numbers are corresponding final number of classes per week.

From the mathematical and algorithmic point of view we obtain interesting results, showing that C-EM can over time sufficiently model evolving real-world data with a relatively stable and high dataset size.

5.4 Interpretations

The results obtained with STMP on this COVID-19 dataset are coherent with public health policy and COVID-19 transmission patterns during this time period. Lockdown in France took place from the 17^{th} of March (beginning of week 12) to the 1^{st} of June. As it takes about two weeks to go from contamination to cytokine storm, no evolution in clusters was expected from week 12 to 14. Thereafter, a decrease in the number of clusters was expected, associated with a moving distribution. Moreover, estimated clusters concentrate closed to Paris periphery, which are low-income neighborhoods, known to favor COVID-19 transmission. The reject on week 15 of the previous time step model can be interpreted as the effect of the lockdown, and we can observe the natural barrier formed by the Seine, as people in France could only move in a perimeter of one kilometer during this lockdown. The numerous clusters during week 18 are residual clusters not solved by the lockdown. They mainly correspond to concentrated positive cases areas, whereas in the rest of the city there are few and scattered cases.

5.5 Improve estimation of overly dispersed datasets

In our research on real datasets of rare diseases, we observed that estimation of mixtures on small and widely scattered data can really be complicated. The joint estimation of mixture models and the number of classes becomes even more difficult, and can lead to pathological, unreliable results. The data we observe are very scattered, but at the same time aggregated to a spatial grid that causes the superposition of several points (Fig. 3.10). From one time step to the next, one can visually tell that the data changes a lot, and that there are small clusters everywhere as some data points are superimposed. But at the same time one wants

correct and stable models, which can approximate small clusters as well as the variability of whole data.

As we will illustrate below, the Modified REM also has difficulties to estimate this type of data. It can lead to estimated models with almost singular covariance matrices, or not be able to keep more than one or two clusters to cover the whole population (Fig. 3.11(a), 3.11(c), 3.11(e)).

One way to improve estimation in this situation is to introduce a regularization on the covariance matrices. Existing works consider an Inverse Wishart prior for the covariance matrices (Fraley and Raftery, 2007; Baudry and Celeux, 2015; Fop et al., 2019), with $\boldsymbol{\Sigma} \sim W^{-1}(w, \boldsymbol{W})$, w degrees of freedom and $\boldsymbol{W}$ scale matrix of prior distribution.

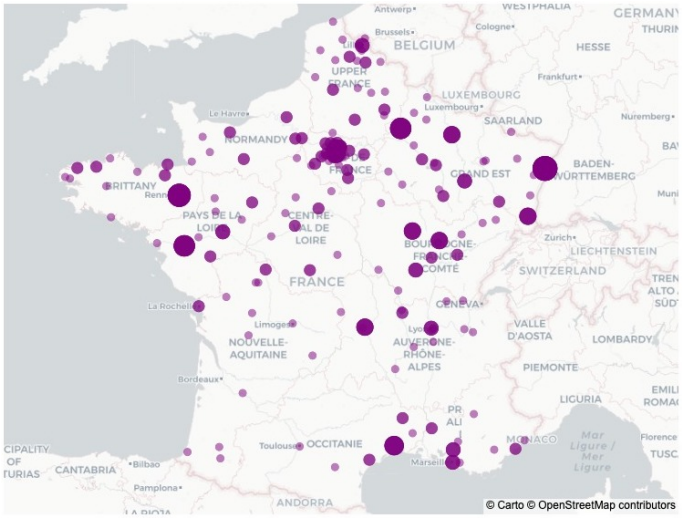
Scale matrices updates $\hat{\Sigma}_k^p$ in the M-step of any EM-like algorithm are replaced at iteration p (see Fraley and Raftery, 2007; Baudry and Celeux, 2015) by

$$\hat{\boldsymbol{\Sigma}}_k^{p,reg} = \frac{\hat{\boldsymbol{\Sigma}}_k^p + \boldsymbol{W}}{n_k + w + d + 1}\,. \tag{3.8}$$

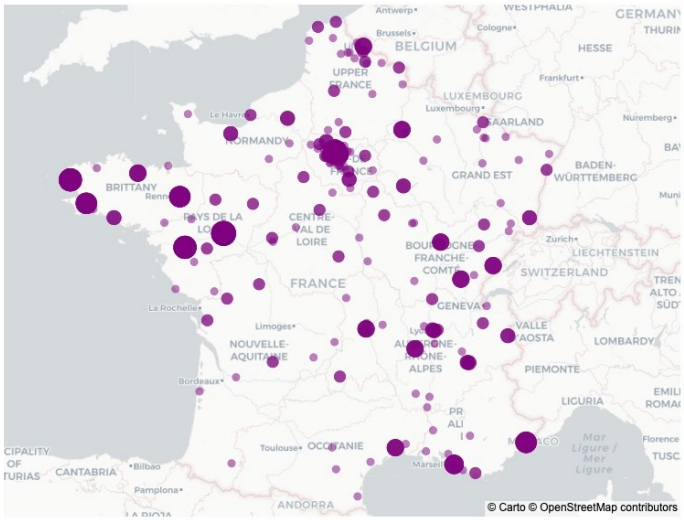
We see that the regularization matrix $\mathbf{W}$ is common to all the clusters. This consideration was adopted by (Fraley and Raftery, 2007; Baudry and Celeux, 2015), and by Fop et al. (2019) on graphs. The remaining point is definition of $\mathbf{W}$ and w, which results in different propositions. Fraley and Raftery proposed $\mathbf{W} = \frac{S}{K^{2/d}}$ with S the overall empirical covariance matrix, and $w = d + 2$. This definition of $\mathbf{W}$ does not rely on any user-chosen value. Later on, Baudry and Celeux (2015) proposed a new regularization which is smoother than this one, but requires to fix an *a priori* weight on its importance.

We introduce here a scale prior for estimation of covariance matrices in Modified REM. In this case, the covariance estimation step in Algorithm 3.2 can be replaced by Eq.(3.8) with $\hat{\boldsymbol{\Sigma}}_k^p = \hat{\boldsymbol{\Sigma}}_{k,\text{MLE}}^p \times \sum_{i=1}^n \tau_i^k$ and $\hat{\boldsymbol{\Sigma}}_{k,\text{MLE}}^p$ being estimated by Eq.(3.12). There is no more use of γ regularization constant and P matrix (see line 5 in Algorithm 3.2).

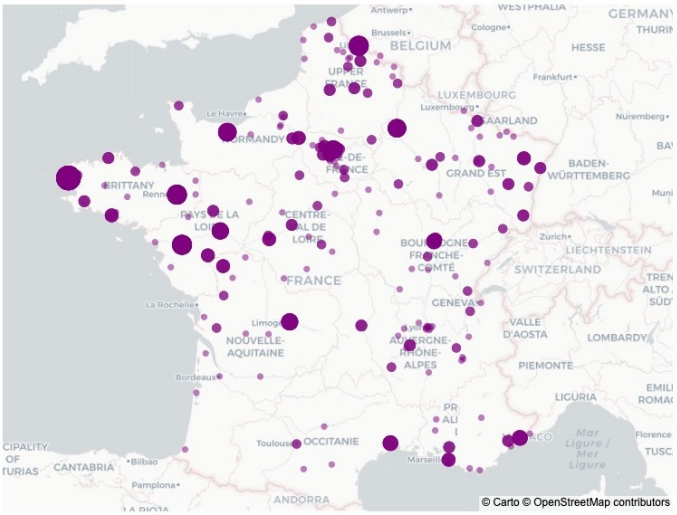
We retrieve on Figure 3.11 estimations on several time steps of MREM with or without regularization. Without regularization, there can be more consistency between the different time steps, but also clusters with little information (Fig. 3.11(a)), or as on Figure 3.11(c) several clusters on a restrictive area without whole variability consideration, while MREM with regularization on the same data gives more variable but also more informative clusters (Fig. 3.11(b),3.11(d)). Another highlighted problematic behavior can be seen on Figure 3.11(e), where we observe a cluster with pathological covariance. Although this obviously raises the question of the relevance of the continuous law considered, the addition of a regularization already allows a better estimate (Fig. 3.11(f)).



(a) Time step 1

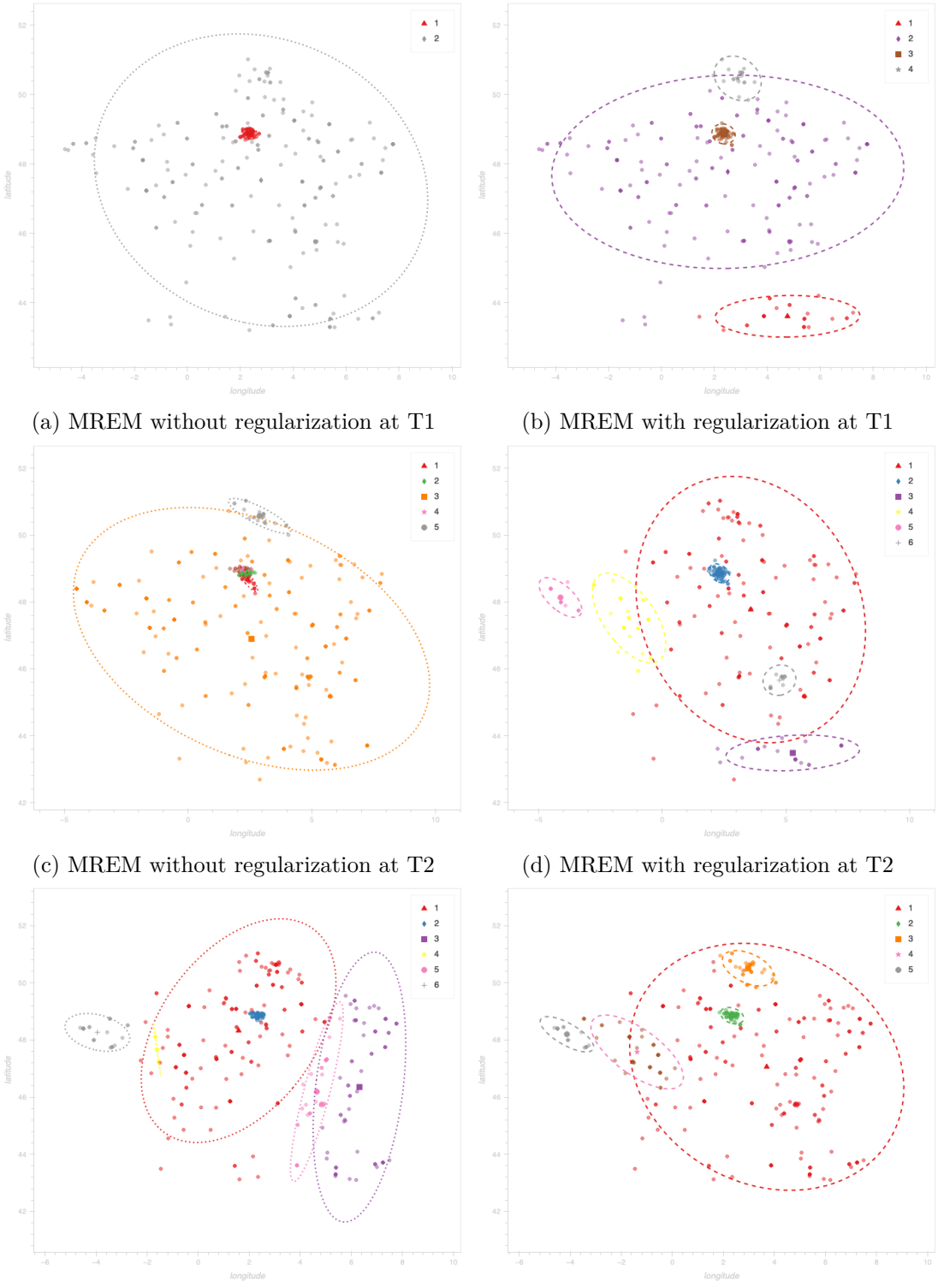


(b) Time step 2



(c) Time step 3

Figure 3.10: Real datasets at three different time steps. The number of data items at the exactly same location gives the size of the point on the figure.



(a) MREM without regularization at T1

(b) MREM with regularization at T1

(c) MREM without regularization at T2

(d) MREM with regularization at T2

(e) MREM without regularization at T3

(f) MREM with regularization at T3

Figure 3.11: Results of estimations by Modified REM (MREM) without (on the left) and with (on the right) covariance regularization for three time steps.

6 Conclusion

We have proposed a complete and generic pipeline for modeling evolution of population distribution, and detecting abnormal changes in this distribution. This STMP was combined with new EM algorithm variants. Our application on public health data shows that this STMP models population distributions well, and raises meaningful alerts.

The STMP for monitoring population distributions and the algorithms to estimate the models are independent objects. This enables future directions of our work when integrating covariates following non-Gaussian distributions in the mixture. We will still be able to use our proposed algorithms as they are blind to the distributions in the mixture.

On the other hand, the performance of the EM algorithms depends on the dataset sizes. We could also try to introduce a temperature parameter in the Modified Robust EM as proposed by Allassonnière and Chevallier (2021) to improve estimations in unstable situations.

Finally, the decision rule was empirically fixed in this work. In future work this decision rule could be modeled as an acceptation probability, taking advantage of Monte Carlo Markov Chains theory.

3.A Appendices

3.A.1 Equations for mixture parameters estimation in the original EM algorithm (Dempster et al., 1977)

The EM algorithm alternates between the two following steps (until convergence criterion is met). At the p^{th} iteration of the algorithm, the update equations are given by:

- E-step : Compute the conditional expectation of the complete log-likelihood. Latent variables z_i^k are discrete, so their conditional expectations are given by

$$\begin{aligned} p_\theta(z_i^k = 1|x_i) &= \frac{\pi_k \mathcal{N}_d(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}_d(x_i|\mu_j, \Sigma_j)} \\ &= \tau_i^k(\theta)\,. \end{aligned} \tag{3.9}$$

- M-step : Update the parameter estimates:

$$\hat{\pi}_{k,\text{MLE}}^p = \frac{1}{n}\sum_{i=1}^n \tau_i^k\,, \tag{3.10}$$

$$\hat{\mu}_{k,\text{MLE}}^p = \frac{\sum_{i=1}^n \tau_i^k x_i}{\sum_{i=1}^n \tau_i^k}\,, \tag{3.11}$$

$$\hat{\Sigma}_{k,\text{MLE}}^p = \frac{\sum_{i=1}^n \tau_i^k (x_i - \mu_i)^\top (x_i - \mu_i)}{\sum_{i=1}^n \tau_i^k}\,. \tag{3.12}$$

3.A.2 Pseudo-Code of the Modified Robust EM presented in Section 3

Algorithm 3.2: Modified Robust EM

Initialization : data set $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\varepsilon > 0$
$K^0 \leftarrow n$
$p \leftarrow 0$, $\beta^0 \leftarrow 1$
$\pi_k^0 \leftarrow 1/n$, $\mu^0 \leftarrow \mathbf{X}$
$\Sigma_k^0 \leftarrow D_{k(\lceil\sqrt{K^{\text{initial}}}\rceil)}\mathbf{I}_d$ with
$D_k = \text{sort}\left\{d^2_{k(i)} = \|x_i - \mu_k\|^2 : d^2_{k(i)} > 0, \quad i \neq k, \quad 1 \leq i \leq n\right\}$
Compute $\tau_i^{k,0}$ with (3.9)
$p \leftarrow 1$
Compute μ_k^p with (3.11)
while $\max_{1\leq k\leq K^p} \|\mu_k^{p+1} - \mu_k^p\|_2 > \varepsilon$ *or Eq.* (3.6) *is verified for some clusters* **do**
  Compute π_k^p with (3.4)
  $\pi_{(1)}^{\text{EM}} \leftarrow \underset{1\leq k\leq K^p}{max}\, \pi_k^{p,\text{EM}}$, $\pi_{(1)}^{(old)} \leftarrow \underset{1\leq k\leq K^p}{max}\, \pi_k^{(old)}$
  $E \leftarrow \sum_{k=1}^{K^p} \pi_k^{(old)} \ln \pi_k^{(old)}$
  $\beta^p \leftarrow \min\left\{\frac{\sum_{k=1}^{K^{p-1}} \exp\left(-\eta n\left|\pi_k^p - \pi_k^{(old)}\right|\right)}{K^{p-1}}, \frac{\left(1-\pi_{(1)}^{\text{EM}}\right)}{\left(-\pi_{(1)}^{(old)} E\right)}\right\}$
  Update class number K^{p-1} to K^p by deleting classes with $\pi_k^p < 1/n$, then adjust π_k^p and $\tau_i^{k,p-1}$
  if $K^{p-1} \neq K^p$ **then**
    $p_{component} \leftarrow 1$ `/* variable to keep in memory the number of iterations with a stable number of components */`
  end

  if $p \geq p_{min}$ *and* $p_{component} \geq 100$ **then**
2   **if** *no superimposed clusters (condition* (3.6) *not fulfilled)* **then**
      $\beta^p \leftarrow 0$
3   **else if** *superimposed clusters and* $p_{component} < 200$ **then** `/* give more time to the algorithm to converge */`
      $p_{min} \leftarrow p_{min} + 50$
4   **else** merge superimposed clusters
      adjust π^p, μ^p, Σ^p and τ^{p-1} by removing redundant clusters and computing τ^{p-1}
    end
  end
5 Compute $\Sigma_{k,\text{MLE}}$ with (3.12) and $\Sigma_k^p = (1-\gamma)\Sigma_{k,\text{MLE}} + \gamma P$ with $\gamma = 0.0001$, $P = d^2_{\min}\mathbf{I}_d$, $d^2_{\min} = \min \|x_i - x_j\|_2^2 > 0$, $1 \leq i, j \leq n$
  Compute $\tau_i^{k,p}$ with (3.9)
  Compute μ_k^{p+1} with (3.11)
  $p \leftarrow p + 1$
  $p_{component} \leftarrow p_{component} + 1$
end

3.A.3 Supplementary analyses of Section 4

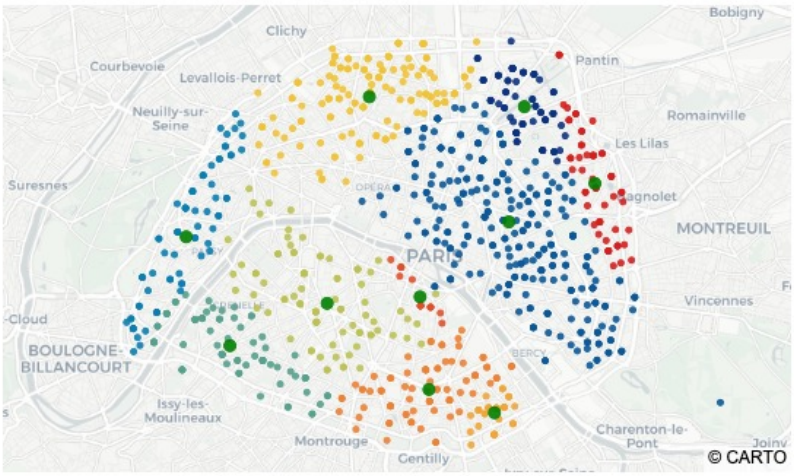
n	Threshold	Case I.	Case II.	Case III.	Case IV.	Case V.	Case VI.	Case VII.	Case VIII.	Case IX.
400	1.00	78%	**100%**	**100%**	**100%**	99%	67%	**99%**	**96%**	55%
	1.05	**2%**	**100%**	**100%**	**100%**	**2%**	21%	4%	69%	1%
	1.10	**2%**	**100%**	**100%**	**100%**	**2%**	**2%**	2%	29%	1%
	1.15	**2%**	**100%**	**100%**	**100%**	**2%**	**1%**	1%	7%	1%
	1.20	**2%**	**100%**	**100%**	**99%**	**2%**	**1%**	1%	4%	1%
	1.25	**2%**	**100%**	**100%**	**92%**	**2%**	**1%**	1%	4%	1%
	1.30	**2%**	**100%**	**100%**	**81%**	**2%**	**1%**	1%	2%	1%
	1.35	**2%**	**100%**	**99%**	73%	**2%**	**1%**	1%	2%	1%
	1.40	**2%**	**100%**	**98%**	72%	**2%**	**1%**	1%	2%	1%
200	1.00	80%	**100%**	**100%**	**100%**	91%	55%	**98%**	**94%**	58%
	1.05	19%	**100%**	**100%**	**100%**	22%	38%	27%	71%	26%
	1.10	19%	**100%**	**100%**	**100%**	18%	20%	16%	47%	21%
	1.15	18%	**100%**	**100%**	**100%**	16%	18%	14%	38%	19%
	1.20	17%	**100%**	**100%**	**96%**	13%	17%	14%	30%	19%
	1.25	17%	**100%**	**100%**	**91%**	13%	17%	13%	29%	19%
	1.30	17%	**100%**	**100%**	**81%**	13%	17%	13%	28%	19%
	1.35	17%	**100%**	**99%**	75%	13%	17%	13%	26%	19%
	1.40	17%	**100%**	**98%**	71%	13%	17%	13%	26%	19%
100	1.00	78%	**100%**	**100%**	**100%**	89%	74%	**96%**	**96%**	**84%**
	1.05	62%	**100%**	**100%**	**100%**	72%	71%	75%	**89%**	74%
	1.10	56%	**100%**	**100%**	**100%**	70%	63%	70%	**80%**	71%
	1.15	56%	**100%**	**100%**	**100%**	69%	54%	70%	77%	67%
	1.20	54%	**100%**	**100%**	**99%**	64%	51%	68%	72%	64%
	1.25	54%	**100%**	**100%**	**97%**	60%	51%	67%	68%	63%
	1.30	53%	**100%**	**100%**	**93%**	60%	51%	64%	67%	63%
	1.35	53%	**100%**	**98%**	**89%**	58%	49%	63%	66%	62%
	1.40	53%	**100%**	**97%**	**86%**	58%	49%	63%	66%	62%

Table 3.A.3.1: Percentage of raised alerts by STMP for each experiment ($S = 100$ runs) on datasets of n points. For each size n and each case is provided the number of alerts for different values of the alert threshold. Bolded numbers are the best percentages: equal or over 80% when the algorithm has to raise an alert, and under 10% when no alert should be raised. Color (red) represents percentages corresponding to selected threshold of 1.1.

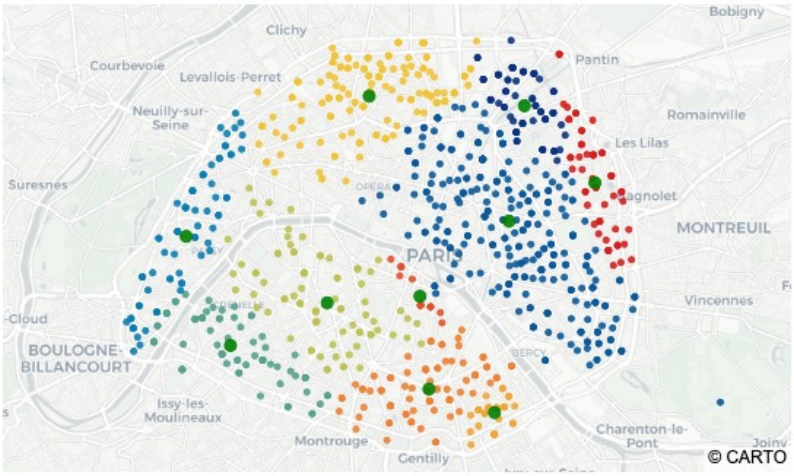
Study case reference	Description at $t=0$	Description at $t=1$	$K_{true}^{(0)}$	$K_{true}^{(1)}$	Parameters at $t=0$	Parameters at $t=1$
Case I.	Setup F.	Same distributions (Setup F.)	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$
Case II.	Setup F.	Emergence of a cluster	3	4	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=\pi_2=\pi_3=\pi_4=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0), \mu_4=(-2.45,6.57)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0, \Sigma_4=\begin{pmatrix}0.88 & 0.\\ 0. & 0.48\end{pmatrix}$
Case III.	Setup F.	Disappearance of a cluster	3	2	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=\pi_2=0.5$,$\mu_1=(-8,3.5), \mu_2=(8,0)$,$\Sigma_1=\Sigma_2=\Sigma_0$
Case IV.	Setup F.	Modification of a cluster	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=\pi_2=\pi_3=1/3$,$\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(6.09,-2.71)$,$\Sigma_1=\Sigma_2=\Sigma_0, \Sigma_3=\begin{pmatrix}1.36 & 0.\\ 0. & 0.92\end{pmatrix}$
Case V.	Setup M.	Setup M.	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-2.5), \mu_2=(-8,2.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$,$\mu_1=(-8,-2.5), \mu_2=(-8,2.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$
Case VI.	Setup C.	Setup C.	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$,$\mu_1=(-8,-1.8), \mu_2=(-8,1.8), \mu_3=(8,0)$,$\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$,$\mu_1=(-8,-1.8), \mu_2=(-8,1.8), \mu_3=(8,0)$,$\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$
Case VII.	Setup F.	Setup M.	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-2.5), \mu_2=(-8,2.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$
Case VIII.	Setup F.	Setup C.	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-3.5), \mu_2=(-8,3.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-1.8), \mu_2=(-8,1.8), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$
Case IX.	Setup M.	Setup C.	3	3	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-2.5), \mu_2=(-8,2.5), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$	$\pi_1=0.5, \pi_2=\pi_3=0.25$, $\mu_1=(-8,-1.8), \mu_2=(-8,1.8), \mu_3=(8,0)$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_0$

Table 3.A.3.2: Different cases of data distribution changes from one time point to the next one (here only considering $t=0$ and $t=1$). Note that $\Sigma_0=\begin{pmatrix}1. & 0.\\ 0. & 1.5\end{pmatrix}$

3.A.4 Results on the COVID-19 data set of Section 5

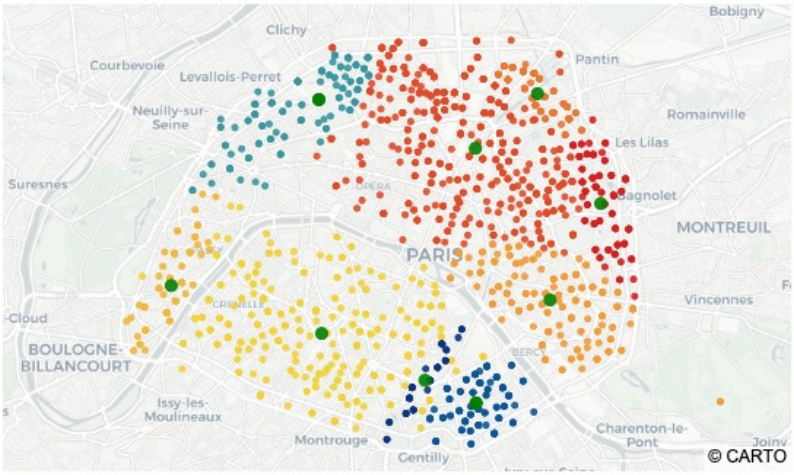


(a) Results of the original Robust EM on week 12.

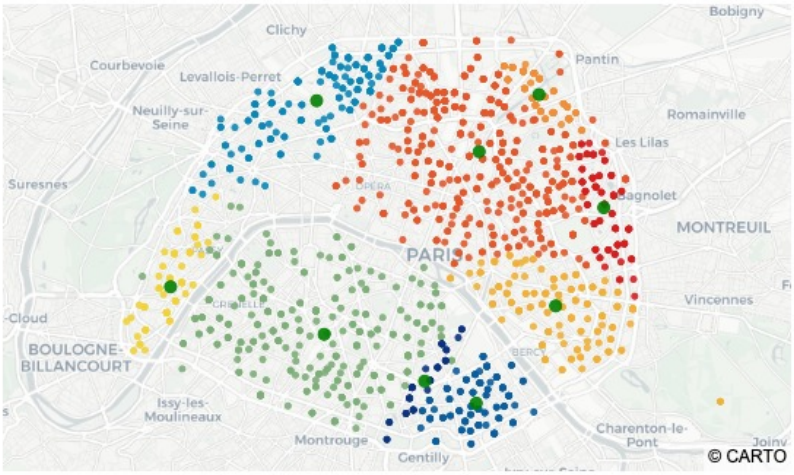


(b) Results of our Modified Robust EM on week 12.

Figure 3.A.1: Estimated GMM labels and centers by Robust EM (Yang et al., 2012) and Modified Robust EM on COVID-19 data set on weeks 12 and 13. Green dots are centers of the clusters. (a) and (b) for week 12, (c) and (d) for week 13.



(c) Results of the original Robust EM on week 13. We obtain here overlapping clusters.



(d) Results of our Modified Robust EM on week 13.

Figure 3.A.1: Estimated GMM labels and centers by Robust EM (Yang et al., 2012) and Modified Robust EM on COVID-19 data set on weeks 12 and 13. Green dots are centers of the clusters. (a) and (b) for week 12, (c) and (d) for week 13.

Parameter	Robust EM (Yang et al., 2012)	Modified Robust EM
$\hat{\pi}$	$\begin{pmatrix} 0.020 \\ \mathbf{0.036} \\ \mathbf{0.036} \\ \mathbf{0.049} \\ \mathbf{0.049} \\ 0.199 \\ 0.039 \\ 0.105 \\ 0.058 \\ 0.346 \\ 0.063 \end{pmatrix}$	$\begin{pmatrix} 0.016 \\ 0.065 \\ 0.093 \\ 0.209 \\ 0.034 \\ 0.090 \\ 0.053 \\ 0.392 \\ 0.049 \end{pmatrix}$
$\hat{\mu}$	$\begin{pmatrix} 2.349 & 48.831 \\ \mathbf{2.365} & \mathbf{48.826} \\ \mathbf{2.365} & \mathbf{48.826} \\ \mathbf{2.316} & \mathbf{48.888} \\ \mathbf{2.316} & \mathbf{48.888} \\ 2.317 & 48.840 \\ 2.270 & 48.850 \\ 2.388 & 48.847 \\ 2.384 & 48.889 \\ 2.364 & 48.878 \\ 2.403 & 48.867 \end{pmatrix}$	$\begin{pmatrix} 2.349 & 48.831 \\ 2.365 & 48.826 \\ 2.315 & 48.888 \\ 2.317 & 48.840 \\ 2.270 & 48.850 \\ 2.389 & 48.846 \\ 2.384 & 48.889 \\ 2.366 & 48.878 \\ 2.404 & 48.866 \end{pmatrix}$
$\hat{\Sigma}$	$\begin{pmatrix} 3.695e{-5} & 3.240e{-5} \\ 3.240e{-5} & 3.290e{-5} \\ - - - - - & - - - - - \\ \mathbf{6.924e\text{-}5} & \mathbf{1.184e\text{-}5} \\ \mathbf{1.184e\text{-}5} & \mathbf{2.077e\text{-}5} \\ - - - - - & - - - - - \\ \mathbf{6.924e\text{-}5} & \mathbf{1.184e\text{-}5} \\ \mathbf{1.184e\text{-}5} & \mathbf{2.077e\text{-}5} \\ - - - - - & - - - - - \\ \mathbf{3.241e\text{-}4} & \mathbf{1.112e\text{-}4} \\ \mathbf{1.112e\text{-}4} & \mathbf{5.957e\text{-}5} \\ - - - - - & - - - - - \\ \mathbf{3.241e\text{-}4} & \mathbf{1.112e\text{-}4} \\ \mathbf{1.112e\text{-}4} & \mathbf{5.957e\text{-}5} \\ - - - - - & - - - - - \\ 5.979e{-4} & -8.670e{-5} \\ -8.670e{-5} & 9.837e{-5} \\ - - - - - & - - - - - \\ 4.250e{-5} & 4.085e{-5} \\ 4.085e{-5} & 6.039e{-5} \\ - - - - - & - - - - - \\ 2.514e{-4} & -7.294e{-5} \\ -7.294e{-5} & 6.280e{-5} \\ - - - - - & - - - - - \\ 6.213e{-5} & -3.206e{-5} \\ -3.206e{-5} & 2.236e{-5} \\ - - - - - & - - - - - \\ 4.057e{-4} & -1.018e{-4} \\ -1.018e{-4} & 1.3259e{-4} \\ - - - - - & - - - - - \\ 3.807e{-5} & -4.028e{-5} \\ -4.028e{-5} & 8.984e{-5} \end{pmatrix}$	$\begin{pmatrix} 3.689e{-5} & 3.244e{-5} \\ 3.244e{-5} & 3.240e{-5} \\ - - - - - & - - - - - \\ 6.842e{-5} & 1.149e{-5} \\ 1.1486e{-5} & 2.044e{-5} \\ - - - - - & - - - - - \\ 3.229e{-4} & 1.126e{-4} \\ 1.126e{-4} & 6.060e{-5} \\ - - - - - & - - - - - \\ 6.399e{-4} & -9.709e{-5} \\ -9.709e{-5} & 9.991e{-5} \\ - - - - - & - - - - - \\ 3.931e{-5} & 3.834e{-5} \\ 3.834e{-5} & 5.746e{-5} \\ - - - - - & - - - - - \\ 2.236e{-4} & -5.983e{-5} \\ -5.983e{-5} & 5.316e{-5} \\ - - - - - & - - - - - \\ 6.068e{-5} & -3.226e{-5} \\ -3.226e{-5} & 2.264e{-5} \\ - - - - - & - - - - - \\ 4.322e{-4} & -1.135e{-4} \\ -1.135e{-4} & 1.462e{-4} \\ - - - - - & - - - - - \\ 3.127e{-5} & -3.657e{-5} \\ -3.657e{-5} & 8.499e{-5} \end{pmatrix}$

Table 3.A.4.1: Estimated parameters by Robust EM (Yang et al., 2012) and Modified Robust EM on week 13 of the COVID-19 dataset. These estimations were performed independently of previous time steps. Bolded numbers indicate overlapping clusters.

Chapter 4

Dynamic Expectation-Maximization Algorithms for Mixed-type Data

This chapter details the methodological developments of mixture models designed for mixed-type data. Component distributions of the continuous attributes can be either Gaussian, Student or Shifted Asymmetric Laplace. Categorical or discrete attributes, assumed independent conditionally on the class membership, can be distributed according to Bernoulli, Multinomial or Poisson distributions. The joint estimation of the number of classes and the parameters is carried out by EM-like algorithms that we have adapted to perform correctly. We show that our different dynamic algorithms allow us to reach the real number of classes and to correctly estimate the parameters of the discrete and continuous laws. We also highlight the benefits of introducing regularization to improve performance in situations where the sample size is insufficient for the complexity of the model. Our various models are then tested on real datasets from the literature, assessing that the objective of jointly estimating the number of components and the model parameters has been achieved.

Contents

1 Introduction

A mixed-type dataset is given by a collection of individual data containing both quantitative and qualitative variables. Mathematically, this corresponds to the representation of each subject by a combination of continuous and discrete variables. Mixed datasets are ubiquitous in many disciplines and, with the era of so-called 'big data', the availability of datasets composed of heterogeneous data sources and types will continue to increase. Statistical analysis of mixed-type data is therefore still a hot topic, whether for clustering, inference or dimension reduction.

1.1 Strategies for mixed-type data

Mixed-type data contain both continuous and categorical (nominal and/or ordinal) variables. As detailed in recent review papers (Foss et al., 2019; Ahmad and Khan, 2019), we can consider several types of strategies to model and cluster mixed-type data.

Firstly, there are methods designed for a single type of data, which require data transformation approaches, such as discretization of continuous variables in order to use categorical data methods (Goodman, 1974; Huang, 1997b), or numerical coding of discrete variables (McCane and Albert, 2008) to be suitable for continuous methods. In this case, the possibilities go from replacing a level by a median to dummy coding, and to more complex methods like copulas.

Copulas Copulas are multivariate cumulative functions with uniform marginals. They imply several restrictive properties but have an interesting flexibility, because we can use a wide variety of copulas to model a set of variables and the copula captures the dependence among the variables, continuous and discrete, independently of their one-dimensional margins. Continuous variables in $\mathbb{R}^d$ and discrete variables are mapped onto a one-dimensional latent space between zero and one by computation of their marginal cumulative distribution functions. Then, the cumulative distribution function of the copula is obtained by joint computation on inverse cumulative distribution functions of latent variables.

As we see, copulas can be used on mixed-type variables, under the guise of defining the dependency structure of the copula. Gaussian copulas, which provide one correlation coefficient per pair of variables, were first associated with mixed-type data by Smith and Khaled (2012) (and Murray et al. (2013)). Then appeared mixture models of copulas for mixed-type data (Kosmidis and Karlis, 2016; Marbac et al., 2017; Sahin and Czado, 2022). The first work on these models (Kosmidis and Karlis, 2016) introduced mixture models of copulas, with separated copulas for continuous variables and discrete variables. With the work of Marbac et al. (2017) mixture models of Gaussian copulas are introduced and formalized for mixed-type data. Continuous variables follow Gaussian distributions, integer variables Poisson distributions and ordinal variables Multinomial distributions. Then the joint model is defined such as discrete variables are conditioned on the continuous ones. A source of difficulty with copulas lies in the lack of uniqueness and identifiability. In fact, if at least one margin is not continuous, the copula is not unique. Moreover, the overall joint multivariate distribution can be difficult to understand when marginal and copula are specified separately. In addition, some copulas have no simple closed-form expressions.

Hybrid distances On the other hand, there are methods involving hybrid distances that can take into account both continuous and categorical variables. The Kullback-Leibler divergence, well-known in statistical learning, is the base of several mixed-type data distances (Barhen and Daudin, 1995; De Leon and Carrière, 2005). A popular hybrid distance is Gower's distance, combining relative absolute difference for continuous variables and indicators for categorical variables, used for example in combination with the partitioning around medoids (PAM) method (Kaufman and Rousseeuw, 1990). Another clustering strategy using the hybrid distance technique is the k-prototypes algorithm (Huang, 1997a, 1998), relying on square Euclidean distance for the continuous variables, and a fixed weight for the whole categorical variable contribution. A frequent limitation of these hybrid distances is the need to use and properly choose weights dictating the relative contribution of each of the variables. Later on, Foss et al. (2016) proposed a method relaxing the parametric hypothesis of mixture models by combining them with k-means algorithms, named KAMILA (package by Foss and Markatou (2018)). Within each cluster, the density function is expressed as a function of a radial kernel density estimate computed on Euclidean distances of continuous values to the nearest center. The categorical variables are assumed independent within a subpopulation and follow multinomial distributions for which the parameters are also estimated. As with other models implying multinomial laws, an important number of multinomial levels requires a commensurate sample size. In addition, they consider a univariate density kernel and rely on hard assignment through partition steps as in the k-means method.

In the same scope, spectral clustering, which uses graph theory to propose a clustering on a graph of similarities between individuals, is adapted for mixed-type data (Mbuga and Tortora, 2021). The traditional Euclidean similarity distance is replaced by a global dissimilarity weighting both continuous dissimilarity and nominal dissimilarity before computing kernel transformation and eigenvalue decomposition for a k-means execution. Spectral clustering for mixed data requires tuning continuous/nominal weight and kernel parameters, but the main limitation is the decomposition of sample size matrices.

Mixture models Among the direct approaches, mixture models are efficient in this mixed-type data context because they can produce generative models, take into account many types of data, manage dependencies between and within variables, and capture a wide range of scenarios. The mixture models allow for obtaining latent classes, which can be used for a clustering goal, but also for some perspectives like dimension reduction, exploration or interpretability of the estimated distributions.

The use of mixture models in this context raises the question of how to jointly model these different types of data, which must all appear in the mixture model. A first approach is to consider that continuous variables are dependent on discrete variables. This leads to considering that we evaluate the continuous variables for each possible realization of all discrete variables. A limitation is the number of combinations increasing exponentially with the number of levels and variables. This may lead to small sample sizes within each categorical class. Moreover, these models can lack identifiability, as proved for the mixture of location models (Willse and Boik, 1999).

A common mixture model for mixed-type data is the normal-multinomial mixture model (Hunt and Jorgensen, 1996; Fraley and Raftery, 2002), where given cluster membership, the data follow a joint distribution with a normal distribution for the continuous variables, and a multinomial distribution per categorical variable, assuming conditional independence between the continuous and categorical variables but also within the categorical variables. This is called local independence and is an important property often required for identifiability of models. A way to characterize dependence within categorical variables is to define new levels corresponding to a combination of the original categorical variables. For dependence between continuous and categorical variables, we can rely as above on numerical coding of categorical variables with a flexible covariance structure. This includes mixtures of factor analyzers (Ghahramani and Hinton, 1996; McLachlan and Peel, 2000; McLachlan et al., 2003), originally combining clustering and dimensional reduction, which were also adapted for mixed-type data through a combination of item response theory models and factor models but also expensive computation (McParland et al., 2014, 2017). In the same decade, Browne and McNicholas (2012) and McParland and Gormley (2016) proposed latent variable mixture models where latent variables follow Gaussian distributions. In the work of Browne and McNicholas (2012), categorical (multiple levels) variables are related to these latent variables by a latent trait model and continuous variables by a factor model. But numerical estimations are complicated by heavy computation of approximations of intractable integrals. With their model named clustMD, McParland and Gormley (2016) developed a mixture of latent Gaussian distributions, with parsimonious covariance structures for the latent variables. Each variable, continuous, ordinal or nominal is, independently of other variables, a manifestation of an underlying continuous latent variable. With the local independence assumption, they do not model any dependencies between mixed-type variables. In the presence of categorical variables, they resort to computationally heavy Monte-Carlo simulations.

Linear mixed models, considered for mixed-type data with longitudinal data and fixed effects, can also be integrated within mixture models when homogeneous regression relationship across subjects is violated (Celeux et al., 2005; Proust-Lima and Jacqmin-Gadda, 2005; Bai et al., 2016; Lee and Chen, 2019).

1.2 Mixture models for continuous distributions

In Chapter 2 we recalled the general expression of mixture models. When the data are multivariate real-valued observations, the usual probability density for each group is the multivariate Gaussian distribution. But in the presence of extreme, scattered, heavy-tailed data, the assumption of normality of the data may no longer be relevant, and we refer here to some works which have proposed mixtures of laws other than Gaussian and which will subsequently be used in our models.

The Student distribution Student's distributions provide a longer-tailed alternative to the normal distribution, and are more robust to atypical observations. First proposition of a mixture of t-distributions was made in the paper of McLachlan and Peel (1998), and Peel

and Mclachlan (2000) gave an algorithm description to estimate the parameters of these models. A t-distributions mixture can be seen as a normal distribution that is embedded in a wider class of elliptically symmetric distributions, with an additional parameter called the degrees of freedom ν. Literature on t-distributions and mixtures of t-distributions especially concentrate on the problem of noisy data, by considering for example a "ghost" class which should capture the outlier points in a clustering context (Lange et al., 1989; McLachlan and Peel, 1998).

Skew distributions Franczak et al. (2014) introduced another non-Gaussian mixture approach that allows for skewness, based on Asymmetric Laplace (AL) distribution (Kotz et al., 2001). The Shifted (or not) Asymmetric Laplace distribution is part of the family of generalized hyperbolic distributions, such as the normal inverse Gaussian distribution. In their work, Franczak et al. (2014) propose a Shifted version of AL distribution, allowing for flexibility in the position of the clusters as in Normal or Student distribution. Moreover, they describe the estimation of their model, based on the Expectation-Maximization (EM) algorithm. In another paper, Franczak et al. (2015) proposed the Multiple Scaled Shifted Asymmetric Laplace distribution, which guarantees convex level sets, similar to elliptical distributions such as the Gaussian.

Besides the Shifted Asymmetric Laplace distribution, there exist skewed versions of the Normal (Azzalini, 1985; Azzalini and Valle, 1996; Arellano-Valle and Azzalini, 2006) and Student (Jones and Faddy, 2003; Azzalini and Capitanio, 2003) distributions to deal with asymmetric behaviors. Later works introduced mixture with these skew Normal (Lin et al., 2007; Lin, 2009) or Student (Lin, 2010; Lee and McLachlan, 2012) distributions.

1.3 Estimation of mixture models and model selection

The classical algorithm to estimate mixture models is the Expectation-Maximization (EM) algorithm introduced by Dempster et al. (1977). It was applied to Gaussian mixture models, and later on, was extended to other continuous distributions (Peel and Mclachlan, 2000; Franczak et al., 2014; Lin et al., 2007; Lin, 2010; Vrbik and McNicholas, 2012; Lee and McLachlan, 2012), still coming with the well-known limits as described in the Chapter 2, which are sensitivity to the initialization, convergence towards the space boundaries, and selection of the number of components.

To solve the sensitivity drawback, several strategies rely on repetitions of a random initialization step or initialization with K-means algorithm (Baudry and Celeux, 2015). Recently, Lartigue et al. (2022) introduced an annealing E-step to better stride the support and become almost independent of the initialization in a Gaussian context. Franczak et al. (2014) also considered deterministic annealing in a SAL context, only during their initialization part.

Another challenge posed by EM-type algorithms is the selection of the number of classes, and intrinsically the choice of model. Beyond the classical model selection criteria (Akaike, 1973; Schwarz, 1978; Birgé and Massart, 2007), dynamic algorithms appeared in the last years, to simultaneously overcome the need for a collection of models, find the optimal number of classes, and avoid bad local maxima: from penalization of the objective function (Figueiredo and Jain, 2002; Law et al., 2004; Yang et al., 2012) to dynamic slope heuristic criterion (Birgé and Massart, 2007) inside EM algorithm (Derman and Pennec, 2017), and split-and-merge EM algorithm (Wang et al., 2004; Zhang et al., 2004). From this collection of methods, few reduce the estimation process from a collection of models to a single one. This is the case of the work of Yang et al. (2012), who associates a single-run EM-like algorithm with an estimation of the number of classes and robustness to initialization. Later on, we solved in Chapter 3 two weaknesses of their method, which are inadequate

early stopping of the algorithm, and lack of superimposed clusters detection, leading to surely incorrect local maxima.

In this present work we seek to deal with the same two problems of the EM algorithm: find the optimal number of classes and avoid sensitivity at Initialization. These objectives, combined with the estimation of mixture models on mixed-type data lead us to consider dynamic estimation algorithms for various continuous laws as described above and discrete variables also as appearing in mixed-type data literature.

1.4 Contributions

In this chapter, we propose three main algorithms, to estimate respectively Gaussian, Student and Shifted Asymmetric Laplace (SAL) distributions, in association with any set of discrete variables simulated from the following distributions: Bernoulli, Multinomial or Poisson. Our general framework, from the EM-type, combines the estimation of parameters of the defined mixture models with the estimation of the number of classes in this mixture, which is an important objective in many statistical applications. For each continuous law considered, we have adapted the general framework, taking inspiration from various proposals in the literature, to correctly estimate the parameters of the mixtures using a dynamic algorithm. We perform simulations on synthetic data, as well as comparisons on model selection and parameter estimation, and show that our algorithms estimated on mixed data find the right number of classes and estimate the numerous model parameters correctly. We apply our methods to two datasets from the literature, each containing different types of variables: a Prostate Cancer dataset and an Australian Institute of Sport dataset. Section 2 contains definitions of considered continuous distributions and how they are estimated by existing EM-like algorithms. Section 3 introduces the general model for mixed-type data and its variants for each continuous distribution. Section 4 details the different dynamic algorithms and associated algorithmic assumptions. Section 5 presents numerical results on synthetic data and Section 6 shows results on the two real datasets. Section 7 provides concluding remarks and perspectives of this work.

2 Background on mixture models

In this section, we describe the formalization of mixture models for the continuous distributions under consideration, *i.e.* Gaussian, Student and SAL. We also present the EM equations and the different variants of this algorithm for estimating these numerous mixtures.

Definition of a mixture model Sample $\boldsymbol{x} \in \mathbb{R}^g$ is drawn by a mixture model with K components if its probability density function (pdf) is written as

$$p(\boldsymbol{x}; \Theta) = \sum_{k=1}^{K} \pi_k p_g(\boldsymbol{x}; \theta_k)\,, \tag{4.1}$$

where $\Theta = (\boldsymbol{\xi}, \boldsymbol{\pi})$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ vector of class proportions, which sum to one and $\boldsymbol{\xi}$ contains elements of θ_k (distribution parameters of each class k) for all $k \in \{1, \ldots, K\}$.

2.1 Gaussian Mixture Models

We now consider n independent draws of a mixture of Gaussian distributions, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathbb{R}^g$. Latent variables $(z_i)_{i=1,\ldots,n}$ are introduced, such that each z_i is following a categorical distribution of parameter $\boldsymbol{\pi}$. This information is then encoded

as a K-dimensional binary variable $\boldsymbol{z}_i$ where $z_i^k = 1$ corresponds to observation $\boldsymbol{x}_i$ belonging to cluster k. For the sake of brevity, we will write in the future for all models that z_i follows a categorical distribution of parameter $\boldsymbol{\pi}$ and is directly 1-of-K encoded.

Complete model The complete Gaussian mixture model writes

$$\begin{cases} z_i & \sim \text{Categorical}(\boldsymbol{\pi})\,, \\ \boldsymbol{x}_i | z_i^k = 1 & \sim \mathcal{N}_g(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_K)\,. \end{cases} \tag{4.2}$$

2.1.1 EM equations

The complete-data likelihood for a Gaussian mixture model (GMM) defined by Model (4.2) is

$$\mathcal{L}^c(\Theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \left[\pi_k \phi_g(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_i^k}\,, \tag{4.3}$$

with ϕ_g multivariate Normal distribution of dimensions g, and the set of parameters is here $\Theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The E-step Compute $Q(\Theta; \Theta^{(t-1)})$ the conditional expectation of the complete-data log-likelihood $\log \mathcal{L}^c(\Theta)$ given observed data and using current fit at iteration t. For a Gaussian mixture, the E-step corresponds to the computation of the current conditional expectation of z_i^k for each component k and individual i:

$$\begin{aligned} p_{\Theta^{(t)}}(z_i^k = 1 | \boldsymbol{x}_i) &= \frac{\pi_k^t \phi_g(\boldsymbol{x}_i | \theta_k^t)}{\sum_{j=1}^{K} \left(\pi_j^t \phi_g(\boldsymbol{x}_i | \theta_j^t)\right)} \\ &= \tau_{ik}^t\,. \end{aligned} \tag{4.4}$$

The M-step Maximize $Q(\Theta; \Theta^{(t-1)})$ to find $\Theta^{(t)}$, which gives the following update equations:

$$\hat{\pi}_k^t = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^t\,, \tag{4.5}$$

$$\hat{\boldsymbol{\mu}}_k^t = \frac{\sum_{i=1}^{n} \tau_{ik}^t \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{ik}^t}\,, \tag{4.6}$$

$$\hat{\boldsymbol{\Sigma}}_k^t = \frac{\sum_{i=1}^{n} \tau_{ik}^t (\boldsymbol{x}_i - \boldsymbol{\mu}_i)^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_i)}{\sum_{i=1}^{n} \tau_{ik}^t}\,. \tag{4.7}$$

Note Computation of $p_{\Theta^{(t)}}(z_i^k = 1 | \boldsymbol{x}_i)$ for other continuous mixture models will provide similar equations of τ_{ik}^t, with the pdf of a Gaussian $\mathcal{N}(\cdot)$ replaced by the pdf of the corresponding law.

2.2 Student Mixture Models

We recall in this subsection definitions and properties of a Student mixture model (McLachlan and Peel, 1998; Peel and Mclachlan, 2000).

Definition The family of t-distributions provides heavy-tailed alternatives to the normal family. A g-dimensional random variable $\mathbf{X}$ follows a multivariate t-distribution of pdf $p_g(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with center $\boldsymbol{\mu}$, positive definite inner product matrix Σ and degrees of freedom $\nu \in (0; \infty]$:

$$p_g(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+g}{2})|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{g/2}\Gamma(\frac{\nu}{2})\{1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma})/\nu\}^{\frac{(\nu+g)}{2}}}, \tag{4.8}$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance.

Another characterization Given a scaling variable $\mathrm{U} \in \mathbb{R}$, $\mathbf{X}$ has a multivariate normal distribution, and $\nu\mathbf{U}$ is $\mathcal{X}^2_\nu$ with

$$\mathbf{X}|\mathrm{U} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}/\mathrm{U}) \text{ and } \mathrm{U} \sim \mathcal{X}^2/\nu\,.$$

We can also write $\mathbf{U}$ as a Gamma distributed random variable: $\mathrm{U} \sim \Gamma(\frac{1}{2}\nu, \frac{1}{2}\nu)$, and this characterization of variable X corresponds to the normal scale model (McLachlan and Peel, 1998), with pdf

$$p(x|\theta) = \int \phi(x; \mu, \boldsymbol{\Sigma}/u) dH(u)\,,$$

where H is in our case the pdf of U defined above.

As $\nu \to \infty$, then $U \to 1$ with probability 1 and $\mathbf{X}$ becomes marginally $\mathcal{N}_g(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Mixture of multivariate t-distributions We consider n independent draws of a mixture of t-distributions, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Following the general definition of a mixture model in Chapter 2, and as for GMM, latent variables $(z_i)_{i=1,\ldots,n}$ are introduced, such that each z_i is following a categorical distribution of parameter $\boldsymbol{\pi}$. With this characterization, the observed data augmented by $(z_i)_{i=1,\ldots,n}$ are still incomplete, requiring to introduce additional missing values $u_1, \ldots, u_n$, which are defined so that for each observation i, given $z_i^k = 1$,

$$\mathbf{X}_i|u_i, z_i^k = 1 \sim \mathcal{N}_g(\mu_k, \Sigma_k/u_i) \tag{4.9}$$

and

$$\mathrm{U}_i|z_i^k = 1 \sim \Gamma(\frac{1}{2}\nu_k, \frac{1}{2}\nu_k)\,. \tag{4.10}$$

Complete model Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ the missing variables $U_1, \ldots, U_n$ are independently distributed according to Eq.(4.10). The complete-data vector for a mixture model of Student distributions is then given by $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, u_1, \ldots, u_n)$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are observed continuous data, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ component-label vectors and $u_1, \ldots, u_n$ latent variables. The complete model is then

$$\begin{cases} z_i & \sim \text{Categorical}(\boldsymbol{\pi})\,, \\ u_i|z_i^k = 1 & \sim \Gamma(\frac{1}{2}\nu_k, \frac{1}{2}\nu_k)\,, \\ \boldsymbol{x}_i|u_i, z_i^k = 1 & \sim \mathcal{N}_g(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/u_i)\,. \end{cases} \tag{4.11}$$

2.2.1 EM equations

The complete-data likelihood for a Student mixture model defined by Model (4.11) is

$$\mathcal{L}^c(\Theta) = \prod_{i=1}^{n}\sum_{k=1}^{K}\left[\pi_k \mathcal{N}_g(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/u_i) \times \Gamma(u_i, \frac{\nu_k}{2}, \frac{\nu_k}{2})\right]^{z_i^k}. \tag{4.12}$$

The E-step As defined in the complete model, we have here two sets of latent variables: $\boldsymbol{z}$ and $\boldsymbol{u}$. The current conditional expectation of latent variables $\mathbf{z}$ leads to the same τ_{ik} equation as in a Gaussian mixture model, with Student pdf in numerator and denominator of Eq.(4.4). Computation of current conditional expectation of U_i for each individual i leads to

$$\mathbb{E}_{\Theta^{(t)}}[U_i|\boldsymbol{x}_i, z_i^k = 1] = \frac{\nu_k^t + g}{\nu_k^t + \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k^t; \boldsymbol{\Sigma}_k^t)} = E_{u,ik}^t \tag{4.13}$$

with $\delta(x_i, \nu_k; \Sigma_k)$ the Mahalanobis distance.

The M-step The M-step maximizes $Q(\Theta; \Theta^{(t-1)})$ to find $\Theta^{(t)}$. In the Student case, the proportions $\boldsymbol{\pi}$ are estimated as in Gaussian Mixture Models, so by the Eq.(4.5), and the centers and covariances parameters by the following equations:

$$\hat{\boldsymbol{\mu}}_k^t = \frac{\sum_{i=1}^n \tau_{ik}^t E_{u,ik}^t \boldsymbol{x}_i}{\sum_{i=1}^n \tau_{ik}^t E_{u,ik}^t}\,, \tag{4.14}$$

$$\hat{\boldsymbol{\Sigma}}_k^t = \frac{\sum_{i=1}^n \tau_{ik}^t E_{u,ik}^t (\boldsymbol{x}_i - \boldsymbol{\mu}_i)^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_i)}{\sum_{i=1}^n \tau_{ik}^t}\,. \tag{4.15}$$

As described by Peel and Mclachlan (2000), the divisor $\sum_{i=1}^n \tau_{ik}^t$ in Eq.(4.15) can be replaced by $\sum_{i=1}^n \tau_{ik}^t E_{u,ik}^t$. They affirm that it converges faster than the conventional EM algorithm, following the proposition of Kent et al. (1994) for a single component t distribution. However, as we are already on a particular adaptation of the EM algorithm, and using multicycling as described in the next part, we do not consider this modification in covariances update. We also noticed that several implementations of Student mixture models did not include this change in covariances estimation.

One advantage of t-distribution is that the degree of robustness, controlled by ν, can be inferred from the data. As shown in the paper of Lange et al. (1989), the degrees of freedom are solutions of fixed point equations. Each $\hat{\nu}_k^t$ is solution of the equation

$$\begin{aligned}\left\{-\psi\left(\frac{1}{2}\nu\right) + \log\left(\frac{1}{2}\nu\right) + 1 + \frac{1}{n_k^t}\sum_{i=1}^n \tau_{ik}^t \left(\log E_{u,ik}^t - E_{u,ik}^t\right)\right.\\ \left.+\psi\left(\frac{\hat{\nu}_k^{t-1} + g}{2}\right) - \log\left(\frac{\hat{\nu}_k^{t-1} + g}{2}\right)\right\} = 0\,,\end{aligned} \tag{4.16}$$

where $n_k^t = \sum_{i=1}^n \tau_{ik}^t$.

The solutions of these equations can be found using a one-dimensional point search, such as Newton's method. This last method was used in several works on Student distributions and then mixtures (Lange et al., 1989; Peel and Mclachlan, 2000), but from the variations of the left-hand equation, it can be replaced as we will show in Subsubsection 4.2.2.

ECM/ECME The Expectation/Conditional-Maximization (ECM) algorithm was first mentioned by Meng and Rubin (1993). The ECM algorithm consists in replacing the original M-step with a sequence of conditional maximization steps (CM-steps). In the case of two CM-steps in a Student mixture model, they are defined as follows at iteration t:

- CM-Step 1. Fix $\hat{\nu}^{t-1}$ and calculate $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$ by maximizing $Q(\Theta; \Theta^{(t-1)})$
- CM-Step 2. Fix $\hat{\boldsymbol{\mu}}^t, \hat{\boldsymbol{\Sigma}}^t$ and calculate ν^t by maximizing $Q(\Theta; \tilde{\Theta}^{(t)})$

Their proposition was primarily to take advantage of possible simple CM-steps instead of a complicated M-step, for example in the presence of missing data, and Liu and Rubin (1995) noted that for unknown ν, the ECM algorithm speeds up compared to EM algorithm. It was also proposed to introduce an intermediate E-step between the CM-steps, and this becomes a multicycle ECM algorithm (Meng and Rubin, 1993; Liu and Rubin, 1995). Peel and Mclachlan (2000) consider the ECM algorithm in their paper, but as parameters are estimated independently, these two CM-steps are equivalent to the M-step in the EM algorithm. They ended up using the multicycle version. In the ECME proposed by Liu and Rubin (1994), the ECM's CM-steps are replaced by steps that maximize constrained actual log-likelihood function $\mathcal{L}(\Theta)$ instead of constrained expected complete-data log-likelihood. However, in many applications, the ECME algorithm relies on complex computations for the parameters equations, or a high-complexity optimization problem with maximization of a function involving K variables to obtain one degree of freedom, and is not applied here.

2.3 Shifted Asymmetric Laplace Mixture Models

We recall in this subsection definitions and properties of the Shifted Asymmetric Laplace (SAL) mixture model as proposed by Franczak et al. (2014).

Definition The probability density function of a random variable $\mathbf{X}$ distributed according to a SAL distribution is

$$\frac{2\exp\left\{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right\}}{(2\pi)^{g/2};\boldsymbol{\Sigma}|^{1/2}}\times\left(\frac{\delta(\boldsymbol{x},\boldsymbol{\mu}|\boldsymbol{\Sigma})}{2+\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\nu/2}K_\nu(u)\,, \tag{4.17}$$

with $\boldsymbol{\alpha}\in\mathbb{R}^g$ skewness parameter, $\boldsymbol{\mu}\in\mathbb{R}^g$ shift parameter, $\boldsymbol{\Sigma}\in\mathbb{R}^{g\times g}$ scale matrix, K_ν modified Bessel function of third kind, with index $\nu=\frac{2-g}{2}$, δ is the Mahalanobis distance, and $u=\sqrt{(2+\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma})}$.

As detailed by Kotz et al. (2001), an Asymmetric Laplace random variable $\mathbf{X}$ admits the representation

$$X=W\boldsymbol{\alpha}+\sqrt{W}\mathbf{Y},\ W\sim\mathcal{E}(1),\ \mathbf{Y}\sim\mathcal{N}_g(0,\boldsymbol{\Sigma})\,, \tag{4.18}$$

with W a random variable from an exponential distribution with rate 1, and $\mathbf{Y}\in\mathbb{R}^{g\times g}$ a random variable from a Normal distribution with mean $\mathbf{0}^g$ and covariance matrix $\boldsymbol{\Sigma}$. When including a shift parameter as described by Franczak et al. (2014), the random variable $\mathbf{X}$ can be generated through

$$X=\boldsymbol{\mu}+W\boldsymbol{\alpha}+\sqrt{W}Y,\ W\sim\mathcal{E}(1),\ Y\sim\mathcal{N}_g(0,\boldsymbol{\Sigma})\,. \tag{4.19}$$

And so, $X|W=w\sim\mathcal{N}_g(\boldsymbol{\mu}+w\boldsymbol{\alpha},w\boldsymbol{\Sigma})$.

This parametrization of Shifted Asymmetric Laplace random variables requires additional latent variables w in the mixture model, in order to obtain a complete model computable by EM algorithms. The distribution of W conditional on the data is obtained using Bayes' theorem:

$$f_W(w|x)=\frac{w^{\nu-1}}{2}\left(\frac{\delta(x,\boldsymbol{\mu};\boldsymbol{\Sigma})}{2+\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{-\nu/2}\times\frac{\exp\{-\frac{1}{2w}\delta(x,\boldsymbol{\mu};\boldsymbol{\Sigma})-\frac{w}{2}(2+\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\}}{K_\nu\left(\sqrt{(2+\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\delta(x,\boldsymbol{\mu};\boldsymbol{\Sigma})}\right)}\,, \tag{4.20}$$

with $\nu,\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}$ as defined for the SAL density in Eq.(4.17). It follows that $f_W(w|X=x)$ is a Generalized Inverse Gaussian density. As proposed by Franczak et al. (2015), multidimensional W latent variables could be considered, which yields convex level sets, useful in a classification perspective.

Complete model Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ the missing variables $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ are independently distributed according to an exponential distribution of rate 1. The complete-data vector is given by $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, w_1, \ldots, w_n)$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are observed continuous data, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are component-label vectors and $w_1, \ldots, w_n$ other latent variables needed for the representation. The complete SAL mixture model is

$$\begin{cases} z_i & \sim \text{Categorical}(\boldsymbol{\pi})\,, \\ w_i | z_i^k = 1 & \sim \mathcal{E}(1)\,, \\ \boldsymbol{x}_i | w_i, z_i^k = 1 & \sim \mathcal{N}_g(\boldsymbol{\mu}_k + w_i \boldsymbol{\alpha}_k, w_i \boldsymbol{\Sigma}_k)\,. \end{cases} \tag{4.21}$$

2.3.1 EM equations

The complete-data likelihood for a SAL mixture model defined by Model (4.21) is

$$\mathcal{L}^c(\Theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} [\pi_k \mathcal{N}_g(\boldsymbol{\mu}_k + w_i \boldsymbol{\alpha}_k, w_i \boldsymbol{\Sigma}_k) h(w_i)]^{z_i^k}\,. \tag{4.22}$$

The E-step In the presence of SAL distributed variables, the expectation step leads to the computation of τ_{ik} with Eq.(4.4), as in Gaussian and Student mixture models, but also of current conditional expectations of latent variables W given by $\mathbb{E}(W_i | \boldsymbol{x}_i, z_i^k = 1) = E_{1ik}$ and $\mathbb{E}(W_i^{-1} | \boldsymbol{x}_i, z_i^k = 1) = E_{2ik}$. Computations of these expectations lead to

$$\mathbb{E}(W_i | \boldsymbol{x}_i, z_i^k = 1) = \sqrt{\frac{b_{ik}}{a_k}} R_\nu(\sqrt{a_k b_{ik}})\,, \tag{4.23}$$

$$\mathbb{E}(W_i^{-1} | \boldsymbol{x}_i, z_i^k = 1) = \sqrt{\frac{a_k}{b_{ik}}} R_\nu(\sqrt{a_k b_{ik}}) - \frac{2\nu}{b_{ik}}\,. \tag{4.24}$$

With ν as defined in Eq.(4.17), $R_\nu(x) = \frac{K_{\nu+1}(x)}{K_\nu(x)}$, $a_k = 2 + \boldsymbol{\alpha}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\alpha}_k$ and $b_{ik} = \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)$.

The M-step The M-step maximizes $Q(\Theta; \Theta^{(t-1)})$ to find $\Theta^{(t)}$ to get the maximum likelihood estimates. The proportions are updated as in Gaussian and Student mixtures, thus by Eq.(4.5).

The skewness parameters are updated by :

$$\hat{\boldsymbol{\alpha}}_k^t = \frac{(\sum_{i=1}^n \tau_{ik}^t E_{2ik}^t)(\sum_{i=1}^n \tau_{ik}^t \boldsymbol{x}_i) - n_k^t (\sum_{i=1}^n \tau_{ik}^t E_{2ik}^t \boldsymbol{x}_i)}{(\sum_{i=1}^n \tau_{ik}^t E_{1ik}^t)(\sum_{i=1}^n \tau_{ik}^t E_{2ik}^t) - (n_k^t)^2}\,, \tag{4.25}$$

with $n_k^t = \sum_{i=1}^n \tau_{ik}^t$.

The shift parameters are updated by:

$$\hat{\boldsymbol{\mu}}_k^t = \frac{(\sum_{i=1}^n \tau_{ik}^t E_{1ik}^t)(\sum_{i=1}^n \tau_{ik}^t E_{2ik} \boldsymbol{x}_i) - n_k^t (\sum_{i=1}^n \tau_{ik}^t \boldsymbol{x}_i)}{(\sum_{i=1}^n \tau_{ik}^t E_{1ik}^t)(\sum_{i=1}^n \tau_{ik}^t E_{2ik}^t) - (n_k^t)^2}\,. \tag{4.26}$$

And the scale parameters are updated by:

$$\hat{\boldsymbol{\Sigma}}_k^t = S_k - \hat{\boldsymbol{\alpha}}_k^t r_k^\top - r_k (\hat{\boldsymbol{\alpha}}_k^t)^\top + \frac{1}{n_k^t} \hat{\boldsymbol{\alpha}}_k^t (\hat{\boldsymbol{\alpha}}_k^t)^\top \sum_{i=1}^n \tau_{ik}^t E_{1ik}^t\,, \tag{4.27}$$

with

$$S_k = \frac{1}{n_k}\sum_{i=1}^{n}\tau_{ik}^t E_{2ik}^t(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^t)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^t)^\top \text{ and } r_k = \frac{1}{n_k^t}\sum_{i=1}^{n}\tau_{ik}^t(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^t).$$

As we can see here, computations of the latent expectations and then estimation of parameters imply Mahalanobis distance between data points and centers. When there is even one center that is too close to a point, it can lead to the infinite likelihood problem, described by Franczak et al. (2014), and then skewness parameters $\boldsymbol{\alpha}$ are not computable. To overcome this problem, Franczak et al. (2014) proceed by taking the value of $\hat{\boldsymbol{\mu}}_k$ at the last iteration before it becomes too close to any data point. This estimated center, noted $\boldsymbol{\mu}_k^{t^\star}$, becomes the actual estimate of the center. Then the skewness parameter α_k is estimated by the following formula:

$$\hat{\boldsymbol{\alpha}}_k^t = \frac{\sum_i^n \tau_{ik}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{t^\star})^\top}{\sum_i^n \tau_{ik}E_{1ik}}. \tag{4.28}$$

This process is summarized in Algorithm 4.1.

Algorithm 4.1: Check superimposed centers and data points for $\boldsymbol{\alpha}$ estimation

```
for k = 1, ..., K do
    if μ_k^t = x_i then
        Find last iteration t* such as μ_k^t ≠ x_i, and assess μ_k^t = μ_k^{t*}
        Compute α̂_k^t with (4.28)
    end
    else
        Compute α̂_k^t with (4.25)
    end
end
```

Not satisfied of this solution for estimation of skewness parameters, Fang et al. (2023) proposed very recently to estimate SAL mixture models with a Bayesian parameter estimation scheme based on a Gibbs sampling framework instead of an EM algorithm, in particular to improve parameter recovery. As they themselves have pointed out, their main drawback is the runtime required for simulations.

In the next section, we will define our mixture models for mixed-type datasets, *i.e.* for data including variables following discrete probability distributions, and variables following one of the continuous distributions defined previously. We will use the set of parameter estimation equations defined above, which we summarize in Tables A.1 and A.2.

3 Mixture models for mixed-type datasets

3.1 Motivation and assumptions on the model

In this chapter, we seek to infer populations with mixed variables, enabling the integration of such models in many areas where the data are heterogeneous and require flexibility and interpretability. Using mixture models allows this flexibility, particularly in the choice of laws, parametrization and use of the models obtained. Drawing on a large body of literature, we have developed a mixture model which, although simple, allows rapid and

robust estimation of various combinations of laws. Rather than relying solely on Gaussian-multinomial combinations, we propose here the use of Student or SAL distributions in place of the Gaussian distribution and Bernoulli, Multinomial and Poisson distributions for discrete data. Our models and algorithms have been designed so that they can be integrated into a pipeline such as the one presented in Chapter 3.

Local independence assumption A major consideration in the specification of a multivariate mixture model is to define whether the variables are independent within a cluster, a property called local independence. This is simply expressed in the Gaussian case by the form of the covariance matrix. In the case of mixed-type data, the considered models are more complex. In all the models that we will define we make the assumption that continuous variables are independent of discrete ones knowing the latent membership variables. In addition, we also make the assumption that all discrete variables are independent knowing the latent variables z. These assumptions may fail to capture some dependencies patterns in the dataset. But they allow an accurate, quick and interpretable estimate of mixture models, as already shown in the literature on mixed data presented in 1.

3.2 Model description

Consider an observation i of mixed variables, given by $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \dots, \boldsymbol{x}_i^g, x_i^{g+1}, \dots, x_i^{g+D}) \in \mathbb{R}^g \times \mathcal{X}$. The g first variables are continuous variables defined on $\mathbb{R}^g$. The vector of these continuous variables is denoted $\boldsymbol{x}_i^c$. The vector of the D discrete (integer, nominal, binary, ...) is defined on $\mathcal{X}$ and denoted $\boldsymbol{x}_i^D$ with x^d being the dth discreetly distributed variable. If x^d is a nominal variable with m_d modalities, then it uses a numeric coding $\{1, \dots, m_d\}$.

The mixture model An observation i of g continuous variables and D categorical/discrete variables is supposed to be a realization of a random variable $\mathbf{X}_i$ distributed according to a mixture model of K classes, whose pdf is written as

$$p(\boldsymbol{x}_i; \Theta) = \sum_{k=1}^{K} \pi_k p_g(\boldsymbol{x}_i^c; \theta_k^c) \prod_{d=1}^{D} p_{X_i^d}(x_i^d; \theta_k^d)\,, \tag{4.29}$$

with $\Theta = (\boldsymbol{\pi}, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K)$ the whole parameters set, and $\boldsymbol{\xi}_k$ contains $(\theta_k^c, \theta_k^1, \dots, \theta_k^D)$, parameters of continuous and discrete distributions of each component $k \in \{1, \dots, K\}$. The vector $\boldsymbol{\pi}$ groups the proportions of all classes, with $\sum_{k=1}^{K} \pi_k = 1$. Each discrete component of observation $\boldsymbol{x}_i$ is accessed by its index $d = 1, \dots, D$. If there is no discrete component in the variables, this reduces to a continuous mixture model as presented in previous sections and chapters. Here continuous data coordinates are drawn from continuous distributions with probability density function (pdf) $p_g(\cdot; \theta^c)$ of dimension g and parameters θ^c. Each discrete variable $d \in D$ is drawn from a discrete distribution, for which the associate probability mass function (pmf) is given by $p_\cdot(\cdot; \theta^d)$ with θ^d the associated value or vector of parameters.

Thereafter, we consider that we have a set of mixed-type observations, of size n, given by $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$.

Complete model As described previously for continuous mixture models, latent variables z are used to complete the data, and at the same time ensure local independence. Let $(z_i)_{i=1,\dots,n}$, latent variables such that each z_i follows a categorical distribution of parameter $\boldsymbol{\pi}$. This information is then 1-of-K encoded under variable $\boldsymbol{z}_i$ with $z_i^k = 1$ if data $\boldsymbol{x}_i$ belongs to cluster k, 0 otherwise. Knowing $z_i^k = 1$, each discrete attribute d for individual i, given by x_i^d, follows a discrete law of parameter θ_k^d.

Our complete model is then described by

$$\begin{cases} z_i & \sim & \text{Categorical}(\boldsymbol{\pi})\,, \\ \boldsymbol{x}_i^c | z_i^k = 1 & \sim & F_g^c(\theta_k^c)\,, \\ x_i^d | z_i^k = 1 & \sim & F^d(\theta_k^d)\ \forall\, d = 1, \dots, D\,, \end{cases} \tag{4.30}$$

with F_g any continuous probability distribution of dimensions g with parameters θ_k^c, and F^d corresponds to the discrete probability distribution for attribute d, of parameter θ_k^d.

Application In this work, we consider three possible continuous distributions, described in Section 2: the Gaussian distribution, the Student distribution or the Shifted Asymmetric Laplace (SAL) distribution. As we saw earlier, Student and SAL distributions require additional latent variables, which can be added by combining Model (4.30) with any of the continuous Models (4.2), (4.11) or (4.21). For the numerical discrete or categorical variables, we consider three different distributions, starting with the Bernoulli distribution, which is very simple, but also the Multinomial distribution, a distant generalization of the Bernoulli distribution. The last distribution considered is the Poisson distribution. Bernoulli and Poisson distributions only require one numerical parameter, so that means K parameters to estimate each. A Multinomial distribution has M modalities, bounded by the following constraint: $\sum_{m=1}^{M} p_m^d = 1$, therefore there are $(M-1)K$ parameters to estimate. We will give the corresponding equations to estimate any parameter of one of these distributions in Subsection 4.3.

Graphical representation of our generic Model (4.30) is on Figure 4.1 with the numerous parameters and latent variables corresponding to the laws under consideration.

3.3 Identifiability

As mentioned previously, a key consideration in specifying a mixture model is local independence. Previous works on categorical and non-parametric distributions have proved its importance to reach identifiability (Allman et al., 2009). Our model assumes within-cluster dependence between continuous variables and conditional independence between continuous and nominal variables. Relaxing the local independence strategy could risk a failure of identifiability, as in location models (Willse and Boik, 1999).

The study of the identifiability of finite mixtures was initiated by Teicher (1963) and further developed by Yakowitz and Spragins (1968), in particular for the finite mixtures of multivariate normal distributions with variable mean vectors and covariance matrices. Identifiability of the finite mixtures of t-distributions with variable degrees of freedom was proven by Holzmann et al. (2006), and for generalized hyperbolic distributions by Browne and McNicholas (2015).

In the next section, we will detail our algorithmic considerations to estimate such models with different continuous distributions and any discrete distribution. Resulting algorithms are named Dynamic EM for Mixed-type Data, and allow estimating properly mixture models for mixed-type data.

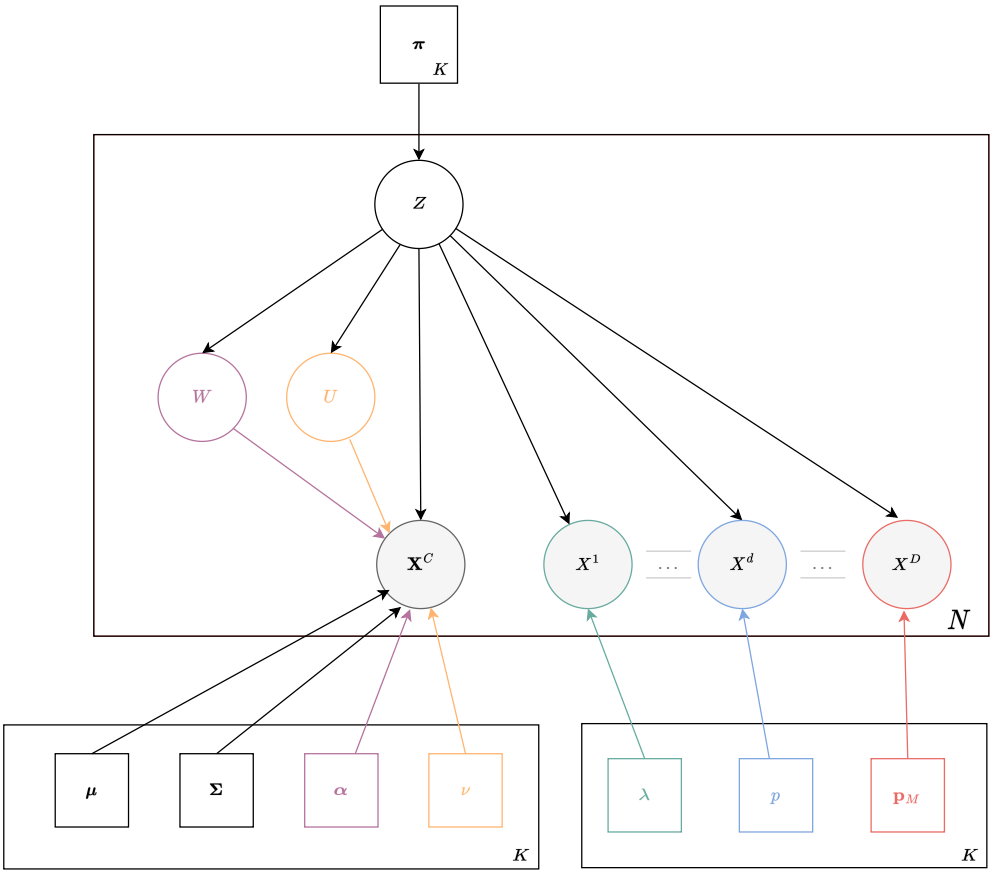


Figure 4.1: Graphical representation of our Model (4.30), with colors for additions/variants from Gaussian distribution. One color per continuous law for additional latent variables and parameters, and one color per type of discrete law: Student distribution (orange), SAL distribution (purple), Poisson (green), Bernoulli (blue) and Multinomial (red).

4 Dynamic EM for Mixed-type Data: Algorithms for mixed-type data

We now turn to the objective of estimating model parameters for a given mixed dataset $(x_i)_{i=1}^n$. First, we will describe our generic algorithm, which we name Dynamic EM for Mixed-type Data (DEM-MD), proposed to estimate parameters of mixture models for mixed-type data. Starting from an existing algorithm originally developed for Gaussian Mixture Models, we adapt and improve it. Then we will detail individual considerations relative to each continuous distribution, which lead to particular adaptations of DEM-MD to perform on all considered models. Finally, we will present the updating equations for discrete random variables to estimate during the M-step of DEM-MD.

4.1 A Dynamic EM algorithm for Mixed-type Data

We propose here a dynamic EM algorithm for the estimation of mixture models on mixed-type data, which implies categorical/ordinal/nominal variables. As a first step to constructing our algorithms, we propose an adaptation of the Modified REM (MREM), developed in Chapter 3, to estimate Model (4.30). With the Modified REM, the number of classes was dynamically selected along with the parameter estimation of a continuous Gaussian

mixture model. This algorithm was an amelioration of the original Robust EM work by Yang et al. (2012). We highlighted two weaknesses of the Robust EM (Yang et al., 2012): an inadequate early stopping of the algorithm, and the lack of superimposed clusters detection. Therefore, we solved these two problems by changing the original REM algorithm to become even more dynamic and detect the wrong maxima.

The Modified Robust EM (as the Robust EM) relies on the addition of an entropy term on the proportions in the objective function of the Expectation-Maximization algorithm and the construction of a weight enhancing the classes' competition. This additional penalization, combined with a competition weight and a pruning condition on the classes, makes it possible to reduce the number of classes when running the EM algorithm, initialized at $\hat{K} = n$.

4.1.1 A Generic Algorithm

Objective function in Dynamic EM for Mixed-type Data algorithm From the generic Model (4.30), the estimated function on complete-data is changed from Eq.(3.3) in the MREM algorithm to the following one, still including a penalization term on the proportions of the mixture:

$$\begin{aligned}\tilde{Q}(\Theta;\Theta^{(t)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} p_{\Theta^{(t)}}(z_i^k = 1|\boldsymbol{x}_i^c, x_i^D)\log\left[\pi_k p_g(\boldsymbol{x}_i^c;\theta_k^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_k^d)\right] \\ &+ \beta\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k\log\pi_k, \text{ with } \beta \geq 0\,.\end{aligned} \tag{4.31}$$

As we can see, there is a hyperparameter β, which comes as a penalty weight in Eq.(4.31). It helps to control the competition between clusters. Acting on the evolution of proportions with β enables one to check at each iteration that all the proportions of the components are above a given threshold, and therefore to delete components of proportion $\pi_k < \frac{1}{n}$.

E-step We compute the conditional expectation of the complete log-likelihood $\mathbb{E}_{p(z|x,\Theta^{(t)})}[\ell(\theta, \boldsymbol{y})]$ with $\boldsymbol{y}$ the complete data vector, including necessary latent variables for the considered continuous distribution. This results in conditional expectations for all the considered latent variables thanks to the exponential form of the complete likelihood.

Computation of conditional expectation of latent variables $\boldsymbol{z}$ leads to the following expression to update latent probabilities:

$$\begin{aligned}p_{\Theta^{(t)}}(z_i^k = 1|\boldsymbol{x}_i^c, x_i^D) &= \frac{\pi_k p_g(\boldsymbol{x}_i|\theta_k^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_k^d)}{\sum_{j=1}^{K}\pi_j p_g(\boldsymbol{x}_i|\theta_j^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_j^d)} \\ &= \tau_{ik}^t\,.\end{aligned} \tag{4.32}$$

For the Student and Shifted Asymmetric Laplace mixture models, additional latent variables are needed, as described in Models (4.11) and (4.21) respectively. Expectations of these latent variables are given in Subsubsection 2.2.1 for Student distribution and Subsubsection 2.3.1 for Shifted Asymmetric Laplace distribution. As discrete and continuous variables are independent knowing class memberships $\boldsymbol{z}$, the conditional expectation of these variables is not changed by the existence or not of discrete variables.

M-step With the objective function in Eq.(4.31) to maximize, the update equation of components proportions $\boldsymbol{\pi}$ inside DEM-MD algorithm is

$$\hat{\pi}_k^t = \hat{\pi}_{k,EM} + \beta \hat{\pi}_k^{t-1} \left(\ln \hat{\pi}_k^{t-1} - \sum_{s=1}^{K} \hat{\pi}_s^{t-1} \ln \hat{\pi}_s^{t-1} \right) , \tag{4.33}$$

with $\hat{\pi}_{k,EM}$ computed by Eq.(4.5), and $\hat{\pi}_k^{t-1}$ being the component weight estimate at previous iteration. The equations to estimate other continuous parameters remain unchanged whatever the continuous distribution.

The M-step is extended here with the estimation of parameters corresponding to discrete distributions of random variables $\mathbf{X}^d \; \forall \; d = 1, \ldots, D$. The parameters of each discrete variable are estimated after or before the continuous parameters, the order does not affect any of the estimated parameters. Corresponding equations for Bernoulli, Poisson or Multinomial laws will be given in Subsection 4.3.

Aitken's convergence Frequent stopping criteria in EM-like algorithms lean on absolute differences between centers at actual and previous iteration, or on the absolute differences of log-likelihoods, which correspond more to a "lack of progress" as said by Böhning et al. (1994) than to actual convergence. In our DEM-MD algorithm as presented above, this is meaningless to compare means of continuous distributions, as they may not be relative from one iteration to another. Moreover, in the case of mixed laws models now, the number of different parameters is increasing, which raises the question of the legitimacy of taking centers into account. In addition, as the number of components decreases during estimation, the objective function is no longer strictly increasing at each iteration.

Application of Aitken's acceleration to a sequence of log-likelihood first appeared in the work of Böhning et al. (1994), under the assumption that the sequence is linearly convergent to some value $l^\star$. Under this assumption, we can obtain the Aitken accelerated estimate of $l^\star$:

$$l_\infty^{t+1} = l^t + \frac{l^{t+1} - l^t}{1 - a^t} ,$$

where $a^t = \frac{l^{t+1} - l^t}{l^t - l^{t-1}}$. With this estimate Böhning et al. (1994) proposed the following stopping criterion for the EM algorithm at iteration $t+1$:

$$|l_\infty^{t+1} - l_\infty^t| < \varepsilon \, . \tag{4.34}$$

There exist other expressions of Aitken's acceleration in the work of Lindsay (1995) and then in the more recent work of McNicholas et al. (2010). In these two works, the stopping criterion is slightly changed, relying more on previous iterations or using only one asymptotic estimate of $l^\star$ instead of two. These different expressions give in general similar results. Therefore, we use the original criterion by Böhning et al. (1994).

Aitken's based criterion may lead to algorithms running longer than really necessary in cases where the log-likelihood is constantly increasing. Moreover, they require additional computations, which leads to their relatively infrequent use when they are more coherent.

As we argued previously the evolution of log-likelihood in a DEM-MD algorithm is complex, and this criterion is relevant to assess the stability of the convergence. However, the following assumptions are necessary to use Aitken's acceleration: a slow convergence rate of the objective function and linear convergence of the algorithm. These conditions to use Aitken's acceleration are easily validated by Expectation-Maximization algorithms (as detailed in (McLachlan and Krishnan, 2008, Chapter 3, Section 9, p.99)), but not by the dynamic versions which also estimate the number of components. These last methods have a conditional expected log-likelihood which is not constantly increasing.

We assume that, for a number of classes K which is constant, the objective function maximized by a dynamic EM algorithm is equivalent to that of an EM algorithm. In fact, as K is constant, the objective function is piecewise-increasing, joining the convergence theory of the EM algorithm.

Application in Dynamic EM algorithms With the idea spelled out above, if the number of classes is constant for at least four consecutive iterations (required to compute the Aitken's criterion), the Aitken's acceleration criterion can be computed and applied to assess the convergence of our Dynamic EM algorithm for Mixed-type Data.

Pseudocodes of DEM-MD and EM-MD algorithms In Algorithm 4.2 we present a generic version of DEM-MD algorithm, which will be adapted to the different combinations of continuous and discrete laws. We also include the EM-MD pseudocode within Algorithm 4.3, which correspond to a classical EM version with an *a priori* fixed number of classes K.

Algorithm 4.2: Pseudocode of "Generic" Dynamic EM for Mixed-type Data algorithm

Input : $\varepsilon > 0$, γ, dataset $\mathbf{X} = [\mathbf{X}^c, \mathbf{X}^D]$ with $\mathbf{X}_i \in \mathbb{R}^g \times \mathcal{X}$
Initialization : $K^0 \leftarrow n$, $\beta^0 \leftarrow 1$
$\pi_k^0 \leftarrow 1/n$, $\boldsymbol{\mu}^0 \leftarrow \mathbf{X}^c$
$\boldsymbol{\Sigma}_k^0 \leftarrow d^2_{k(\lceil\sqrt{K^0}\rceil)}\mathbf{I}_d$
Initialize other continuous parameters
Initialize $p_k^{d,0}$ $\forall\, d \in [\![1, \ldots, D]\!]$ with (4.36)
$t \leftarrow 1$
Compute τ_{ik}^t with (4.32)
1 **while** $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  M-Step
  Compute π_k^t with (4.33)
  Compute $\boldsymbol{\mu}_k^t$
  $\beta^t \leftarrow$ Algo. 4.4
  case *delete classes with* $\pi_k^t < 1/n$ **do**
    update class number K^t, adjust π_k^t and τ_{ik}^t
  otherwise do
    $K^t \leftarrow K^{t-1}$
  end
  Compute $\boldsymbol{\Sigma}_k^t$
  Compute other continuous parameters
  Compute discrete probabilities $p_k^{d,t}$ with (4.37) , (4.38) , (4.39)
  E-Step
  Compute τ_{ik}^{t+1} with (4.32)
  Compute other latent variables
  $t \leftarrow t + 1$
end

4.2 Adaptations in the DEM-MD algorithm for different continuous distributions

4.2.1 The Gaussian case

With a Gaussian assumption on the continuous variables, our Dynamic EM for Mixed-type Data algorithm is similar to the considered case of a Modified REM (MREM) as described in Chapter 3. The difference lies in the presence here of discrete variables in the model, and therefore the equations in DEM-MD to estimate these additional variables. From Algorithm 4.2, adaptation for estimation of models with Gaussian continuous variables and any discrete variable is given by Algorithm 4.5. Initialization and estimations of discrete parameters are given in Subsection 4.3.

4.2.2 The Student case

Initialization The Student distribution does not present particular constraints on the Initialization of its parameters. The only question is how to initialize the degrees of freedom ν. Other estimated parameters follow the existing rules. Moreover, expectations of latent variables $\boldsymbol{u}_i$ are computed as τ_{ik}^0 with initial parameters. Initialization of degrees of freedom does not have an impact on the next steps and we fix initial values to a unique constant, here $\nu_k = 10$ for each cluster k. There are methods in the literature that also initialize the degrees of freedom with constant values (Andrews et al., 2011; Lin, 2010).

Estimation of the degrees of freedom According to Eq.(4.16), the degrees of freedom are, at each iteration of an EM-like algorithm, the solution of a fixed point equation. This equation can be solved by a one-dimensional line search method. In some articles mentioned previously, the considered algorithm to solve this equation is Newton's (or Halley's) method, which requires first (and second) derivatives of the function whose zero we are finding. But as the considered function is monotonically decreasing for $x \geq 2$, it allows considering simpler algorithms such as Brent's method (Brent, 2013), which solves the fixed point equation on a bounded domain of x. Moreover, as we constrain the degrees of freedom to be greater or equal to 2, if the sign of $f(x_{min}) = f(2)$ is the same as the sign of $f(x_{max})$, then we fix $\nu^{new} = 2$ or $\nu^{new} = x_{max}$, depending on the sign of the function values. Previous works relied on this numerical method and even restricted the $\hat{\nu}$ estimates (Andrews et al., 2011).

Multicycle ECM Following the multicycle Expectation/Conditional-Maximization (ECM) proposition of Peel and Mclachlan (2000) on the estimation of mixtures of Student distributions, we inserted an intermediate E-step between a CM-step to estimate $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a CM-step to estimate ν. However, this multicycle version did not bring real improvement in the estimation or computation time of our Student DEM-MD on all our experiments, so we decided to drop it out.

Pseudocode From generic Algorithm 4.2, adaptation for estimation of models with Student continuous variables is given by Algorithm 4.6.

4.2.3 The Shifted Asymmetric Laplace case

Initialization As indicated in Subsubsection 2.3.1, estimation of the skewness parameters involves the Mahalanobis distance in the denominator of the equation. Initialization of the DEM-MD algorithm originally involves starting with each data point as its own cluster, so basically $\boldsymbol{\mu}^0 = \mathbf{X}^c$. This initialization, associated with the computation of skewness parameters can quickly lead to computation errors at the beginning of the algorithm, leading

to its early stop. A simple solution we consider is to add a very small noise to the initial centers $\boldsymbol{\mu}^0 = \mathbf{X}^c + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Multicycle ECM Following the ideas introduced for the Student mixture model with Expectation/Conditional-Maximization (ECM) algorithms, as explained in Subsubsection 2.2.1, we decide to introduce an intermediate E-step just before the estimation of scales parameters in the SAL DEM-MD. As the original Robust EM for Gaussian mixtures was already built on the dynamic changes of the number of components inside the algorithm, the estimated parameters may lose their "meaning" during the estimation process, such as the latent probabilities. This is even more true when the set of parameters is wide and complex, such as with SAL distributions. By adding an intermediate E-step, we recompute latent probabilities τ_{ik} and expectations of W after the estimation of the centers, proportions and skewness parameters, *i.e.*, before the estimation of the scale matrices. This intermediate E-step avoids estimation errors, particularly during scale computations which can lead to singular matrix problems.

Deterministic Annealing Deterministic annealing for EM algorithms (Ueda and Nakano, 1994, 1998) relies on the introduction of annealing (or also named temperature) into the membership probabilities derived in the E-step. According to this principle, which is inspired by the theory of thermodynamics and free energy, the corresponding annealed version of Eq.(4.32) is

$$\tau_{ik}^t = \frac{\left[\pi_k p_g(\boldsymbol{x}_i|\theta_k^c) \prod_{d=1}^D p_{X_i^d}(x_i^d; \theta_k^d)\right]^{1/T}}{\sum_{j=1}^K \left[\pi_j p_g(\boldsymbol{x}_i|\theta_j^c) \prod_{d=1}^D p_{X_i^d}(x_i^d; \theta_j^d)\right]^{1/T}} \,. \tag{4.35}$$

We consider here using deterministic annealing in SAL DEM-MD algorithms. As a matter of fact, SAL DEM-MD without annealing struggles to converge. The annealing usually helps the algorithm to explore the solution space, and in this case, it also helps to avoid estimation errors which lead to non-convergence of the algorithm. Our experiments were drastically different between without and with temperature, leading us to keep it.

Here we consider temperature schemes inspired by works of Allassonnière and Chevallier (2021) and Lartigue et al. (2022). The idea is to consider an oscillating tempered pattern with decreasing amplitude towards 1, leading to the classical expectation computation after a certain number of iterations. We define our sequence of temperatures implemented in SAL DEM-MD algorithm by

$$T_t = 1 + a * \frac{sin(t/b)}{t/b} \; \forall \, t \in \mathbb{N} \,.$$

Deterministic annealing also appears in Franczak et al. (2014) but only in the first initial steps of their algorithm, to select the correct initial values. This use also leads them to consider a fixed sequence of temperatures, determining the number of iterations in the Initialization part.

Estimation of scale matrices Although present in several implementations of the EM algorithm, and in the recent REM and MREM versions of Yang et al. (2012) and Chapter 3, small regularization of scale parameters is not present in SAL mixture of Franczak et al. (2014). As a result, we question the legitimacy of this regularization, originally designed to

avoid the calculation of singular matrices. Moreover, we have already proposed improvements for the convergence of the SAL DEM-MD algorithm and to avoid calculation errors in the previous paragraphs.

In the Modified Robust EM, and our generic DEM-MD, scale matrices $\mathbf{\Sigma}_k^t$ are by default computed as $\mathbf{\Sigma}_k^t = (1-\gamma)\mathbf{\Sigma}_k^{\mathrm{EM}} + \gamma\mathbf{P}$ for each k, with $\mathbf{\Sigma}_k^{\mathrm{EM}}$ computed according to the considered continuous distribution, and $\mathbf{P}$ a diagonal matrix containing very low coefficients. We led a comparison study on our DEM-MD for SAL continuous laws, with different γ (the regularization parameter in DEM-MD algorithms), of values $0.0, 10e{-}9, 10e{-}5$. We observed no significant difference in the convergence of simulations, the number of correct $\hat{K}$ and even the estimation errors of the different parameters. So we decided to fix $\gamma = 0.0$ as the default value, and therefore eliminate noise regulation, to avoid a future question of changing it according to an arbitrary criterion.

Pseudocode From Algorithm 4.2, adaptation for estimation of models with Shifted Asymmetric Laplace continuous variables is given by Algorithm 4.7.

4.3 Adaptations in the DEM-MD algorithm for discrete distributions

From our definition of Model (4.30), the estimated parameters of each discrete variable are independently computed at each M-step, and independently of the continuous parameters described previously.

4.3.1 Initialization

The DEM-MD algorithms are initialized by considering each data point as the center of its own cluster, so $K^0 = n$. Concerning the initialization of the parameters of discrete distributions in any DEM-MD, it is simply done by considering, for each discrete variable d, at the beginning of the algorithm, that

$$\hat{p}^{d,0} = \begin{cases} x^d \in \{0,1\}^n \text{ if } d \text{ is Bernoulli}\,, \\ x^d \in \mathbb{R}_+^{\star,n} \text{ if } d \text{ is Poisson}\,, \\ [\mathbb{1}_{x^d=1}^T, \ldots, \mathbb{1}_{x^d=M}^T] \; with \; \mathbb{1}_{x^d=m} \in \{0,1\}^n \text{ if } d \text{ is Multinomial}\,. \end{cases} \tag{4.36}$$

4.3.2 EM equations for discrete parameters

Bernoulli case The equation for updating the parameter of a Bernoulli distribution in the M-step of DEM-MD algorithm is given by

$$\hat{p}_k^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_i^d}{\sum_{i=1}^n \tau_{ik}^t}\,. \tag{4.37}$$

Multinomial case A variable following a Multinomial distribution generally arrives encoded as $\boldsymbol{x}^d \in \{1, \ldots, M\}^n$ for an attribute with M modalities. At the beginning of DEM-MD we transform it as a one-hot encoding matrix, so $\boldsymbol{x}^d \in \{0,1\}^{n \times M}$, and x_{im}^d correspond to sample i and column/modality m.

The equation for updating the parameter of a Multinomial distribution in the M-step of DEM-MD algorithm is given by

$$\hat{p}_{k,m}^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_{im}^d}{\sum_{i=1}^n \tau_{ik}^t}\,, \tag{4.38}$$

with m a modality of the Multinomial law with M modalities.

Poisson case The equation for updating the parameter λ of a Poisson distribution in M-step of DEM-MD algorithm is given by

$$\hat{p}_k^{d,t} = \hat{\lambda}_k^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_i^d}{\sum_{i=1}^n \tau_{ik}^t}\,. \tag{4.39}$$

Remark To avoid computation errors on the estimation of discrete distribution factors in the computation of posterior probabilities, when one discrete distribution parameter (for a given class k) is equal to one, its associated log-probabilities are not event computed as it should be zero.

Implementation details In this entire section, we hope to have covered all the algorithmic points. All the DEM-MD or EM-MD algorithms presented in this chapter are implemented in Python 3.9.

5 Experiments on simulated data

In this section we present various results from our Dynamic EM for Mixed-type Data algorithm for estimating many mixture models on mixed-type data, validating the different algorithms. We begin with the description of ten different settings, combining continuous distributions with different combinations of discrete distributions. For all of these given ten settings by continuous law, we simulate $S = 100$ datasets that will be used to analyze convergence, estimation of classes and parameters of both DEM-MD and EM-MD algorithms. Then, we look at the convergence success of DEM-MD algorithms and the average iteration number to reach convergence across all ten simulation studies. We continue this section with a comparison of the estimation of K by DEM-MD algorithms and model selection criteria with EM algorithms. Thereafter, for each simulation study, we compare the parameter estimates returned by DEM-MD algorithms and EM-MD algorithms to the true parameter values through the computation of relative errors. We also include comparisons with the literature on the estimation of Student or SAL mixture models, assessing that our algorithms estimate the different parameters just as well and that challenges remain on some parameters. We finally conclude this section with a study on the inclusion of covariance regularizations when the complexity of the model is significant.

5.1 Description of experiments

We consider several settings where we let vary the number of clusters K, the continuous dimensions g, the discrete dimensions D and the type of associated discrete variables.

For each configuration defined in Table 4.1, we simulated $S = 100$ datasets, with a fixed number of points $n = 600$ each. In practice, EM-MDs are initialized with a short k-means computation, and a fixed maximal number of iterations is considered. Gaussian and Student DEM-MD have $\epsilon = 10e{-}7$ and $\gamma = 10e{-}5$, and SAL DEM-MD has $\epsilon = 10e{-}5$, $\gamma = 0.0$, $a = 1$ and $b = 3$. Gaussian, Student and SAL EM-MD have $\epsilon = 10e{-}5$. In the next parts, we assess the performance of our various algorithms on the estimation of the number of components and the parameters.

5.2 Convergence of the DEM-MD

In Table 4.1, we see that on the majority of configurations, the Dynamic EM for Mixed-type Data algorithm converges for 100% of runs. Convergence is lower for SAL DEM-MD, especially on settings with a higher continuous dimensional space as C_{451}^M which has

Parameters				
K	g	D	Discrete Parameters	Setting abbreviation
2	2	0	None	C_{220}
2	2	1	Poisson	C^{P}_{221}
2	2	1	Bernoulli	C^{B}_{221}
2	2	1	Multinomial	C^{M}_{221}
4	5	1	Multinomial	C^{M}_{451}
2	2	3	Bernoulli Poisson Multinomial	C_{223}
2	2	3	Bernoulli Poisson Poisson	C^{P}_{223}
4	2	3	Bernoulli Poisson Poisson	C_{423}
5	2	4	Bernoulli Poisson Poisson Multinomial	C_{524}
3	4	3	Bernoulli Poisson Multinomial	C_{343}

Table 4.1: Description of simulated configurations.

a 18% convergence rate and C_{343} with 74%. Convergence in a DEM-MD is assessed by stabilization of $\hat{K}$ and by stopping the algorithm using the Aitken criterion. Non-convergent executions, therefore, correspond to executions where the number of clusters is reduced to one or generates calculation errors if the algorithm still reaches the space boundaries, which happens here for a high-complexity setting.

Convergence is usually not a difficulty for a not dynamic EM algorithm, with enough iterations, and therefore all EM-MD runs have 100% rates.

		C_{220}	C^{B}_{221}	C^{M}_{221}	C^{M}_{451}	C^{P}_{221}	C_{223}	C_{343}	C^{P}_{223}	C_{423}	C_{524}
EM-MD	Gaussian	100	100	100	100	100	100	100	100	100	100
	Student	100	100	100	100	100	100	100	100	100	100
	SAL	100	100	100	100	100	100	100	100	100	100
DEM-MD	Gaussian	100	100	100	100	100	100	100	100	100	100
	Student	100	100	100	100	100	100	100	100	100	100
	SAL	100	99	98	18	100	98	74	98	97	98

Table 4.1: Proportions of converged runs for each setting, each algorithm and each continuous distribution over $S = 100$ runs for each experiment.

The number of iterations Additionally, we show in Tables 4.2 and 4.3 the average (and standard deviation) number of iterations. In Chapter 3 we only compared our Modified Robust EM to the original Robust EM, showing that we were closed in terms of iterations and run-times, but no comparison was made with a classical EM algorithm. As expected, the number of iterations is greater for DEM-MD algorithms than for EM-MD algorithms. EM versions are run directly with the correct number of classes and initialized with a short k-means algorithm. It is worth noting that from low continuous and/or discrete dimensional configurations to larger ones, given a considered continuous distribution, the number of iterations stays in the same scale, the number of iterations varies, but not so drastically. Additionally, the number of iterations varies not only with the number of dimensions but also with the complexity of the model to be recovered, in terms of parameter values.

		C_{220}	C^B_{221}	C^M_{221}	C^M_{451}	C^P_{221}
EM-MD	Gaussian	5 (0.5)	5 (0.5)	4 (0.5)	4 (2.1)	4 (0.0)
	Student	24 (5.8)	24 (6.4)	24 (4.6)	27 (4.1)	25 (4.7)
	SAL	32 (6.5)	31 (7.0)	31 (6.9)	27 (8.3)	31 (6.0)
DEM-MD	Gaussian	91 (48.2)	85 (43.0)	77 (41.5)	125 (39.2)	46 (12.6)
	Student	157 (85.0)	142 (69.5)	146 (60.9)	219 (50.1)	242 (97.4)
	SAL	388 (79.6)	409 (138.3)	451 (77.5)	97 (171.1)	464 (41.3)

Table 4.2: Mean (std) number of iterations over $S = 100$ runs.

		C_{223}	C_{343}	C^P_{223}	C_{423}	C_{524}
EM-MD	Gaussian	4 (0.2)	4 (0.0)	4 (0.2)	5 (3.0)	7 (7.4)
	Student	26 (4.7)	22 (4.5)	26 (4.6)	26 (5.1)	24 (8.0)
	SAL	31 (6.8)	27 (8.3)	30 (7.0)	30 (10.3)	38 (15.9)
DEM-MD	Gaussian	49 (19.8)	36 (28.5)	53 (15.8)	113 (50.8)	153 (22.9)
	Student	234 (82.6)	181 (68.3)	231 (88.6)	311 (65.7)	250 (69.3)
	SAL	446 (90.5)	349 (227.6)	456 (60.5)	427 (126.2)	439 (100.8)

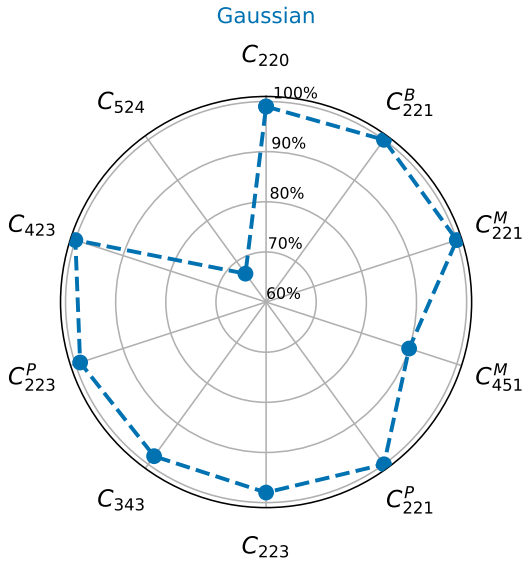
Table 4.3: Mean (std) number of iterations over $S = 100$ runs.

5.3 Estimation of the number of components

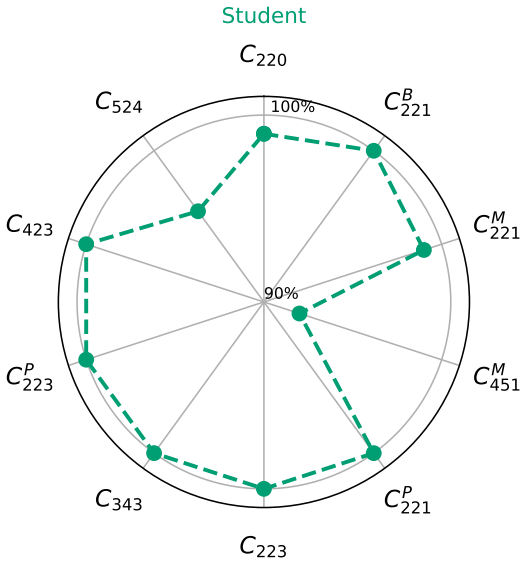
5.3.1 In mixed-type data context

We note C the number of correctly estimated number of components for a given configuration. This gives $C = \#\{\hat{K} = K^\star\}$ with $K^\star$ varying following the set of considered parameters. Figure 4.2 gives us the radar charts of C for all the configurations and the different continuous distributions on executions of DEM-MD. We see that we reach high percentages of C, between 90 and 100% for almost all configurations with Gaussian continuous laws (Fig. 4.2(a)). The only exception is the setting C_{524} with $D = 4$ discrete variables and $K = 5$ clusters, which obtains a C rate of 67%. On Student continuous configurations (Fig. 4.2(b)), we reach at least 92% of correct K for all configurations, with several ones at 100%. DEM-MD with SAL continuous distributions leads to lower performances, with several estimation percentages above 80%, but also obtains 67% for C_{343} and even a problematic 18% for C^M_{451}. These poor estimates are expected as complexity increases rapidly with the number of continuous dimensions. The number of samples is still $n = 600$ as for all considered settings, and it is not sufficient for a SAL DEM-MD to estimate correctly all the

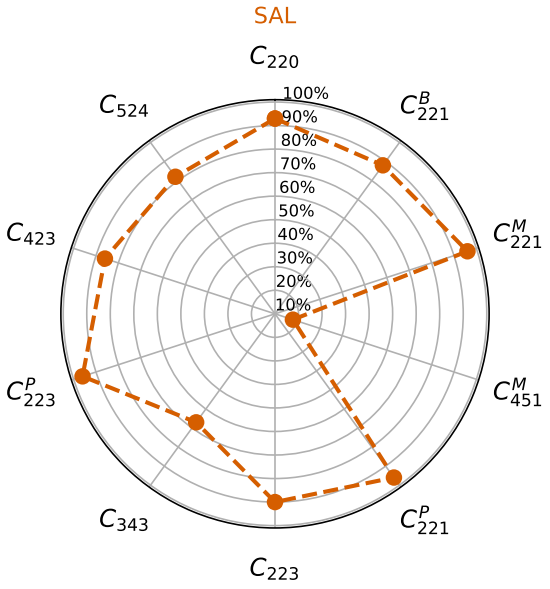
parameters. Table 4.1 gives also C values for all configurations, across the 100 simulated datasets.



(a) Radar chart of relative values $C = \#\{\hat{K} = K^\star\}$ for Gaussian DEM-MD.



(b) Radar chart of relative values $C = \#\{\hat{K} = K^\star\}$ for Student DEM-MD.



(c) Radar chart of relative values $C = \#\{\hat{K} = K^\star\}$ for SAL DEM-MD.

Figure 4.2: Radar charts of estimated number of classes $C = \#\{\hat{K} = K^\star\}$ by DEM-MD algorithm per continuous distribution. Each radar chart contains settings defined in Table 4.1.

Remark As noticed previously, the SAL DEM-MD, which has the highest continuous distribution complexity, cannot correctly estimate C_{451}^M with only $n = 600$ samples. We illustrate here that, as for a lot of models, a higher number of points is required to converge correctly and obtain correct performances on the estimation of $\hat{K}$. Figure 4.3 confirms

		C_{220}	C^B_{221}	C^M_{221}	C^M_{451}	C^P_{221}	C_{223}	C_{343}	C^P_{223}	C_{423}	C_{524}
EM-MD	Gaussian	100	100	100	100	100	100	100	100	100	100
	Student	100	100	100	100	100	100	100	100	100	100
	SAL	100	100	100	100	100	100	100	100	100	100
DEM-MD	Gaussian	99	100	100	90	100	98	98	99	100	67
	Student	99	100	99	92	100	100	100	100	100	96
	SAL	93	88	96	18	96	90	67	96	86	82

Table 4.1: Percentages of $C = \#\{\hat{K} = K^\star\}$ for each continuous distribution and each setting over $S = 100$ runs.

this phenomenon. We quickly see that increasing the dataset size allows us to improve the algorithm's performance.

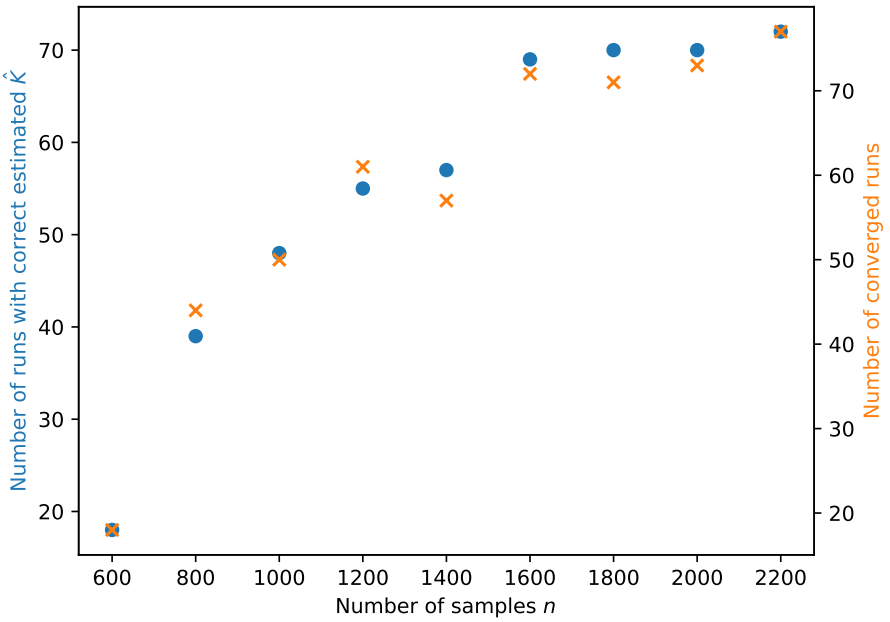


Figure 4.3: Performances of SAL DEM-MD on C^M_{451} according to the dataset size. Dots are numbers of correctly estimated $\hat{K}$ and crosses are numbers of converged runs, over 100 runs.

5.3.2 Comparison with model selection criteria

In this part, we compare the estimation of the number of components by our DEM-MD algorithms with model selection on mixture models, which requires choosing a criterion. We experiment, on all the settings, DEM-MD and EM-MD associated with one of these three criteria: BIC (Schwarz, 1978), ICL (Biernacki et al., 1998) or NEC (Celeux and Soromenho, 1996), detailed in Subsection 2.2. In recent works on mixtures of Shifted Asymmetric Laplace distributions (Franczak et al., 2014; Fang et al., 2023), BIC and/or ICL criteria were used for a model selection goal. While works on mixtures of Student (Peel and Mclachlan, 2000) or skew-student distributions (Lin et al., 2007; Lin, 2009) considered AIC and BIC criteria. For mixed-type methods such as KAMILA (Foss et al., 2016) or clustMD (McParland and Gormley, 2016), it is complicated by the absence of likelihood for the first one or calculation of intractable integrals for the second one. In KAMILA, model selection

was based on a prediction strength criterion, and they openly discussed an approximation of log-likelihood but not developed here. McParland and Gormley (2016) computed an approximation of the observed likelihood, leading to the possibility of using an approximated BIC criterion for model selection.

As for all the experiments, we simulated $S = 100$ datasets of each configuration, with $n = 600$ each time. On each dataset, we ran a DEM-MD and several EM-MDs, also from our codes, with a fixed K from a range of values. For each dataset (associated with a set of estimated models), an *a posteriori* selection is done by computing BIC, ICL and NEC on all EM-MD models, which have different K. For each one of these criteria, the best model is the one with the smaller value. Finally, over $S = 100$ runs we have the number of correctly estimated $\hat{K}$ for each method (Tables 4.2,4.3,4.4).

For Gaussian distributions, DEM-MD and EM-MD-BIC, EM-MD-ICL criteria give very good results. Whereas the NEC criterion is leading to very bad model selection for some configurations. It gives very extreme results, either it perfectly selects K, or it has 0 correct selection for C^M_{451}, C_{423} and C_{524}. Only once it gives an intermediate result, with $C = 38$ for C_{343}. DEM-MD has a 81% rate on C_{524}, which is the worst performance here, while the EM-MD-BIC and EM-MD-NEC criteria are hardly better.

Student DEM-MD and model selection have similar performances to the Gaussian case. The NEC criterion has high difficulties to select the correct model on the same settings as for Gaussian DEM-MD, with only a better score on C_{343} with a 74% rate. BIC and ICL criteria perform worse on C_{524}, as for the Gaussian distribution. But here the DEM-MD gives 100% of correct estimation for C_{524}.

When all model selection criteria for SAL distribution perform well on C^P_{221}, C_{223}, C^P_{223}, SAL DEM-MD also obtain correct rates, from 88% to 94%. In addition, SAL DEM-MD obtains a 76% rate on C_{343}, while the model selection criteria obtain very good results (above 90%). The same applies to C^M_{451} setting where SAL DEM-MD obtains 26% in contrary to BIC and ICL criteria which are 94%. As for Student and Gaussian distributions, the NEC criterion has difficulty for this setting as well as for C_{524} and C_{423}. Conversely, SAL DEM-MD has the best results on these two last settings. On the less complex settings, C_{220}, C^B_{221}, C^M_{221} and C^P_{221}, DEM-MD as well as EM-MD-ICL and EM-MD-NEC have very good performances, while the EM-MD-BIC obtains between 53% and 73% of correct $\hat{K}$.

NEC is a classification criterion, and as explained by Celeux and Soromenho (1996) themselves, the NEC criterion was designed to "choose the mixture model providing the greatest evidence for partitioning data". Difficulties emerge when clusters are not well separated, and this leads to difficulties in model selection as we can see in our different settings for the three continuous distributions.

Overall, DEM-MD algorithms are as correct as model selection criteria for each continuous law to find the true number of classes $K^\star$. The challenging configurations in high dimensions for our Dynamic EM for Mixed-type Data are also difficult for the NEC, but not for the BIC and ICL criteria. However, correct model selection does not guarantee good parameter estimations, and the model selection process is complicated if the optimal (unknown) K is not in the tested list. Moreover, difficulties for model selection criteria appear with an increasing number of clusters, as they require an arbitrary number of runs depending on the values of K tested. DEM-MD algorithms, meanwhile, save calculation time by only making a single run.

5.4 Performances on the estimation of parameters

The results presented in the next parts are relative errors calculated on the set of experiments where $\hat{K} = K^\star$. For each setting, the corresponding number of experiments with correct K is available in Table 4.1 above.

	DEM-MD	EM-MD-BIC	EM-MD-ICL	EM-MD-NEC
C_{220}	99	100	100	100
C^B_{221}	98	100	100	100
C^M_{221}	97	100	100	100
C^M_{451}	89	98	98	0
C^P_{221}	100	100	100	100
C_{223}	99	100	100	100
C_{343}	100	99	99	38
C^P_{223}	99	100	100	100
C_{423}	100	86	86	0
C_{524}	81	84	84	0

Table 4.2: Percentages of correctly estimated or selected K for configurations with Gaussian continuous distribution.

	DEM-MD	EM-MD-BIC	EM-MD-ICL	EM-MD-NEC
C_{220}	98	100	100	100
C^B_{221}	100	100	100	100
C^M_{221}	99	100	100	100
C^M_{451}	94	99	99	12
C^P_{221}	100	100	100	100
C_{223}	100	100	100	100
C_{343}	100	100	100	74
C^P_{223}	100	100	100	100
C_{423}	99	97	97	0
C_{524}	100	72	72	1

Table 4.3: Percentages of correctly estimated or selected K for configurations with Student continuous distribution.

5.4.1 Estimation of the continuous parameters

Figures 4.4, 4.5, 4.6, 4.7 and 4.8 give us estimated parameters for all our simulated configurations on experiments with correct $\hat{K}$. For each type of parameter, the associated figure shows boxplots of relative errors by DEM-MD and EM-MD algorithms.

At a glance, it appears that both schemes return good estimates of the true parameter values for centers, scales and proportions parameters.

For each continuous distribution we observe correct median of relatives errors on centers parameters (Fig. 4.4), but mean values sometimes explode, affected by extreme values (that are not plotted here), which seem to occur more with the EM-MD algorithm.

Errors on scale matrices were computed with the Frobenius norm, and as we can see relative errors may be higher than centers or proportions ones (Fig. 4.5). For Gaussian simulations DEM-MD and EM-MD errors are similar while, on Student simulations, DEM-MDs presents frequently lower median and mean errors compared to EM-MDs. SAL DEM-MD has more difficulties to estimate scale matrices, as median values are all higher than for Gaussian and Studnet DEM-MD, but median values for DEM and EM versions are similar for the majority of settings. Mean values can be really for DEM-MD, driven by extreme values (not represented here).

	DEM-MD	EM-MD-BIC	EM-MD-ICL	EM-MD-NEC
C_{220}	91	53	100	100
C^B_{221}	90	68	100	100
C^M_{221}	94	73	100	99
C^M_{451}	26	94	94	2
C^P_{221}	92	73	99	100
C_{223}	88	94	100	100
C_{343}	76	91	91	94
C^P_{223}	94	85	100	100
C_{423}	86	76	77	3
C_{524}	80	29	34	1

Table 4.4: Percentages of correctly estimated or selected K for configurations with SAL continuous distribution.

The median relative errors are also small for the proportions (Fig. 4.6), for all settings of the three continuous laws, for DEM-MD and EM-MD algorithms. On the other hand, they frequently explode on average. But since the errors are relative to the true value, which is always rather small for the proportions, this leads to large errors that do not necessarily reflect real extreme values in the estimates. DEM-MD presents high average errors for SAL distributions on seven settings, but for the other three, it is the EM-MD that is not any good. In the simplest settings, Gaussian and Student DEM-MD results present higher dispersion and medians than EM-MD results, while in the other settings, it is more variable, and overall the DEM-MD algorithm is doing just as well as the EM-MD algorithm.

Additionally, we can notice from all these results that the integration of discrete variables into our models does not affect the estimation of the continuous parameters by Dynamic EM for Mixed-type Data or EM-MD algorithms. Overall, DEM-MD estimates, as well as EM-MD the different continuous parameters, and both algorithms encounter the same difficulties on the degrees of freedom or skewness parameters as we will detail next.

Estimation of degrees of freedom From Figure 4.7 we observe that relative errors are pretty high for EM-MD and DEM-MD estimates. Estimation of the degrees of freedom stays a challenge with EM-like algorithms, despite the efforts and propositions of recent works on this problem. As mentioned in Subsubsection 4.2.2, we divided the M-step into two parts, and we inserted an intermediate E-step, which corresponds to the multicycle ECM (Peel and Mclachlan, 2000). But comparisons of algorithms with and without intermediate E-step did not lead to significative differences in parameter estimations on our ten configurations.

Comparing EM-MD and DEM-MD on the estimation of degrees of freedom, the errors for the DEM-MD version are more dispersed than the EM-MD ones. For settings C_{220}, C^B_{221}, C^M_{221}, C^P_{221}, C_{223} and C^P_{223}, DEM-MDs present lower median values and frequently lower mean values. For the other settings, the median values of DEM-MD are still close to the EM-MD ones even if the DEM-MD results are dispersed.

Although the errors obtained with DEM-MD and EM-MD algorithms seem pretty high, we will see in Subsubsection 5.4.3 that the obtained estimates with our algorithms are similar to those in the literature, which leads us to think this is due to the model itself. Although there may be a zero-search solution to the equation (4.16), each iteration of the algorithm can lead to a defined equation that reflects estimation errors on the latent variables. This can result in solutions to the equation that are far from the true values, or even outside the limits of the domain.

Estimation of skewness parameters From Figure 4.8, we observe the boxplots of relative errors on the estimation of the skewness parameter $\boldsymbol{\alpha}$ with SAL DEM-MD and SAL EM-MD algorithms. On a general level, the relative errors for the SAL DEM-MD are slightly more dispersed and the third quartile is higher. Mean values are also slightly higher but in the same scales. In addition, the SAL DEM-MD algorithm presents as good medians as EM-MD one. Globally, the DEM-MD algorithm presents as the same results as the EM-MD algorithm, which is encouraging, given the complexity of SAL mixture models and existing challenges estimating skewness parameters in particular.

We also observe that medians of relative errors can reach 10 at 20% of true $\boldsymbol{\alpha}$ values, in the presence or not of discrete variables, which are relative errors also obtained by similar implementations on SAL distributions, such as the MIXSAL R package (Franczak et al., 2018), implementation of the method of Franczak et al. (2014). Comparisons with results from this package can be found in Subsubsection 5.4.3.

5.4.2 Estimation of the parameters of discrete distributions

Now we look at the performances of DEM-MD and EM-MD algorithms on the estimation of the different parameters of discrete distributions. The results are obtained from the same experiments as above. The implemented discrete distributions are Bernoulli, Multinomial or Poisson, which correspond to binary, ordinal/nominal or integer variables. Several settings have only one discrete random variable, and the other ones are different combinations of three or four variables. As described in Model (4.30), discrete variables are independent conditional on class memberships. Figures 4.9, 4.10 and 4.11 show the relative errors for each setting and each discrete distribution parameter over the executions with a correct $\hat{K}$.

In Figure 4.9 we can observe relative errors of discrete distribution parameters for configurations containing only one discrete variable. We see in each column relative errors of respectively Gaussian, Student and SAL DEM-MD on (vertically in this order) C^B_{221}, C^M_{221}, C^M_{451} and C^P_{221}. Firstly, relative errors for each discrete distribution parameter are in the same intervals for both Gaussian, Student and SAL models. In addition, for each setting and model, EM-MD and DEM-MD give similar results for their averages, medians and whiskers. Comparing models on C^M_{221} and C^M_{451} for each continuous distribution shows that relative errors of the multinomial parameter vector are higher in C^M_{451} configuration, which has the highest complexity. As we saw earlier, this has led to more difficult computations, particularly for the SAL DEM-MD.

On configuration C_{524}, we observe explosion in the average of several parameters for EM-MD simulations (Fig. 4.10(a)), and even overall the average values are rather high, whereas the DEM-MD reveals estimates that are rather stable and less dispersed, unlike the EM-MD algorithm. Medians and means for DEM-MD results are generally low, as are several medians and means for EM-MD results.

We observe similar trends on C_{423} setting for both Gaussian, Student and SAL distributions (Fig. 4.10(b)), especially on the two Poisson parameters. Errors on the Bernoulli parameter are relatively high for each class and each algorithm, but this is partially due to the same relative effect as on proportion parameters explained before.

We observe on the setting C^P_{223} that the errors are similar for DEM-MD and EM-MD results (Fig. 4.10(c)). As for C_{423} setting, medians of Poisson errors are low, around 0.75% and 0.5% for the first Poisson parameter and around 2% and 1% median for the second Poisson parameter. The Bernoulli parameter errors are more dispersed, and medians are around $6-7\%$ for the first component and 4% for the second component.

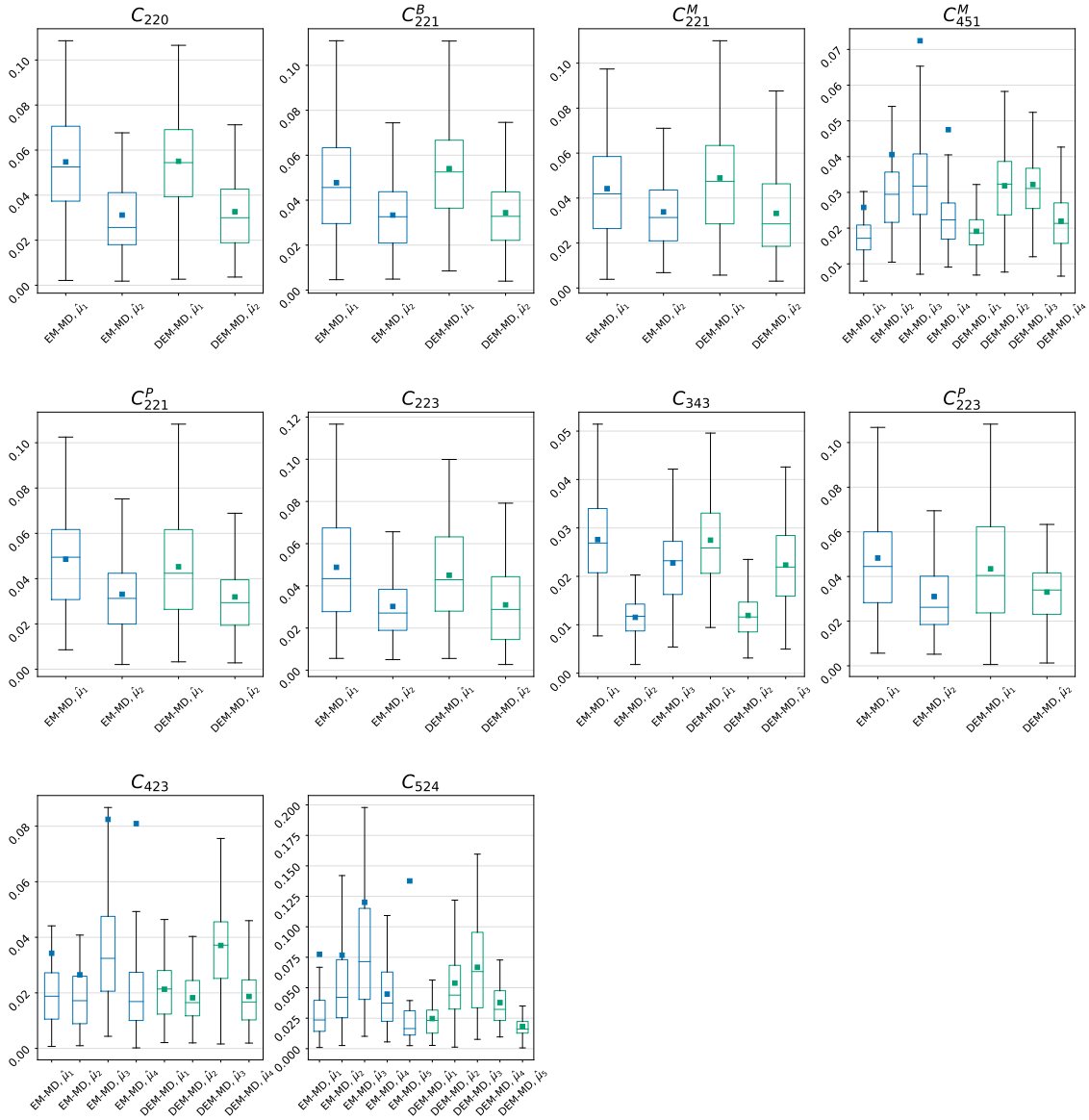
For configurations C_{223} and C_{343}, which differ by the number of classes and of continuous dimensions, the trends are as above (Fig. 4.11). Errors are low for Poisson parameters on both DEM-MD and EM-MD. Relative errors on multinomial parameters are lower for C_{223}

setting, maybe due to a lower overall model complexity, in terms of parameters to estimate and class distinctions. Bernoulli relative errors are here very different depending on the class, for both settings. Again, as for multinomial parameters, errors are higher for C_{343} setting. Generally speaking, the results are correct and the DEM-MD algorithm performs well, compared with an EM-MD that estimates the parameters with the right number of classes from the start.

5.4.3 Comparisons to other methods on continuous data

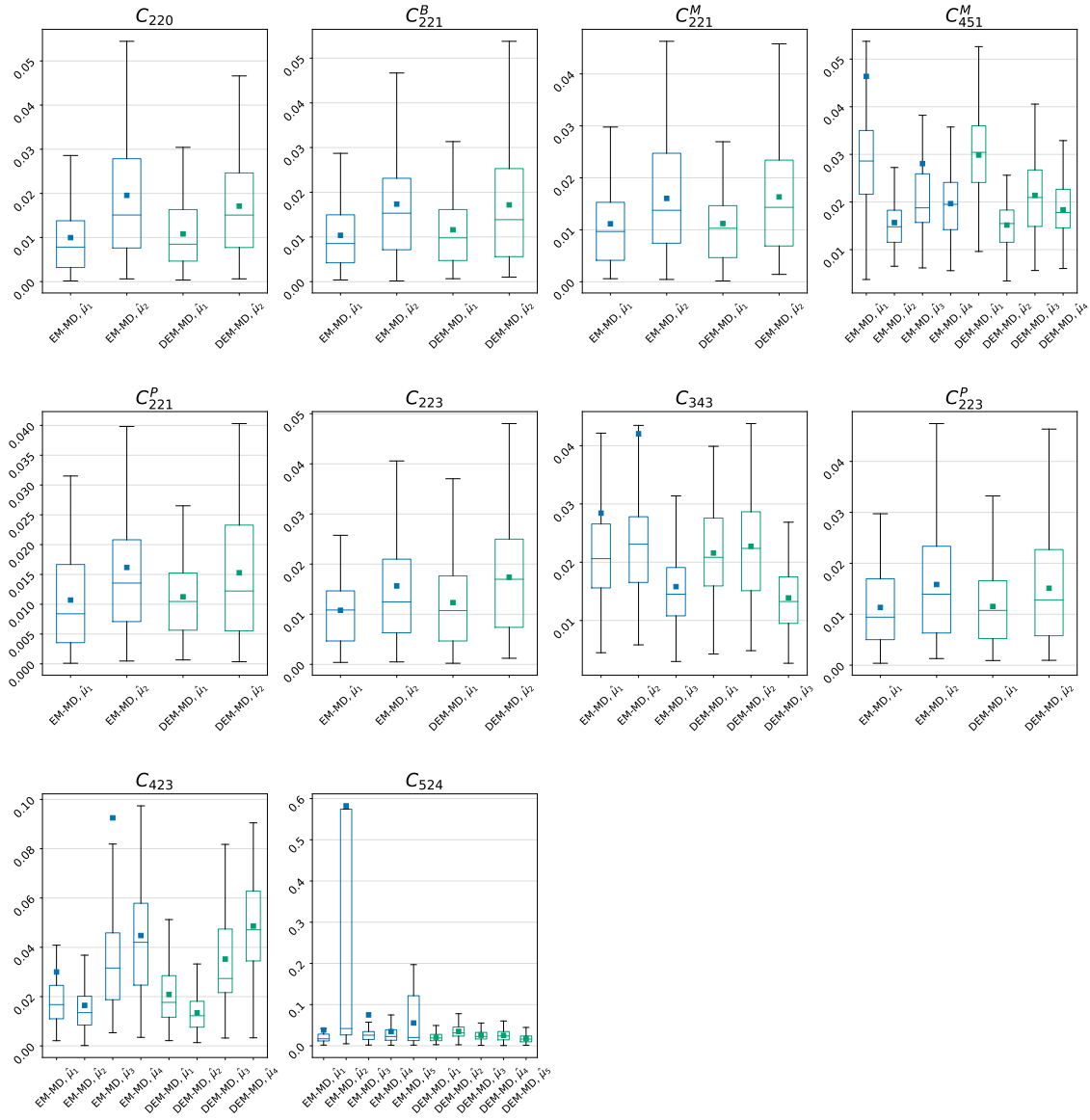
The trillium dataset In the very recent paper of Fang et al. (2023), the authors compare themselves with an EM method of Franczak et al. (2014) to estimate SAL mixture models. They test in particular both methods on a simulated setting named *the inversed trillium*, composed of three SAL clusters, which we reproduce here to compare our estimated parameters with theirs. We run our SAL DEM-MD on 100 simulated datasets of *the inverse trillium* setting. As we also estimate K in our method, we have $C = 92$ over the 100 runs, and all the runs converged. From a model selection perspective, in their article, Fang et al. (2023) had between 96 and 100% correct estimates of the number of clusters.

Table 4.1 contains the true parameter, as well as the average estimates, with standard deviation, returned by DEM-MD algorithm, and the ones obtained by Fang et al. (2023). We have directly reported the estimated parameters given in their article (see Fang et al., 2023, Table 3). We recall that estimated parameters for DEM-MD are over 92 runs with correct $\hat{K}$, while the MSAL-EM and MSAL-Bayes results are over 100 runs. These results show that our algorithm retrieves correctly the different parameters, as well as the other methods. The averages with DEM-MD are similar to the other ones, sometimes closer to true parameters, sometimes farther but never drastically. However, the standard deviations are in the majority not lower than the MSALD-Bayes ones from Fang et al. (2023), but equivalent to the MSALD-EM ones.



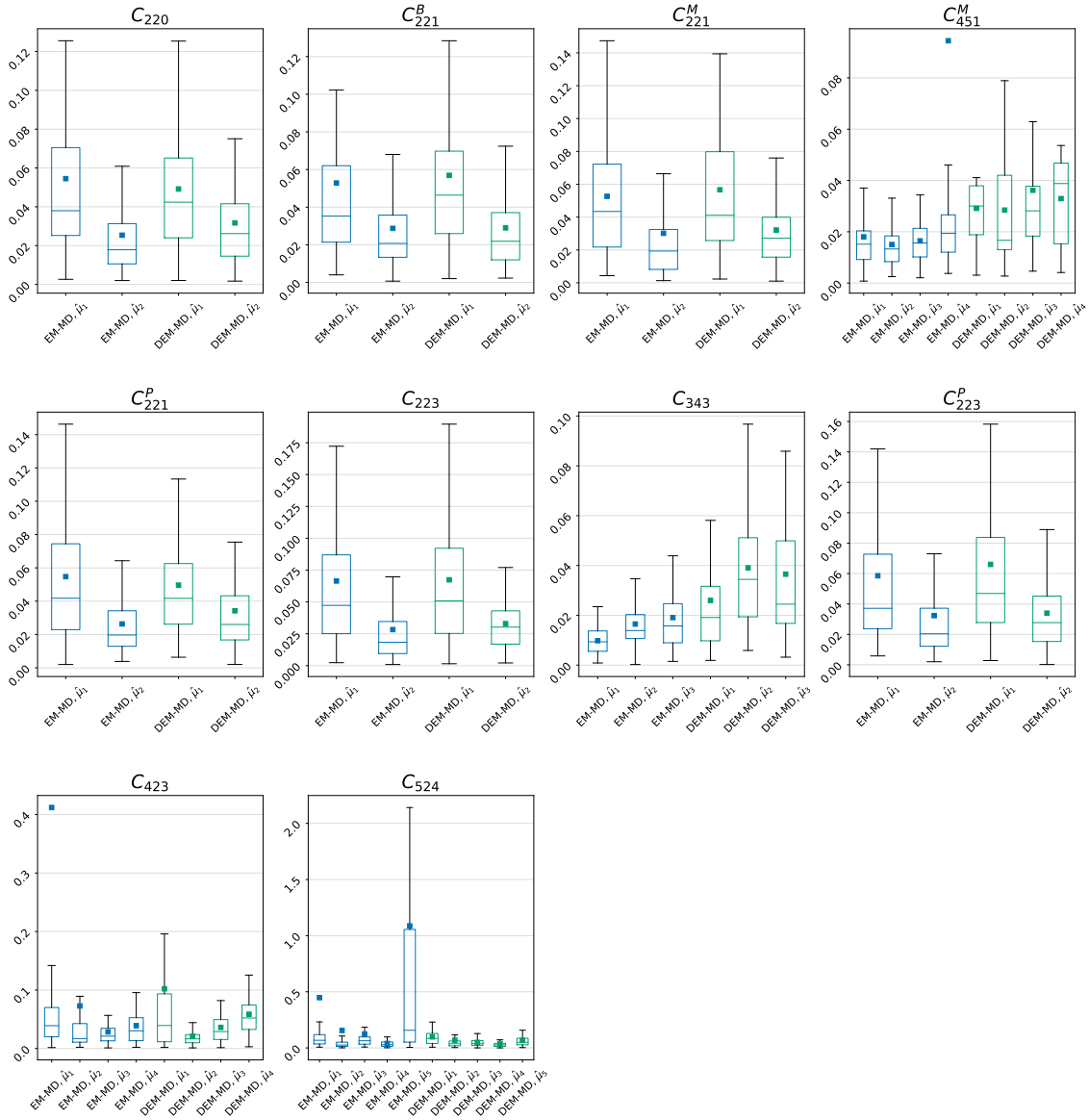
(a) Gaussian settings

Figure 4.4: Boxplots of centers relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



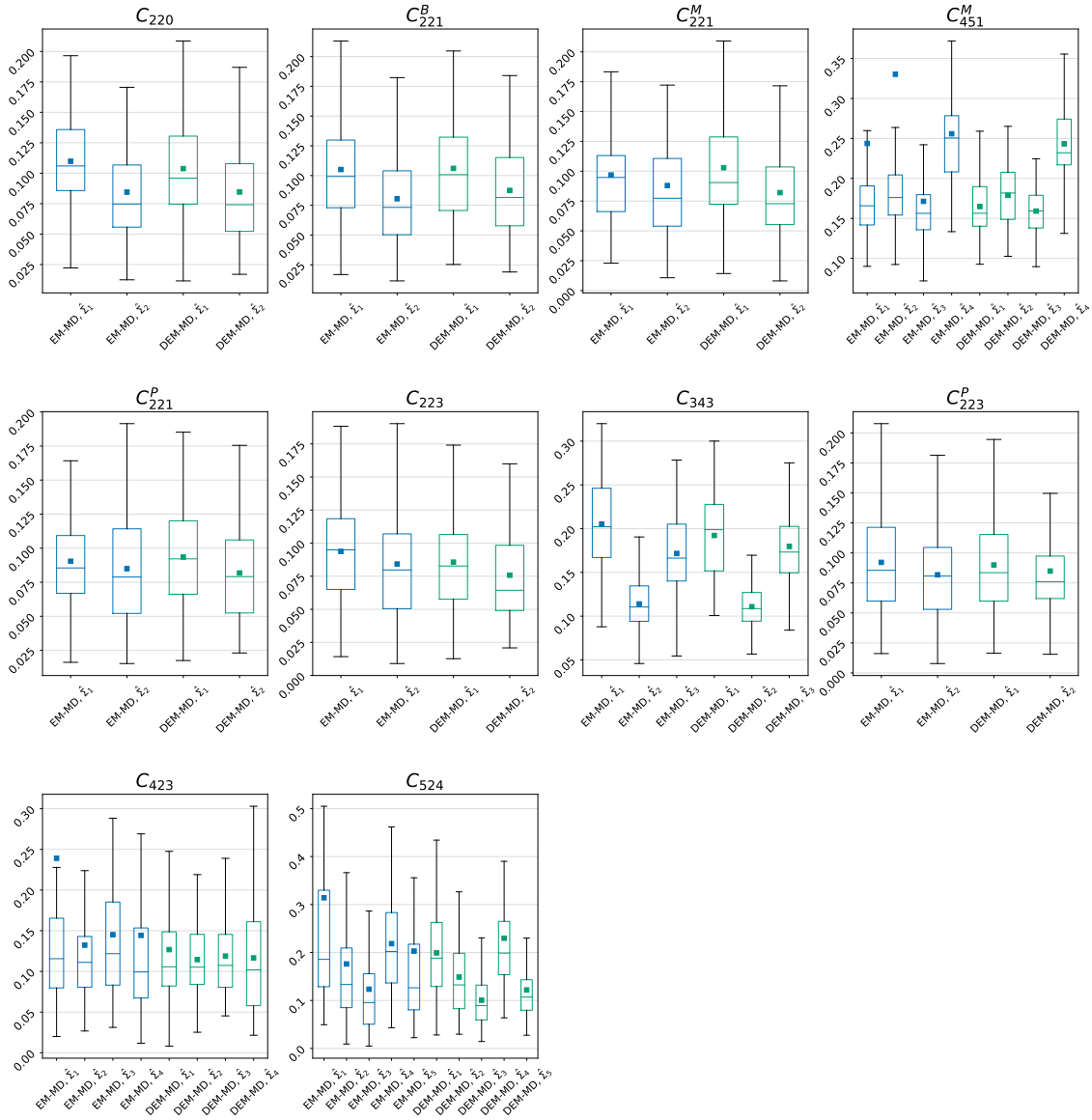
(b) Student settings

Figure 4.4: Boxplots of centers relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



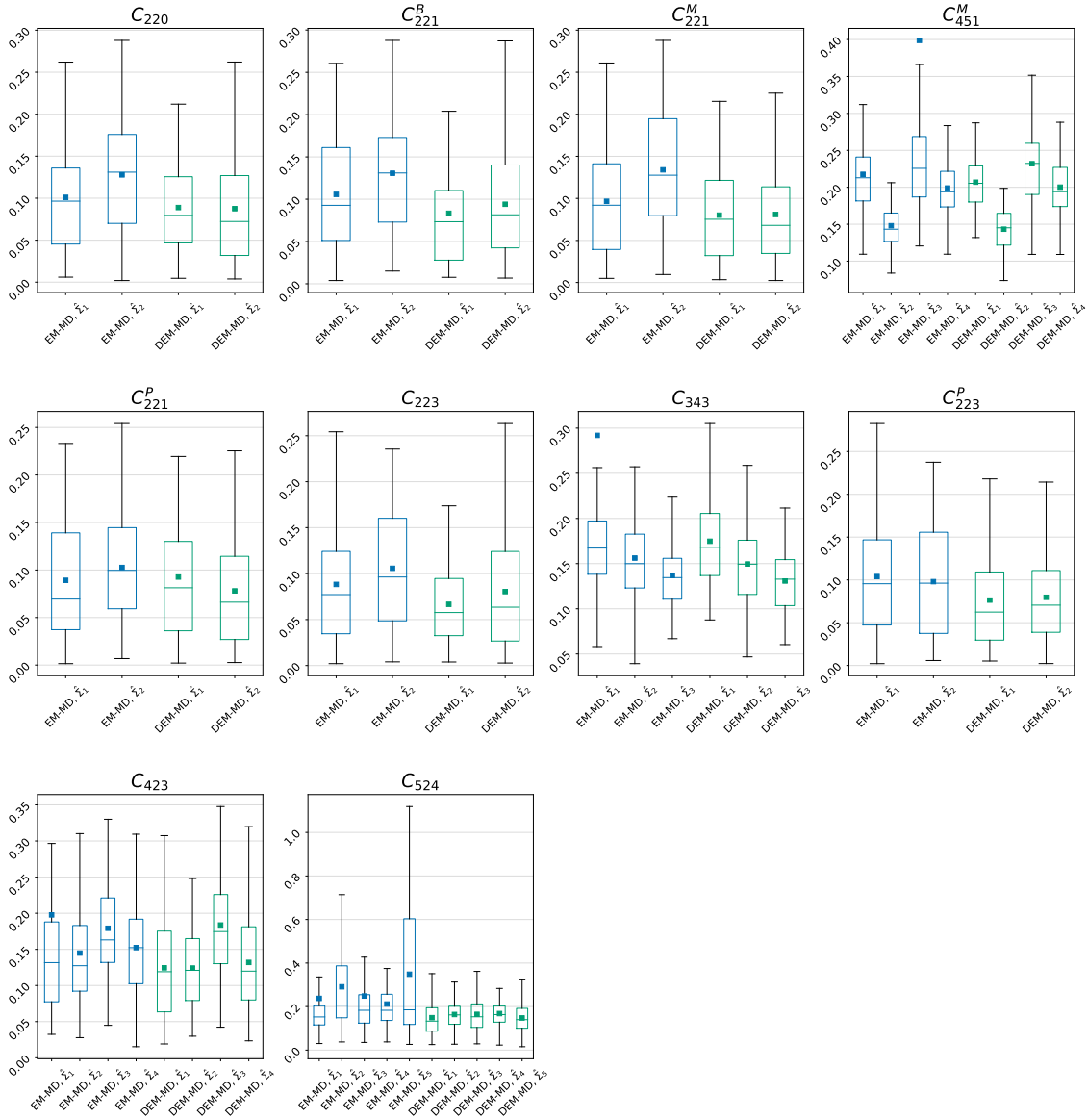
(c) SAL settings

Figure 4.4: Boxplots of centers relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



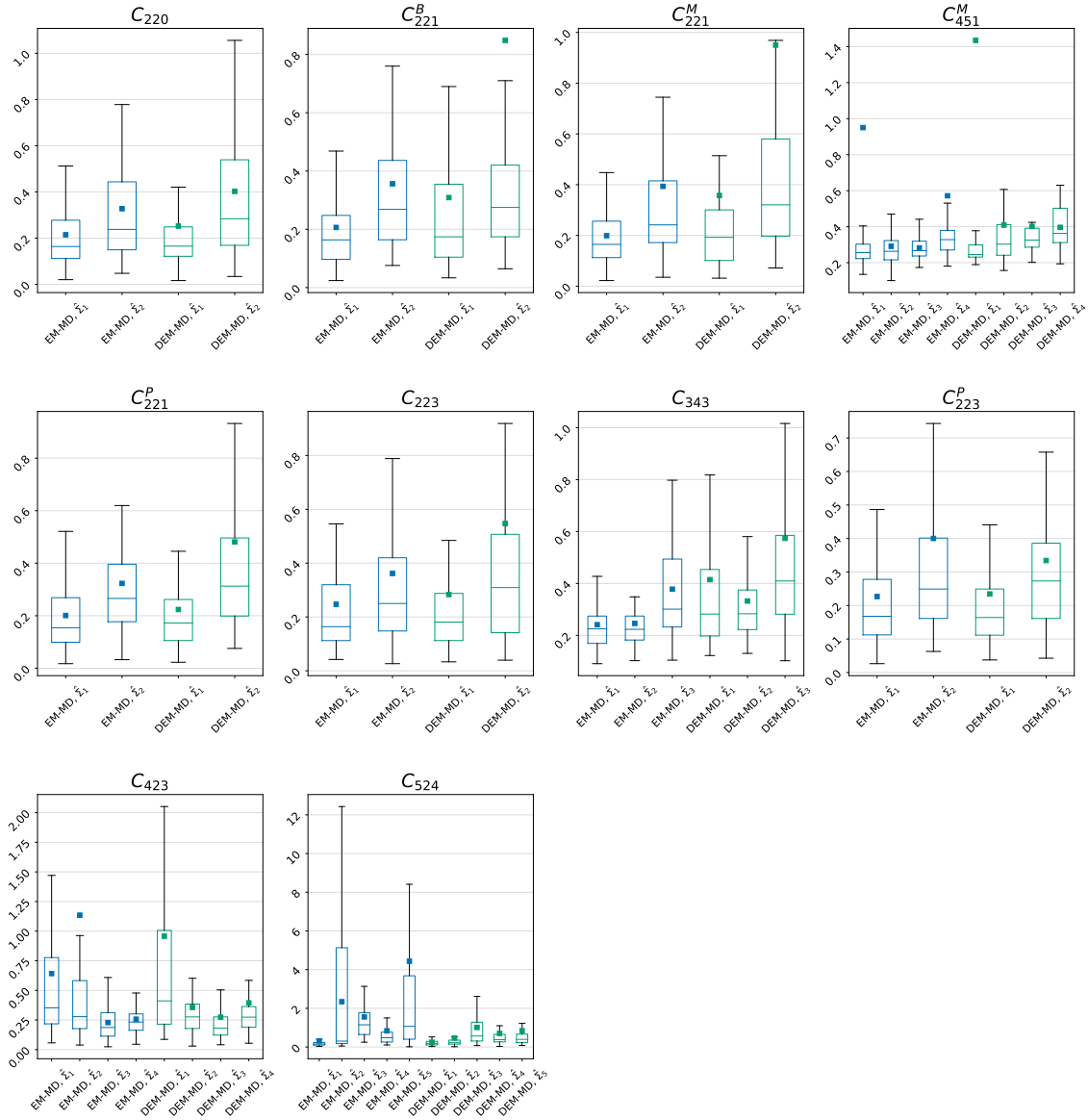
(a) Gaussian configurations

Figure 4.5: Boxplots of scales relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



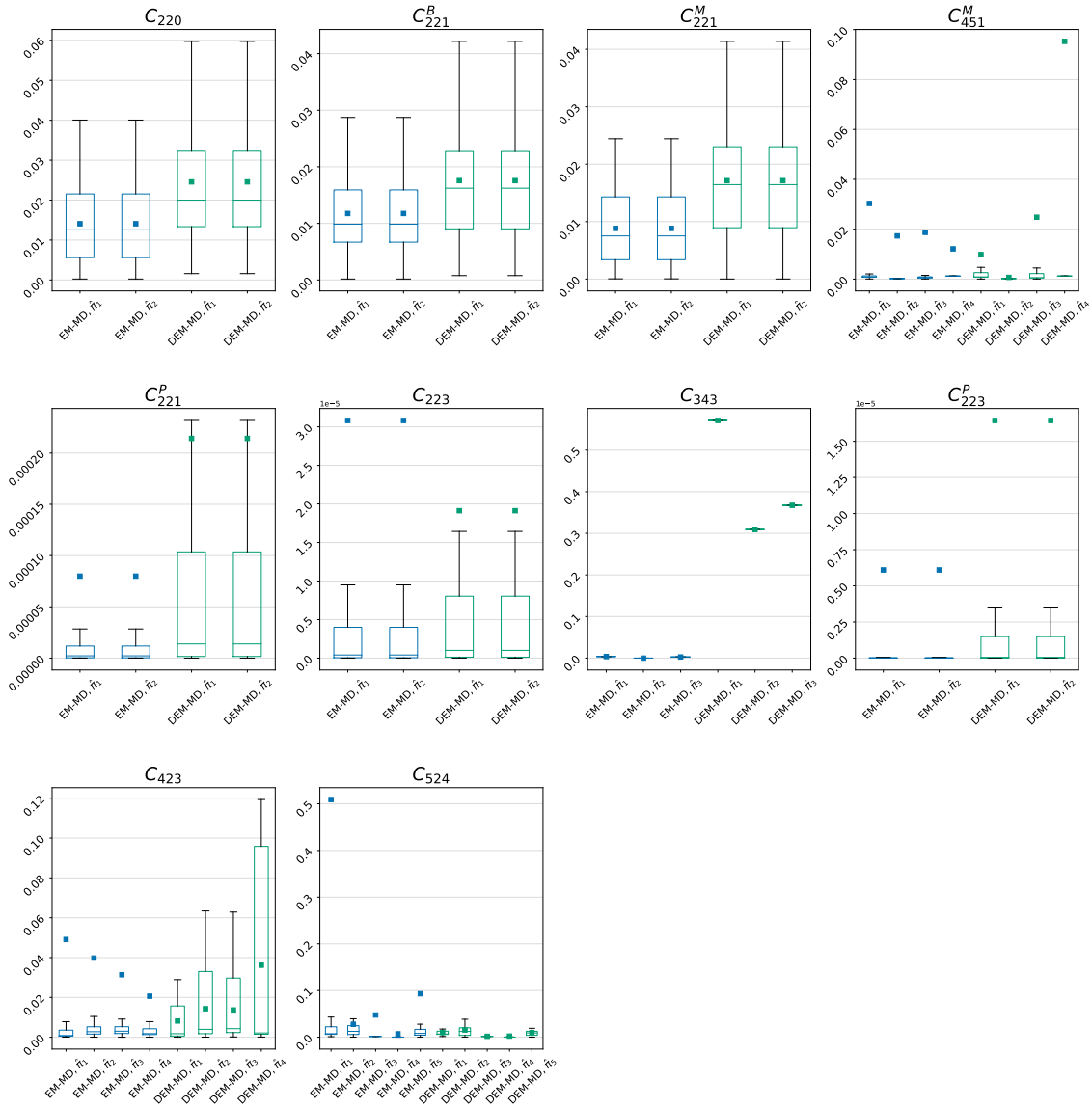
(b) Student configurations

Figure 4.5: Boxplots of scales relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



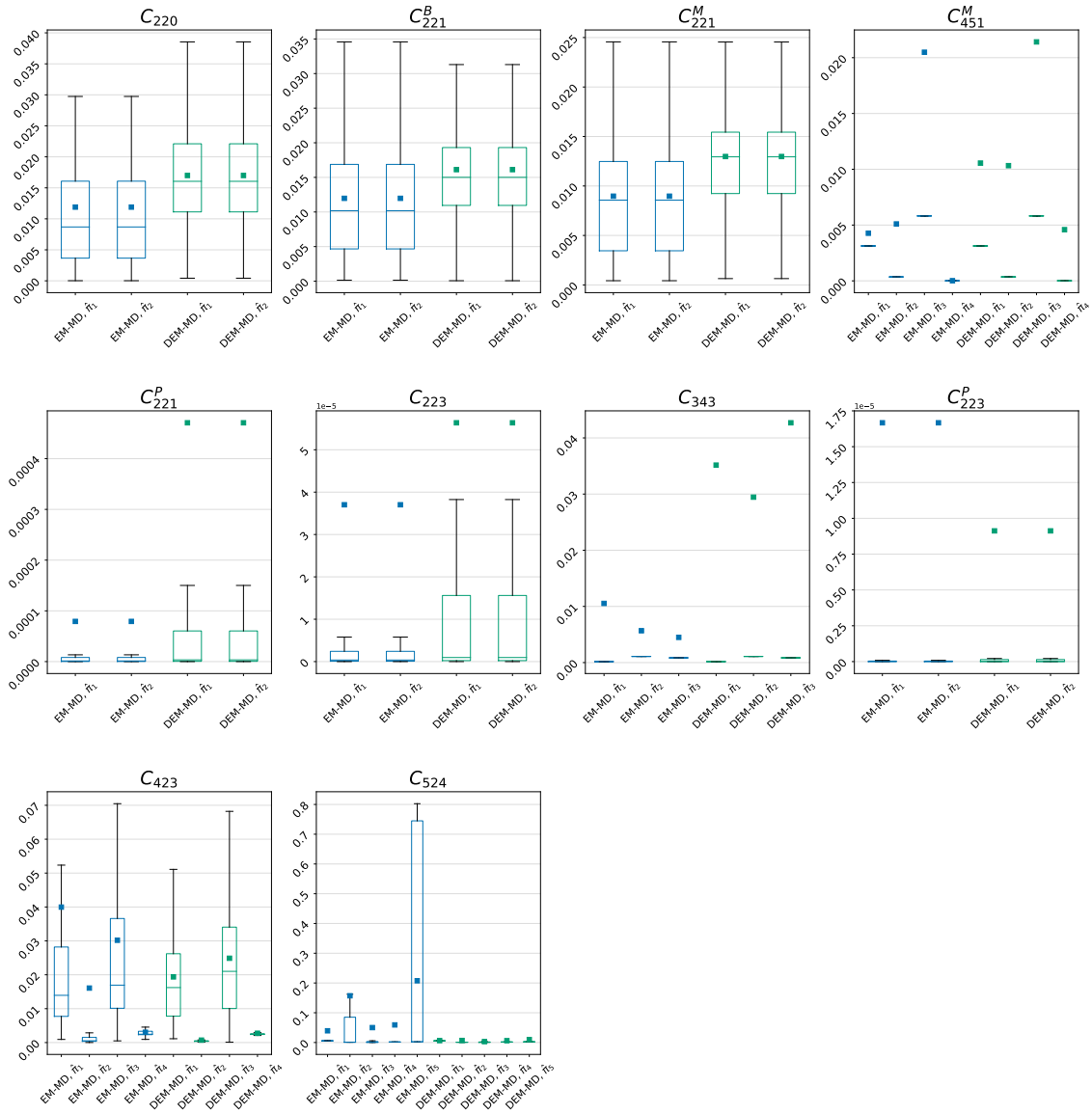
(c) SAL configurations

Figure 4.5: Boxplots of scales relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



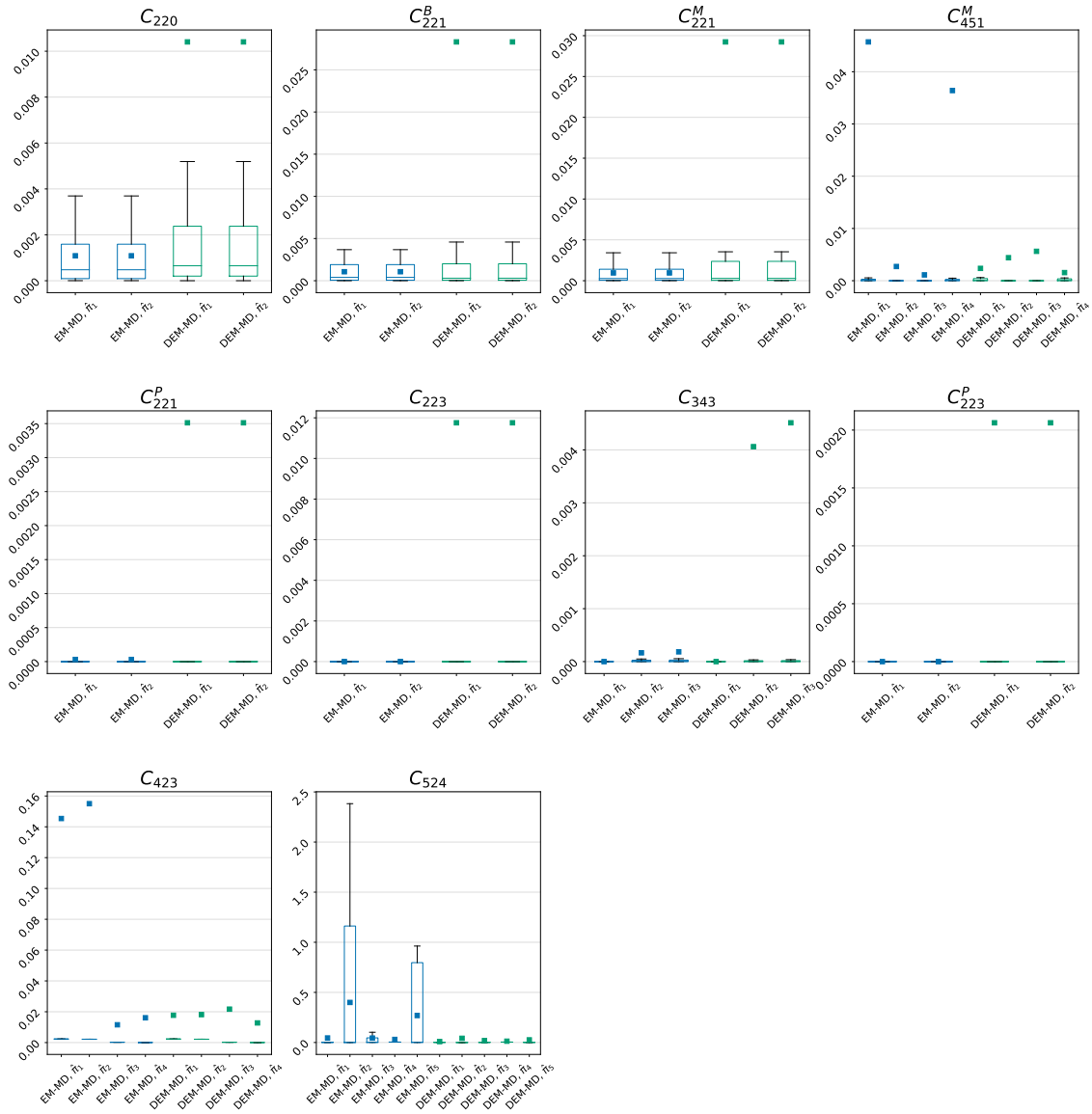
(a) Gaussian settings

Figure 4.6: Boxplots of proportions relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



(b) Student settings

Figure 4.6: Boxplots of proportions relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).



(c) SAL settings

Figure 4.6: Boxplots of proportions relative errors for all settings on both DEM-MD (green) and EM-MD (blue). Each subplot corresponds to a continuous distribution: Gaussian (a), Student (b), SAL (c).

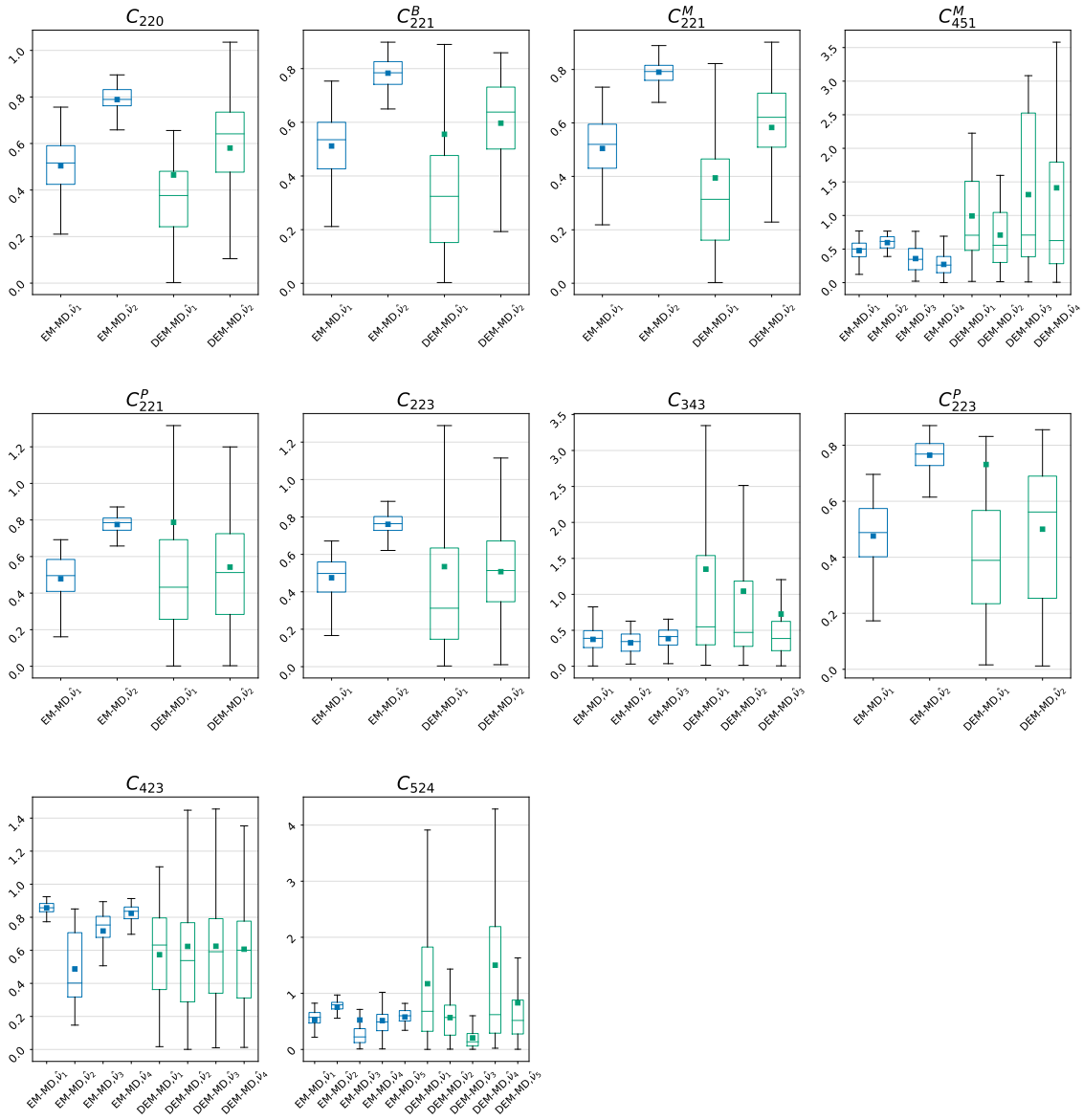


Figure 4.7: Boxplots of degrees of freedom relative errors for all settings on both Student DEM-MD (green) and Student EM-MD (blue).

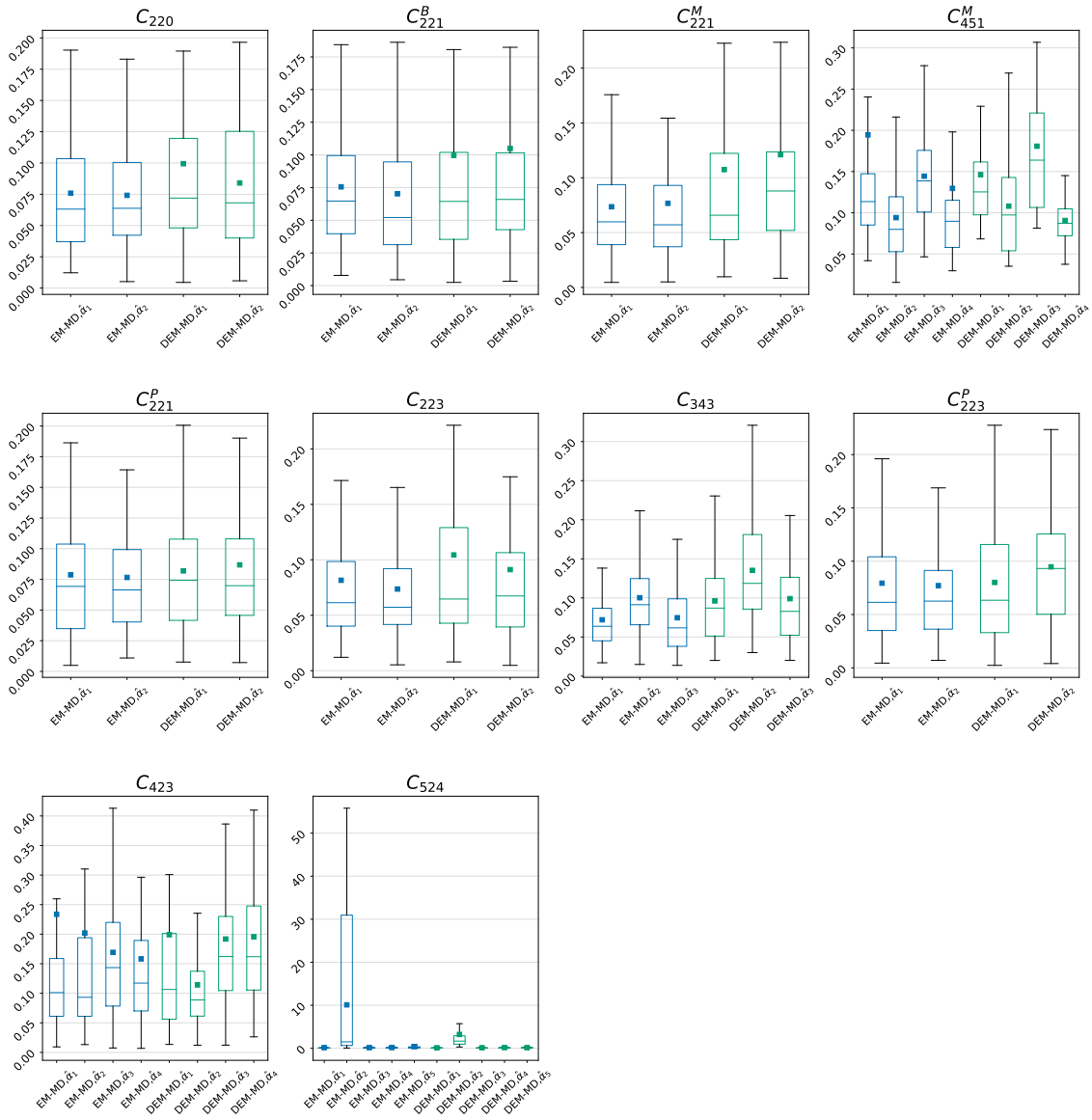


Figure 4.8: Boxplots of skewness relative errors for all settings on both SAL DEM-MD (green) and SAL EM-MD (blue).

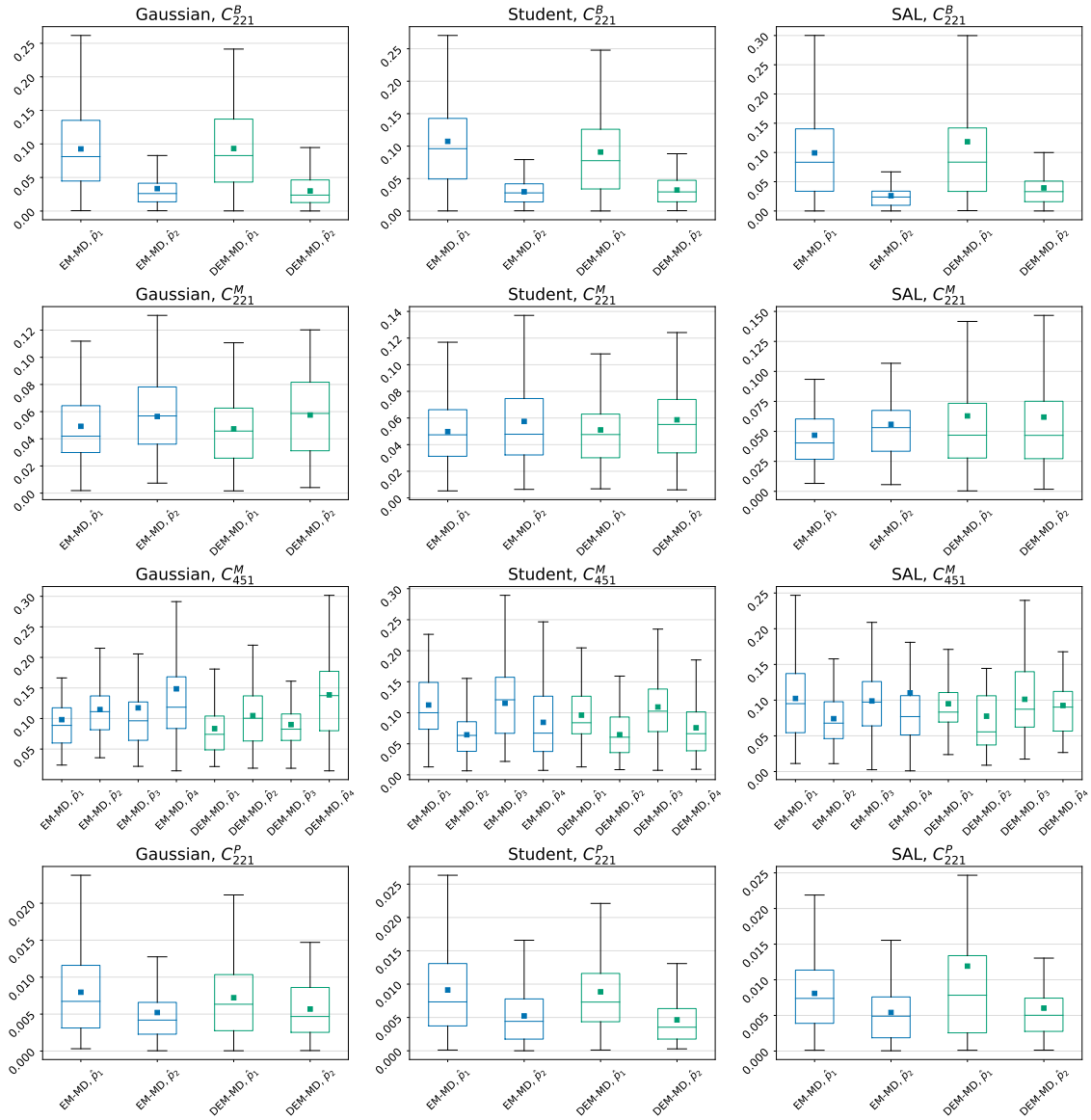
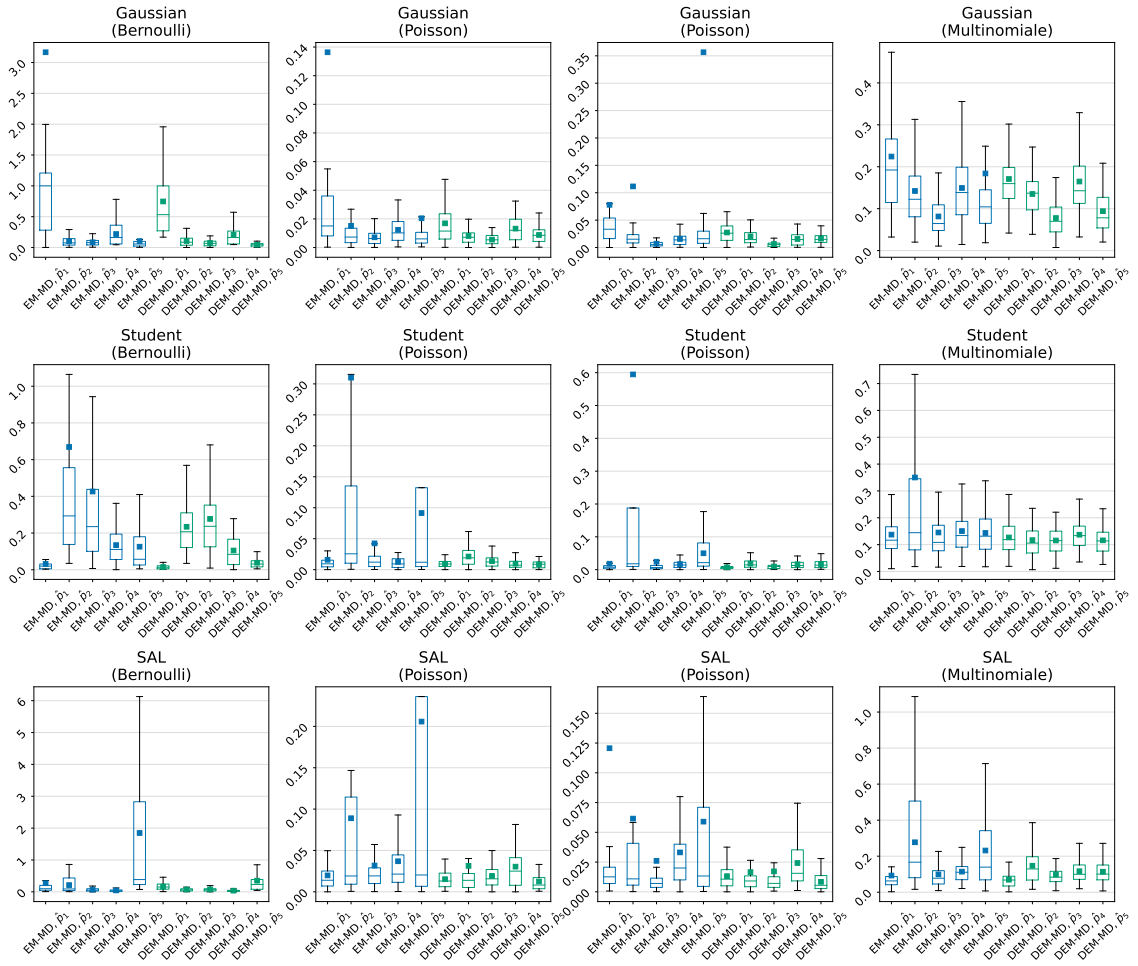
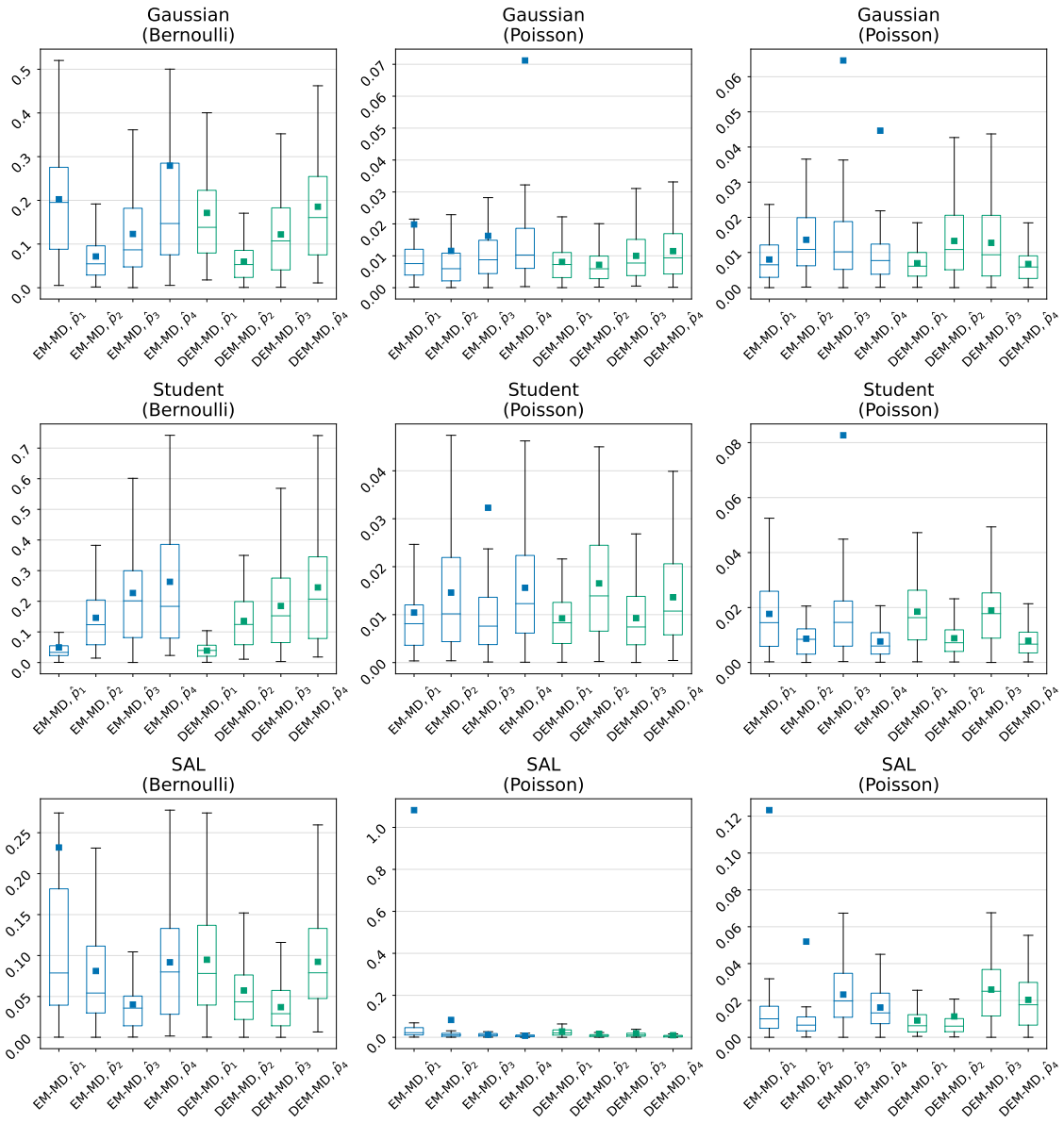


Figure 4.9: Boxplots of relative discrete distribution parameter errors for settings C^B_{221}, C^P_{221}, C^M_{221} and C^M_{451}, for the three continuous distributions on both DEM-MD (green) and EM-MD (blue). Each column corresponds to a continuous distribution, in this order: Gaussian, Student, SAL. Each row corresponds to a setting, in this order: C^B_{221}, C^M_{221}, C^M_{451} and C^P_{221}.



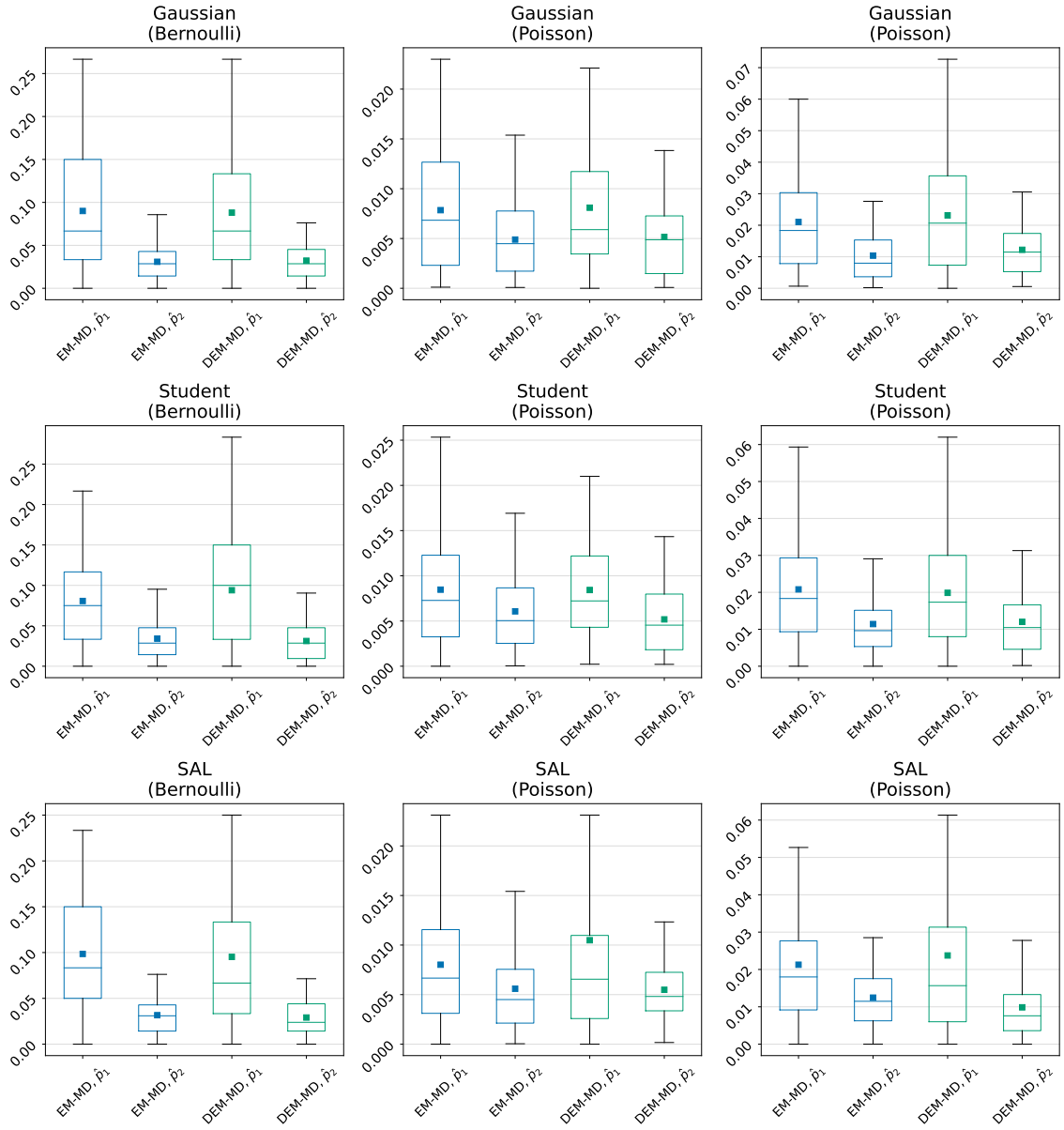
(a) Configuration C_{524}

Figure 4.10: Boxplots of relative discrete distribution parameter errors for settings C_{524} (a), C_{423} (b) and C^P_{223} (c) on both DEM-MD (green) and EM-MD (blue). In each subplot, a row corresponds to a continuous distribution (Gaussian, Student, SAL in this order).



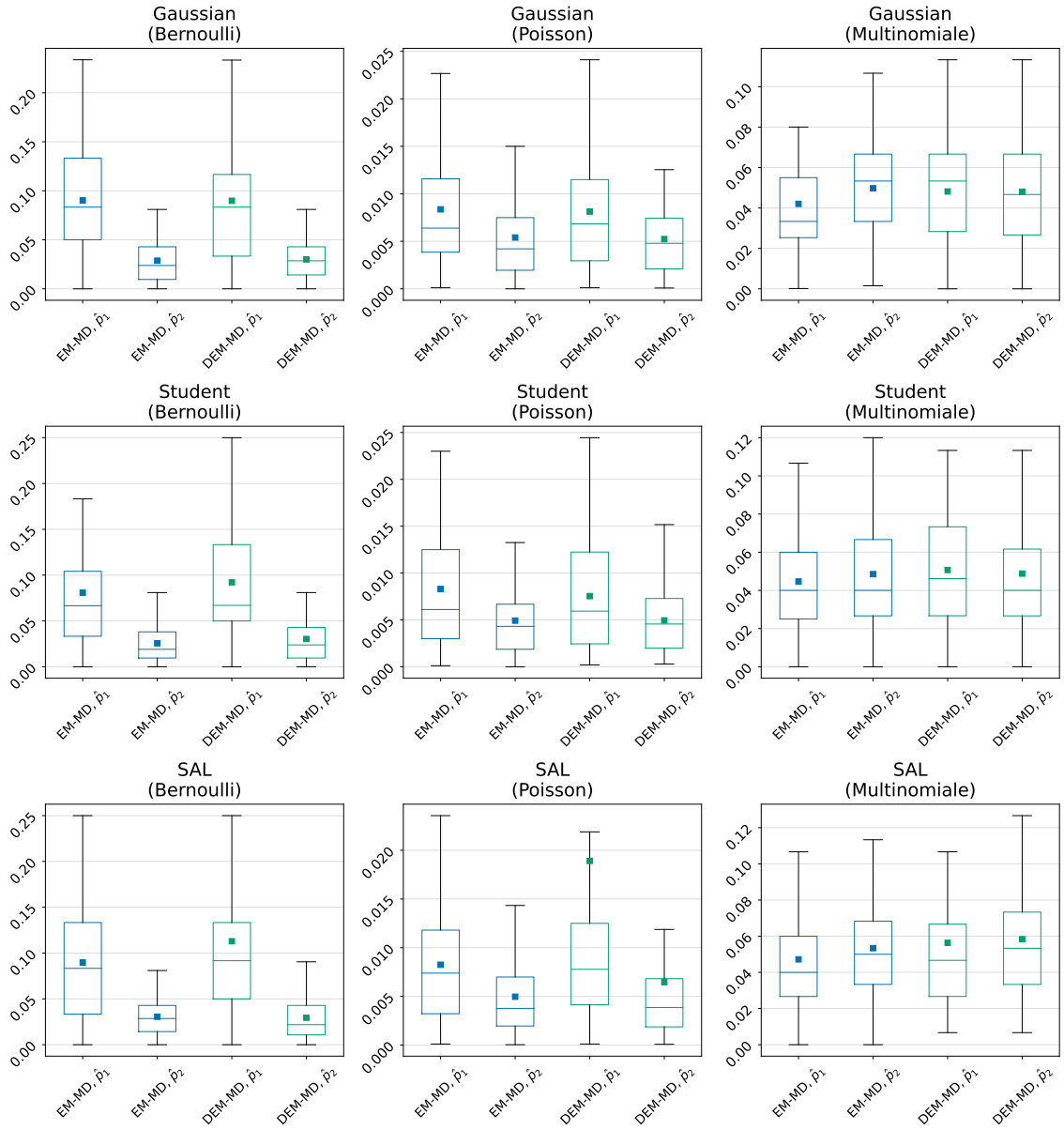
(b) Configuration C_{423}

Figure 4.10: Boxplots of relative discrete distribution parameter errors for settings C_{524} (a), C_{423} (b) and C^P_{223} (c) on both DEM-MD (green) and EM-MD (blue). In each subplot, a row corresponds to a continuous distribution (Gaussian, Student, SAL in this order).



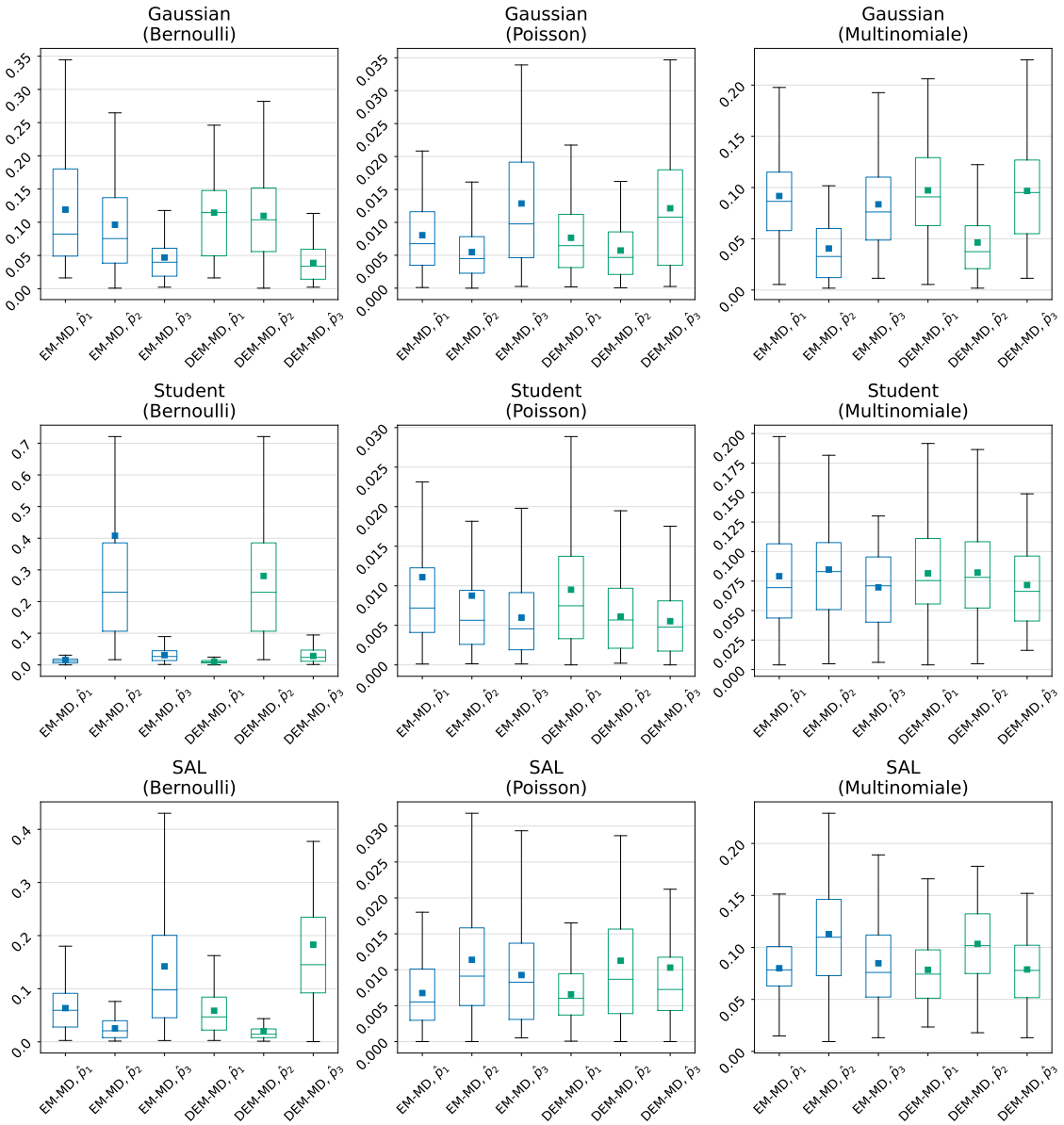
(c) Configuration C^P_{223}

Figure 4.10: Boxplots of relative discrete distribution parameter errors for settings C_{524} (a), C_{423} (b) and C^P_{223} (c) on both DEM-MD (green) and EM-MD (blue). In each subplot, a row corresponds to a continuous distribution (Gaussian, Student, SAL in this order).



(a) Configuration C_{223}

Figure 4.11: Boxplots of relative discrete distribution parameter errors for settings C_{223} (a) and C_{343} (b), for the three DEM-MD on both DEM-MD (green) and EM-MD (blue). In each subplot, a row corresponds to a continuous distribution (Gaussian, Student, SAL in this order).



(b) Configuration C_{343}

Figure 4.11: Boxplots of relative discrete distribution parameter errors for settings C_{223} (a) and C_{343} (b), for the three DEM-MD on both DEM-MD (green) and EM-MD (blue). In each subplot, a row corresponds to a continuous distribution (Gaussian, Student, SAL in this order).

Parameter	True value	DEM-MD	MSAL-BAYES	MSAL-EM
α_1	$\begin{pmatrix} 0 \\ -3 \end{pmatrix}$	$\begin{pmatrix} 0.00 \pm 0.20 \\ -2.98 \pm 0.61 \end{pmatrix}$	$\begin{pmatrix} -0.01 \pm 0.13 \\ -3.30 \pm 0.50 \end{pmatrix}$	$\begin{pmatrix} -0.00 \pm 0.13 \\ -2.78 \pm 0.44 \end{pmatrix}$
μ_1	$\begin{pmatrix} 0 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 0.02 \pm 0.09 \\ 9.85 \pm 0.17 \end{pmatrix}$	$\begin{pmatrix} 0.01 \pm 0.08 \\ 10.02 \pm 0.09 \end{pmatrix}$	$\begin{pmatrix} 0.00 \pm 0.09 \\ 9.77 \pm 0.24 \end{pmatrix}$
Σ_1	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\begin{pmatrix} 1.11 \pm 0.5 & 0.51 \pm 0.28 \\ 0.51 \pm 0.28 & 1.42 \pm 0.55 \end{pmatrix}$	$\begin{pmatrix} 1.21 \pm 0.33 & 0.53 \pm 0.26 \\ 0.53 \pm 0.26 & 1.00 \pm 0.51 \end{pmatrix}$	$\begin{pmatrix} 1.01 \pm 0.20 & 0.48 \pm 0.21 \\ 0.48 \pm 0.21 & 1.68 \pm 0.87 \end{pmatrix}$
π_1	1./3.	0.33 ± 0.02	0.35 ± 0.03	0.33 ± 0.03
α_2	$\begin{pmatrix} 3 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 2.92 \pm 0.36 \\ 2.95 \pm 0.53 \end{pmatrix}$	$\begin{pmatrix} 3.02 \pm 0.35 \\ 3.05 \pm 0.34 \end{pmatrix}$	$\begin{pmatrix} 2.85 \pm 0.37 \\ 2.84 \pm 0.36 \end{pmatrix}$
μ_2	$\begin{pmatrix} -10 \\ -10 \end{pmatrix}$	$\begin{pmatrix} -9.91 \pm 0.14 \\ -9.90 \pm 0.11 \end{pmatrix}$	$\begin{pmatrix} -10.01 \pm 0.08 \\ -10.03 \pm 0.07 \end{pmatrix}$	$\begin{pmatrix} -9.82 \pm 0.19 \\ -9.82 \pm 0.20 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1.23 \pm 0.54 & 0.26 \pm 0.39 \\ 0.26 \pm 0.39 & 1.34 \pm 0.74 \end{pmatrix}$	$\begin{pmatrix} 1.05 \pm 0.47 & -0.21 \pm 0.22 \\ -0.21 \pm 0.22 & 0.91 \pm 0.35 \end{pmatrix}$	$\begin{pmatrix} 1.54 \pm 0.65 & 0.53 \pm 0.65 \\ 0.53 \pm 0.65 & 1.53 \pm 0.68 \end{pmatrix}$
π_2	1./3.	0.33 ± 0.01	0.33 ± 0.03	0.33 ± 0.03
α_3	$\begin{pmatrix} -3 \\ 3 \end{pmatrix}$	$\begin{pmatrix} -2.96 \pm 0.36 \\ 2.98 \pm 0.50 \end{pmatrix}$	$\begin{pmatrix} -2.96 \pm 0.30 \\ 2.94 \pm 0.30 \end{pmatrix}$	$\begin{pmatrix} -2.83 \pm 0.31 \\ 2.80 \pm 0.32 \end{pmatrix}$
μ_3	$\begin{pmatrix} 10 \\ -10 \end{pmatrix}$	$\begin{pmatrix} 9.89 \pm 0.12 \\ -9.88 \pm 0.11 \end{pmatrix}$	$\begin{pmatrix} 10.04 \pm 0.07 \\ -10.02 \pm 0.07 \end{pmatrix}$	$\begin{pmatrix} 9.84 \pm 0.17 \\ -9.82 \pm 0.18 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}$	$\begin{pmatrix} 1.30 \pm 0.49 & -0.06 \pm 0.4 \\ -0.06 \pm 0.4 & 1.44 \pm 0.78 \end{pmatrix}$	$\begin{pmatrix} 0.94 \pm 0.31 & 0.33 \pm 0.16 \\ 0.33 \pm 0.16 & 0.88 \pm 0.30 \end{pmatrix}$	$\begin{pmatrix} 1.55 \pm 0.66 & -0.29 \pm 0.60 \\ -0.29 \pm 0.60 & 1.50 \pm 0.65 \end{pmatrix}$
π_3	1./3.	0.33 ± 0.02	0.33 ± 0.03	0.34 ± 0.03

Table 4.1: True parameter values and mean estimates with standard deviations returned by our DEM-MD algorithm and extracted results from Fang et al. (2023)'s paper (see Table 3).

The Peel dataset To assess the performance of our DEM-MD on Student continuous distributions, we consider here a simulated two-dimensional Student mixture model with three clusters, originally from Peel and Mclachlan (2000). Given this mixture model, from which the true parameters are available in Table 4.2, we simulate 100 datasets of size $n = 200$, as in the literature. On each dataset, we run DEM-MD algorithm, which we compare to estimation by a classical EM algorithm with R package `mixture` (Pocuca et al., 2022). Student mixture models can be estimated in `mixture` with the function `tpcm` for which we kept all the default parameter values, with a completely unconstrained covariance structure. In their package, Pocuca et al. (2022) also consider Aitken's convergence as a criterion for stopping their algorithm.

Firstly, Student DEM-MD obtains $C = 84$ over the 100 runs, which all converged. We retrieve in Table 4.2 the average (and standard deviation) estimated parameters from these 84 correct DEM-MD runs and the 100 runs by EM-`mixture`.

These results show that Student DEM-MD retrieves correctly the different parameters, better than the Student EM algorithm from `mixture`. On degrees of freedom estimation, as noticed in Subsubsection 5.4.1 on all our simulations, both algorithms are far from the true values, confirming that this stays a challenge with EM-like algorithms. For all the parameters of scales, centers and degrees of freedom, the average estimates with DEM-MD algorithm are closer to the real values than those obtained by the EM-`mixture` algorithm. Moreover, the majority of the DEM-MD estimates have smaller standard deviations than those returned by EM-`mixture` algorithm. Only proportions parameters are slightly better in average and dispersion with EM-`mixture` algorithm.

Conclusion Our DEM-MD algorithms for Student and Shifted Asymmetric Laplace distributions perform as well as the algorithms in the literature on simulated Trillium and Peel datasets. In addition, DEM-MD algorithms have to estimate the number of classes and perform very well, reducing computation time.

5.5 Penalize covariances with Inverse-Wishart priors

We observed limitations of SAL DEM-MD in the presence of an undersized dataset in Subsection 5.3. This was particularly noticeable on setting C^M_{451} where $G = 5$ and the algorithm only obtained 18% of correct $\hat{K}$. In this problematic case, a possibility to improve and/or stabilize the estimation is to regularize the covariances.

Estimation of scale matrices with an Inverse Wishart prior We rely here on regularization as proposed in Subsection 5.5, by the introduction of a prior on scales matrices (Fraley and Raftery, 2007; Fop et al., 2019; Baudry and Celeux, 2015) in the objective function $\tilde{Q}$. The considered prior is an Inverse Wishart prior with $\boldsymbol{\Sigma} \sim W^{-1}(w, \boldsymbol{W})$ with w degrees of freedom and $\boldsymbol{W}$ scale matrix of the prior distribution. We shall refer the reader to Subsection 5.5 for further details on the choice of W and w. This prior was also adopted in the very recent work of Fang et al. (2023) which estimated parameters by a Gibbs sampling method and therefore needed to define parameter priors.

We recall that, when regularized by this prior, the scale matrix updates $\hat{\boldsymbol{\Sigma}}^t_k$ in the M-step of any EM-like algorithm are replaced at t iteration (Fraley and Raftery, 2007; Baudry and Celeux, 2015) by

$$\hat{\boldsymbol{\Sigma}}^{t,reg}_k = \frac{\hat{\boldsymbol{\Sigma}}^t_k + \boldsymbol{W}}{n_k + w + g + 1}. \tag{4.40}$$

Motivation on DEM-MD In a SAL DEM-MD or EM-MD, as in other EM-like algorithms, when the number of points per mixture is not high enough, estimation becomes

	True parameters	DEM-MD	EM-`mixture`
ν_1	5	9.52 ± 13.11	20.36 ± 31.90
μ_1	$\begin{pmatrix} 0 \\ 3 \end{pmatrix}$	$\begin{pmatrix} -0.02 \pm 0.38 \\ 2.96 \pm 0.34 \end{pmatrix}$	$\begin{pmatrix} 0.05 \pm 1.01 \\ 2.68 \pm 0.93 \end{pmatrix}$
Σ_1	$\begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 2.06 \pm 0.55 & 0.51 \pm 0.2 \\ 0.51 \pm 0.2 & 0.49 \pm 0.34 \end{pmatrix}$	$\begin{pmatrix} 1.57 \pm 0.6 & -0.01 \pm 0.41 \\ -0.01 \pm 0.41 & 0.34 \pm 0.22 \end{pmatrix}$
π_1	1./3.	0.33 ± 0.02	0.33 ± 0.01
ν_2	30	54.84 ± 53.69	56.07 ± 42.53
μ_2	$\begin{pmatrix} 3 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2.96 \pm 0.31 \\ 0.04 \pm 0.34 \end{pmatrix}$	$\begin{pmatrix} 2.53 \pm 1.48 \\ 0.15 \pm 0.65 \end{pmatrix}$
Σ_2	$\begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 0.98 \pm 0.30 & 0.01 \pm 0.08 \\ 0.01 \pm 0.08 & 0.1 \pm 0.05 \end{pmatrix}$	$\begin{pmatrix} 1.73 \pm 0.7 & 0.09 \pm 0.46 \\ 0.09 \pm 0.46 & 0.37 \pm 0.22 \end{pmatrix}$
π_2	1./3.	0.34 ± 0.01	0.33 ± 0.01
ν_3	10	23.51 ± 34.79	31.88 ± 38.33
μ_3	$\begin{pmatrix} -3 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -3.04 \pm 0.2 \\ 0.0 \pm 0.11 \end{pmatrix}$	$\begin{pmatrix} -2.66 \pm 1.24 \\ 0.18 \pm 0.72 \end{pmatrix}$
Σ_3	$\begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 1.86 \pm 0.47 & -0.5 \pm 0.19 \\ -0.5 \pm 0.19 & 0.51 \pm 0.13 \end{pmatrix}$	$\begin{pmatrix} 1.7 \pm 0.66 & -0.05 \pm 0.48 \\ -0.05 \pm 0.48 & 0.40 \pm 0.22 \end{pmatrix}$
π_3	1./3.	0.34 ± 0.02	0.33 ± 0.01

Table 4.2: True parameter values and average estimates with standard deviations returned by our DEM-MD algorithm and EM algorithm from R package `mixture` on Student distributions (function `tpcm`).

harder, and this is particularly true for the scales matrices which can become singular and cause the algorithm to diverge. Regularization parameters are employed to avoid this problem, frequently artificial ones in classical EM algorithms, and regularizations based on prior information in variational methods. Thus, fewer pathological special cases can be obtained but with subtle bias and worse parameter estimations. In cases where the number of points is sufficient with regard to the number of parameters, EM-type algorithms without regularization or with a small regularization value generally perform both as well.

Prior regularizations become valuable when dealing with high-dimensional datasets. In DEM-MD with SAL continuous distributions, which correspond to the highest complexity, this scale matrix prior helps to achieve greater convergence and better results. With a not-so-large continuous space dimension, DEM-MD frequently diverges without this type of regularization.

With the prior regularization, the scale matrix estimation step in Algorithm 4.7 can be replaced by Eq.(4.40) with $\hat{\mathbf{\Sigma}}_k^t = n_k^t \times \hat{\mathbf{\Sigma}}_{k,\text{EM}}^t$ and $\hat{\mathbf{\Sigma}}_{k,\text{EM}}^t$ estimated by Eq.(4.27). This consideration is particularly important in real cases where it is frequent to have several variables and a not-so-large dataset, as we will see in Section 6.

Application on setting C_{451}^M Here, the addition of a regularization such as Fraley's in the SAL DEM-MD for setting C_{451}^M leads to a clear improvement in the convergence of

the algorithm and the estimation of the number of classes. As said above, changes in the algorithm and implementation are lights, only requiring a few additional steps to compute $W = \frac{S}{K^{2/g}}$ with S the overall empirical covariance matrix and then $\hat{\mathbf{\Sigma}}_k^{t,reg}$ with Eq.(4.40) for each cluster k. Our simulations on $S = 100$ new datasets of size $n = 600$ simulated from C_{451}^M and modeled with **regularized** SAL DEM-MD drastically improved convergence of the algorithm, going from 18% to 100%. The rate of correct $\hat{K}$ is now 72% instead of 18%, reaching performances obtained with $n = 2200$ points without scale regularization (see Fig. 4.3). Figure 4.12 gives us the relative errors of these simulations computed similarly to the previous ones. In addition, relative errors of SAL DEM-MD without regularization are also featured. We observe that the errors are sometimes more dispersed but globally similar to the ones without regularization on the configuration C_{451}^M.

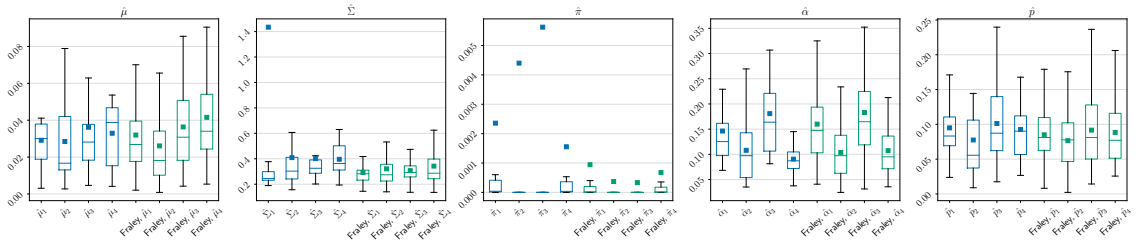


Figure 4.12: Relative errors of SAL DEM-MD with/out Fraley regularization, on setting C_{451}^M. In blue are given the errors without regularization, from the same 18 previous runs. In green are the errors with scale regularization. Results are over 72 runs with $n = 600$.

These experiments open up possibilities for better estimating this type of model at high complexity and with a low volume of data. Particularly for real use cases where there may be many variables and where DEM-MD, without regularization, struggles to estimate the mixture model. This applies also to the problem discussed in Chapter 3, where there are few, widely dispersed data, which also complicates the estimation of the model.

6 Experiments on real datasets

6.1 A Prostate Cancer dataset

Description This dataset was firstly analyzed by Byar and Green (1980), and then by Hunt and Jorgensen (1996). Recently it was analyzed in papers on mixed-type data models (McParland and Gormley, 2016; Foss and Markatou, 2018). Fifthteen variables are available for $n = 475$ prostate cancer patients who were diagnosed as having either stage 3 or stage 4 prostate cancer. The variables are as detailed in the following table.

The outputs variables, which should be spread apart are Stage and SurvStat, as well as Observation which are patient IDs. Depending on the existing works, some may include Stage in the observed dataset and try to explain the SurvStat variable (Foss and Markatou, 2018) while others consider it as an output to be explained by the estimated clustering (McParland and Gormley, 2016). Following precedent works and univariate analysis, the variables size of primary tumor and serum prostatic acid phosphatase were transformed, with respectively a square root transformation and a logarithmic transformation.

In brief, we have 8 continuous variables and 4 categorical variables which are: Performance rating (Multinomial), Cardiovascular disease history (Bernoulli), Bone metastasis (Bernoulli) and Electrocardiogram code (Multinomial). As we saw in the previous section, the SAL DEM-MD algorithm quickly encounters difficulties in estimating a model with a

Variable	Type	Description
Stage	Output variable	Stage of patient's prostate cancer: 3 or 4
SurvStat	Output variable	Post-trial survival statut: 0-alive, 1-dead from prostatic cancer, 2-dead from heart or cardiovascular disease, 3-dead from cerebrovascular accident, 4-dead from a pulmonary embolus, 5-dead from other cancer, 6-dead from respiratory disease, 7-dead from other specific non-cancer cause, 8-dead from other unspecified non-cancer cause, 9-dead from unknown cause
Age	Continuous	Age of the patient
Weight	Continuous	Weight of the patient
Performance rating	Ordinal	how active the patient is: 0-normal activity, 1-in bed less than 50% of daytime, 2-in bed more than 50% of daytime, 3-confined to bed.
Cardiovascular disease history	Binary	cardiovascular disease history: 0-no, 1-yes
Systolic Blood Pressure	Continuous	Systolic blood pressure of the patient in units of ten
Diastolic Blood Pressure	Continuous	Diastolic blood pressure of the patient in units of ten
Electrocardiogram code	Categorical	0-normal, 1-benign, 2-rhythmic disturbances, 3-heart blocks or conduction defects, 4-heart strain, 5-old myocardial infarct, 6-recent myocardial infarct
Serum haemoglobin	Continuous	Serum haemoglobins levels in g/100ml
Size of primary tumor	Continuous	Estimated size of primary tumor in cm^2
Index of tumor stage and histolic grade	Continuous	Combined index of tumor stage and histolic grade of the patient
Serum prostatic acid phosphatase	Continuous	Serum prostatic acid phosphatase levels in King-Armstrong units
Bone Metastase	Binary	Presence of bone metastasis: 0-no, 1-yes

Table 4.1: Description of variables of the Prostate Cancer dataset.

high-dimensional dataset. For its estimation on the prostate cancer dataset, we therefore apply a Fraley regularization on the scale matrices as presented in the Subsection 5.5.

Results With this dataset in hand and variables treated as explained above, Student DEM-MD and SAL DEM-MD found $\hat{K} = 2$, while Gaussian DEM-MD estimated $\hat{K} = 3$ clusters. The selection strategy for clustMD (McParland and Gormley, 2016) returned 3 classes as the best model, while the one for KAMILA (Foss and Markatou, 2018) gave 2 classes. Both solutions may be acceptable for a clustering objective as Stage output variable has 2 modalities and SurvStat 3 modalities. What is more interesting is that clustMD method aims at retrieve Stage variable (which has two modalities) while KAMILA aims at

retrieve SurvStat (which has 3 modalities after aggregation of modalities as applied by Foss and Markatou (2018)).

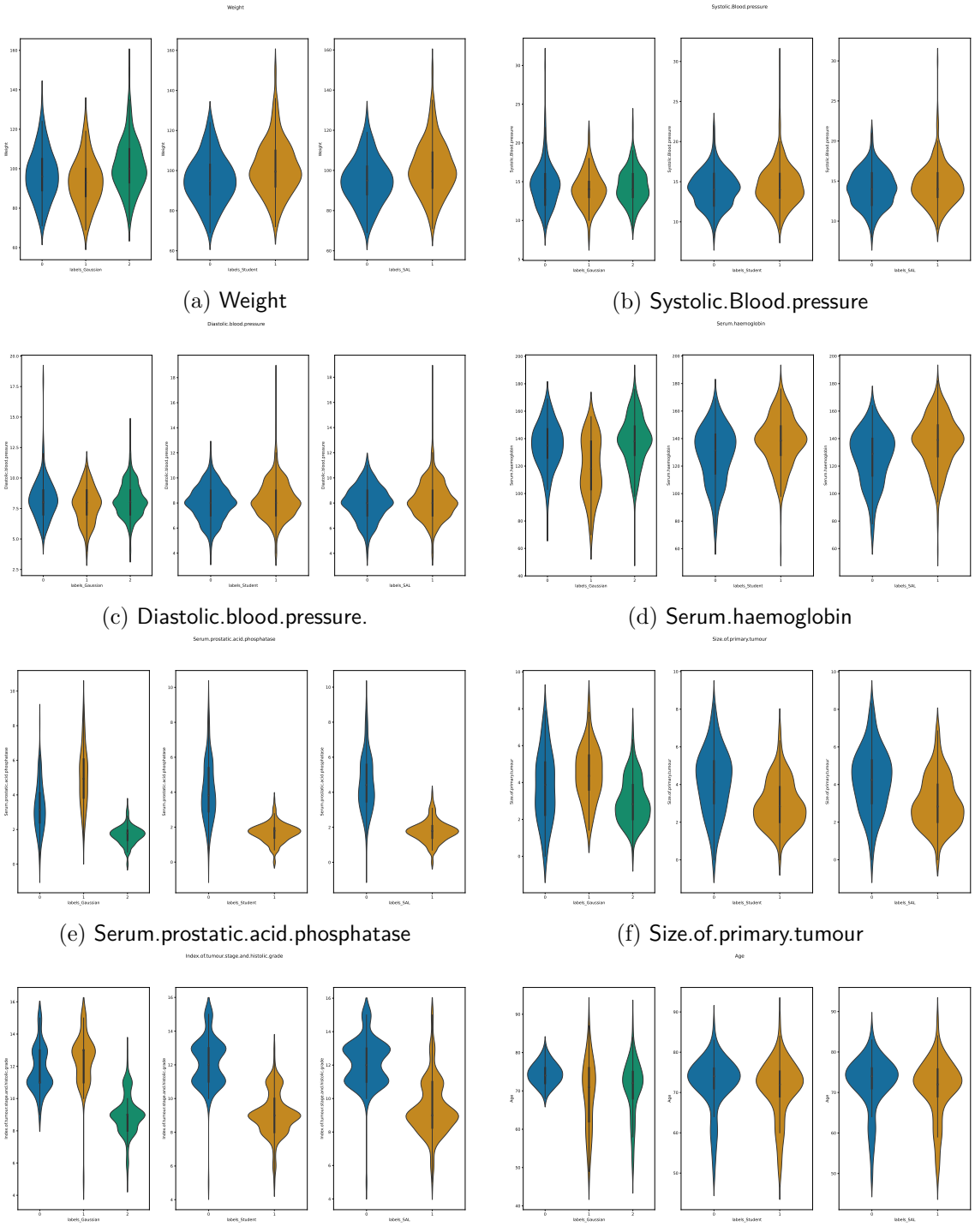
A cross-tabulation of the cluster labels versus the cancer stage diagnosis is given in Table 4.2. Estimated classes by Student and SAL DEM-MD retrieve correctly the Stage of cancer (Stage 3 or 4). As the Gaussian DEM-MD estimates three classes, they cannot directly correspond to the cancer stage. But we see that the third cluster is mainly containing Stage 3 patients, while classes 1 and 2 contain more Stage 4 patients. Student and SAL DEM-MD succeed in characterizing the two groups of stages, even though this is a complex variable, based on the subjective separation of cancer progression. On the contrary, the Gaussian DEM-MD is insufficient to separate Stage groups correctly and lacks flexibility.

Model	Cluster	Stage 3	Stage 4
Gaussian	1	25	100
	2	2	75
	3	246	27
Student	1	17	170
	2	256	32
SAL	1	6	143
	2	267	59

Table 4.2: Cross-tabulation of estimated cluster labels for each model versus the diagnosed prostate cancer stage

Violinplots on Figure 4.13 show the distribution of each continuous variable for each estimated model, by component. For Student and SAL models, differences between clusters are only clearly visible on variables Size.of.primary.tumor, Index.of.tumor.stage and.histolic.grade and Serum.prostatic.acid.phosphatase. For the Gaussian DEM-MD, additional differences can be observed between the first and second classes on the Age and Serum.haemoglobin variables. The patients in the first class have lower serum prostatic acid phosphatase, in average lower index of tumor stage, higher serum hemoglobin and are less dispersed in Age, compared to patients in the second class, all of whom have Stage 4 cancer.

Comparing center vectors for clusters of each model (Fig. 4.14(a)), it can be seen that the classes for the three models are differentiated by the Serum.haemoglobin, Size.of.primary.tumour, Index.tumour.stage.histolic.grade and Serum.prostatic.acid.phosphatase variables. Whereas Age and Diastolic.blood.pressure only differentiate well for Gaussian and SAL models. On the other hand, Weight separates Gaussian and Student models. As the Gaussian model has three classes, we can look at the parameters that best characterize its classes, in particular the first and second ones, which both contain Stage 4 patients. Patients in cluster 1 (against cluster 2) are on average older and heavier, with higher diastolic pressure, higher serum hemoglobin level, lower size of the primary tumor and lower serum prostatic acid phosphatase level. In addition, estimated parameters of binary variables (Fig. 4.14(c)) indicate (still for 1^{st} class compared to 2^{nd} one's) a higher probability of having cardiovascular disease history and lower risk of bone metastases. Multi-modalities variable indicate: from electrocardiogram code (Fig. 4.14(b)) a lower probability for modality 2, corresponding to rhythmic disturbance, the highest probability for 1^{st} class being a normal code, and rhythmic disturbance for 2^{nd} class. The multimodal performance variable indicates a high probability of normal activity for first class against second class (Fig. 4.14(d)). Overall, the probabilities for each modality of the multinomial variables are in similar ranges.



(a) Weight (b) Systolic.Blood.pressure

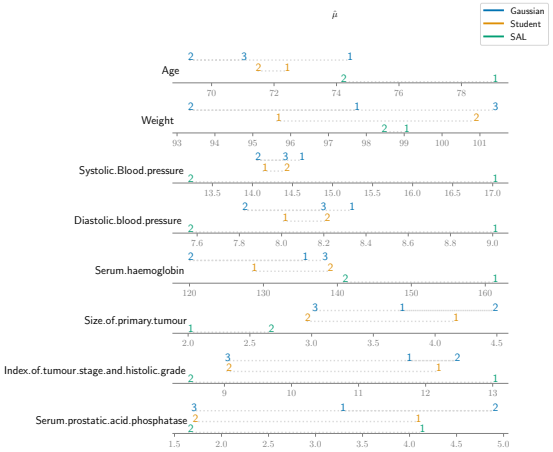
(c) Diastolic.blood.pressure. (d) Serum.haemoglobin

(e) Serum.prostatic.acid.phosphatase (f) Size.of.primary.tumour

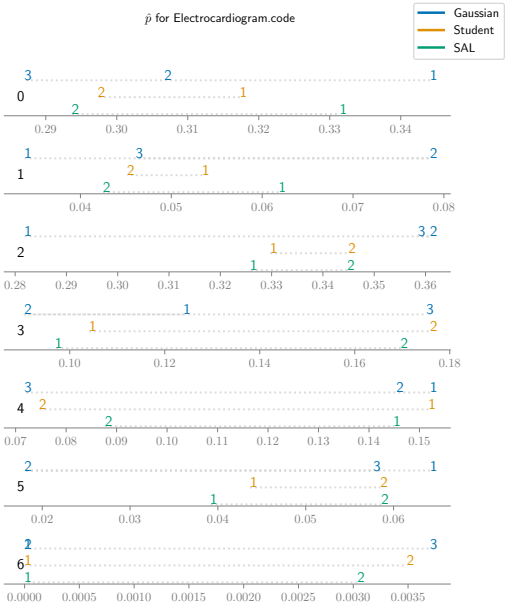
(g) Index.of.tumour.stage.and.histolic.grade (h) Age

Figure 4.13: Violin plots for each continuous variable in prostate cancer dataset, according to clusters assignments by each estimated model. In each subplot, the first column is with Gaussian assignments, the second one with Student assignments and the last one with SAL assignments.

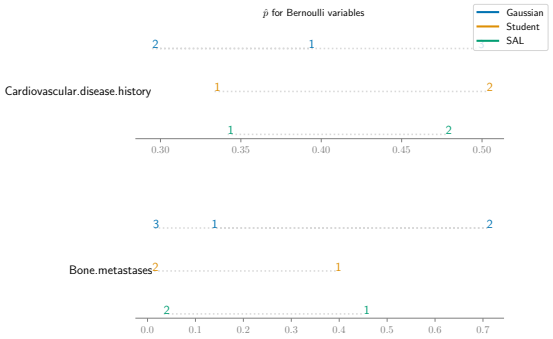
Conclusion Gaussian DEM-MD is not the most appropriate model here to estimate and cluster the prostate cancer dataset. On the other hand, Student and SAL DEM-MD can



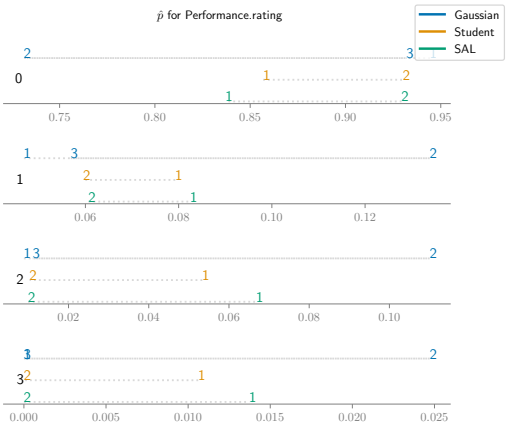
(a) Center parameter estimates



(b) Electrocardiogram.code parameter, with each row being a modality



(c) Bernoulli parameter estimates



(d) Performance.rating parameter, with each row being a modality

Figure 4.14: Estimated centers and discrete parameters for all DEM-MD on the prostate cancer dataset. Each color is associated with a model (see figure legends) and numbers indicate clusters. Rows indicate variable names in (a) and (c). Rows indicate modality numbers in (b) and (d).

dynamically find two classes sufficiently separating the cancer stages. By examining the marginal distributions per class or the various estimated parameters, we note that certain variables differentiate the cancer stages or the Gaussian classes, which improves interpretability.

6.2 The Australian Institute of Sport dataset

Description We now illustrate DEM-MD algorithms on the Australian Institute of Sport (AIS) dataset (Telford and Cunningham, 1991). This dataset was also analyzed in (Lee and McLachlan, 2012, 2014; Lin, 2010) where the authors only use a subsample of continuous variables to try to cluster individuals, which are here athletes, by sex. Thirteen variables are available for $n = 202$ Australian athletes, male and female in ten different sports. Apart from sex and sport variables which are respectively binary and nominals, we have 11 continuous variables, corresponding to physical and blood measurements (Table 4.1).

Variable	Description
sex	the sex of the athlete
sport	the sport of the athlete, one of BBall (basketball), Field, Gym (gymnastics), Netball, Rowing, Swim, T400m, (track, further than 400m), Tennis, TSprint (track sprint events), WPolo (waterpolo)
Ht	height in cm
Wt	weight in kg
LBM	lean body mass in kg
RCC	red blood cell count
WCC	white cell count
HCT	hematocrit in percent
HGB	hemoglobin concentration, in grams per decilitre
Ferr	plasma ferritins in ng per decilitre
SSF	sum of skin folds
Bfat	percentage body fat
BMI	body mass index, in kg per m2

Table 4.1: Description of Australian Institute of Sport dataset

Several models can be considered with this dataset. Since the literature mainly attempts to separate men and women, we will consider a similar framework, excluding sex from the set of estimates, but including sport since we have a model capable of handling categorical variables. Since our main objective is to infer the data using mixture models, without any major clustering objective, we could also decide to use all the variables to estimate a model, including the variable sex. As we saw in the previous section, the SAL DEM-MD algorithm quickly encounters difficulties in estimating a model with a high-dimensional dataset. Therefore, for its estimation on the AIS dataset, we apply a Fraley regularization on the scale matrices as presented in the Subsection 5.5.

Results Running DEM-MD with 12 variables (all except sex variable) on the $n = 202$ athletes, Gaussian DEM-MD estimates $\hat{K} = 3$ classes while a Student DEM-MD estimates $\hat{K} = 4$ and a SAL DEM-MD $\hat{K} = 2$.

The class assignments of the different models are cross-tabulated with sex and sport variables, obtaining interesting results (Tables 4.2 and 4.3). We can see that all three models tend to separate men and women as can be seen in Table 4.2. But where the SAL DEM-MD estimates two classes that correspond pretty well to the sex of athletes, the Gaussian and Student DEM-MD estimate three and four classes respectively, which reveals other interesting information. With the Gaussian model, one class corresponds to women, one to men but the third class is half men and half women. From Table 4.3, the third cluster contains the majority of track athletes, the four gym athletes, several swimmers and tennis players, and some other sports.

With the Student model, we even have two clusters for female athletes. The mixed class of the Student model (Cluster 4) is pretty similar to the Gaussian one (Cluster 3), with all Track 400m, a majority of Track sprint, several swimmers and tennis players, and some other sports. The smaller feminine class is composed of four gyms, several basketball players and a minority of netball and row. The second feminine cluster mainly contains netball players and rowers. These two classes are hard to interpret in terms of categorical variables only but take into account the physiological and biological variables. Figures 4.15 and 4.16 give us for each continuous variable its distribution according to the attributed class of athletes, and for each estimated DEM-MD. We can observe interesting differences between classes that do not correspond only to male/female separation. For example in the Student model, for the fourth class which contains track sports and some swimmers and tennis players, from both sexes, the sum of skin folds is inferior, in density and median. Moreover, Weight and BMI are also in the low range, as it is the case with Cluster 1, composed essentially of women practicing Basketball, Gym, Netball or Row.

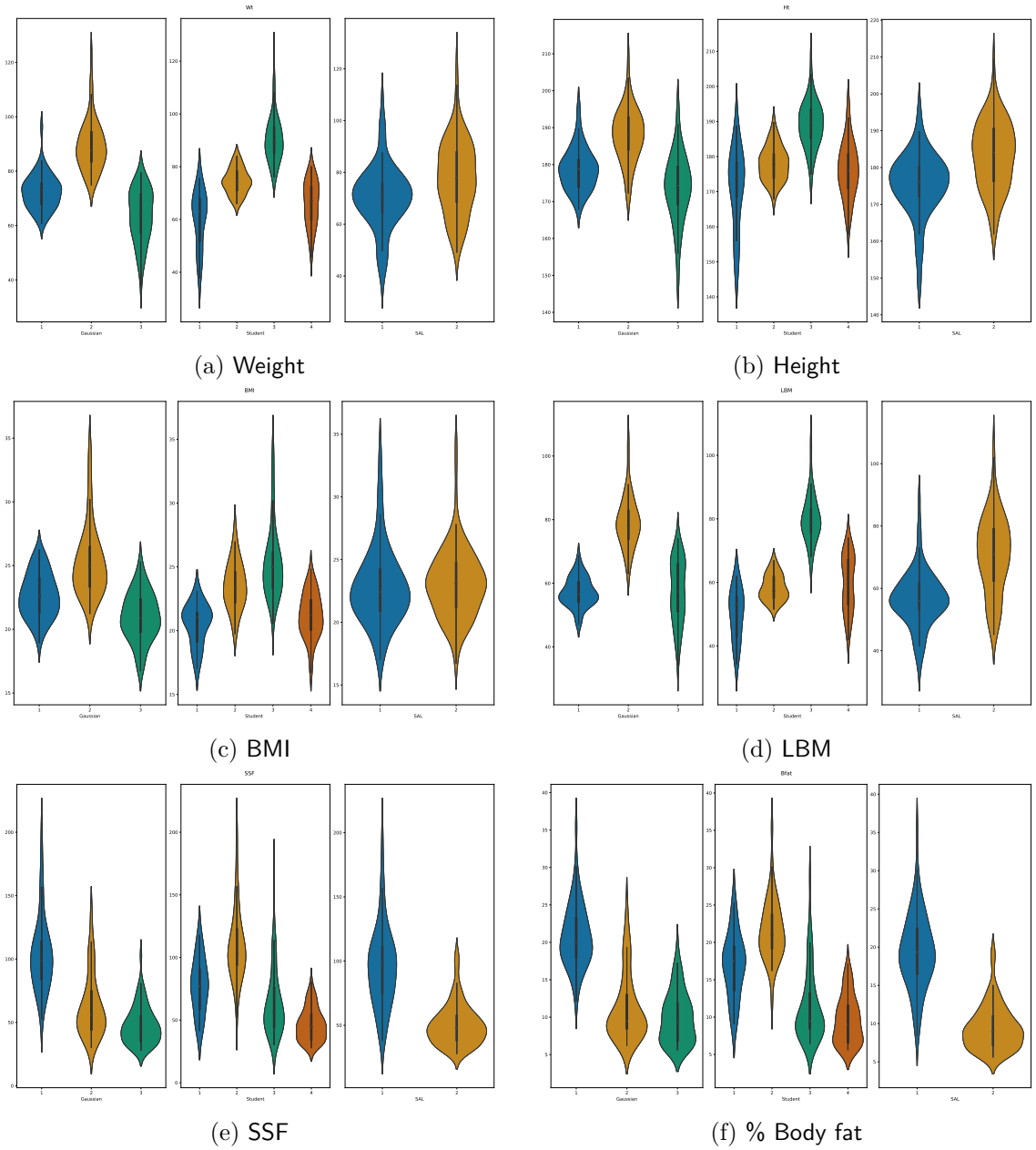
Model	Cluster	Female	Male
Gaussian	1	58	
	2	5	65
	3	37	37
Student	1	21	
	2	44	
	3	4	60
	4	31	42
SAL	1	80	3
	2	20	99

Table 4.2: Cross-tabulation of estimated cluster labels for each model versus the athlete sex

Model	Cluster	Basket Ball	Field	Gym	Netball	Row	Swim	Track 400m	Track Sprint	Tennis	Water Polo
Gaussian	1	12			20	21	3			2	
	2	10	17			14	8	1	3	1	16
	3	3	2	4	3	2	11	28	12	8	1
Student	1	7		4	6	3	1				
	2	3	4		14	19	2			2	
	3	12	13			14	7		3		15
	4	3	2		3	1	12	29	12	9	2
SAL	1	13	8	4	21	22	5	4		6	
	2	12	11		2	15	17	25	15	5	17

Table 4.3: Cross-tabulation of estimated cluster labels for each model versus the practiced sports

Conclusion and perspectives This dataset contains heterogeneous data involving both asymmetric and heavily tailed behaviors. Here each DEM-MD gives a different model with two, three or four clusters. While the SAL DEM-MD well retrieves the sex of the athletes, the Student DEM-MD estimates two classes for women athletes and an additional mixed



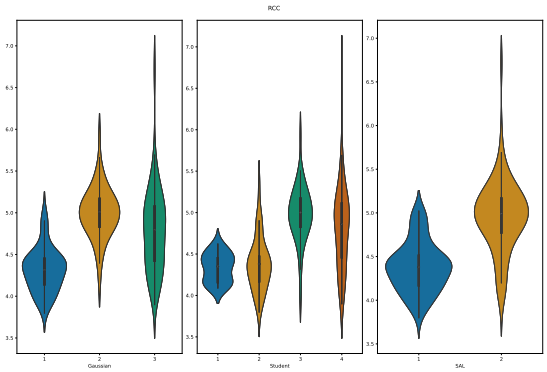
(a) Weight (b) Height

(c) BMI (d) LBM

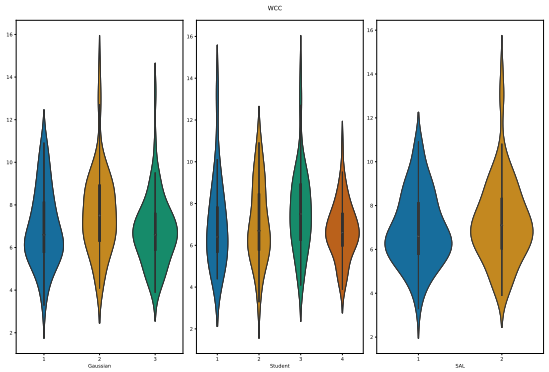
(e) SSF (f) % Body fat

Figure 4.15: Violin plots for each continuous variable in the AIS dataset, according to cluster assignments by each estimated model. In each subplot, the first column is with Gaussian assignments, the second one with Student assignments and the last one with SAL assignments.

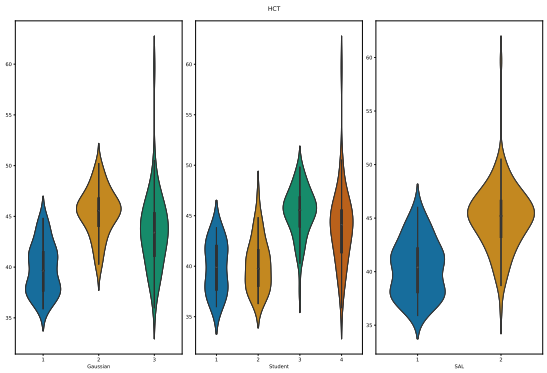
class, which requires looking at sport repartition for a good interpretation. Similarly, the Gaussian DEM-MD returned three clusters which require looking also at assigned classes per sport. All these models give interesting results characterized by different variables and splitting along sex and/or sport with meaningful results.



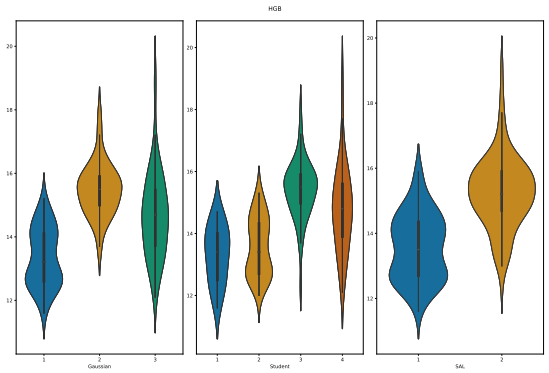
(a) RCC



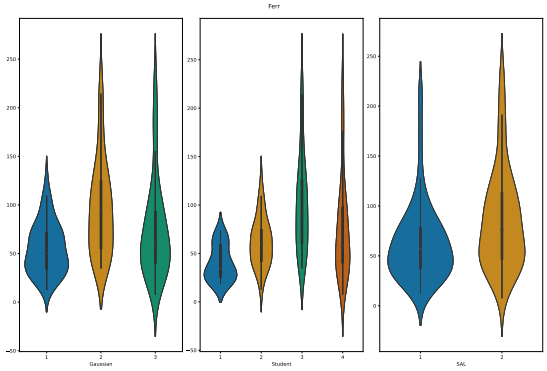
(b) WCC



(c) HCT



(d) HGB



(e) Ferr

Figure 4.16: Violin plots for each continuous variable in the AIS dataset, according to cluster assignments by each estimated model. In each subplot, the first column is with Gaussian assignments, the second one with Student assignments and the last one with SAL assignments.

7 Conclusion and perspectives

Mixture models for mixed-type data with different possible continuous and discrete variables were introduced in this chapter, following a well-established principle of local independence, ensuring the identifiability of models. We proposed Dynamic EM for Mixed-type Data algorithms for these models, allowing us to jointly estimate the number of classes and the various parameters, for both continuous and discrete variables, of these models. We introduced improvements compared to the EM versions, to ensure algorithm convergences and estimations. Especially, our SAL DEM-MD relies on multicycle ECM (Meng and

Rubin, 1993) and deterministic annealing (Ueda and Nakano, 1994) concepts. We also considered Aitken's acceleration for all DEM-MD, which makes more sense than comparing only mixture centers, especially in a dynamic context.

Dynamic EM for Mixed-type Data algorithms were illustrated on both simulated and real datasets. Firstly, we assessed the class estimation performances, including comparisons with some existing model selection criteria. Then we provided parameter recovery results on several simulated settings and comparisons with EM-like algorithms that we also implemented. Estimation capacities are limited by the number of observations and the complexity of the model, as we have observed in certain simulated situations and real scenarios. We also compared ourselves with existing algorithms/implementations on Student or SAL continuous distributions, showing that parameter recovery with DEM-MD is similar, as well as correctly estimating the number of classes.

Finally, we analyzed the DEM-MD's estimates on two real datasets taken from the literature: the Prostate Cancer Dataset and the Australian Institute of Sport Dataset. On both datasets, the different results achieved with different continuous assumptions did not alter their meaning, as they still can be interpreted, and are coherent with the literature, whereas other works considered different objectives. Our Dynamic EM for Mixed-type Data algorithms with the different continuous laws were able to retrieve meaningful classes, despite small dataset sizes, based if necessary on regularized covariances as introduced earlier.

A clear limitation of our models is the local independence assumption. We saw in the introduction that other families of methods can be used to establish links for all variables, but generally involve either the transformation of certain variables or statistical conditioning, such as factor analyzers or copulas. An extension could be to associate mixtures of copulas for mixed-type data estimated by EM-like algorithms (Zhao and Udell, 2020; Rajan and Bhattacharya, 2016) with dynamic estimation of the number of components as done in this chapter. As a lot of copula models are estimated with Bayesian approaches, it could be interesting to estimate mixtures of copula for mixed-type data with reversible jump Monte-Carlo Markov Chain methods to find the optimal space dimensions.

Secondly, although we have introduced regularization on the covariances in situations with high complexity and low sample size, prior regularization could be built into any DEM-MD permanently. By this, we mean that this regularization can be incorporated into the generic version of the algorithm, instead of adding a small amount of noise. This integration should take into account the dimension of the data in order to adapt the regularization to each situation and not obstruct the estimation of the parameters. In the context of a very large continuous dataset, another idea would be to combine DEM-MD for mixed data with factor analyzers, as Franczak et al. (2013) extended mixture of factor analyzers to include SAL density. Thirdly, additional continuous and discrete laws can be modeled and implemented within the framework of the DEM-MD algorithm.

8 Appendix

A Estimation of continuous distributions in mixture models

Law	Latent variable	Updating equation
Gaussian	τ_{ik}	$\frac{\pi_k p_{\text{Gaussian}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} [\pi_j p_{\text{Gaussian}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]}$
Student	τ_{ik}	$\frac{\pi_k p_{\text{Student}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)}{\sum_{j=1}^{K} [\pi_j p_{\text{Student}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j)]}$
	$E_{u,ik}$	$\frac{\nu_k + g}{\nu_k + \delta(\boldsymbol{x}_i \vert \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$
SAL	τ_{ik}	$\frac{\pi_k p_{\text{SAL}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k)}{\sum_{j=1}^{K} [\pi_j p_{\text{SAL}}(\boldsymbol{x}_i \vert \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\alpha}_j)]}$
	E_{1ik}	$\sqrt{\frac{b_{ik}}{a_k}} R_\nu(\sqrt{a_k b_{ik}})$
	E_{2ik}	$\sqrt{\frac{a_k}{b_{ik}}} R_\nu(\sqrt{a_k b_{ik}}) - \frac{2\nu}{b_{ik}}$

Table A.1: Equations to estimate expectation of latent variables at E-step.

Law	Parameter	Equation
Gaussian	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$
	$\hat{\boldsymbol{\mu}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{n}\tau_{ik}}$
	$\hat{\boldsymbol{\Sigma}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}{\sum_{i=1}^{n}\tau_{ik}}$
Student	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$
	$\hat{\boldsymbol{\mu}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}\boldsymbol{x}_i}{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}}$
	$\hat{\boldsymbol{\Sigma}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}{\sum_{i=1}^{n}\tau_{ik}}$
	$\hat{\nu}_k$	$\left\{-\psi\left(\frac{1}{2}\nu\right)+\log\left(\frac{1}{2}\nu\right)+1+\frac{1}{n_k}\sum_{i=1}^{n}\tau_{ik}\left(\log E_{u,ik}-E_{u,ik}\right)+\psi\left(\frac{\nu_k+g}{2}\right)-\log\left(\frac{\nu_k+g}{2}\right)\right\}=0$
SAL	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$
	$\hat{\boldsymbol{\mu}}_k$	$\frac{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik}\boldsymbol{x}_i)-n_k(\sum_i^n\tau_{ik}\boldsymbol{x}_i)}{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik})-n_k^2}$
	$\hat{\boldsymbol{\Sigma}}_k$	$S_k-\boldsymbol{\alpha}_k r_k^\top-r_k\boldsymbol{\alpha}_k^\top+\frac{1}{n_k}\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^\top\sum_{i=1}^{n}\tau_{ik}E_{1ik}$ with $S_k=\frac{1}{n_k}\sum_i^n\tau_{ik}E_{2ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top$ and $r_k=\frac{1}{n_k}\sum_i^n\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)$
	$\hat{\boldsymbol{\alpha}}_k$	$\frac{(\sum_i^n\tau_{ik}E_{2ik})(\sum_i^n\tau_{ik}\boldsymbol{x}_i)-n_k(\sum_i^n\tau_{ik}E_{2ik}\boldsymbol{x}_i)}{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik})-n_k^2}$ or $\frac{\sum_i^n\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k^\star)^\top}{\sum_i^n(\tau_{ik}E_{1ik})}$

Table A.2: Equations to estimate continuous parameters during M-step.

B Pseudocodes

B.1 Generic EM-MD and β computation pseudocodes

Algorithm 4.3: Generic EM algorithm

Input : $\varepsilon > 0$, K, t^{max}, dataset $\mathbf{X} = [\mathbf{X}^c, \mathbf{X}^D]$ with $\mathbf{X}_i \in \mathbb{R}^g \times \mathcal{X}$
Initialization: Compute $\tau_{ik}^0 \leftarrow$ `K-Means(`$K, \mathbf{X}^c$`,`*maxiter=1*`)`
Compute π_k^0 with (4.5)
Compute $\boldsymbol{\mu}_k^0$
Compute $\boldsymbol{\Sigma}_k^0$
Compute other continuous parameters
Compute $p_k^{d,0}$ $\forall$ $d \in [\![1, \ldots, D]\!]$ with (4.37) , (4.38) , (4.39)
$t \leftarrow 1$
while $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ *and* $t < t^{max}$ **do** `/* Aitken's convergence */`
  E-Step
  Compute τ_{ik}^t with (4.32)
  Compute other latent variables
  M-Step
  Compute π_k^t with (4.5)
  Compute $\boldsymbol{\mu}_k^t$
  Compute $\boldsymbol{\Sigma}_k^t$
  Compute other continuous parameters
  Compute discrete probabilities $p_k^{d,t}$ with (4.37) , (4.38) , (4.39)
  $t \leftarrow t + 1$
end

Algorithm 4.4: Computation of parameter β

Input : $\boldsymbol{\pi}^{\text{EM}}$,$\boldsymbol{\pi}^{(new)}$,$\boldsymbol{\pi}^{(old)}$, K, n
$\pi_{(1)}^{\text{EM}} \leftarrow \underset{1 \leq k \leq K}{max}\, \pi_k^{\text{EM}}$, $\pi_{(1)}^{(old)} \leftarrow \underset{1 \leq k \leq K}{max}\, \pi_k^{(old)}$
$E \leftarrow \sum_{k=1}^K \pi_k^{(old)} \log(\pi_k^{(old)})$
$\beta \leftarrow \min \left\{ \frac{\sum_{k=1}^K \exp\left(-\eta n \left|\pi_k^{(new)} - \pi_k^{(old)}\right|\right)}{K}, \frac{\left(1-\pi_{(1)}^{EM}\right)}{\left(-\pi_{(1)}^{(old)} E\right)} \right\}$
Output: β

B.2 DEM-MD pseudocodes

Algorithm 4.5: DEM-MD for Gaussian Mixtures

Input: $\varepsilon > 0$, γ, dataset $\mathbf{X} = [\mathbf{X}^c, \mathbf{X}^D]$ with $\mathbf{X}_i \in \mathbb{R}^g \times \mathcal{X}$
Initialization : $K^0 \leftarrow n$, $\beta^0 \leftarrow 1$
$\pi_k^0 \leftarrow 1/n \ \forall \ k$, $\boldsymbol{\mu}^0 \leftarrow \mathbf{X}^c$
$\boldsymbol{\Sigma}_k^0 \leftarrow D_{k(\lceil\sqrt{K^{\text{initial}}}\rceil)}\mathbf{I}_g$ with
$D_k = \text{sort}\left\{d_{ki}^2 = \|x_i - \mu_k\|_2^2 : d_{ki}^2 > 0, \quad i \neq k, \quad 1 \leq i \leq n\right\}$
Initialize $p_k^{d,0} \ \forall \ d \in [\![1, \ldots, D]\!]$ with (4.36)
$t \leftarrow 1$
Compute τ_{ik}^t with (4.32)
while $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  M-Step
  Compute π_k^t with (4.33)
  Compute $\boldsymbol{\mu}_k^t$ with (4.6)
  $\beta^t \leftarrow$ Algorithm 4.4
  case *delete classes with* $\pi_k^t < 1/n$ **do**
    update class number K^t, adjust π_k^t and τ_{ik}^t
    $t_{component} \leftarrow 1$ `/* variable to keep in memory the number of iterations with a stable number of components */`
  otherwise do
    $K^t \leftarrow K^{t-1}$
  end
  if $t \geq t_{min}$ *and* $t_{component} \geq 100$ **then**
    if *no superimposed clusters* **then**
      $\beta^t = 0$
    else if *superimposed clusters and* $t_{component} < 200$ **then** `/* give more time to the algorithm to converge */`
      $t_{min} \leftarrow t_{min} + 50$
    else merge superimposed clusters
      adjust $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$, $\boldsymbol{\Sigma}^t$ and $\boldsymbol{\tau}^t$
    end
  end
  Compute $\boldsymbol{\Sigma}_k^{\text{EM}}$ with (4.7) and $\boldsymbol{\Sigma}_k^t = (1-\gamma)\boldsymbol{\Sigma}_k^{\text{EM}} + \gamma\mathbf{P}$ with
  $\mathbf{P} = d_{\min}^2\mathbf{I}_g$, $d_{\min}^2 = \min\{d_{ij}^2 : d_{ij}^2 = \|x_i - x_j\|_2^2 > 0, \ 1 \leq i, j \leq n\}$
  Compute discrete probabilities $p_k^{d,t}$ with (4.37) , (4.38) , (4.39)
  E-Step
  Compute τ_{ik}^{t+1} with (4.32)
  $t \leftarrow t + 1$
  $t_{component} \leftarrow t_{component} + 1$
end

Algorithm 4.6: DEM-MD for Student Mixtures

Input: $\varepsilon > 0$, γ, dataset $\mathbf{X} = [\mathbf{X}^c, \mathbf{X}^D]$ with $\mathbf{X}_i \in \mathbb{R}^g \times \mathcal{X}$
Initialization: $K^0 \leftarrow n$, $\beta^0 \leftarrow 1$
$\pi_k^0 \leftarrow 1/n \ \forall \ k$, $\boldsymbol{\mu}^0 \leftarrow \mathbf{X}^c$, $\nu_k^0 \leftarrow 10 \ \forall \ k$
$\boldsymbol{\Sigma}_k^0 \leftarrow D_{k(\lceil\sqrt{K^{\text{initial}}}\rceil)}\mathbf{I}_g$ with
$D_k = \text{sort}\left\{d_{ki}^2 = \|x_i - \mu_k\|_2^2 : d_{ki}^2 > 0, \quad i \neq k, \quad 1 \leq i \leq n\right\}$
initialize $p_k^{d,0} \ \forall \ d \in [\![1, \ldots, D]\!]$ with (4.36)
$t \leftarrow 1$
Compute τ_{ik}^t with (4.32)
Compute $E_{u,ik}^t$ with (4.13)

1 **while** $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  M-Step
  Compute π_k^t with (4.33)
  Compute $\boldsymbol{\mu}_k^t$ with (4.14)
  $\beta^t \leftarrow$ Algorithm 4.4
  case *delete classes with* $\pi_k^t < 1/n$ **do**
    update class number K^t, adjust π_k^t and τ_{ik}^t
    $t_{component} \leftarrow 1$ `/* variable to keep in memory the number of iterations with a stable number of components */`
  otherwise do
    $K^t \leftarrow K^{t-1}$
  end
  if $t \geq t_{min}$ *and* $t_{component} \geq 100$ **then**
2   **if** *no superimposed clusters* **then**
      $\beta^t = 0$
3   **else if** *superimposed clusters and* $t_{component} < 200$ **then** `/* give more time to the algorithm to converge */`
      $t_{min} \leftarrow t_{min} + 50$
4   **else** merge superimposed clusters
      adjust $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$, $\boldsymbol{\Sigma}^{t-1}$, ν^{t-1} and $\boldsymbol{\tau}^t$
    end
  end
  Compute $\boldsymbol{\Sigma}_k^{\text{EM}}$ with (4.15) and $\boldsymbol{\Sigma}_k^t = (1-\gamma)\boldsymbol{\Sigma}_k^{\text{EM}} + \gamma\mathbf{P}$ with
  $\gamma = 0.0001$, $\mathbf{P} = d_{\min}^2\mathbf{I}_g$, $d_{\min}^2 = \min\{d_{ij}^2 : d_{ij}^2 = \|x_i - x_j\|_2^2 > 0, \ 1 \leq i, j \leq n\}$
  Compute ν_k^t by solving (4.16) with Brent's method on the interval $[2, 200]$
  Compute discrete probabilities $p_k^{d,t}$ with (4.37) , (4.38) , (4.39)
  E-Step
  Compute τ_{ik}^{t+1} with (4.32)
  Compute $E_{u,ik}^{t+1}$ with (4.13)
  $t \leftarrow t + 1$
  $t_{component} \leftarrow t_{component} + 1$
end

Algorithm 4.7: DEM-MD for SAL Mixtures

Input: $\varepsilon > 0$, dataset $\mathbf{X} = [\mathbf{X}^c, \mathbf{X}^D]$ with $\mathbf{X}_i \in \mathbb{R}^g \times \mathcal{X}$, a and b for temperature

Initialization: $K^0 \leftarrow n$, $\beta^0 \leftarrow 1$

$\pi_k^0 \leftarrow 1/n \;\forall\; k$, $\boldsymbol{\alpha}_k^0 \leftarrow [0, \ldots, 0]_g \;\forall\; k$

$\boldsymbol{\mu}^0 \leftarrow \mathbf{X}^c + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$

$\boldsymbol{\Sigma}_k^0 \leftarrow D_{k(\lceil\sqrt{K^{\text{initial}}}\rceil)}\mathbf{I}_g$ with

$D_k = \text{sort}\left\{d_{ki}^2 = \|x_i - \mu_k\|_2^2 : d_{ki}^2 > 0, \quad i \neq k, \quad 1 \leq i \leq n\right\}$

Initialize $p_k^{d,0}\ \forall\ d \in [\![1, \ldots, D]\!]$ with (4.36)

$t \leftarrow 1$

Compute τ_{ik}^t with (4.35)

Compute $E_{1i}^{k,t}$ with (4.23)

Compute $E_{2i}^{k,t}$ with (4.24)

1 **while** $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`

CM-Step 1

Compute π_k^t with (4.33)

Compute $\boldsymbol{\mu}_k^t$ with (4.26)

$\beta^t \leftarrow$ Algorithm 4.4

case *delete classes with* $\pi_k^t < 1/n$ **do**

update class number K^t, adjust π_k^t and τ_{ik}^t

$t_{component} \leftarrow 1$ `/* variable to keep in memory the number of iterations with a stable number of components */`

otherwise do

$K^t \leftarrow K^{t-1}$

end

if $t \geq t_{min}$ *and* $t_{component} \geq 100$ **then**

2 **if** *no superimposed clusters* **then**

$\beta^t = 0$

3 **else if** *superimposed clusters and* $t_{component} < 200$ **then** `/* give more time to the algorithm to converge */`

$t_{min} \leftarrow t_{min} + 50$

4 **else** merge superimposed clusters

adjust $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$, $\boldsymbol{\alpha}^{t-1}$, $\boldsymbol{\Sigma}^{t-1}$ and $\boldsymbol{\tau}^t$

end

end

Compute discrete probabilities $p_k^{d,t}$ with (4.37) , (4.38) , (4.39)

Check $\boldsymbol{\mu}^t = \boldsymbol{x}_i^c$ and compute α_k^t with Algorithm 4.1

Intermediate E-step

Compute $\tilde{\tau}_{ik}^t$, $\tilde{E}_{1i}^{k,t}$ and $\tilde{E}_{2i}^{k,t}$ with respectively (4.35), (4.23) and (4.24)

CM-Step 2

Compute $\boldsymbol{\Sigma}_k^t$ with (4.27)

E-Step

Compute τ_{ik}^{t+1} with (4.35)

Compute $E_{1i}^{k,t+1}$ with (4.23)

Compute $E_{2i}^{k,t+1}$ with (4.24)

$t \leftarrow t + 1$

$t_{component} \leftarrow t_{component} + 1$

end

Conclusion

In this thesis, we have proposed several contributions to improve and diversify the estimation of mixture models, particularly on mixed-type data. We have also proposed a simple pipeline that can integrate these mixture models for temporal monitoring, and this is particularly interesting for spatial datasets. The objective with this pipeline and the associated estimation components was to rely on model-based approaches to propose a spatio-temporal monitoring tool based on flexible, configurable mixtures. This part concludes the thesis and provides possible research prospects that can be expected as a follow-up to this thesis.

In Chapter 3 we have proposed a complete and generic pipeline for modeling the evolution of the distribution of a population, and detecting abnormal changes in this distribution. This pipeline, named spatio-temporal mixture process (STMP), relies on a simple evaluation of the likelihood ratio between two models. At time t these two models are estimated on the same data but with different algorithmic constraints, making it possible to build a temporal link throughout the process. The spatio-temporal mixture process for monitoring population distributions and the algorithms to estimate the models are two independent objects. As first components to estimate the mixtures, we proposed two EM algorithm variants to estimate Gaussian mixture models.

The first variant, a Modified Robust EM, was robust to initialization, space boundaries and overfitting. More importantly, it has made it possible to estimate the number of components in a mixture model without external selection or *a posteriori*, by using the entropy of the proportions as a penalization during execution.

The second component was a limited EM algorithm, implemented in such a way as to constrain the evolution of parameters in a neighborhood of the initial parameters. The parameters were constrained in their evolution by relative limitation of their values, easy projections, or by *a posteriori* cosine dissimilarity computation for covariances. We are therefore aware that improvements are possible by changing the evolution constraint strategies in this EM algorithm. Constrained evolution is a challenge that could also be addressed from the whole pipeline perspective, by changing the way the past model (from previous time step) is considered during an actual time step and especially within the evolution test with the likelihood ratio.

This ties in with another limitation, which is that the decision rule has been established empirically, albeit by experiments leading to the computation of theoretical likelihoods. Modeling this decision rule as an acceptation probability, similarly to Monte Carlo methods, would be a good improvement. This could lead to a different way of treating new data at time t, by still keeping mixture models to represent data, and combine EM algorithms and MCMC methods for estimation. For now, as the STMP and mixture estimation algorithms are independent, this enables future directions to consider different mixtures such as the ones developed in our second work.

In Chapter 4, we have proposed mixture models for mixed-type data with different possi-

ble continuous and discrete variables, following a well-established principle of local independence, ensuring the model identifiability. We proposed Dynamic EM for Mixed-type Data (DEM-MD) algorithms for these models, allowing to jointly estimate, for each model, the number of classes and the various parameters of the continuous and discrete distributions. Associated with Aitken's acceleration for testing algorithm convergence, and deterministic annealing for Shifted Asymmetric Laplace distribution, the improved algorithms performed correctly on the three considered distributions.

A clear extension would be to consider and implement DEM-MD algorithms for other continuous and discrete laws, such as the Generalized Hyperbolic distribution which has in fact several continuous distributions as limiting cases, such as the ones considered here. Considering mixture of Generalized Hyperbolic distributions has already been proposed by Browne and McNicholas (2015).

Another debatable point of our mixture models for mixed-type date is the local independence assumption. We saw in the introduction that other families of methods can be used to establish links for all variables, but generally involve either the transformation of certain variables or statistical conditioning, as factor analyzers or copulas. An extension could be to associate a mixture of copulas for mixed-type data estimated by EM-like algorithms (Zhao and Udell, 2020; Rajan and Bhattacharya, 2016) with a dynamic estimation of the number of components as in our work.

Although we have introduced regularization on the covariance matrices in situations with high complexity and low sample size, prior regularization could be built into any DEM-MD on a permanent basis. By this we mean that this regularization could be incorporated into the generic version of the algorithm, instead of adding a small amount of noise as usually done in a lot of EM algorithm implementations. This integration should take into account the dimension of the data in order to adapt the regularization to each situation and not obstruct the estimation of the parameters.

In the context of high-dimensional data, other possibilities would be to consider parsimonious models and subspace clustering methods such as factor analyzers (Bouveyron and Brunet-Saumard, 2014; Franczak et al., 2013).

Bibliography

J. J. Abellan, S. Richardson, and N. Best. Use of Space–Time Models to Investigate the Stability of Patterns of Disease. *Environ Health Perspect*, 116(8):1111–1119, Aug. 2008. doi: 10.1289/ehp.10814. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.10814.

A. Ahmad and S. S. Khan. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access*, 7:31883–31902, 2019. doi: 10.1109/ACCESS.2019.2903568. URL https://ieeexplore.ieee.org/document/8662561.

A. C. Aitken. On Bernoulli's Numerical Solution of Algebraic Equations. *Proc. R. Soc. Edinb.*, 46:289–305, 1927. doi: 10.1017/S0370164600022070. URL https://www.cambridge.org/core/product/identifier/S0370164600022070/type/journal_article.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, 1973. URL https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

S. Allassonnière and J. Chevallier. A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling. *Computational Statistics & Data Analysis*, 159:17, July 2021. doi: 10.1016/j.csda.2020.107159. URL https://www.sciencedirect.com/science/article/abs/pii/S0167947320302504.

S. Allassonnière and E. Kuhn. Stochastic algorithm for Bayesian mixture effect template estimation. *ESAIM: PS*, 14:382–408, 2010. doi: 10.1051/ps/2009001. URL http://www.esaim-ps.org/10.1051/ps/2009001.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, Dec. 2009. URL https://doi.org/10.1214/09-AOS689.

J. L. Andrews, P. D. McNicholas, and S. Subedi. Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics & Data Analysis*, 55(1):520–529, Jan. 2011. doi: 10.1016/j.csda.2010.05.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947310002203.

R. B. Arellano-Valle and A. Azzalini. On the Unification of Families of Skew-normal Distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006. doi: 10.1111/j.1467-9469.2006.00503.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2006.00503.x.

A. Azzalini. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985. URL https://www.jstor.org/stable/4615982.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003. doi: 10.1111/1467-9868.00391. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00391.

A. Azzalini and A. D. Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4): 715–726, Dec. 1996. doi: 10.1093/biomet/83.4.715. URL https://doi.org/10.1093/biomet/83.4.715.

X. Bai, K. Chen, and W. Yao. Mixture of linear mixed models using multivariate t distribution. *Journal of Statistical Computation and Simulation*, 86(4):771–787, Mar. 2016. doi: 10.1080/00949655.2015.1036431. URL https://escholarship.org/uc/item/2cz925kv.

A. Barhen and J. Daudin. Generalization of the Mahalanobis Distance in the Mixed Case. *Journal of Multivariate Analysis*, 53(2):332–342, May 1995. doi: 10.1006/jmva.1995.1040. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X85710408.

J.-P. Baudry and G. Celeux. EM for mixtures: Initialization requires special care. *Stat Comput*, 25(4):713–726, July 2015. doi: 10.1007/s11222-015-9561-x. URL http://link.springer.com/10.1007/s11222-015-9561-x.

J.-P. Baudry, C. Maugis, and B. Michel. Slope Heuristics: Overview and Implementation. *Statistics and Computing*, 22(2):455–470, Mar. 2012. doi: 10.1007/s11222-011-9236-1. URL https://link.springer.com/article/10.1007/s11222-011-9236-1.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. *Institut national de recherche en informatique et en automatique*, 37(1):55–57, Dec. 1998. doi: 10.1177/075910639203700105. URL http://journals.sagepub.com/doi/10.1177/075910639203700105.

C. Biernacki, G. Celeux, and G. Govaert. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272, Mar. 1999. doi: 10.1016/S0167-8655(98)00144-5. URL https://www.sciencedirect.com/science/article/pii/S0167865598001445.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, Jan. 2003. doi: 10.1016/S0167-9473(02)00163-9. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947302001639.

L. Birgé and P. Massart. Minimal Penalties for Gaussian Model Selection. *Probab. Theory Relat. Fields*, 138(1-2):33–73, Feb. 2007. doi: 10.1007/s00440-006-0011-8. URL http://link.springer.com/10.1007/s00440-006-0011-8.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann Inst Stat Math*, 46(2):373–388, June 1994. doi: 10.1007/BF01720593. URL http://link.springer.com/10.1007/BF01720593.

C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, Mar. 2014. doi: 10.1016/j.csda.2012.12.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947312004422.

M. Brand. Structure Learning in Conditional Probability Models via an Entropic Prior and Parameter Extinction. *Neural Computation*, 11(5):1155–1182, July 1999. doi: 10.1162/089976699300016395. URL https://direct.mit.edu/neco/article/11/5/1155-1182/6274.

R. P. Brent. *Algorithms for Minimization Without Derivatives*. Dover Books on Mathematics Series. Dover Publications, June 2013. ISBN 978-0-486-14368-2.

R. P. Browne and P. D. McNicholas. Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, 142(11):2976–2984, Nov. 2012. doi: 10.1016/j.jspi.2012.05.001. URL https://www.sciencedirect.com/science/article/pii/S0378375812001838.

R. P. Browne and P. D. McNicholas. A mixture of generalized hyperbolic distributions. *Can. J. Statistics*, 43(2):176–198, June 2015. doi: 10.1002/cjs.11246. URL https://onlinelibrary.wiley.com/doi/10.1002/cjs.11246.

C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007. doi: 10.1016/j.compenvurbsys.2005.07.009. URL https://www.sciencedirect.com/science/article/pii/S0198971506000354.

D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information. *Bull Cancer*, 67(4):477–490, 1980.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, Sept. 1996. doi: 10.1007/BF01246098. URL http://link.springer.com/10.1007/BF01246098.

G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A Component-Wise EM Algorithm for Mixtures. *Journal of Computational and Graphical Statistics*, 10(4):697–712, Dec. 2001. doi: 10.1198/106186001317243403. URL https://www.tandfonline.com/doi/abs/10.1198/106186001317243403.

G. Celeux, O. C. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5(3):243–267, Oct. 2005. doi: 10.1191/1471082x05st096oa. URL https://journals.sagepub.com/doi/10.1191/1471082X05st096oa.

A. S. Cheam, M. Fredette, M. Marbac, and F. Navarro. Translation-invariant functional clustering on COVID-19 deaths adjusted on population risk factors. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2):387–413, May 2023. doi: 10.1093/jrsssc/qlad014. URL http://arxiv.org/abs/2012.10629.

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

J. Chen and X. Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383, Aug. 2009. doi: 10.1016/j.jmva.2008.12.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X08002728.

G. Ciuperca, A. Ridolfi, and J. Idier. Penalized Maximum Likelihood Estimator for Normal Mixtures. *Scand J Stat*, 30(1):45–59, Mar. 2003. doi: 10.1111/1467-9469.00317. URL https://onlinelibrary.wiley.com/doi/10.1111/1467-9469.00317.

L. Clarotto, D. Allard, T. Romary, and N. Desassis. The SPDE approach for spatio-temporal datasets with advection and diffusion, Jan. 2023. URL http://arxiv.org/abs/2208.14015.

A. De Leon and K. Carrière. A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, 92(1):174–185, Jan. 2005. doi: 10.1016/j.jmva.2003.08.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X03001507.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, 27(1):94–128, 1999. URL https://www.jstor.org/stable/120120.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x.

E. Derman and E. L. Pennec. Clustering and Model Selection via Penalized Likelihood for Different-sized Categorical Data Vectors, 2017. URL https://arxiv.org/abs/1709.02294.

P. Elliott and D. Wartenberg. Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*, 112(9):998–1006, June 2004. doi: 10.1289/ehp.6735. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.6735.

M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, June 1995. doi: 10.1080/01621459.1995.10476550. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550.

Y. Fang, B. C. Franczak, and S. Subedi. Tackling the infinite likelihood problem when fitting mixtures of shifted asymmetric Laplace distributions, Mar. 2023. URL http://arxiv.org/abs/2303.14211.

M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002. doi: 10.1109/34.990138. URL https://ieeexplore.ieee.org/document/990138.

M. Fop, T. B. Murphy, and L. Scrucca. Model-based Clustering with Sparse Covariance Matrices. *Statistics and Computing*, 29(4):791–819, 2019. doi: 10.1007/s11222-018-9838-y. URL https://doi.org/10.1007/s11222-018-9838-y.

F. Forbes, M. Charras-Garrido, L. Azizi, S. Doyle, and D. Abrial. Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Ann. Appl. Stat.*, 7(2), June 2013. doi: 10.1214/13-AOAS629. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-7/issue-2/Spatial-risk-mapping-for-rare-disease-with-hidden-Markov-fields/10.1214/13-AOAS629.full.

A. Foss, M. Markatou, B. Ray, and A. Heching. A semiparametric method for clustering mixed data. *Mach Learn*, 105(3):419–458, Dec. 2016. doi: 10.1007/s10994-016-5575-7. URL http://link.springer.com/10.1007/s10994-016-5575-7.

A. H. Foss and M. Markatou. Kamila: Clustering Mixed-Type Data in R and Hadoop. *Journal of Statistical Software*, 83:1–44, Feb. 2018. doi: 10.18637/jss.v083.i13. URL https://doi.org/10.18637/jss.v083.i13.

A. H. Foss, M. Markatou, and B. Ray. Distance Metrics and Clustering Methods for Mixed-type Data. *International Statistical Review*, 87(1):80–109, Apr. 2019. doi: 10.1111/insr.12274. URL https://onlinelibrary.wiley.com/doi/10.1111/insr.12274.

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. URL https://www.jstor.org/stable/3085676.

C. Fraley and A. E. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 24(2):155–181, 2007. doi: 10.1007/s00357-007-0004-5. URL https://link.springer.com/article/10.1007/s00357-007-0004-5.

S. P. France. Point épidémiologique hebdomadaire du 25 juin 2020. Technical report, Santé Publique France, June 2020. URL https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-et-infections-respiratoires/infection-a-coronavirus/documents/bulletin-national/covid-19-point-epidemiologique-du-25-juin-2020.

B. C. Franczak, P. D. McNicholas, R. P. Browne, and P. M. Murray. Parsimonious Shifted Asymmetric Laplace Mixtures, Nov. 2013. URL http://arxiv.org/abs/1311.0317.

B. C. Franczak, R. P. Browne, and P. D. McNicholas. Mixtures of Shifted Asymmetric Laplace Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1149–1157, June 2014. doi: 10.1109/TPAMI.2013.216. URL https://ieeexplore.ieee.org/document/6654117.

B. C. Franczak, C. Tortora, R. P. Browne, and P. D. McNicholas. Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters*, 58:69–76, June 2015. doi: 10.1016/j.patrec.2015.02.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0167865515000598.

B. C. Franczak, R. P. Browne, P. D. McNicholas, and K. L. Burak. MixSAL: Mixtures of Multivariate Shifted Asymmetric Laplace (SAL) Distributions, May 2018. URL https://cran.r-project.org/web/packages/MixSAL/index.html.

C. Frévent, M.-S. Ahmed, S. Dabo-Niang, and M. Genin. Investigating spatial scan statistics for multivariate functional data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2):450–475, May 2023. doi: 10.1093/jrsssc/qlad017. URL https://doi.org/10.1093/jrsssc/qlad017.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-32909-3 978-0-387-35768-3.

Z. Ghahramani and G. E. Hinton. The EM Algorithm for Mixtures of Factor Analyzers. Technical Report CRG-TR-96-1, University of Toronto, Toronto, Canada, May 1996. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=766f4465747394d304d162197e091f1ae8f7f577.

L. A. Goodman. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61(2):215–231, 1974. doi: 10.2307/2334349. URL https://www.jstor.org/stable/2334349.

P. J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, Dec. 1995. doi: 10.2307/2337340. URL https://www.jstor.org/stable/2337340.

P. J. Green and S. Richardson. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, Dec. 2002. doi: 10.1198/016214502388618870. URL http://www.tandfonline.com/doi/abs/10.1198/016214502388618870.

M. Herdin, N. Czink, H. Ozcelik, and E. Bonek. Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels. In *2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 136–140, Stockholm, Sweden, May 2005. IEEE. ISBN 0-7803-8887-9. doi: 10.1109/VETECS.2005.1543265. URL https://ieeexplore.ieee.org/document/1543265.

T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(1):1–14, Jan./1997. doi: 10.1109/34.566806. URL http://ieeexplore.ieee.org/document/566806/.

H. Holzmann, A. Munk, and T. Gneiting. Identifiability of Finite Mixtures of Elliptical Distributions. *Scand J Stat*, 33(4):753–763, Dec. 2006. doi: 10.1111/j.1467-9469.2006.00505.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2006.00505.x.

Z. Huang. Clustering Large Data Sets With Mixed Numeric And Categorical Values. In *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*, pages 21–34, Singapore, 1997a. World Scientific. ISBN 978-981-02-3072-2. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d42bb5ad2d03be6d8fefa63d25d02c0711d19728.

Z. Huang. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, The University of British Columbia, 1997b.

Z. Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998. doi: 10.1023/A:1009769707641. URL http://link.springer.com/10.1023/A:1009769707641.

L. A. Hunt and M. A. Jorgensen. Mixture Model Clustering of Data Sets with Categorical and Continuous Variables. In *Information, Statistics and Induction in Science*, volume 96, pages 375–284, July 1996. ISBN 978-981-4547-26-0. doi: 10.1142/9789814530637.

G. M. Jacquez. A k Nearest Neighbour Test for Space-Time Interaction. *Statistics in Medicine*, 15(18):1935–1949, 1996. doi: 10.1002/(SICI)1097-0258(19960930)15:18<1935::AID-SIM406>3.0.CO;2-I. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819960930%2915%3A18%3C1935%3A%3AAID-SIM406%3E3.0.CO%3B2-I.

M. C. Jones and M. J. Faddy. A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):159–174, 2003. doi: 10.1111/1467-9868.00378. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00378.

Z. I. Kalantan and J. Einbeck. Quantile-Based Estimation of the Finite Cauchy Mixture Model. *Symmetry*, 11(9):1186, Sept. 2019. doi: 10.3390/sym11091186. URL https://www.mdpi.com/2073-8994/11/9/1186.

D. Karlis and A. Santourian. Model-based clustering with non-elliptically contoured distributions. *Stat Comput*, 19(1):73–83, Mar. 2009. doi: 10.1007/s11222-008-9072-0. URL http://link.springer.com/10.1007/s11222-008-9072-0.

L. Kaufman and P. J. Rousseeuw, editors. *Finding Groups in Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, Mar. 1990. ISBN 978-0-470-31680-1 978-0-471-87876-6. doi: 10.1002/9780470316801. URL http://doi.wiley.com/10.1002/9780470316801.

J. T. Kent, D. E. Tyler, and Yahuda. Vard. A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics - Simulation and Computation*, 23(2):441–453, Jan. 1994. doi: 10.1080/03610919408813180. URL http://www.tandfonline.com/doi/abs/10.1080/03610919408813180.

R. S. Kirby, E. Delmelle, and J. M. Eberth. Advances in spatial epidemiology and geographic information systems. *Annals of Epidemiology*, 27(1):1–9, Jan. 2017. doi: 10.1016/j.annepidem.2016.12.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1047279716304951.

K. Kleinman. A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism. *American Journal of Epidemiology*, 159(3):217–224, Feb. 2004. doi: 10.1093/aje/kwh029. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwh029.

E. G. Knox and M. S. Bartlett. The Detection of Space-Time Interactions. *Applied Statistics*, 13(1):25, 1964. doi: 10.2307/2985220. URL https://www.jstor.org/stable/2985220?origin=crossref.

A. B. Koehler and E. S. Murphree. A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order. *Applied Statistics*, 37(2):187, 1988. doi: 10.2307/2347338. URL https://www.jstor.org/stable/10.2307/2347338?origin=crossref.

S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer, New York, 2008. ISBN 978-0-387-71886-6.

I. Kosmidis and D. Karlis. Model-based clustering using copulas with applications. *Stat Comput*, 26(5):1079–1099, Sept. 2016. doi: 10.1007/s11222-015-9590-5. URL http://arxiv.org/abs/1404.4077.

S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace Distribution and Generalizations*. Birkhäuser Boston, Boston, MA, 2001. ISBN 978-1-4612-6646-4 978-1-4612-0173-1. doi: 10.1007/978-1-4612-0173-1. URL http://link.springer.com/10.1007/978-1-4612-0173-1.

W. J. Krzanowski. The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1):25–49, Jan. 1993. doi: 10.1007/BF02638452. URL http://link.springer.com/10.1007/BF02638452.

E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: PS*, 8:115–131, Aug. 2004. doi: 10.1051/ps:2004007. URL http://www.esaim-ps.org/10.1051/ps:2004007.

M. Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, Jan. 1997. doi: 10.1080/03610929708831995. URL http://www.tandfonline.com/doi/abs/10.1080/03610929708831995.

M. Kulldorff and N. Nagarwalla. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995. doi: 10.1002/sim.4780140809. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140809.

M. Kulldorff, W. F. Athas, E. J. Feurer, B. A. Miller, and C. R. Key. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Jealth*, 88(9):1377–1380, 1998. doi: 10.2105/ajph.88.9.1377. URL https://doi.org/10.2105/AJPH.88.9.1377.

M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Med*, 2(3):e59, Feb. 2005. doi: 10.1371/journal.pmed.0020059. URL https://dx.plos.org/10.1371/journal.pmed.0020059.

K. Lange. A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995. URL http://www.jstor.org/stable/2345971.

K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989. doi: 10.2307/2290063. URL https://www.jstor.org/stable/2290063.

T. Lartigue, S. Durrleman, and S. Allassonnière. Deterministic Approximate EM Algorithm; Application to the Riemann Approximation EM and the Tempered EM. *Algorithms*, 15 (3):78, Feb. 2022. doi: 10.3390/a15030078. URL https://www.mdpi.com/1999-4893/15/3/78.

M. H. Law, M. A. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166, 2004. doi: 10.1109/TPAMI.2004.71. URL https://ieeexplore.ieee.org/abstract/document/1316850.

K. J. Lee and R.-B. Chen. Bayesian variable selection in a finite mixture of linear mixed-effects models. *Journal of Statistical Computation and Simulation*, 89(13):2434–2453, May 2019. doi: 10.1080/00949655.2019.1620746. URL https://doi.org/10.1080/00949655.2019.1620746.

S. Lee and G. J. McLachlan. Finite mixtures of multivariate skew t-distributions: Some recent and new results. *Stat Comput*, 24(2):181–202, Mar. 2014. doi: 10.1007/s11222-012-9362-4. URL http://link.springer.com/10.1007/s11222-012-9362-4.

S. X. Lee and G. J. McLachlan. On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm, Sept. 2012. URL http://arxiv.org/abs/1109.4706.

T. I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257–265, Feb. 2009. doi: 10.1016/j.jmva.2008.04.010. URL https://www.sciencedirect.com/science/article/pii/S0047259X08001152.

T.-I. Lin. Robust mixture modeling using multivariate skew t distributions. *Stat Comput*, 20 (3):343–356, July 2010. doi: 10.1007/s11222-009-9128-9. URL http://link.springer.com/10.1007/s11222-009-9128-9.

T. I. Lin, J. C. Lee, and S. Y. Yen. Finite Mixture Modelling Using the Skew Normal Distribution. *Statistica Sinica*, 17(3):909–927, 2007. URL https://www.jstor.org/stable/24307705.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: 10.1111/j.1467-9868.2011.00777.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x.

B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Assocation, 1995. ISBN 0-940600-32-3. URL http://www.jstor.org/stable/4153184.

C. Liu and D. B. Rubin. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648, 1994. doi: 10.1093/biomet/81.4.633. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/81.4.633.

C. Liu and D. B. Rubin. ML Estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):21, 1995. URL http://www.jstor.org/stable/24305551.

F. Louzada, D. C. Nascimento, and O. A. Egbon. Spatial Statistical Models: An overview under the Bayesian Approach. *arXiv:2009.14371 [stat]*, Sept. 2020. URL http://arxiv.org/abs/2009.14371.

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

C. Mavridis and J. Baras. Online Deterministic Annealing for Classification and Clustering. *IEEE Trans. Neural Netw. Learning Syst.*, pages 1–10, 2022. doi: 10.1109/TNNLS.2021.3138676. URL http://arxiv.org/abs/2102.05836.

F. Mbuga and C. Tortora. Spectral Clustering of Mixed-Type Data. *Stats*, 5(1):1–11, Dec. 2021. doi: 10.3390/stats5010001. URL https://www.mdpi.com/2571-905X/5/1/1.

B. McCane and M. Albert. Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, 29(7):986–993, May 2008. doi: 10.1016/j.patrec.2008.01.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0167865508000524.

G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4):379–388, Jan. 2003. doi: 10.1016/S0167-9473(02)00183-4. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947302001834.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, Hoboken, second edition, 2008. ISBN 978-0-471-20170-0.

G. J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 658–666, Berlin, Heidelberg, 1998. Springer. ISBN 978-3-540-68526-5. doi: 10.1007/BFb0033290. URL https://doi.org/10.1007/BFb0033290.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics Applied Probability and Statistics Section. Wiley, New York, 2000. ISBN 978-0-471-00626-8.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, Princeton Oxford, 2015. ISBN 978-0-691-16627-8.

P. McNicholas, T. Murphy, A. McDaid, and D. Frost. Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis*, 54(3):711–723, Mar. 2010. doi: 10.1016/j.csda.2009.02.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947309000632.

D. McParland and I. C. Gormley. Model based clustering for mixed data: clustMD. *Adv Data Anal Classif*, 10(2):155–169, June 2016. doi: 10.1007/s11634-016-0238-x. URL http://link.springer.com/10.1007/s11634-016-0238-x.

D. McParland, I. C. Gormley, T. H. McCormick, S. J. Clark, C. W. Kabudula, and M. A. Collinson. Clustering South African Households Based on Their Asset Status Using Latent Variable Models. *The Annals of Applied Statistics*, 8(2):747–776, 2014. URL https://www.jstor.org/stable/24522075.

D. McParland, C. M. Phillips, L. Brennan, H. M. Roche, and I. C. Gormley. Clustering high-dimensional mixed data to uncover sub-phenotypes: Joint analysis of phenotypic and genotypic data. *Statist. Med.*, 36(28):4548–4569, Dec. 2017. doi: 10.1002/sim.7371. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.7371.

P. Meinicke and H. Ritter. Resolution-Based Complexity Control for Gaussian Mixture Models. *Neural Computation*, 13(2):453–475, Feb. 2001. doi: 10.1162/089976601300014600. URL https://direct.mit.edu/neco/article/13/2/453-475/6478.

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993. doi: 10.1093/biomet/80.2.267. URL https://doi.org/10.1093/biomet/80.2.267.

J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas. Bayesian Gaussian Copula Factor Models for Mixed Data. *Journal of the American Statistical Association*, 108(502):656–665, June 2013. doi: 10.1080/01621459.2012.762328. URL https://doi.org/10.1080/01621459.2012.762328.

I. Naim and D. Gildea. Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients. In *Proceedings of the 29th International Conference on Machine Learning*, page 8, Edinburgh, Scotland, 2012. doi: 10.48550/arXiv.1206.6427. URL https://arxiv.org/abs/1206.6427.

T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010. doi: 10.1111/j.1467-9671.2010.01194.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2010.01194.x.

R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249, June 2000. doi: 10.2307/1390653. URL https://www.jstor.org/stable/1390653?origin=crossref.

J. Park, S. Yi, W. Chang, and J. Mateu. A Spatio-Temporal Dirichlet Process Mixture Model for Coronavirus Disease-19, July 2022. URL http://arxiv.org/abs/2207.06587.

K. Pearson. Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. URL https://www.jstor.org/stable/90667.

D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000. doi: 10.1023/A:1008981510081. URL https://link.springer.com/article/10.1023/A:1008981510081.

A. Pervez and D. Lee. A Componentwise Simulated Annealing EM Algorithm for Mixtures. In S. Hölldobler, R. Peñaloza, and S. Rudolph, editors, *KI 2015: Advances in Artificial Intelligence*, volume 9324, pages 287–294, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24488-4 978-3-319-24489-1. doi: 10.1007/978-3-319-24489-1_25. URL http://link.springer.com/10.1007/978-3-319-24489-1_25.

N. Pocuca, R. P. Browne, and P. D. McNicholas. Mixture: Mixture Models for Clustering and Classification, Sept. 2022. URL https://cran.r-project.org/web/packages/mixture/index.html.

C. Proust-Lima and H. Jacqmin-Gadda. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput Methods Programs Biomed*, 78(2):165–173, May 2005. doi: 10.1016/j.cmpb.2004.12.004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1913221/.

S. Pruilh, A.-S. Jannot, and S. Allassonnière. Spatio-temporal mixture process estimation to detect dynamical changes in population. *Artificial Intelligence in Medicine*, 126:102258, Apr. 2022. doi: 10.1016/j.artmed.2022.102258. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365722000239.

V. Rajan and S. Bhattacharya. Dependency Clustering of Mixed Data with Gaussian Mixture Copulas. In *International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1967–1973, Palo Alto, California, July 2016. AAAI Press. ISBN 978-1-57735-770-4. URL https://www.ijcai.org/Proceedings/16/Papers/281.pdf.

Sylvia. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997. doi: 10.1111/1467-9868.00095. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00095.

C. Robertson, T. A. Nelson, Y. C. MacNab, and A. B. Lawson. Review of methods for space–time disease surveillance. *Spatial and Spatio-temporal Epidemiology*, 1(2-3):105–116, July 2010. doi: 10.1016/j.sste.2009.12.001. URL https://linkinghub.elsevier.com/retrieve/pii/S187758451000002X.

P. A. Rogerson. Surveillance systems for monitoring the development of spatial patterns. *Stat Med*, 16(18):2081–2093, Sept. 1997. doi: 10.1002/(sici)1097-0258(19970930)16:18<2081::aid-sim638>3.0.co;2-w. URL https://pubmed.ncbi.nlm.nih.gov/9308133/.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, Nov. 1998. doi: 10.1109/5.726788. URL https://ieeexplore.ieee.org/document/726788.

D. L. Sackett. Bias in analytic research. *Journal of Chronic Diseases*, 32(1-2):51–63, Jan. 1979. doi: 10.1016/0021-9681(79)90012-2. URL https://linkinghub.elsevier.com/retrieve/pii/0021968179900122.

O. Sahin and C. Czado. Vine copula mixture models and clustering for non-Gaussian data. *Econometrics and Statistics*, 22:136–158, Apr. 2022. doi: 10.1016/j.ecosta.2021.08.011. URL https://www.sciencedirect.com/science/article/pii/S2452306221001052.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136. URL https://www.jstor.org/stable/2958889.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1986. ISBN 978-0-412-24620-3.

M. S. Smith and M. A. Khaled. Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association*, 107(497):290–303, Mar. 2012. doi: 10.1080/01621459.2011.644501. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2011.644501.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, 28(1), Feb. 2000. doi: 10.1214/aos/1016120364. URL https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-1/Bayesian-analysis-of-mixture-models-with-an-unknown-number-of/10.1214/aos/1016120364.full.

H. Teicher. Identifiability of Finite Mixtures. *Ann. Math. Statist*, 34(4):1265–1269, 1963. doi: 10.1214/aoms/1177703862. URL https://doi.org/10.1214/aoms/1177703862.

R. D. Telford and R. B. Cunningham. Sex, sport, and body-size dependency of hematology in highly trained athletes. *Med Sci Sports Exerc*, 23(7):788–794, July 1991. URL https://journals.lww.com/acsm-msse/Abstract/1991/07000/Sex,_sport,_and_body_size_dependency_of_hematology.4.aspx.

N. Ueda and R. Nakano. Mixture density estimation via EM algorithm with deterministic annealing. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pages 69–77, Ermioni, Greece, Sept. 1994. IEEE. ISBN 0-7803-2026-3. doi: 10.1109/NNSP.1994.366062. URL https://ieeexplore.ieee.org/abstract/document/366062.

N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11 (2):271–282, Mar. 1998. doi: 10.1016/s0893-6080(97)00133-0. URL https://www.sciencedirect.com/science/article/abs/pii/S0893608097001330?via%3Dihub.

I. Vrbik and P. McNicholas. Analytic calculations for the EM algorithm for multivariate skew- mixture models. *Statistics & Probability Letters*, 82(6):1169–1174, June 2012. doi: 10.1016/j.spl.2012.02.020. URL https://linkinghub.elsevier.com/retrieve/pii/S0167715212000673.

C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999. doi: 10.1093/comjnl/42.4.270. URL https://ieeexplore.ieee.org/document/8138704.

C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(3):240–252, 1987. doi: 10.1111/j.2517-6161.1987.tb01695.x. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1987.tb01695.x.

B. Wang, F. Wan, P. U. Mak, P. I. Mak, and M. I. Vai. Entropy penalized learning for Gaussian mixture models. In *The 2011 International Joint Conference on Neural Networks*, pages 2067–2073, San Jose, CA, USA, July 2011. IEEE. ISBN 978-1-4244-9635-8. doi: 10.1109/IJCNN.2011.6033481. URL http://ieeexplore.ieee.org/document/6033481/.

H. Wang, B. Luo, Q. bing Zhang, and S. Wei. Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. *Pattern Recognition Letters*, 25(16):1799–1809, 2004.

A. Willse and R. J. Boik. Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9:111–121, 1999. doi: 10.1023/A:1008842432747. URL https://link.springer.com/article/10.1023/A:1008842432747.

C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *Ann. Statist.*, 11(1), Mar. 1983. doi: 10.1214/aos/1176346060. URL https://projecteuclid.org/journals/annals-of-statistics/volume-11/issue-1/On-the-Convergence-Properties-of-the-EM-Algorithm/10.1214/aos/1176346060.full.

S. J. Yakowitz and J. D. Spragins. On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968. URL https://www.jstor.org/stable/2238925.

P. Yan and M. K. Clayton. A cluster model for space–time disease counts. *Statist. Med.*, 25 (5):867–881, Mar. 2006. doi: 10.1002/sim.2424. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.2424.

M.-S. Yang, C.-Y. Lai, and C.-Y. Lin. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, Nov. 2012. doi: 10.1016/j.patcog.2012.04.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320312002117.

B. Zhang, C. Zhang, and X. Yi. Competitive EM algorithm for finite mixture models. *Pattern recognition*, 37(1):131–144, 2004. doi: 10.1016/S0031-3203(03)00140-7. URL https://www.sciencedirect.com/science/article/abs/pii/S0031320303001407.

Y. Zhao and M. Udell. Missing Value Imputation for Mixed Data via Gaussian Copula. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 636–646. ACM, Aug. 2020. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403106. URL https://dl.acm.org/doi/10.1145/3394486.3403106.

H. Zhou and K. L. Lange. On the Bumpy Road to the Dominant Mode: On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics*, 37(4):612–631, Dec. 2010. doi: 10.1111/j.1467-9469.2009.00681.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2009.00681.x.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Modèles de mélange dynamiques et suivi longitudinal pour l'inférence de données mixtes et spatio-temporelles : application en Santé Publique

Mots clés : Modélisation statistique, Modèles de mélange, Données mixtes, pipeline spatio-temporel, Données de Santé Publique

Résumé : Dans cette thèse, nous nous intéressons à des méthodes d'apprentissage statistique sur données spatio-temporelles et données mixtes. En effet, la croissance rapide des systèmes d'informations en santé publique permet aujourd'hui de disposer de données variées en temps réel pour de nombreuses maladies. L'objectif est de développer des méthodes pour utiliser ces données afin de construire des systèmes d'aide à la décision exploitables.
Nous proposons d'abord un pipeline spatio-temporel pour estimer la distribution de la population et mettre en évidence des différences temporelles. Ce pipeline est une première étape vers un système d'aide à la décision et d'alerte pour l'analyse spatio-temporelle de l'évolution d'une population. Ce pipeline est conçu de manière à ce que différentes distributions et donc différents algorithmes puissent être envisagés. Pour une première application, ce pipeline est combiné avec des algorithmes EM robustes permettant l'estimation de modèles de mélange gaussiens. Il est éprouvé sur des données d'hôpitaux parisiens correspondant aux personnes testées positives à l'infection par le SARS-CoV-2 sur onze semaines en 2020.
Dans une deuxième partie nous proposons un ensemble d'algorithmes pour l'estimation de modèles de mélange sur données mixtes. Nous décrivons d'abord des modèles de mélange pour diverses lois continues et discrètes, en supposant une indépendance conditionnelle entre les variables discrètes et continues. Nous proposons ensuite des algorithmes dynamiques de type EM, permettant l'estimation de tous les paramètres du mélange ainsi que l'estimation du nombre de classes. Nous montrons que nos différents algorithmes dynamiques permettent d'atteindre le nombre réel de classes et d'estimer correctement les paramètres des lois discrètes comme continues. Nous soulignons aussi l'intérêt d'introduire des régularisations sur des paramètres particuliers afin d'améliorer les performances dans des situations où la taille de l'échantillon n'est pas suffisante en regard de la complexité du modèle. Ces algorithmes dynamiques sont ensuite validés sur des données réelles issues de la littérature.

Title: Dynamic mixture models and longitudinal monitoring for mixed-type and spatio-temporal data inference: application in Public Health

Keywords: Statistical Modeling, Mixture Models, Mixed-type data, Spatio-Temporal Pipeline, Public Health Data

Abstract: In this thesis, we focus on statistical learning methods for spatio-temporal and mixed-type data. With the rapid growth of public health information systems, a wide range of real-time data is now available for many diseases. The aim is to develop methods for using this data to build operational decision support systems.
We first propose a spatio-temporal pipeline for estimating the distribution of a population and highlighting temporal differences. This pipeline is a first step towards a decision support and alert system for the spatio-temporal analysis of population trends. This pipeline is designed so that different distributions and therefore different algorithms can be considered. For an initial application, this pipeline is combined with robust EM algorithms for estimating Gaussian mixture models. It is evaluated using data from Paris hospitals corresponding to people who tested positive for SARS-CoV-2 infection over eleven weeks in 2020.
In the second part, we propose a set of algorithms for estimating statistical models on mixed data. We consider that mixed-type data are distributed according to mixtures of laws. We first describe mixture models for various continuous and discrete laws, assuming local independence between discrete and continuous variables. We then propose dynamic algorithms of the EM type, allowing the estimation of all the parameters of the mixture as well as the estimation of the number of classes. We show that our different dynamic algorithms allow us to reach the real number of classes and to correctly estimate the parameters of the discrete and continuous laws. We also highlight the benefits of introducing regularizations to improve performance in situations where the sample size is insufficient for the complexity of the model. These dynamic algorithms are then validated on real data from the literature.

Institut Polytechnique de Paris
91120 Palaiseau, France