Dynamic Expectation-Maximization algorithms for Mixed-type Data

Solange Pruilh, Stéphanie Allassonnière

Dynamic Expectation-Maximization algorithms for Mixed-type Data

Solange Pruilh[1,2] and Stéphanie Allassonnière[2]

[1]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France
[2]HeKA team, PariSantéCampus, Centre de Recherche des Cordeliers, Université Paris Cité, INRIA, INSERM, Sorbonne Université

Abstract

Modelling mixed-type data is still complex because of the heterogeneity of encountered data. With clustering as the objective, many methods are already doing well, but the inference of models and a posteriori exploitation is made difficult if not impossible. In this article we propose 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.

1 Introduction and related works

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. Although there are many clustering methods, grouped under different data exploitation strategies, as we will see below, few involve statistical models, enabling inference and a posteriori reuse. In addition, few assumptions on continuous variables have been proposed. Then, difficulty remains for correctly inferring mixed-type data with an interpretable and easily computable model, and is the subject of this article.

1.1 Challenges on mixed-type data

Mixed-type data contain both numerical and categorical (nominal and/or ordinal) variables. As detailed in recent review papers (Foss et al., 2019; Ahmad and Khan, 2019), especially for a clustering goal, several types of strategies exist on 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 (Smith and Khaled, 2012; Murray et al., 2013), and later mixture of copulas (Kosmidis and Karlis, 2016; Marbac et al., 2017; Sahin and Czado, 2022).

Secondly, a whole range of literature involves hybrid distances that can take into account both continuous and categorical variables. 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). 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. Another method, named KAMILA, relaxed the parametric hypothesis of mixture models by combining them with k-means algorithms, and categorical variables were assumed independent within a subpopulation ((Foss et al., 2016; Foss and Markatou, 2018)). In the same scope, spectral clustering was adapted for mixed-type data (Mbuga and Tortora, 2021). The main limitation of spectral clustering is the decomposition of sample size matrices, and as previous methods, it may require tuning continuous/nominal weight and kernel parameters.

Transforming variables or using hybrid distances really limit cluster analysis, especially with methods loosing original space of variables. In addition, outside of a clustering goal, and sometimes dimensionality reduction, they provide limited insights and usable results. On the other hand, model-based methods can be used for different goals, not only clustering one, like dimension reduction, exploration or interpretability of the estimated distributions. They do not rely on parameter tuning for the importance of discrete variables. But require adequate distributions, and assumptions on within-cluster variable interactions.

Among the direct approaches in a mixed-type data context, mixture models are efficient 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). Another family includes the normal-multinomial mixture model (Hunt and Jorgensen, 1996; Fraley and Raftery, 2002), where given cluster membership, 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 model identifiability. Numerical coding of categorical variables with a flexible covariance structure allow for explicit dependencies between continuous and categorical variables. This includes mixtures of factor analyzers (Ghahramani and Hinton, 1996; McLachlan and Peel, 2000; McLachlan et al., 2003), originally combining clustering and dimensional reduction, 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. Linear mixed models can also be integrated within mixture models when homogeneous regression relationship across subjects is violated (Celeux et al., 2005; Bai et al., 2016; Lee and Chen, 2019).

1.2 Mixture models for continuous data

In finite mixture models of parametric distributions, a class is defined as a subset of points arising from the same mixture component. When the data are multivariate real-valued observations, the usual probability density function for each component is the multivariate Gaussian distribution. But in the presence of extreme, scattered, heavy-tailed data, the assumption of data normality may no longer be relevant, and we refer here to some works which have proposed mixtures of laws other than Gaussian and may subsequently be used in our proposed models.

Mixture of t-distributions provide longer-tailed alternative to the normal distribution, and are more robust to atypical observations (McLachlan and Peel, 1998; Peel and Mclachlan, 2000). Literature on t-distributions and associated mixtures 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).

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. Moreover, they described an estimation method for their model, based on the Expectation-Maximization (EM) algorithm. Later, Franczak et al. (2015) proposed the Multiple Scaled Shifted Asymmetric Laplace distribution, which guarantees convex level sets. 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 models with skew Normal (Lin et al., 2007; Lin, 2009) or skew Student (Lin, 2010; Lee and McLachlan, 2012) distributions.

1.3 Estimation and selection of mixture models

The Expectation-Maximization (EM) algorithm Dempster et al. (1977) was proposed to estimate models on incomplete data, such as mixture models. 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 well-known limits, which include sensitivity to the initialization, selection of the number of component, and convergence towards the space boundaries.

To solve the sensitivity drawback, several strategies rely on repeted runs with random initializations 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 for SAL mixture models, only during their initialization part.

Another challenge posed by EM-type algorithms is the selection of the number of components, and intrinsically selection of the best 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 components, and avoid bad local maxima. Proposed methods go 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 works of Yang et al. (2012); Pruilh et al. (2022), who associate a single-run EM-like algorithm with an estimation of the number of components. In addition, their proposed algorithms make estimates more robust to initialization.

Additionnally, mixture model estimation may be jeopardized with degeneracies. This obstructs statistical inference and incorrect estimates are obtained. These degeneracies are more likely to occur with overparametrized initial models, or with small dataset sizes.

In this work we seek to deal with the same problems of the EM algorithm: find the optimal number of components, avoid sensitivity at initialization, and convergence towards the space boundaries. 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. In addition, to guarantee the performances of certain models on small-scale data, we introduce regularizations by using conjugate priors on some parameters.

1.4 Contributions

In this paper, we propose three main algorithms, to estimate mixture models with respectively Gaussian, Student and Shifted Asymmetric Laplace (SAL) multivariate distributions, associated with any set of discrete variables simulated from the following distributions: Bernoulli, Multinomial or Poisson. Our general framework, adapted from (Pruilh et al., 2022), combines the estimation of parameters of the defined mixture models with the estimation of the number of components in this mixture, which is an important objective in many statistical applications. For each continuous law considered, we transform 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 outperform these methods both on estimating the right number of components and the numerous model parameters. Performances of our methods named DEM-MD algorithms are then demonstrated on two real datasets with mixed-type data (Byar and Green, 1980; Telford and Cunningham, 1991).

2 Materials and methods

Formally, a sample $\boldsymbol{x} \in \mathbb{R}^g$ from a parametric finite mixture distribution with K components has the following probability density function (pdf)

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

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 $k \in \{1, \ldots, K\}$.

This model is highly flexible and one can imagine many pdf p_g for continuous data. Below we provide the three main ones we will consider as they carry different properties.

Firstly, we will consider multivariate Gaussian distributions (Eq. (29)). Being symmetric, it is often the very first approximation made on continuous data, and corresponding estimation methods are the most proven. However, the estimates of component mean and covariance parameters are not robust to outliers, atypical observations. As a solution to deal with these type of data, Student distributions (Eq. (30)), which have longer tails than normal distributions, are a suitable alternative to obtain estimates robust to outliers. Hence, Student distributions are considered as an hypothesis for continuous variables in our considered mixture models. Finally, we consider a third continuous multivariate distribution to cover different assumptions on continuous variables in mixture models. This third distribution is the Shifted Asymmetric Laplace distribution (Eq. (32)), which is useful in situations when data are heterogeneous and present large errors. Shifted Asymmetric Laplace distributions present asymmetric and heavy tails, in contrary to Gaussian distributions. With these three continuous distributions, we can handle different types of continuous data, that we will model with mixture models. We now recall how these distributions are incorporated into continuous mixture models, and the associated estimating equations.

2.1 Mixture models and parameter estimation

2.1.1 Complete models

For any finite mixture model on continuous data, the complete-data comprise n independent observed $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathbb{R}^g$, and missing component membership variables $(z_i)_{i=1,\ldots,n}$. 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 class 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.

With these data in hand, the complete Gaussian mixture model is given by

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

Student mixture models (McLachlan and Peel, 1998; Peel and Mclachlan, 2000) require additional latent variables to be able to estimate easily with EM-like algorithms. Regarding the representation of $\boldsymbol{X}$ given in Eq.(31), we introduce $u_1, \ldots, u_n$, latent variables, such as given $z_i^k = 1$, $X_i | u_i, z_i^k = 1$ follows a Gaussian distribution. Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, the $u_1, \ldots, u_n$ are independently distributed. The complete-data vector for a mixture model of Student distributions is then given by $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, u_1, \ldots, u_n)$. The complete model is

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

SAL mixture models (Franczak et al., 2014) also require additional latent variables in order to estimate a complete model with an EM-like algorithm. From the representation in Eq.(33), $\mathbf{X}$ can be generated from a Gaussian distribution, knowing an exponential variable W. Thus, given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ in a mixture model, latent variables $w_1, \ldots, w_n$ enable this generation of $\mathbf{X}$. Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, $w_1, \ldots, w_n$ are independently distributed according to an exponential distribution of rate 1. The complete-data vector is given by $\boldsymbol{y} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, w_1, \ldots, w_n)$. The complete SAL mixture model is given by

$$\left\{ \begin{array}{ll} z_i & \sim \text{Categorical}(\boldsymbol{\pi})\,, \\ w_i | z_i^k = 1 & \sim \mathcal{E}(1)\,, \\ \boldsymbol{x}_i | w_i, z_i^k = 1 & \sim \mathcal{N}_g(\boldsymbol{\mu}_k + w_i \boldsymbol{\alpha}_k, w_i \boldsymbol{\Sigma}_k)\,. \end{array} \right. \tag{3}$$

2.1.2 Parameter estimation

Using an EM-like algorithm for finding maximum likelihood require a complete-data model as the ones detailed above. At E-step, computation of the current conditional expectation of z_i^k for each component k and individual i leads to the equations (34), (35) or (37) in Table 12. Update equations of current conditional expectations for other latent variables are given respectively by Eq.(36) for Student models and Eq.(38) and (39) for SAL models.

At M-step of EM-type algorithms, expectation of the complete-data log-likelihood is maximized, leading to newly estimated parameters. For Gaussian. Student or SAL distributions, update equations of parameters and mixture proportions are given in Table 1. Additional details are provided on estimation of some parameters for Student or SAL mixture models.

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

$$\begin{aligned}\left\{-\psi\left(\frac{1}{2}\nu\right)+\log\left(\frac{1}{2}\nu\right)+1+\frac{1}{n_k^t}\sum_{i=1}^{n}\tau_{ik}^t\left(\log E_{u,ik}^t-E_{u,ik}^t\right)\right.\\ \left.+\psi\left(\frac{\hat{\nu}_k^{t-1}+g}{2}\right)-\log\left(\frac{\hat{\nu}_k^{t-1}+g}{2}\right)\right\}=0\,,\end{aligned} \tag{4}$$

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 or Brent's methods.

For Shifted Asymmetric Laplace distributions, computation of the conditional 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

$$\hat{\boldsymbol{\alpha}}_k^t=\frac{\sum_i^n\tau_{ik}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k^{t^\star})^\top}{\sum_i^n\tau_{ik}E_{1ik}}\,. \tag{5}$$

This process is summarized in Algorithm 1.

Algorithm 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 (5)
    end
    else
        Compute α̂_k^t with (15)
    end
end
```

Law	Parameter	Updating Equation	
Gaussian	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$	(6)
	$\hat{\boldsymbol{\mu}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{n}\tau_{ik}}$	(7)
	$\hat{\boldsymbol{\Sigma}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}{\sum_{i=1}^{n}\tau_{ik}}$	(8)
Student	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$	(9)
	$\hat{\boldsymbol{\mu}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}\boldsymbol{x}_i}{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}}$	(10)
	$\hat{\boldsymbol{\Sigma}}_k$	$\frac{\sum_{i=1}^{n}\tau_{ik}E_{u,ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}{\sum_{i=1}^{n}\tau_{ik}}$	(11)
	$\hat{\nu}_k$	$\left\{-\psi\left(\frac{1}{2}\nu\right)+\log\left(\frac{1}{2}\nu\right)+1+\frac{1}{n_k}\sum_{i=1}^{n}\tau_{ik}\left(\log E_{u,ik}-E_{u,ik}\right)+\psi\left(\frac{\nu_k+g}{2}\right)-\log\left(\frac{\nu_k+g}{2}\right)\right\}=0$	
SAL	$\hat{\pi}_k$	$\frac{1}{n}\sum_{i=1}^{n}\tau_{ik}$	(12)
	$\hat{\boldsymbol{\mu}}_k$	$\frac{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik}\boldsymbol{x}_i)-n_k(\sum_i^n\tau_{ik}\boldsymbol{x}_i)}{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik})-n_k^2}$	(13)
	$\hat{\boldsymbol{\Sigma}}_k$	$S_k-\boldsymbol{\alpha}_k r_k^\top-r_k\boldsymbol{\alpha}_k^\top+\frac{1}{n_k}\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^\top\sum_{i=1}^{n}\tau_{ik}E_{1ik}$ 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 $r_k=\frac{1}{n_k}\sum_i^n\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)$	(14)
	$\hat{\boldsymbol{\alpha}}_k$	$\frac{(\sum_i^n\tau_{ik}E_{2ik})(\sum_i^n\tau_{ik}\boldsymbol{x}_i)-n_k(\sum_i^n\tau_{ik}E_{2ik}\boldsymbol{x}_i)}{(\sum_i^n\tau_{ik}E_{1ik})(\sum_i^n\tau_{ik}E_{2ik})-n_k^2}$	(15)
		or $\frac{\sum_i^n\tau_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k^\star)^\top}{\sum_i^n(\tau_{ik}E_{1ik})}$	(16)

Table 1: Equations to estimate continuous distribution parameters during M-step of an EM-type algorithm.

3 Mixture models for mixed-type datasets

3.1 Motivation and assumptions on the model

In this paper, 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, parameterization 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 mixture models, we propose here the use of Student or SAL distributions in place of the Gaussian distribution, and Bernoulli, Multinomial and Poisson distributions for discrete variables. These different continuous or discrete considerations allow to model data with a panel of assumptions and then be able to cluster, generate or interpret correctly.

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 Section 1.

3.2 Model description

3.2.1 Generic description

Consider an observation i of mixed variables, given by $\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^g, x_i^{g+1}, \ldots, x_i^{g+D}) \in \mathbb{R}^g \times \mathcal{X}$. 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 $\{1, \ldots, m_d\}$.

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

$$p(\boldsymbol{x}_i; \Theta) = \sum_{k=1}^{K} \pi_k p_g(\boldsymbol{x}_i^c; \theta_k^c) \prod_{d=1}^{D} p_{X_i^d}(x_i^d; \theta_k^d)\,, \tag{17}$$

with $\Theta = (\boldsymbol{\pi}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)$ the whole parameters set, and $\boldsymbol{\xi}_k$ contains $\left(\theta_k^c, \theta_k^1, \ldots, \theta_k^D\right)$, parameters of continuous and discrete distributions of each component $k \in \{1, \ldots, K\}$. 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 $d = 1, \ldots, D$. 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 $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

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 $(z_i)_{i=1,\ldots,n}$, 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

$$\begin{cases} z_i & \sim & \text{Categorical}(\boldsymbol{\pi})\,, \\ \boldsymbol{x}_i^c | z_i^k = 1 & \sim & F_g^c(\theta_k^c)\,, \\ x_i^d | z_i^k = 1 & \sim & F^d(\theta_k^d)\ \forall\, d = 1, \ldots, D\,, \end{cases} \tag{18}$$

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.

3.2.2 Application to different distributions

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 (18) with any of the continuous Models (1), (2) or (3). 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 (18) is on Figure 1 with the numerous parameters and latent variables corresponding to the considered laws in this paper.

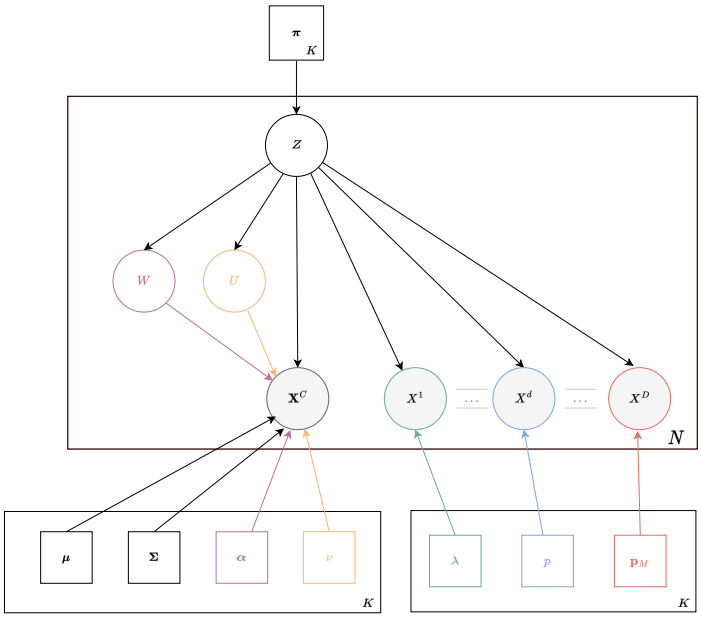


Figure 1: Graphical representation of our Model (18), with colors for additions/variants from Gaussian distribution. One color per continuous law for additional latent variables and parameters, and one color per type of discrete law: Student distribution (orange), SAL distribution (purple), Poisson (green), Bernoulli (blue) and Multinomial (red).

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.

4 Dynamic EM 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. Then we will detail individual considerations relative to each continuous distribution, which lead to particular adaptations of DEM-MD to perform correctly. Finally, we will present the updating equations to estimate parameters of discrete random variables 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) (Pruilh et al., 2022), to estimate Model (18). With the Modified REM, the number of classes was dynamically selected jointly 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) which had two weaknesses: an inadequate early stopping of the algorithm, and the lack of superimposed clusters detection, leading to surely wrong local 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

From the generic Model (18), the estimated function on complete-data is the following one, still including a penalization term on the mixture proportions:

$$\begin{aligned}\tilde{Q}(\Theta;\Theta^{(t)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} p_{\Theta^{(t)}}(z_i^k = 1; \boldsymbol{x}_i^c, x_i^D) \log\left[\pi_k p_g(\boldsymbol{x}_i^c;\theta_k^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_k^d)\right] \\ &+ \beta\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k\log\pi_k, \text{with } \beta \geq 0\,.\end{aligned} \tag{19}$$

Hyperparameter β comes as a penalty weight in Eq.(19). 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 $\pi_k < \frac{1}{n}$.

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:

$$\begin{aligned}p_{\Theta^{(t)}}(z_i^k = 1|\boldsymbol{x}_i^c, x_i^D) &= \frac{\pi_k p_g(\boldsymbol{x}_i|\theta_k^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_k^d)}{\sum_{j=1}^{K}\pi_j p_g(\boldsymbol{x}_i|\theta_j^c)\prod_{d=1}^{D} p_{X_i^d}(x_i^d;\theta_j^d)} \\ &= \tau_{ik}^t\,.\end{aligned} \tag{20}$$

For the Student and Shifted Asymmetric Laplace mixture models, additional latent variables are needed, as described in Models (2) and (3) respectively. 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.

With the objective function in Eq.(19) to maximize, the update equation of components proportions $\boldsymbol{\pi}$ inside DEM-MD algorithm is:

$$\hat{\pi}_k^t = \hat{\pi}_{k,\text{EM}} + \beta\hat{\pi}_k^{t-1}\left(\ln\hat{\pi}_k^{t-1} - \sum_{s=1}^{K}\hat{\pi}_s^{t-1}\ln\hat{\pi}_s^{t-1}\right), \tag{21}$$

with $\hat{\pi}_{k,\text{EM}}$ computed by Eq.(6), 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.

4.1.2 Aitken's acceleration criterion

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.

The Aitken's acceleration (Böhning et al., 1994) can be used to assess convergence, through asymptotic estimates of log-likelihood:

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

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 $t + 1$:

$$|l_\infty^{t+1} - l_\infty^t| < \varepsilon\,. \tag{22}$$

There exist other expressions of Aitken's acceleration in (Lindsay, 1995; McNicholas et al., 2010).

As we argued previously the evolution of log-likelihood in a DEM-MD algorithm is complex, and the Aitken's acceleration is relevant to assess the stability of the convergence. However, the following assumptions are necessary to use Aitken's acceleration: linear convergence of the algorithm and a slow convergence rate of the objective function. 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. Thus, if the number of classes is constant for at least four consecutive iterations (required to compute the Aitken's acceleration), the Aitken's acceleration criterion can be computed and applied to assess the convergence of our Dynamic EM algorithm for Mixed-type Data.

4.1.3 The generic DEM-MD algorithm

In Algorithm 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 3, which correspond to a classical EM version with an *a priori* fixed number of classes K.

4.2 Adaptations to different continuous distributions

4.2.1 A Gaussian continuous assumption

With a Gaussian assumption on the continuous variables, our Dynamic EM for Mixed-type Data is similar to the Modified REM (MREM) (Pruilh et al., 2022).

4.2.2 A Student continuous assumption

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. Initialize the degrees of freedom with constant values is a common strategy (Andrews et al., 2011; Lin, 2010).

Algorithm 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 (24)
$t \leftarrow 1$
Compute τ_{ik}^t with (20)
while $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  M-Step
  Compute π_k^t with (21)
  Compute $\boldsymbol{\mu}_k^t$
  $\beta^t \leftarrow$ Algo. 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 (25) , (26) , (27)
  E-Step
  Compute τ_{ik}^{t+1} with (20)
  Compute other latent variables
  $t \leftarrow t + 1$
end

According to Eq.(4), 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).

From generic Algorithm 2, the adaptation for estimation of models with Student continuous distributions is given by Algorithm 6.

4.2.3 A Shifted Asymmetric Laplace continuous assumption

Estimation of the skewness parameters involves the Mahalanobis distance in the denominator of (15). 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 $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We consider here the Expectation/Conditional-Maximization (ECM) algorithm, well-known for estimation of Student distributions (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 $\hat{\Theta}_2^{t-1}$ and calculate ${\Theta_1}^t$ by maximizing $Q(\Theta; \Theta^{(t-1)})$
- CM-Step 2. Fix $\hat{\Theta}_2^t$ and calculate Θ_2^t by maximizing $Q(\Theta; \tilde{\Theta}^{(t)})$.

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). Following this idea, we introduce an intermediate E-step just before the estimation of scales parameters in the SAL DEM-MD algorithm. 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.

We consider here using deterministic annealing in SAL DEM-MD algorithms. As a matter of fact, SAL DEM-MD without annealing struggles to converge. 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, the corresponding annealed version of Eq.(20) is

$$\tau_{ik}^t = \frac{\left[\pi_k p_g(\boldsymbol{x}_i|\theta_k^c) \prod_{d=1}^D p_{X_i^d}(x_i^d; \theta_k^d)\right]^{1/T}}{\sum_{j=1}^K \left[\pi_j p_g(\boldsymbol{x}_i|\theta_j^c) \prod_{d=1}^D p_{X_i^d}(x_i^d; \theta_j^d)\right]^{1/T}} \,. \tag{23}$$

Deterministic 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.

We consider temperature schemes inspired by works of Allassonnière and Chevallier (2021) and Lartigue et al. (2022). The idea is to consider an oscillating temperating 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 by:

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

Estimation of scale matrices requires special care, and usually small regularizations are used to estimate them (Yang et al., 2012; Pruilh et al., 2022). But these regularizations are not present in EM algorithms of Franczak et al. (2014), to estimate SAL mixture models. 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 computation errors in the previous paragraphs.

In our generic DEM-MD, scale matrices $\mathbf{\Sigma}_k^t$ are by default computed as $\mathbf{\Sigma}_k^t = (1-\gamma)\mathbf{\Sigma}_k^{EM} + \gamma\mathbf{P}$ for each k, with $\mathbf{\Sigma}_k^{EM}$ 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 on covariance matrices in DEM-MD algorithms), of values $0.0, 10e{-}9, 10e{-}5$. We observed no significative 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.

From Algorithm 2, adaptation for estimation of models with Shifted Asymmetric Laplace continuous variables is given by Algorithm 7.

4.3 Estimation of discrete distribution parameters in DEM-MD algorithms

With our defined Model (18), 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

$$\hat{p}^{d,0} = \begin{cases} x^d \in \{0,1\}^n \text{ if } d \text{ is Bernoulli}\,, \\ x^d \in \mathbb{R}_+^{\star,n} \text{ if } d \text{ is Poisson}\,, \\ [\mathbb{1}_{x^d=1}^T, \ldots, \mathbb{1}_{x^d=M}^T] \; \textit{with} \; \mathbb{1}_{x^d=m} \in \{0,1\}^n \text{ if } d \text{ is Multinomial}\,. \end{cases} \tag{24}$$

4.3.2 Equations to update discrete distribution parameters

At M-step, parameters for discrete distributions are also estimated, independently of parameters for continuous distributions. The equations to update parameters of the different discrete distributions are gathered in Table 2.

At the beginning of any DEM-MD algorithm, we transform a Multinomial variable as a one-hot encoding matrix, so $\boldsymbol{x}^d \in \{0,1\}^{n\times M}$, and x_{im}^d corresponds to sample i and column/modality m.

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.

All the DEM-MD and EM-MD algorithms presented in this article were implemented in Python 3.9.

5 Experiments on simulated data

We present here results to validate our Dynamic EM for Mixed-type Data algorithms on estimation of many mixture models on mixed-type data. After the description of numerous settings with different continuous and discrete distributions, we look at the convergence success of DEM-MD algorithms and the average iteration number to reach convergence across all simulation studies. We continue with a comparison of the estimation of K by DEM-MD algorithms and model selection criteria with EM algorithms. Thereafter, we provide performances on

Discrete distribution	Update equation at M-step	
Bernoulli	$\hat{p}_k^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_i^d}{\sum_{i=1}^n \tau_{ik}^t}$	(25)
Multinomial	$\hat{p}_{k,m}^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_{im}^d}{\sum_{i=1}^n \tau_{ik}^t}$ with m a modality of the Multinomial law with M modalities	(26)
Poisson	$\hat{p}_k^{d,t} = \hat{\lambda}_k^{d,t} = \frac{\sum_{i=1}^n \tau_{ik}^t x_i^d}{\sum_{i=1}^n \tau_{ik}^t}$	(27)

Table 2: Equations at M-step to estimate discrete distribution parameters.

estimation of discrete distribution parameters, and continuous distribution parameters through comparisons with literature methods. We finally conclude with a study on the inclusion of covariance matrix 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 3, 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.

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. The DEM-MD algorithms converge for 100% of runs for each setting with Gaussian or Student continuous distributions. Convergence is lower for SAL DEM-MD, especially on settings with an higher continuous dimensional space as C^M_{451} which has a 18% convergence rate and C_{343} with 74%. The other settings with SAL distributions have at least 97% convergence rates.

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

5.2 Estimation of the number of components

5.2.1 In mixed-type data context

We note C the number of correctly estimated number of components for a given configuration over a certain number of runs. This gives $C = \#\{\hat{K} = K^\star\}$ with $K^\star$ varying following the set of considered parameters. Radar charts 2 give C values estimated by DEM-MD algorithms for each continuous distribution and associated configurations. For each configuration, 100 datasets were simulated. We see that we reach high percentages of C, between 90 and 100% for almost all configurations with Gaussian continuous laws. 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, 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.

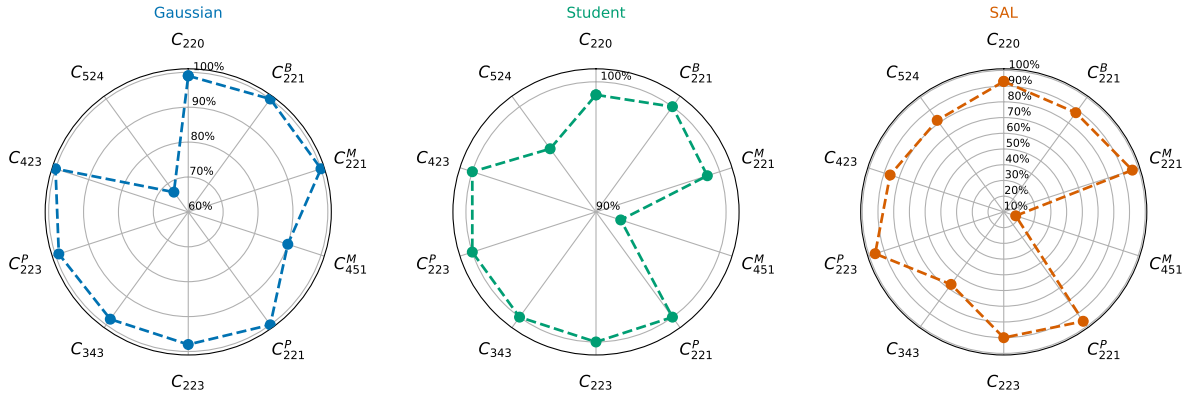
5.2.2 Comparison with model selection criteria

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 presented above, on the one hand DEM-MD algorithm, and on the other hand EM-MD algorithm associated with one of these three criteria: BIC (Schwarz, 1978), ICL (Biernacki et al., 1998) or NEC (Celeux and Soromenho, 1996). 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.

As for all the experiments, we simulated $S = 100$ datasets of each setting, 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 estimated models, which have different K. For each one of these criteria, the best model is the one with the lowest value. Finally, over $S = 100$ runs we have the number of correctly estimated $\hat{K}$ for each method (Tables 4a,4b,4c).

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 3: Description of simulated configurations.



(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 2: Radar charts of estimated number of classes $C = \#\{\hat{K} = K^{\star}\}$ by DEM-MD algorithms per continuous distribution. Each radar chart contains settings defined in Table 3.

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. Here the DEM-MD gives 100% of correct estimates 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.

5.3 Performances on the estimation of mixture parameters

Firstly, we present relative errors of DEM-MD and EM-MD on the estimation of discrete distribution parameters for the various settings described above. The results presented in the next parts are calculated on the set of experiments where $\hat{K} = K^\star$. For each setting, the corresponding number of experiments with correct K can be seen on Figure 2 above. Secondly, we compare DEM-MD algorithms with existing methods in the literature, in order to assess performances of Student and SAL DEM-MD algorithms on estimation of the continuous distribution parameters. A Gaussian DEM-MD algorithm without any discrete variable only differ of Modified REM (Pruilh et al., 2022) by the stopping criterion. Performances on estimation of Gaussian mixture parameters have therefore been confirmed by Pruilh et al. (2022) and are not shown here.

5.3.1 Performances on the estimation of discrete distribution parameters

We consider and implement Bernoulli, Multinomial or Poisson distributions, 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 (18), discrete features are independent conditional on class memberships. Figures 3, 4 and 5 show the relative errors for each setting and each discrete distribution parameter over the runs with a correct $\hat{K}$.

In Figure 3 we can observe relative errors of discrete distribution parameters for configurations containing only one discrete variable. 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(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.

Errors are similar for DEM-MD and EM-MD algorithms on the setting C^P_{223} (Fig. 4(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. 5). 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.Again, as for multinomial parameters, errors are higher for C_{343}

	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

a Gaussian continuous distributions.

	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

b Student continuous distributions.

	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

c SAL continuous distributions.

Table 4: Percentages of correctly estimated or selected K for DEM-MD algorithms with different continuous distributions.

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.

5.3.2 The trillium dataset

In their recent paper, Fang et al. (2023) proposed a Gibbs sampling framework to estimate SAL mixture models, and compared themselves to the EM-type algorithm of Franczak et al. (2014). They tested 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, Fang et al. (2023) had between 96 and 100% correct number of clusters with selection by BIC or ICL criterion.

Table 5 gives the true parameters, as well as the average estimates, with standard deviation, returned by our DEM-MD algorithm, and the ones obtained in Fang et al. (2023), directly reported from their article (Fang et al., 2023, see Table 3). 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.

5.3.3 The Peel dataset

To assess the performance of DEM-MD algorithm on Student continuous distributions, we consider here a simulated two-dimensional Student mixture model with three clusters, originally defined in Peel and Mclachlan (2000). We simulate 100 datasets of size $n = 200$, as in the literature. On each dataset, we run a Student 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 6 the true parameters, the average (and standard deviation) estimated parameters from these 84 correct DEM-MD runs and from 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, both algorithms are far from the true values, confirming that estimation of this parameter stays a challenge with EM-like algorithms. The average estimates of means, covariance matrices and degrees of freedom 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.

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.

5.4 Penalize covariances with Inverse-Wishart priors

We observed limitations of SAL DEM-MD in the presence of an undersized dataset in Subsection 5.2. This was particularly noticeable on setting C^M_{451} where $G = 5$ and the algorithm only obtained 18% of correct $\hat{K}$. 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 harder, and this is particularly true for the scales matrices which can become singular and cause the algorithm to diverge. Regularization is employed to avoid this problem, frequently artificial one in classical EM algorithms, and regularization 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. In this problematic case, a possibility to improve and/or stabilize the estimation is to regularize the covariance matrices.

We introduce here regularization using a prior on scale 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 $w = g + 2$ degrees of freedom, $\boldsymbol{W} = \frac{S}{K^{2/d}}$ scale matrix of the prior distribution, with S the

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

	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 6: 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`).

overall empirical covariance matrix. This prior was also adopted in the very recent work of Fang et al. (2023) which estimated parameters with a Gibbs sampling method and therefore needed to define parameter priors.

The scale matrix updates $\hat{\boldsymbol{\Sigma}}_k^t$ in the M-step of any EM-like algorithm are replaced at t iteration (Fraley and Raftery, 2007; Baudry and Celeux, 2015) by

$$\hat{\boldsymbol{\Sigma}}_k^{t,reg} = \frac{\hat{\boldsymbol{\Sigma}}_k^t + \boldsymbol{W}}{n_k + w + g + 1}\,. \tag{28}$$

Prior regularization 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 introduced prior, the scale matrix estimation step in Algorithm 7 can be replaced by Eq.(28) with $\hat{\boldsymbol{\Sigma}}_k^t = n_k^t \times \hat{\boldsymbol{\Sigma}}_{k,EM}^t$ and $\hat{\boldsymbol{\Sigma}}_{k,EM}^t$ estimated by Eq.(14). 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.

The addition of a regularization such as Fraley's one in the SAL DEM-MD leads to a clear improvement in the convergence of the algorithm and estimation of the number of classes for setting C_{451}^M . 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 $\hat{\boldsymbol{\Sigma}}_k^{t,reg}$ with Eq.(28) 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%. Figure 6 gives 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.

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 mixture models.

6 Experiments on real datasets

6.1 A Prostate Cancer dataset

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). 15 mixed-type 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 shortly described in Table 7.

Variable	Type	Variable	Type
Stage	Output variable (2 levels)	SurvStat (10 levels)	Output variable
Serum haemoglobin	Continuous	Size of primary tumor	Continuous
Age	Continuous	Index of tumor stage and histolic grade	Continuous
Weight	Continuous	Serum prostatic acid phosphatase	Continuous
Systolic Blood Pressure	Continuous	Diastolic Blood Pressure	Continuous
Performance rating	Ordinal (4 levels)	Electrocardiogram code	Nominal (7 levels)
Cardiovascular disease history	Binary	Bone Metastase	Binary

Table 7: Variables in the Prostate Cancer dataset.

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).

In brief, we have 8 continuous variables and 4 categorical variables. As we saw in the previous section, the SAL DEM-MD algorithm quickly encounters difficulties in estimating a model with a high-dimensional dataset. For its estimation on the prostate cancer dataset, we therefore apply regularization on the scale matrices as presented in the Subsection 5.4.

Student DEM-MD and SAL DEM-MD found $\hat{K} = 2$, while Gaussian DEM-MD estimated $\hat{K} = 3$ clusters. In comparison, clustMD's selection strategy (McParland and Gormley, 2016) returned three classes as the best model, while KAMILA's one (Foss and Markatou, 2018) returner two classes. Both solutions may be acceptable for a clustering objective as Stage output variable has two modalities and SurvStat three modalities (after aggregation of modalities (Foss and Markatou, 2018)).

A cross-tabulation of the cluster labels versus the cancer stage diagnosis is given in Table 8. Estimated classes by Student and SAL DEM-MD algorithms 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.

Comparing center vectors for clusters of each model (Fig. 7(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 Gaus-

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

Table 8: For Prostate cancer dataset, cross-tabulation of estimated cluster labels for each model versus the diagnosed prostate cancer stage.

sian and SAL models. On the other hand, Weight separates Gaussian and Student models. As the Gaussian model has three classes, looking at estimated parameters help to differientiate its classes, which improves interpretability.

6.2 The Australian Institute of Sport dataset

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 used 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 eleven continuous variables, corresponding to physical and blood measurements (Table 9).

Variable	Description
sex	the sex of the athlete
sport	the sport of the athlete, one of BBall, Field, Gym, Netball, Rowing, Swim, T400m, Tennis, TSprint, WPolo
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 9: Description of Australian Institute of Sport 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. As we saw in the section on simulated data, 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 5.4.

Running DEM-MD algorithms with 12 variables (all except sex variable) on the $n = 202$ athletes, DEM-MD on a Gaussian model estimates $\hat{K} = 3$ classes while DEM-MD on a Student model 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 10 and 11). We can see that all three models tend to separate men and women. 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 class, which requires looking at sport distribution for a good interpretation. Similarly, the Gaussian DEM-MD returned three classes 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. Differences between classes that do not correspond only to sex and sport separation can be observed by looking at marginal distributions of variables with respect to the assigned cluster.

7 Conclusion and perspectives

We have proposed Dynamic EM for Mixed-type Data algorithms to estimate mixture models for mixed-type data with different possible continuous and discrete variables, allowing to jointly estimate the number of classes and the various parameters. We introduced improvements compared to the EM versions, to ensure algorithm convergences and correct 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 as a stopping criterion, especially in a dynamic context. Comparisons with existing algorithms on Student or SAL continuous distributions shown good parameter recovery with DEM-MDalgorithms. Additionally, our algorithms obtained good performances on estimation of the number of components, for both simulated and real datasets. Finally, on the two real datasets, the Dynamic EM for

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

Table 10: Cross-tabulation of estimated cluster labels for each model versus the athlete sex

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

Table 11: Cross-tabulation of estimated cluster labels for each model versus the practiced sports

Mixed-type Data algorithms were able to retrieve meaningful classes, despite small dataset sizes, based if necessary on regularized covariance matrices as introduced earlier.

A clear limitation of our proposed mixture 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 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, additional continuous and discrete laws can still be modeled and estimated within the framework of the DEM-MD algorithm.

Declaration of Competing Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Fundings

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article.
This work was supported by a grant from Région Île-de-France. This work was supported by a grant from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir " program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

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

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. URL https://www.sciencedirect.com/science/article/abs/pii/S0167947320302504.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 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. 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. 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. 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. 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. URL https://escholarship.org/uc/item/2cz925kv.

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

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. URL http://journals.sagepub.com/doi/10.1177/075910639203700105.

L. Birgé and P. Massart. Minimal Penalties for Gaussian Model Selection. *Probability Theory and Related Fields*, 138(1-2):33–73, Feb. 2007. 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. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388, June 1994. URL http://link.springer.com/10.1007/BF01720593.

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. URL https://www.sciencedirect.com/science/article/pii/S0378375812001838.

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

D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information. *Bulletin Du 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. URL http://link.springer.com/10.1007/BF01246098.

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. URL https://journals.sagepub.com/doi/10.1191/1471082X05st096oa.

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. 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.

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. 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. URL https://doi.org/10.1007/s11222-018-9838-y.

A. Foss, M. Markatou, B. Ray, and A. Heching. A semiparametric method for clustering mixed data. *Machine Learning*, 105(3):419–458, Dec. 2016. 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. 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. 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.

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. 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. URL https://linkinghub.elsevier.com/retrieve/pii/S0167865515000598.

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. URL https://www.jstor.org/stable/2334349.

H. Holzmann, A. Munk, and T. Gneiting. Identifiability of Finite Mixtures of Elliptical Distributions. *Scandinavian Journal of Statistics*, 33(4):753–763, Dec. 2006. 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. 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.

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. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00378.

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. URL http://doi.wiley.com/10.1002/9780470316801.

I. Kosmidis and D. Karlis. Model-based clustering using copulas with applications. *Statistics and Computing*, 26 (5):1079–1099, Sept. 2016. 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. URL http://link.springer.com/10.1007/978-1-4612-0173-1.

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. 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. 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. 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. 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. *Statistics and Computing*, 24(2):181–202, Mar. 2014. 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. URL https://www.sciencedirect.com/science/article/pii/S0047259X08001152.

T.-I. Lin. Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, 20(3): 343–356, July 2010. 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.

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. 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.

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. URL https://www.tandfonline.com/doi/full/10.1080/03610926.2016.1277753.

F. Mbuga and C. Tortora. Spectral Clustering of Mixed-Type Data. *Stats*, 5(1):1–11, Dec. 2021. 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. 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. 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, 2nd 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. 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.

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. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947309000632.

D. McParland and I. C. Gormley. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2):155–169, June 2016. 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. *Statistics in Medicine*, 36(28): 4548–4569, Dec. 2017. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.7371.

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993. 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. URL https://doi.org/10.1080/01621459.2012.762328.

D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10: 339–348, 2000. URL https://link.springer.com/article/10.1023/A:1008981510081.

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.

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. 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.

O. Sahin and C. Czado. Vine copula mixture models and clustering for non-Gaussian data. *Econometrics and Statistics*, 22:136–158, Apr. 2022. 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. URL https://www.jstor.org/stable/2958889.

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. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2011.644501.

H. Teicher. Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963. 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. *Medicine and Science in Sports and Exercise*, 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. 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. 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. URL https://linkinghub.elsevier.com/retrieve/pii/S0167715212000673.

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. URL https://link.springer.com/article/10.1023/A:1008842432747.

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.

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. 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. 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. URL https://dl.acm.org/doi/10.1145/3394486.3403106.

A Appendix

A.1 Material on continuous distributions

A.1.1 Gaussian distribution

A g-dimensional random variable $\mathbf{X}$ following a multivariate Gaussian distribution has the following density

$$p_g(\boldsymbol{x}_i;\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{g/2}}\exp(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})) \tag{29}$$

A.1.2 Student distribution

A g-dimensional random variable $\mathbf{X}$ following a multivariate t-distribution has the following density

$$p_g(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma(\frac{\nu+g}{2})|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{g/2}\Gamma(\frac{\nu}{2})\{1+\delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma})/\nu\}^{\frac{(\nu+g)}{2}}}, \tag{30}$$

with center $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, degrees of freedom $\nu \in (0; \infty]$ and $\delta(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance.

Given a scaling variable $\mathrm{U} \in \mathbb{R}$, $\mathbf{X}$ has a multivariate normal distribution, and $\mathbf{U}$ is Γ with

$$\mathbf{X}|\mathrm{U} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}/\mathrm{U}) \text{ and } \mathrm{U} \sim \Gamma(\frac{1}{2}\nu, \frac{1}{2}\nu)\,. \tag{31}$$

A.1.3 Shifted Asymmetric Laplace distribution

The probability density function of a g-dimensional random variable $\mathbf{X}$ distributed according to a Shifted Asymmetric Laplace distribution is

$$p_g(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \frac{2 \exp\left\{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\right\}}{(2\pi)^{g/2} |\boldsymbol{\Sigma}|^{1/2}} \times \left(\frac{\delta(\boldsymbol{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{2 + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}} \right)^{\nu/2} K_\nu(u)\,, \tag{32}$$

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 $u = \sqrt{(2 + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}) \delta(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma})}$.

From Kotz et al. (2001) for Asymmetric Laplace distributions and Franczak et al. (2014) for Shifted AL distributions, the random variable $\mathbf{X}$ admits the representation

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}Y,\ W \sim \mathcal{E}(1),\ Y \sim \mathcal{N}_g(0, \boldsymbol{\Sigma})\,, \tag{33}$$

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 $\boldsymbol{\Sigma}$. And so, $X|W = w \sim \mathcal{N}_g(\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$.
Conditional on the data, W follows a Generalized Inverse Gaussian distribution.

A.2 Estimation of continuous distributions in mixture models

Law	Expectations of latent variables	Updating equation	
Gaussian	τ_{ik}	$\frac{\pi_k p_{Gaussian}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}[\pi_j p_{Gaussian}(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)]}$	(34)
Student	τ_{ik}	$\frac{\pi_k p_{Student}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k)}{\sum_{j=1}^{K}[\pi_j p_{Student}(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\nu_j)]}$	(35)
	$E_{u,ik}$	$\frac{\nu_k+g}{\nu_k+\delta(\boldsymbol{x}_i\|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$	(36)
SAL	τ_{ik}	$\frac{\pi_k p_{SAL}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\alpha}_k)}{\sum_{j=1}^{K}[\pi_j p_{SAL}(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\boldsymbol{\alpha}_j)]}$	(37)
	E_{1ik}	$\sqrt{\frac{b_{ik}}{a_k}}R_\nu(\sqrt{a_k b_{ik}})$	(38)
	E_{2ik}	$\sqrt{\frac{a_k}{b_{ik}}}R_\nu(\sqrt{a_k b_{ik}})-\frac{2\nu}{b_{ik}}$	(39)

Table 12: Equations to compute conditional expectations of latent variables at E-step of an EM-type algorithm.

A.3 Pseudocodes

A.3.1 Generic EM-MD and β computation pseudocodes

Algorithm 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 (6)
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 (25) , (26) , (27)
$t \leftarrow 1$
while $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ *and* $t < t^{max}$ **do** `/* Aitken's convergence */`
  E-Step
  Compute τ_{ik}^t with (20)
  Compute other latent variables
  M-Step
  Compute π_k^t with (6)
  Compute $\boldsymbol{\mu}_k^t$
  Compute $\boldsymbol{\Sigma}_k^t$
  Compute other continuous parameters
  Compute discrete probabilities $p_k^{d,t}$ with (25) , (26) , (27)
  $t \leftarrow t+1$
end

A.3.2 DEM-MD pseudocodes

Algorithm 4: Computation of parameter β

Input : $\boldsymbol{\pi}^{EM}$,$\boldsymbol{\pi}^{(new)}$,$\boldsymbol{\pi}^{(old)}$, K, n

$\pi_{(1)}^{EM} \leftarrow \underset{1\leq k\leq K}{max}\, \pi_k^{EM}$, $\pi_{(1)}^{(old)} \leftarrow \underset{1\leq k\leq K}{max}\, \pi_k^{(old)}$

$E \leftarrow \sum_{k=1}^{K} \pi_k^{(old)} \log(\pi_k^{(old)})$

$\beta \leftarrow \min\left\{ \frac{\sum_{k=1}^{K} \exp\left(-\eta n \left|\pi_k^{(new)} - \pi_k^{(old)}\right|\right)}{K}, \frac{\left(1-\pi_{(1)}^{EM}\right)}{\left(-\pi_{(1)}^{(old)} E\right)} \right\}$

Output: β

Algorithm 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 (24)

$t \leftarrow 1$

Compute τ_{ik}^t with (20)

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

  M-Step

  Compute π_k^t with (21)

  Compute $\boldsymbol{\mu}_k^t$ with (7)

  $\beta^t \leftarrow$ Algorithm 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 $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$, $\boldsymbol{\Sigma}^t$ and $\boldsymbol{\tau}^t$

    end

  end

  Compute $\boldsymbol{\Sigma}_k^{EM}$ with (8) and $\boldsymbol{\Sigma}_k^t = (1-\gamma)\boldsymbol{\Sigma}_k^{EM} + \gamma\mathbf{P}$ 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 (25) , (26) , (27)

  E-Step

  Compute τ_{ik}^{t+1} with (20)

  $t \leftarrow t + 1$

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

end

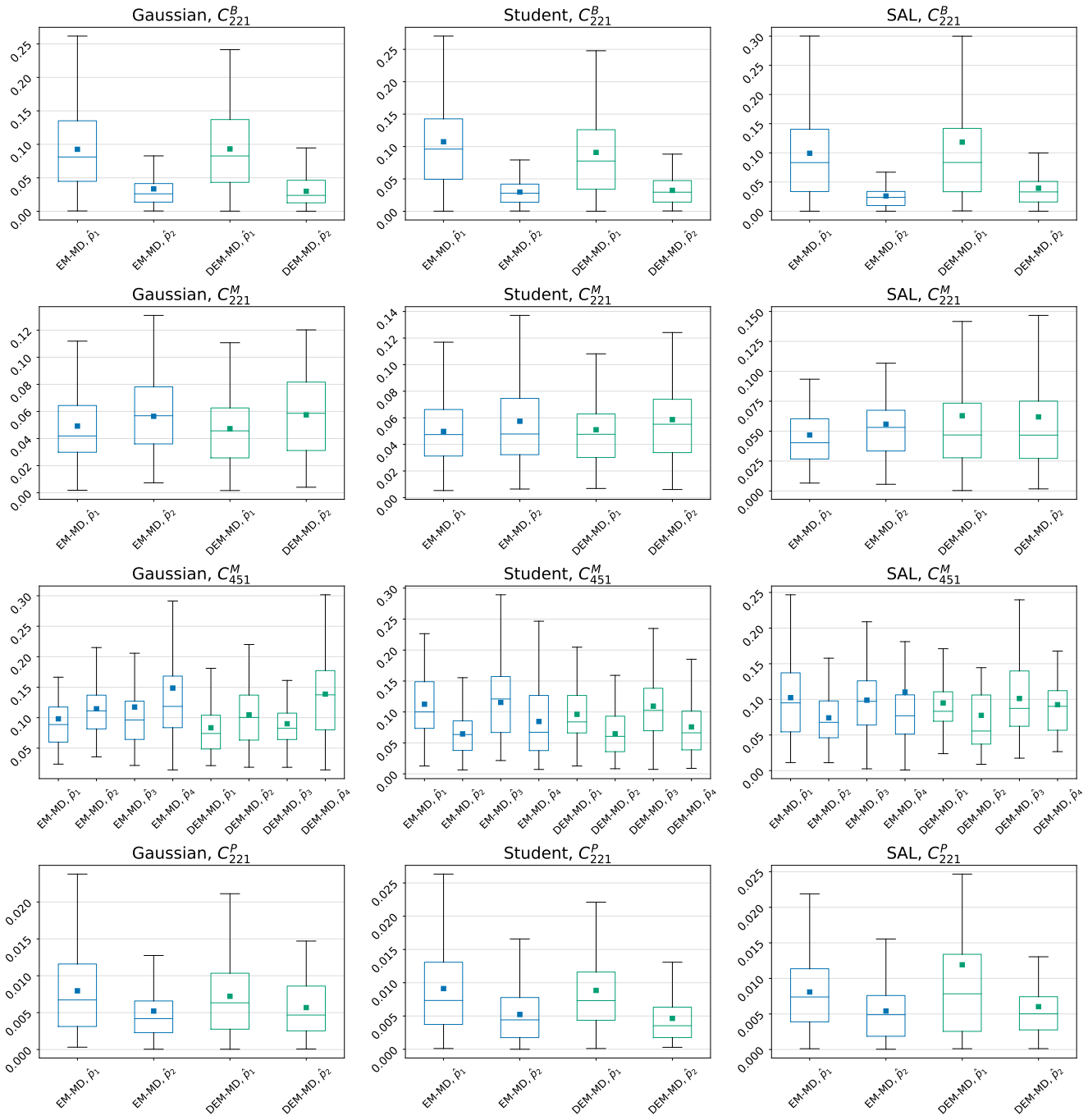
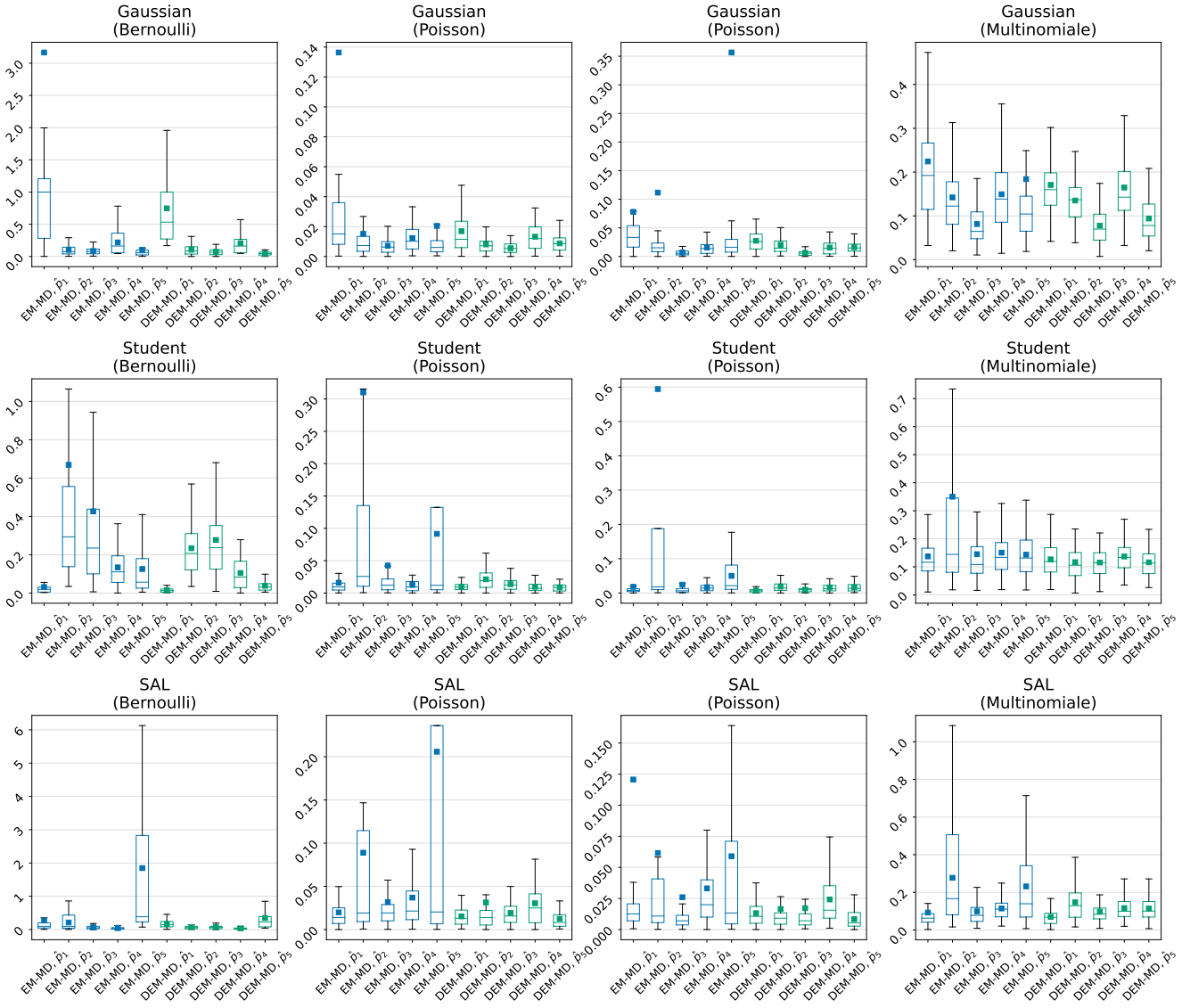
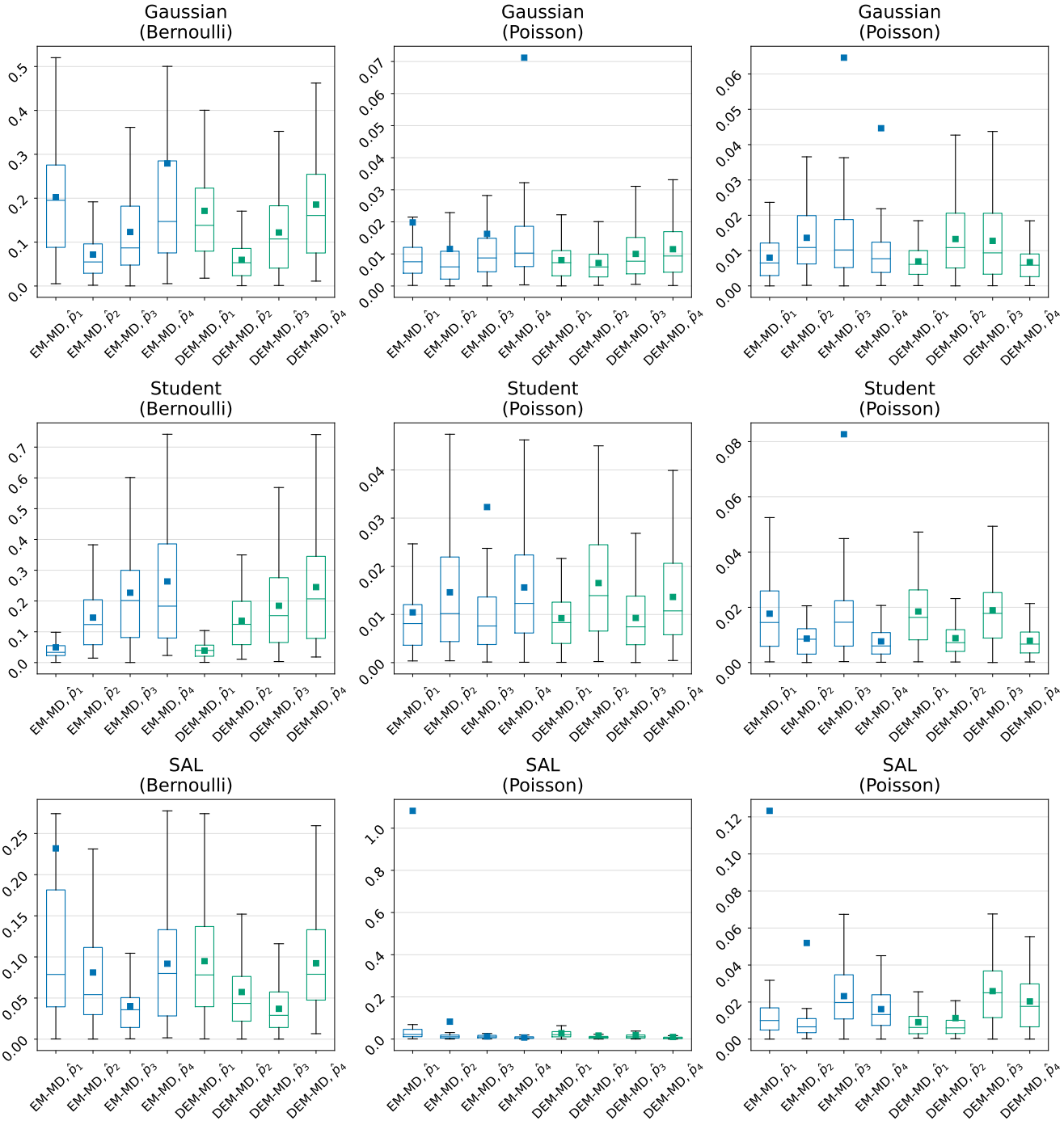


Figure 3: Boxplots of relative discrete distribution parameter errors for settings C^B_{221}, C^P_{221}, C^M_{221} and C^M_{451}, for the three continuous distributions on both DEM-MD (green) and EM-MD (blue). Each column corresponds to a continuous distribution, in this order: Gaussian, Student, SAL. Each row corresponds to a setting, in this order: C^B_{221}, C^M_{221}, C^M_{451} and C^P_{221}.



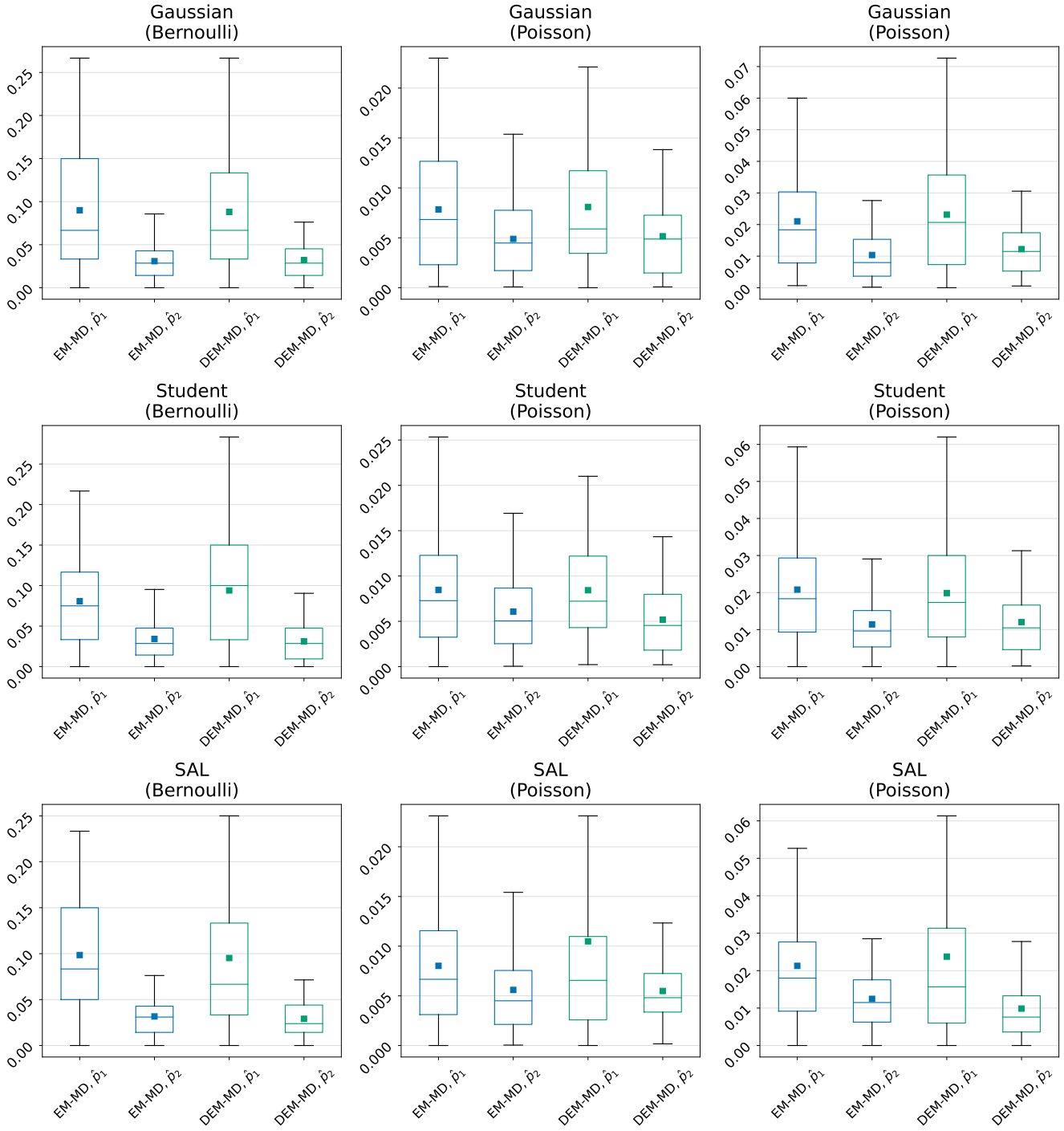
(a) Configuration C_{524}

Figure 4: 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).



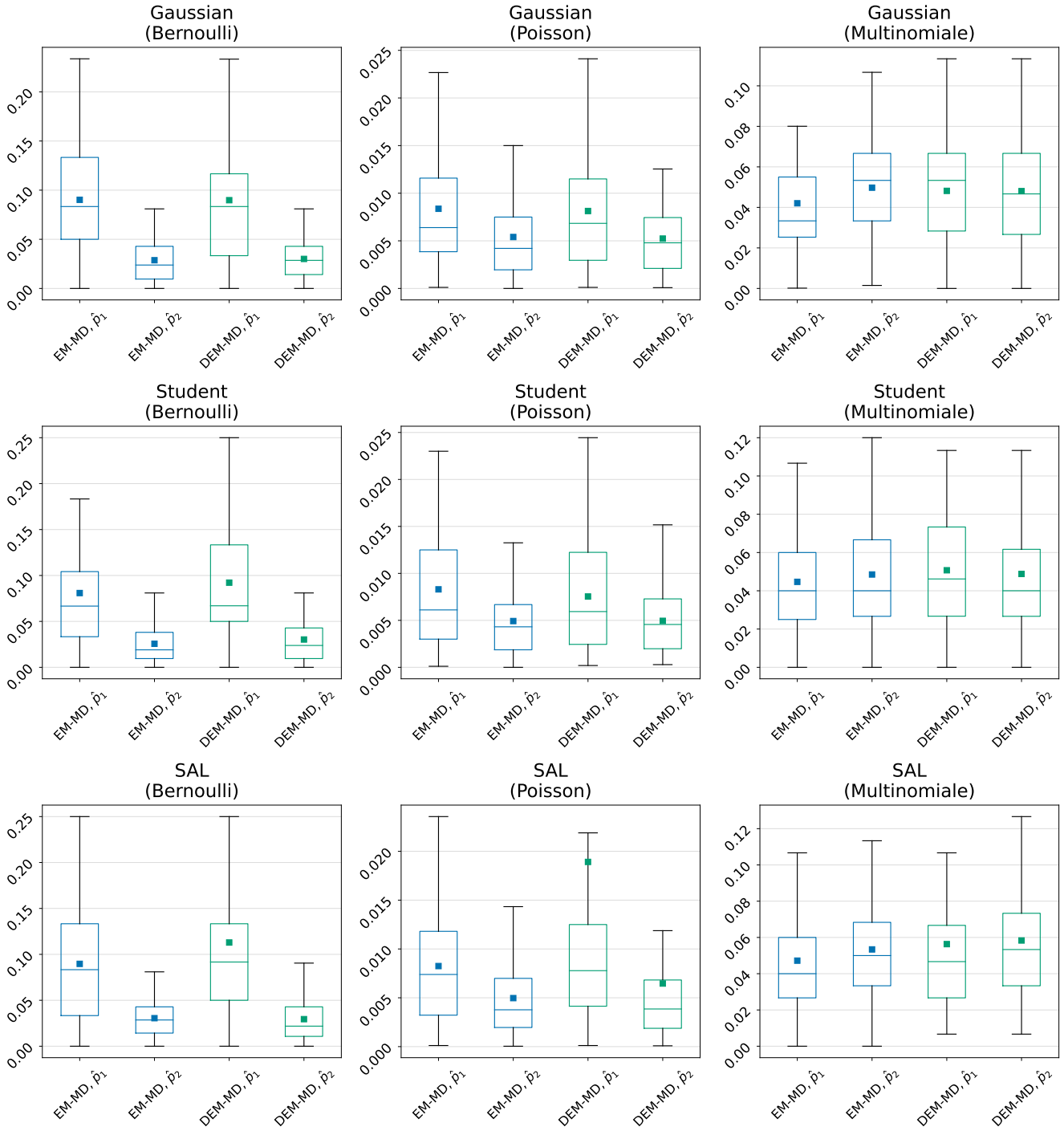
(b) Configuration C_{423}

Figure 4: 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).



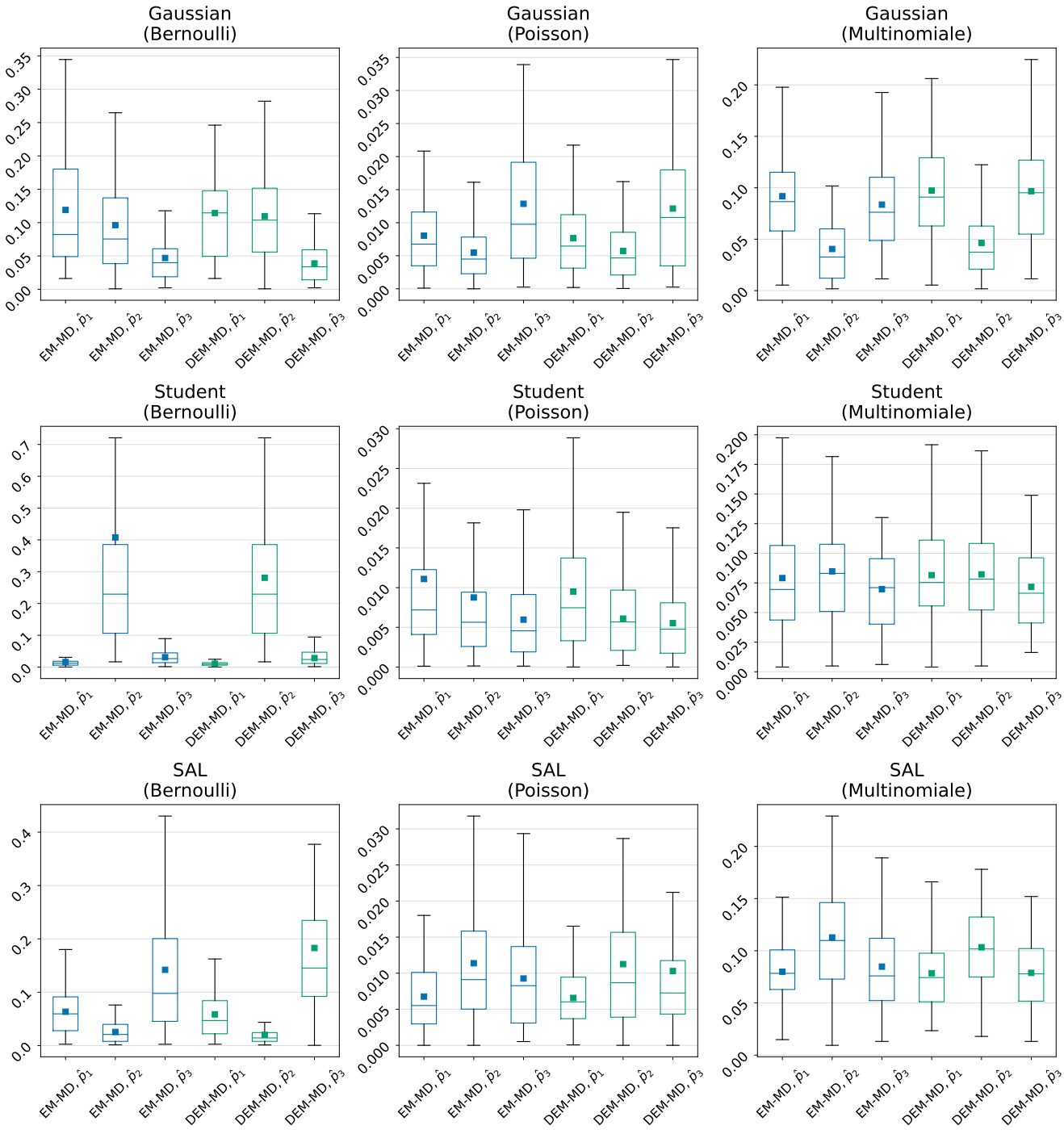
(c) Configuration C^P_{223}

Figure 4: 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 5: 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 5: 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).

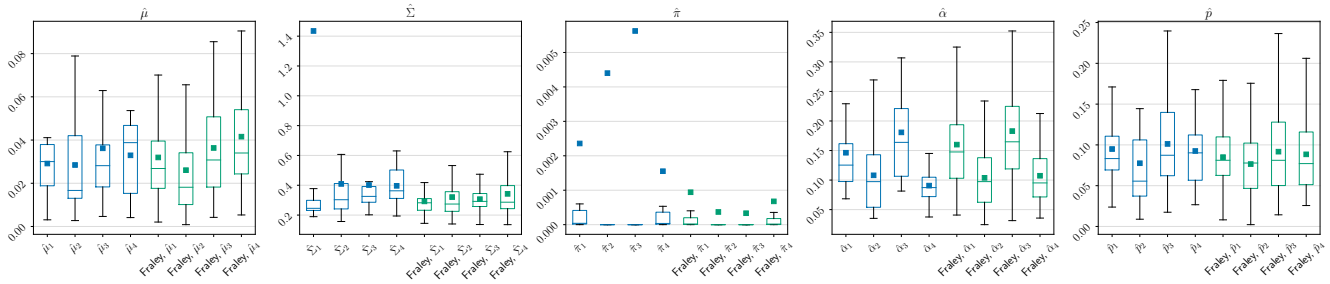
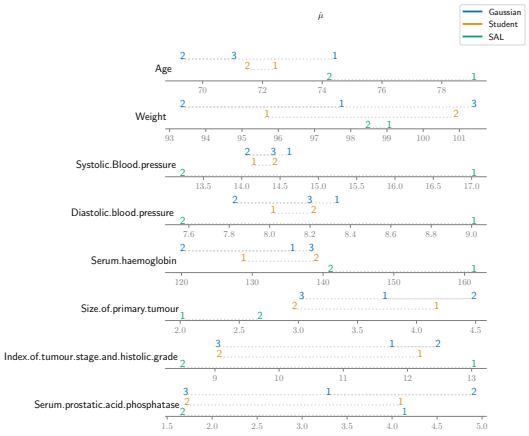
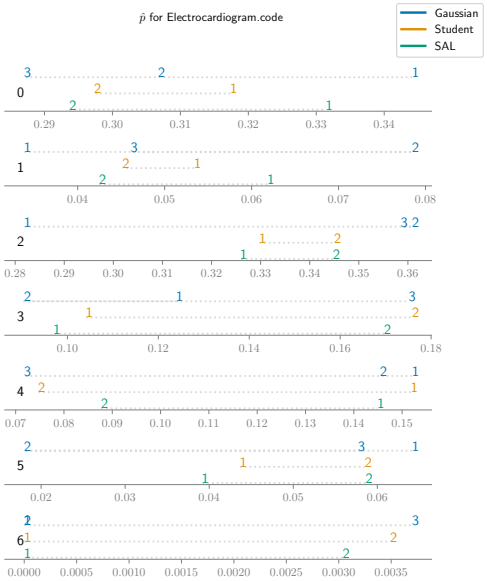


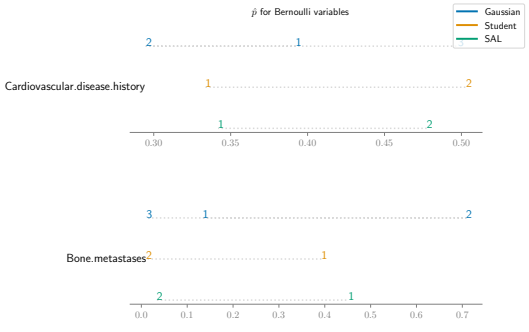
Figure 6: Relative errors of SAL DEM-MD with/out Fraley regularization, on setting C_{451}^{M}. In blue are given the errors without regularization, from the same eighteen previous runs. In green are the errors with scale regularization. Results are over 72 runs with $n = 600$.



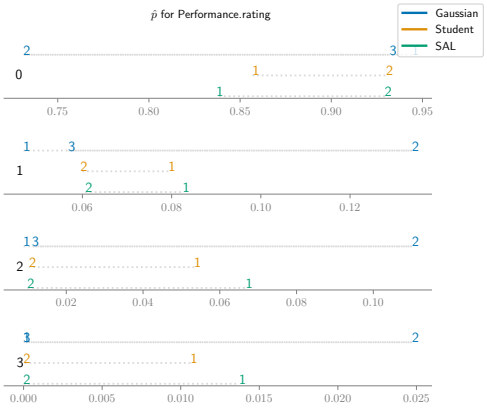
(a) Center parameter estimates



(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 7: 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).

Algorithm 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 (24)
$t \leftarrow 1$
Compute τ_{ik}^t with (20)
Compute $E_{u,ik}^t$ with (36)

while $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  M-Step
  Compute π_k^t with (21)
  Compute $\boldsymbol{\mu}_k^t$ with (10)
  $\beta^t \leftarrow$ Algorithm 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 $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$, $\boldsymbol{\Sigma}^{t-1}$, ν^{t-1} and $\boldsymbol{\tau}^t$
    end
  end
  Compute $\boldsymbol{\Sigma}_k^{EM}$ with (11) and $\boldsymbol{\Sigma}_k^t = (1-\gamma)\boldsymbol{\Sigma}_k^{EM} + \gamma\mathbf{P}$ 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) with Brent's method on the interval $[2, 200]$
  Compute discrete probabilities $p_k^{d,t}$ with (25) , (26) , (27)
  E-Step
  Compute τ_{ik}^{t+1} with (20)
  Compute $E_{u,ik}^{t+1}$ with (36)
  $t \leftarrow t + 1$
  $t_{component} \leftarrow t_{component} + 1$
end

Algorithm 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 (24)
$t \leftarrow 1$
Compute τ_{ik}^t with (23)
Compute $E_{1i}^{k,t}$ with (38)
Compute $E_{2i}^{k,t}$ with (39)

1 **while** $|l_\infty^t - l_\infty^{t-1}| > \varepsilon$ **do** `/* Aitken's convergence */`
  CM-Step 1
  Compute π_k^t with (21)
  Compute $\boldsymbol{\mu}_k^t$ with (13)
  $\beta^t \leftarrow$ Algorithm 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 (25) , (26) , (27)
  Check $\boldsymbol{\mu}^t = \boldsymbol{x}_i^c$ and compute α_k^t with Algorithm 1
  Intermediate E-step
  Compute $\tilde{\tau}_{ik}^t$, $\tilde{E}_{1i}^{k,t}$ and $\tilde{E}_{2i}^{k,t}$ with respectively (23), (38) and (39)
  CM-Step 2
  Compute $\boldsymbol{\Sigma}_k^t$ with (14)
  E-Step
  Compute τ_{ik}^{t+1} with (23)
  Compute $E_{1i}^{k,t+1}$ with (38)
  Compute $E_{2i}^{k,t+1}$ with (39)
  $t \leftarrow t + 1$
  $t_{component} \leftarrow t_{component} + 1$
end