Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach

Filippo Antonazzo, Christophe Biernacki, Christine Keribin

Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach

Filippo Antonazzo[1,2,3*], Christophe Biernacki[1,2,3] and Christine Keribin[1,2,4]

[1]Inria, France.
[2]CNRS, France.
[3]Laboratoire de mathématiques Painlevé, Université de Lille, Villeneuve d'Ascq, 59650, France.
[4]Laboratoire de mathématiques d'Orsay, Université Paris-Saclay, Orsay, 91405, France.

*Corresponding author(s). E-mail(s): filippo.antonazzo@inria.fr;
Contributing authors: christophe.biernacki@inria.fr;
christine.keribin@universite-paris-saclay.fr;

Abstract

Clustering conceptually reveals all its interest when the dataset size considerably increases since there is the opportunity to discover tiny but possibly high value clusters which were out of reach with more modest sample sizes. However, clustering is practically faced to computer limits with such high data volume, since possibly requiring extremely high memory and computation resources. In addition, the classical subsampling strategy, often adopted to overcome these limitations, is expected to heavily failed for discovering clusters in the highly imbalanced cluster case. Our proposal first consists in drastically compressing the data volume by just preserving its bin-marginal values, thus discarding the bin-cross ones. Despite this extreme information loss, we then prove identifiability property for the diagonal mixture model and also introduce a specific EM-like algorithm associated to a composite likelihood approach. This latter is extremely more frugal than a regular but unfeasible EM algorithm expected to be used on our bin-marginal data, while preserving all consistency properties. Finally, numerical experiments highlight that this proposed method outperforms subsampling both in controlled simulations and in various real applications where imbalanced clusters may typically appear, such as image segmentation, hazardous asteroids recognition and fraud detection.

Keywords: Imbalanced clustering, Large size data, Gaussian mixture models, Binned data, Random subsampling, Frugal learning

1 Introduction

In many contexts it is possible to collect data grouped in classes, for which two statistical analyses are of common interest, depending on the fact whether true classification is provided or not. In the first one, named *classification*, the aim is to assign to the right class each data record. If class labels are not available, the objective is to recover a partition of the dataset whose groups are homogeneous according to a certain criterion: this operation is named *clustering*. Both tasks could be difficult if there is a class represented by very few elements in comparison to the others: in this case the dataset is said to be *imbalanced*. This is usual

in several fields, such as credit card fraud detection (Chan and Stolfo 1998), cancer recognition (Yu et al. 2012), fraudulent calls (Fawcett and Provost 1997), where typically very few "abnormal" objects have to be recognised among a large amount of "normal" ones. Due to this presence of various applications, imbalanced datasets are the object of study in the rest of the paper.

Usually imbalanced datasets are analyzed in classification settings, where class labels are known. The most employed techniques consist in the creation of an artificial balanced dataset in the pre-preprocessing stage by oversampling the minority class (Chawla et al. 2002) or undersampling (Tahir et al. 2009) the majority one. Here, we focus on solving the corresponding clustering problem motivated by the fact that labelling records could be sometimes difficult, especially when sample size is large. Our purpose has been ultimately strengthened by the explosion of Big Data, which has made possible the availability of very large datasets, mostly imbalanced (Leevy et al. 2018; Fernández et al. 2017). A quick tour on one of the best-known datasets repository, the UCI Machine Learning Repository, currently shows several dozens of datasets with more than 1 million records. In presence of these huge datasets, classical methods are difficult to use because of the dramatic increase of time, memory and energy consumption. Although a possible solution consists in employing powerful computers or distributed architectures, such as MapReduce (Dean and Ghemawat 2008), here we focus on those strategies, named *frugal*, that use only the resources of a single ordinary laptop.

In this context, a common procedure to save computational resources is *random subsampling*, which consists in analyzing a small randomly selected portion of the original dataset. We find this approach critical in clustering, as the subsample could not contain any information about the small classes, especially if its size is really small. For example, if the subsampling size is fixed to 100 and the real small class proportion is equal to 0.001, the probability of extracting a subsample without one of its representatives is equal to 0.91, which is really hazardous. For this reason, we propose a new data-reduction technique based on a *marginal* construction of *binned data*. Such compressed dataset will consist in the *marginal counts* of the original (or *raw*) observations. They will be a collection of D univariate binned data, where D is the dimension of the original dataset. In virtue of its particular formulation, we name our approach *bin-marginal*.

Clustering is historically performed using geometrical heuristics, such as distances between the data points. A more recent way of clustering, the *model-based clustering*, has become popular because it allows a well-posed mathematical definition of the clusters. Indeed, it is principally based on likelihood estimation of Gaussian Mixture Models by using the EM algorithm (Dempster et al. 1977). Model-based clustering has also proved to be successful in case of moderate size datasets and typically it is applied on huge ones using random subsampling (Banfield and Raftery 1993; Fraley and Raftery 2002; Tsapanos et al. 2016; Xia et al. 2019). However, for the reasons explained before, it is inappropriate if the dataset contains small classes. Therefore, our aim is to make model-based clustering able to frugally detect them by using our bin-marginal approach. However, while subsampling does not need any changes in the EM formulation, the only usage of binned data requests particular versions of it (McLachlan and Jones 1988; Cadez et al. 2002). In addition, the further data-reduction given by marginalization requires a new EM algorithm to optimize bin-marginal likelihood, but this procedure is computationally unfeasible, as we will show in Section 3. This leads us to propose the optimization of the *composite likelihood* (Lindsay 1988) instead of the full one, providing a feasible EM-like algorithm. This final step finally defines our frugal Gaussian clustering proposal based on the bin-marginal approach. However, the usage of composite likelihood is not new in mixture models (Gao and Song 2011) and it has already been used in combination with bivariate marginal binned data in Whitaker et al. (2020) and Ranalli and Rocci (2016). But in our paper we go further, proposing an harder data-reduction where only univariate marginal binned data are used.

In proposing our bin-marginal technique, we will adopt the subsampling method as our main competitor, due to its prominency in the field. We will compare both methods under identical computational constraints, assuring they will use the same amount of computer memory. Full dataset

result will be use as a benchmark, even if it is far from being competitively frugal. Indeed, relatively to our proposal, it will confirm to consume a huge amount of resources without substantially improving the quality of clustering. Moreover, we will also establish theoretical results for binned mixture identifiability. In fact, we noted this problem was omitted, with the exception of Ranalli and Rocci (2016), where a necessary condition on bivariate marginal binned mixtures is provided. We will establish this theoretical property under the hypothesis of diagonal covariance matrices, due to the theoretical impossibility of estimating covariance parameters, as it will be shown in Section 4.3. This diagonal restriction is in fact common in literature, as it is employed in some popular clustering methods, as k-means (MacQueen et al. 1967), or in the so-called parsimonious Gaussian mixture models (McNicholas and Murphy 2008).

The paper is organized as follows: in Section 2, we specify the notations and the computational motivations justifying our proposal. Section 3 presents our bin-marginal data-reduction proposal and the associated results about identifiability of binned mixtures. In Section 4, we describe the composite likelihood EM algorithm necessary to frugally perform clustering on the new highly-compressed dataset. In Section 5, we train our methodology in several simulation settings involving large sample imbalanced data, comparing it to the subsampling strategy. In Section 6, the method is tested on real large data sets, spacing from image segmentation to fraud detection. Finally, in Section 7, we conclude with a discussion giving also ideas for further research. In Appendix A, all proofs regarding the principal theoretical results are provided.

2 Finite mixture models with binned data

2.1 Mixtures with raw data

Gaussian mixtures assume data come from K different Gaussian sub-populations. Let consider a set of n observations with D variables. We suppose that the observations $\mathbf{x} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \dots, n\}$ are i.i.d. and generated according to a D-dimensional Gaussian mixture with K components, whose probability density function is:

$$\begin{aligned} f(\boldsymbol{x};\ \boldsymbol{\psi}) &= \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \sum_{k=1}^{K} \pi_k &= 1, \quad \pi_k > 0 \quad (k = 1, \dots, K), \end{aligned} \tag{1}$$

where, for each component $k = 1, \dots, K$, $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kD})$ is the vector of means and $\boldsymbol{\Sigma}_k$ is the covariance matrix, with diagonal $(\sigma^2_{k1}, \dots, \sigma^2_{kD})$. In addition, $\boldsymbol{\psi} = (\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K)$ is the vector of parameters contained in a real space $\boldsymbol{\Psi}$. The set of all possible vectors of proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ will be denoted as Π_K. Often, an EM algorithm (Dempster et al. 1977) is used for estimating $\boldsymbol{\psi}$.

2.2 Mixtures with binned data

Sometimes, raw data $\boldsymbol{x}_i$ are unobservable and instead of knowing them, the only available information is a vector $\boldsymbol{n} = (n_1, \dots, n_B)$. Here, each n_b represents the number of observations lying inside a certain region $\mathcal{B}_b$, belonging to a partition $\{\mathcal{B}_b \subset \mathbb{R}^d, b = 1, \dots, B\}$ into which the original sample space can be divided. Thus, $n_b = \#\{\boldsymbol{x}_i \in \mathcal{B}_b\}$.

The vector $\boldsymbol{n}$ contains what we call *binned data*. According to Cadez et al. (2002), $\boldsymbol{n}$ arises from a multinomial model with probability mass function

$$p(\boldsymbol{n};\ \boldsymbol{\psi}) \propto \prod_{b=1}^{B} \Big(\sum_{k=1}^{K} \pi_k \int_{\mathcal{B}_b} \phi(\boldsymbol{x};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} \Big)^{n_b}.$$

In the same paper, the authors also provide an estimate of $\boldsymbol{\psi}$ using a binned version of the EM algorithm, whose time and memory complexity depends linearly on B. For this reason, if $B \ll n$, there is a considerable gain in terms of computational time and storage. Thus, a first scalable method involves an artificial generation of binned data, typically obtained by imposing a D-dimensional Cartesian grid $G = G_1 \times \ldots \times G_D$ such that G_d is a univariate grid with $R_d + 2$ cut points $(a_{d0}, \dots, a_{d(R_d+1)})$, where $a_{d0} = -\infty$ and $a_{d(R_d+1)} = \infty$. This grid, whose *refinement* is defined as $R = \prod_{d=1}^{D} R_d$, divides the sample space

into $B = \prod_{d=1}^{D}(R_d + 1)$ real intervals of dimension D. Each interval (or *bin*) is defined as $\mathcal{B}_b = \prod_{d=1}^{D}[a_{db_d}, a_{d(b_d+1)})$, where $(b_1, \ldots, b_D)$ is the vector containing the numerical indices varying in $\prod_{d=1}^{D}\{0, \ldots, R_d\}$ and satisfying the relation

$$b = 1 + b_1 + \sum_{d=2}^{D} b_d \prod_{d'=1}^{d-1}(R_{d'} + 1).$$

2.3 Curse of dimensionality for binned data

In a univariate context this methodology works well if $B = R + 1 \ll n$, where R is the refinement of the only grid considered. But, we have to point out the arising of some issues when D increases. Indeed, as the number of non-empty bins depends exponentially on the dimension D, it is impossible to obtain a manageable amount of binned data, as depicted in Figure 1. Thus, in the D-dimensional context, a *classical* approach with binned data vanishes any kind of gain.

3 Bin-marginal model

3.1 Compressed binned data: bin-marginal solution

In the previous section we pointed out the storage issues linked to a classical use of binned data. Our first idea consists in using what we call *marginal counts*, that are the collection of binned data obtained on each dimension *separately*. In the present section we illustrate a full likelihood estimation of the model generating marginal counts, highlighting its complexity, which motivates completely our final proposal in Section 4 based on an alternative composite likelihood approach.

Let define $\boldsymbol{m} = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_D\}$, where $\boldsymbol{m}_d$ is the binned data vector referring to the projection on the axis d of the observations $\boldsymbol{x}_i$ after imposing the grid G_d, which produces $B_d = R_d + 1$ bins. It means that, for each $d = 1, \ldots, D$, $\boldsymbol{m}_d = (m_{d1}, \ldots, m_{dB_d})$, where each component is defined as $m_{db_d} = \#\{x_{id} : a_{d(b_d-1)} \leq x_{id} < a_{db_d}\}$ and x_{id} is the d-th component of $\boldsymbol{x}_i$. Thus, the collection $\boldsymbol{m}$ contains the *marginal counts* of $\boldsymbol{n}$. To facilitate the comprehension of the specific data compression mechanism and its related notation, a simple bivariate situation is depicted in Figure 2. Here, a 3×3 grid overlaps 20 raw individuals $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{20})$ and both the bivariate binned data $\boldsymbol{n}$ and marginal counts $\boldsymbol{m}$ are highlighted.

The introduction of marginal counts makes resource savings possible: in fact, it is clear that storing them instead of the full grid is convenient for computer memory, as we have to save at most $\sum_{d=1}^{D} B_d$ elements instead of $\prod_{d=1}^{D} B_d$ ones. So, a first attempt could be the estimation of the *bin-marginal* model whose probability mass function is:

$$p_m(\boldsymbol{m};\ \boldsymbol{\psi}) = \sum_{\boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{m}}} p(\boldsymbol{n}';\ \boldsymbol{\psi}), \tag{2}$$

where $\mathcal{F}_{\boldsymbol{m}}$ is the set of tables $\boldsymbol{n}'$ sharing the same marginals $\boldsymbol{m}$. Formally:

$$\mathcal{F}_{\boldsymbol{m}} = \{\boldsymbol{n}' : \boldsymbol{m}' = \boldsymbol{m}\},$$

where $\boldsymbol{m}'$ are the marginal counts of each table $\boldsymbol{n}'$.

But now we need to assess three important issues before proposing this model as a useful frugal method:

- *Identifiability of the model.* We wonder if different parameters index different bin-marginal probability mass functions. This question will be treated in Section 3.2.
- *Mathematical complexity of the log-likelihood* $\ell_m(\boldsymbol{\psi};\ \boldsymbol{m}) = \log p_m(\boldsymbol{m};\ \boldsymbol{\psi})$. From (2) we note that the computation of this log-likelihood is intractable, because we need to calculate a considerable number of complete tables. Section 4 will be dedicated to overcome this specific issue.
- *Optimization of the likelihood.* In Section 3.3 we give a version of the EM algorithm to do this task. We will show it does not solve all the issues appeared in 2.3 and, again, Section 4 will propose a specific solution.

3.2 Requirements for identifiability

Typically, before proceeding with the estimation of any statistical model $\mathcal{P} = \{p(\boldsymbol{x};\ \boldsymbol{\psi}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\psi} \in \Psi\}$, statisticians are interested in knowing if it is *identifiable*, i.e. if any different value of the model parameter $\boldsymbol{\psi}$ indexes different elements in $\mathcal{P}$. In case of continuous model, these elements are

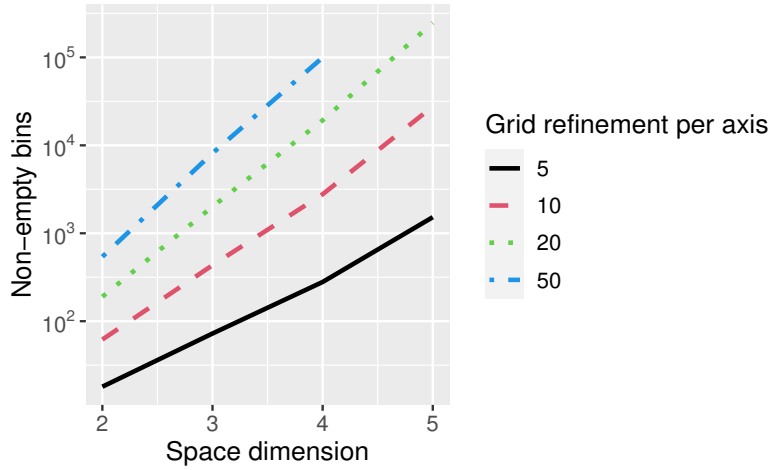


Figure 1: Number of non-empty bins depending on both space dimension D and grid refinement (per axis) generated by a single D-variate standard Gaussian.

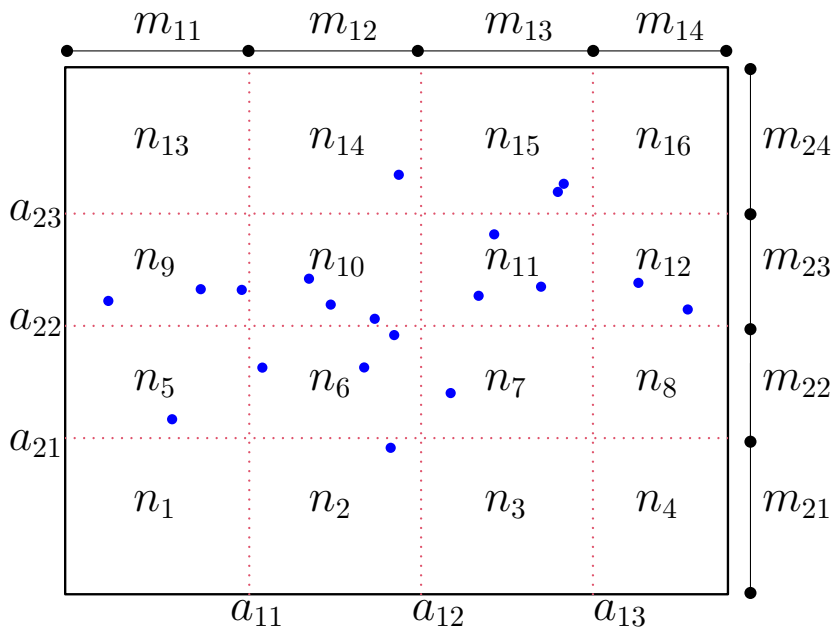


Figure 2: Bivariate representation of a 3×3 grid (red dotted lines) superposing on 20 points $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{20})$ (in blue). Bivariate binned data are $\boldsymbol{n} = (n_1, \ldots, n_{20})$, while marginal counts are $\boldsymbol{m} = \{\boldsymbol{m}_1, \boldsymbol{m}_2\}$, where $\boldsymbol{m}_1 = (m_{11}, \ldots, m_{14})$ and $\boldsymbol{m}_2 = (m_{21}, \ldots, m_{24})$.

densities, while they are probability mass functions if the model is discrete, as in our binned data case. Gaussian mixtures are identifiable up to a labelling permutation (Yakowitz and Spragins 1968), but this is proved only in the raw data case. Surprisingly, to the best of our knowledge, there is no reference to Gaussian mixtures identifiability with binned data, neither in the seminal works of McLachlan and Jones (1988) and Cadez et al. (2002), which pass directly to the estimation phase. In this section we cover partially this lack, giving some conditions on the grid assuring identifiability, both in full binned data and bin-marginal cases and under hypothesis of diagonal covariance matrices. However, this apparent restriction does not affect our proposal, because this assumption is common in several clustering approaches, even for the raw data case, as k-means (MacQueen et al. 1967) and parsimonious Gaussian mixture models (Celeux and Govaert 1995), and because, in Section 4, our proposal will be presented under these conditions. As regarding the identifiability of the complete model, intuitively a quite dense grid could be sufficient, because the binned model tends to the raw one when the number of cut points goes to infinity. However, a precise theoretical result could be useful to know what is the coarsest identifiable grid. This is important, especially in our context, where coarser grids mean bigger memory.

In case of mixture models with full binned data, given the density $p(\cdot;\ \boldsymbol{\psi})$ indexed by a parameter $\boldsymbol{\psi}$ belonging to the parametric space Ψ, generic identifiability is assured if, up to a null measure set:

$$\forall \boldsymbol{\psi}, \boldsymbol{\psi}^* \in \Psi: \quad p(\boldsymbol{n};\ \boldsymbol{\psi}) = p(\boldsymbol{n};\ \boldsymbol{\psi}^*)\ \ \forall G, \boldsymbol{n} \\ \Rightarrow \boldsymbol{\psi} = \boldsymbol{\psi}^*. \tag{3}$$

In case of mixture models with marginal binned data the previous statement becomes

$$\forall \boldsymbol{\psi}, \boldsymbol{\psi}^* \in \Psi: \quad p_m(\boldsymbol{m};\ \boldsymbol{\psi}) = p_m(\boldsymbol{m};\ \boldsymbol{\psi}^*)\ \ \forall G, \boldsymbol{m} \\ \Rightarrow \boldsymbol{\psi} = \boldsymbol{\psi}^*. \tag{4}$$

The identifiability of D-variate binned mixture models is assured by the following proposition, whose proof is given in Appendix A.

Proposition 1 (**Full binned Gaussian diagonal mixtures**). *Under hypothesis of diagonal covariance matrices, binned D-variate mixtures of K components are identifiable if $R_d > 4K - 3$, $d = 1, \ldots, D$, up to labels permutations.*

Proposition 1 is important not only for its prominence in the field, but also because it is crucial for proving identifiability of bin-marginal Gaussian mixtures themselves. Indeed, Proposition 2 establishes below that bin-marginal mixtures are identifiable if binned mixtures are identifiable (proof in Appendix A). This result is of central interest in this work, since we will consider only the bin-marginal data in order to preserve computer memory.

Proposition 2 (**Bin-marginal Gaussian diagonal mixtures**). *Bin-marginal D-variate mixtures of K components are identifiable if binned D-variate mixtures are identifiable. So, under diagonal covariance matrices hypothesis, identifiability is achieved if $R_d > 4K - 3$, $d = 1, \ldots, D$, up to labels permutation.*

3.3 EM algorithm

It is possible to formulate a specific EM algorithm in order to maximize the bin-marginal log-likelihood $\ell_m(\boldsymbol{m};\ \boldsymbol{\psi}) = \log p_m(\boldsymbol{m};\ \boldsymbol{\psi})$ associated to the bin-marginal dataset. Therefore, we introduce the *complete log-likelihood*

$$\ell^c(\boldsymbol{\psi};\ \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log(\pi_k \phi(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)),$$

where $\mathbf{z}$ is an $n \times K$ matrix whose generic element z_{ik} is equal to 1 if $\boldsymbol{x}_i$ belongs to population k, it is 0 otherwise. Thus, $\mathbf{z}$ contains the hidden class membership of the raw data $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. More precisely, at each iteration $j \geq 0$, given the current estimate $\boldsymbol{\psi}^{(j)}$, the complete log-likehood is used in the so-called E-step, where the following quantity is calculated

$$Q_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)}) = \mathbb{E}_{\boldsymbol{\psi}^{(j)}}[\ell^c(\boldsymbol{\psi};\ \mathbf{X}, \mathbf{Z}) | \boldsymbol{m}], \tag{5}$$

taking the expectation with respect to $p(\mathbf{x}, \mathbf{z} | \boldsymbol{m};\ \boldsymbol{\psi}^{(j)})$. Note that $\mathbf{X}$ and $\mathbf{Z}$ denote, respectively, the random variables generating $\mathbf{x}$ and $\mathbf{z}$.

Let rewrite (5):

$$Q_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)}) = \sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} p(\boldsymbol{n}|\boldsymbol{m};\ \boldsymbol{\psi}^{(j)}) \mathbb{E}_{\boldsymbol{\psi}^{(j)}}[\ell^c(\boldsymbol{\psi};\ \mathbf{X}, \mathbf{Z})|\boldsymbol{n}]$$

where

$$p(\boldsymbol{n}|\boldsymbol{m};\ \boldsymbol{\psi}^{(j)}) = \frac{p(\boldsymbol{n};\ \boldsymbol{\psi}^{(j)})}{\sum_{\boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{m}}} p(\boldsymbol{n}';\ \boldsymbol{\psi}^{(j)})} \mathbb{1}_{\{\boldsymbol{n} \in \mathcal{F}_m\}}. \tag{6}$$

After few calculus this expression reduces to:

$$Q_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)}) = \sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} p(\boldsymbol{n}|\boldsymbol{m};\ \boldsymbol{\psi}^{(j)}) \times \sum_{k=1}^{K} \sum_{b=1}^{B} n_b \mathbb{E}_b[\tau_k^{(j)}(\mathbf{X}) \log[\pi_k \phi(\mathbf{X};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]]$$

where $\mathbb{E}_b$ refers to the expectation with respect to the density $g_b^{(j)}(\boldsymbol{x}) = \frac{f(\boldsymbol{x};\, \boldsymbol{\psi}^{(j)})}{\int_{\mathcal{B}_b} f(\boldsymbol{y};\, \boldsymbol{\psi}^{(j)}) d\boldsymbol{y}}$ and $\tau_k^{(j)}(\boldsymbol{x}) = \frac{\pi_k^{(j)} \phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)})}{f(\boldsymbol{x};\, \boldsymbol{\psi}^{(j)})}$. Before proceeding with the M-step, we introduce the following quantities to simplify the notations:

$$\begin{aligned}
\alpha^{(j)}(\boldsymbol{n}) &= p(\boldsymbol{n}|\boldsymbol{m};\ \boldsymbol{\psi}^{(j)}) \\
A_{kb}^{(j)} &= \int_{\mathcal{B}_b} \tau_k^{(j)}(\boldsymbol{x}) g_b^{(j)}(\boldsymbol{x}) d\boldsymbol{x} \\
B_{kb}^{(j)} &= \int_{\mathcal{B}_b} \boldsymbol{x} \tau_k^{(j)}(\boldsymbol{x}) g_b^{(j)}(\boldsymbol{x}) d\boldsymbol{x} \\
C_{kb}^{(j)} &= \int_{\mathcal{B}_b} (\boldsymbol{x} - \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)^t \tau_k^{(j)}(\boldsymbol{x}) g_b^{(j)}(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

Then, in the M-step we maximize $Q_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)})$, obtaining the following update formulas for each component $k = 1, \ldots, K$:

$$\begin{aligned}
\pi_k^{(j+1)} &= \frac{1}{n} \sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} \alpha^{(j)}(\boldsymbol{n}) \sum_{b=1}^{B} n_b A_{kb}^{(j)} \\
\boldsymbol{\mu}_k^{(j+1)} &= \frac{\sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} \alpha^{(j)}(\boldsymbol{n}) \sum_{b=1}^{B} n_b B_{kb}^{(j)}}{\sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} \alpha^{(j)}(\boldsymbol{n}) \sum_{b=1}^{B} n_b A_{kb}^{(j)}} \\
\boldsymbol{\Sigma}_k^{(j+1)} &= \frac{\sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} \alpha^{(j)}(\boldsymbol{n}) \sum_{b=1}^{B} n_b C_{kb}^{(j)}}{\sum_{\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{m}}} \alpha^{(j)}(\boldsymbol{n}) \sum_{b=1}^{B} n_b A_{kb}^{(j)}}.
\end{aligned}$$

Unfortunately, both previous E and M steps involve the computation of all "crossed" tables $\mathcal{F}_m$ sharing the same marginals, coming back to a memory issue (and also a time computation one). Therefore, an estimation based on the full likelihood of the bin-marginal model is not numerically tractable. For this very reason we will provide in the following section estimates following a composite likelihood approach, after having given a brief introduction of this concept.

4 Estimation strategy

In this section we present the estimation part of our contribution, working with diagonal Gaussian mixtures (i.e., matrices $\boldsymbol{\Sigma}_k$ in (1) are diagonal). Before, it is necessary to briefly introduce the *marginal composite likelihood*, on which our estimation proposal is based.

4.1 Marginal composite likelihood

Marginal composite likelihood is a pseudo-likelihood used to obtain asymptotically consistent estimates (see Varin et al. (2011) for instance) when the optimization of the full likelihood is too burdensome. The marginal composite likelihood relies only on univariate marginal likelihoods and it is a special case of *composite likelihood* (Lindsay 1988), where more general multivariate marginal likelihoods can take into account.

Let $\mathbf{x}$ be a D-dimensional sample with n observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})$, $i = 1, \ldots, n$, generated by a Gaussian diagonal mixture model with parameter $\boldsymbol{\psi} \in \Psi$, as in Section 2. Denoting with $\mathbf{x}_d = (x_{1d}, \ldots, x_{nd})$ the component d of the whole raw dataset, $L_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d)$ is the likelihood of the univariate Gaussian mixture at dimension d with parameter $\boldsymbol{\psi}_d = (\pi_1, \ldots, \pi_K, \mu_{1d}, \ldots, \mu_{Kd}, \sigma_{1d}^2, \ldots, \sigma_{Kd}^2)$. Then, the *marginal composite likelihood* is defined as

$$\tilde{L}(\boldsymbol{\psi};\ \mathbf{x}) = \prod_{d=1}^{D} L_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d).$$

Similarly, the *marginal composite log-likelihood* is $\tilde{\ell}(\boldsymbol{\psi};\ \mathbf{x}) = \sum_{d=1}^{D} \ell_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d)$, with $\ell_d(\psi_d;\ \mathbf{x}_d) = \log L_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d)$.

The estimator $\tilde{\boldsymbol{\psi}}$ maximizing $\tilde{L}(\boldsymbol{\psi};\ \mathbf{x})$ is named *maximum marginal composite likelihood estimator*. It has proved to be consistent and asymptotically normally distributed under very

mild conditions about the regularity of the marginal densities (see Molenberghs and Verbeke (2005) for instance).

4.2 Bin-marginal composite likelihood

Having given the necessary notation in the previous paragraphs, we can now complete our proposal, in which we will combine the memory reduction offered by bin-marginal data with the computational advantages of marginal composite likelihood. Actually, more general but less frugal versions of composite likelihood have already been used in the area of mixture models. Indeed, a formalization of EM algorithm with composite likelihood could be seen in Gao and Song (2011). They also established three fundamental properties of the associated so-called CL-EM algorithm: ascent property, convergence to a stationary point and a quantification of its rate of convergence. An application of composite likelihood on binned data appeared in Ranalli and Rocci (2016), where these ones arose from a discrete data problem. This is quite similar to the technique we are about to describe, but it is different as it uses bivariate grids and it does not build artificially binned data as a solution for scalability, because they were already given in the problem statement.

Assuming a marginal D-dimensional Cartesian grid G as defined in Section 3 and diagonal covariance matrices, instead of maximizing the too complex bin-marginal log-likelihood $\ell_m(\boldsymbol{\psi};\ \boldsymbol{m})$, we aim to maximize the following bin-marginal composite log-likelihood:

$$\begin{aligned}\tilde{\ell}_m(\boldsymbol{\psi};\ \boldsymbol{m}) &= \sum_{d=1}^{D} \ell_d(\boldsymbol{\psi}_d;\ \boldsymbol{m}_d) \\ &= \sum_{d=1}^{D}\sum_{b_d=1}^{B_d} m_{db_d} \log\Big(\int_{\mathcal{B}_{b_d}^d} f_d(x_d;\ \boldsymbol{\psi}_d)dx_d\Big).\end{aligned} \tag{7}$$

Here, $\ell_d(\boldsymbol{\psi}_d;\ \boldsymbol{m}_d)$ is the binned log-likelihood for a univariate Gaussian mixture with K components of density $f_d(x_d;\ \boldsymbol{\psi}_d)$ indexed by the parameter $\boldsymbol{\psi}_d$. The expression of (7) motivates why we work with diagonal mixtures: it is impossible to estimate any kind of covariance parameter, since none of them appear in $\boldsymbol{\psi}_d$.

4.3 Properties of the bin-marginal composite likelihood

The asymptotic identifiability of the optimization criterion for the maximum marginal composite likelihood estimator (i.e., the asymptotic criterion is maximized at the unique value of the true parameter) is a necessary condition to prove its consistency (Wald 1949; Lindsay 1988). In this section we prove that this property is fulfilled almost everywhere, except in a null measure set, as the following Proposition 3 assures (proof in Appendix A). Moreover, the same proposition also defines precisely the null measure set, which turns out to be composed by those mixtures with two equal proportions or two components sharing the projection on the same axis.

Proposition 3. *Assuming the true model is outside the null measure set* $\Psi^{**} = \Pi'_K \times \mathbb{R}^{2k} \times \mathbb{R}^{+2k} \cup \Psi'$ *, where* $\Pi'_K = \{\boldsymbol{\pi} \in \Pi_K : \exists i,j : \pi_i = \pi_j\}$ *and* $\Psi' = \{\psi \in \Psi : \exists k,k',d : \mu_{kd} = \mu_{k'd}, \sigma^2_{kd} = \sigma^2_{k'd}\}$*, the optimization criterion of the bin-marginal composite log-likelihood, using a grid* $G = G_1 \times \ldots \times G_d$ *with* $\prod_{d=1}^{D} R_d$ *cut points is asymptotically identifiable if* $R_d > 4K - 3$*,* $d = 1, \ldots, D$ *up to labels permutation.*

4.4 Bin-marginal CL-EM algorithm

We can now maximize (7) using an EM-like approach. At each data $\boldsymbol{m}_d, d = 1, \ldots, D$ we associate the *missing* vectors $(\mathbf{x}_d, \mathbf{z}_d), d = 1, \ldots, D$, where $\mathbf{x}_d$ contains the component d of the raw data $\mathbf{x}$ and $\mathbf{z}_d$ is the indicator membership matrix for $\mathbf{x}_d$. Thus, it is an $n \times K$ matrix whose generic element z_{dik} is equal to 1 if $\boldsymbol{x}_{id}$ belongs to population k, 0 otherwise.

To simplify the notation, we set $\tilde{\mathbf{z}} = \{\mathbf{z}_1, \ldots, \mathbf{z}_D\}$: the couple $(\mathbf{x}, \tilde{\mathbf{z}})$ is named *complete* data. Then, we introduce the *complete marginal composite log-likelihood*:

$$\tilde{\ell}^c_m(\boldsymbol{\psi};\ \mathbf{x}, \tilde{\mathbf{z}}) = \sum_{d=1}^{D} \ell^c_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d, \mathbf{z}_d), \tag{8}$$

where $\ell^c_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d, \mathbf{z}_d)$ denotes the complete log-likelihood for the d-th marginal couple of data $(\mathbf{x}_d, \mathbf{z}_d)$.

Algorithm 1 Bin-marginal CL-EM algorithm for D-dimensional Gaussian diagonal mixtures

1. Initialization phase: provide an initial guess $\boldsymbol{\psi}^{(0)}$.
2. For $j \geq 0$:

- **Binned CL-E Step:** Given the estimate $\boldsymbol{\psi}^{(j)}$, calculate $\tilde{Q}_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)})$;
- **Binned CL-M Step:** Obtain the new estimate $\boldsymbol{\psi}^{(j)}$, maximizing $\tilde{Q}_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)})$.

For $d = 1, \ldots, D$:
For $b_d = 1, \ldots, B_d$:

$$g^{(j)}_{db_d}(x_d) = \frac{f(x_d;\ \psi^{(j)}_d)}{\int_{\mathcal{B}^d_{b_d}} f(y_d;\ \psi^{(j)}_d) dy_d}$$

For $k = 1, \ldots, K$ and $d = 1, \ldots, D$:

$$\tau^{(j)}_{kd}(x_d) = \frac{\pi^{(j)}_k \phi(x_d;\ \mu^{(j)}_{kd}, \sigma^{2(j)}_{kd})}{f(x_d;\ \psi^{(j)}_d)}$$

$$\pi^{(j+1)}_k = \frac{\sum_{d=1}^{D} \sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}^d_{b_d}} \tau^{(j)}_{kd}(x_d) g^{(j)}_{db_d}(x_d) dx_d}{Dn}$$

$$\mu^{(j+1)}_{kd} = \frac{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}^d_{b_d}} x_d \tau^{(j)}_{kd}(x_d) g^{(j)}_{db_d}(x_d) dx_d}{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}^d_{b_d}} \tau^{(j)}_{kd}(x_d) g^{(j)}_{db_d}(x_d) dx_d}$$

$$\sigma^{2^(j+1)}_{kd} = \frac{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}^d_{b_d}} (x_d - \mu^{(j)}_{kd})^2 \tau^{(j)}_{kd}(x_d) g^{(j)}_{db_d}(x_d) dx_d}{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}^d_{b_d}} \tau^{(j)}_{kd}(x_d) g^{(j)}_{db_d}(x_d) dx_d}$$

Stop if (9) is verified, continue otherwise.

At iteration $j \geq 0$, $\boldsymbol{\psi}^{(j)}$ denotes the current estimate for $\boldsymbol{\psi}$. Then, denoting respectively with $\mathbf{X}_d$ and $\mathbf{Z}_d$ the random variables generating $\mathbf{x}_d$ and $\mathbf{z}_d$, we now define the quantity:

$$\tilde{Q}_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)}) = \sum_{d=1}^{D} \mathbb{E}_{\boldsymbol{\psi}^{(j)}_d}[\ell^c_d(\boldsymbol{\psi}_d;\ \mathbf{X}_d, \mathbf{Z}_d)|\boldsymbol{m}_d],$$

where the expectations are taken with respect to the conditional densities $f(\mathbf{x}_d, \mathbf{z}_d|\boldsymbol{m}_d;\ \boldsymbol{\psi}^{(j)}_d)$, $d = 1, \ldots, D$.

Let re-write $\tilde{Q}_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)})$, indicating with $\mathcal{X}_d \times \mathcal{Z}_d$ the integration domain of $(\mathbf{x}_d, \mathbf{z}_d)$. We have

$$\tilde{Q}_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j)}) = \sum_{d=1}^{D} \int_{\mathcal{X}_d \times \mathcal{Z}_d} \ell^c_d(\boldsymbol{\psi}_d;\ \mathbf{x}_d, \mathbf{z}_d) \times f(\mathbf{x}_d, \mathbf{z}_d|\boldsymbol{m}_d;\ \boldsymbol{\psi}^{(j)}_d) d\mathbf{x}_d d\mathbf{z}_d.$$

Now, we can define our bin-marginal CL-EM algorithm, whose fundamental steps are resumed in Algorithm 1. Therein $\mathcal{B}^d_{b_d}$ indicates the b_d-th interval bin on the $d-$th dimension.

Initialization

We adopt a uniform random initialization for proportions, means and variances. In particular, for each dimension, means are values extracted from the range of values of the data and variances are

positive uniform values lower than the variance of the data.

Stopping rule

Binned CL-EM algorithm stops as soon as

$$\left| \frac{\tilde{\ell}_m(\boldsymbol{\psi}^{(j)};\, \boldsymbol{m}) - \tilde{\ell}_m(\boldsymbol{\psi}^{(j)};\, \boldsymbol{m})}{\tilde{\ell}_m(\boldsymbol{\psi}^{(j)};\, \boldsymbol{m})} \right| < \epsilon, \qquad (9)$$

where ϵ is a chosen threshold.

Obtaining the final clustering partition

Once obtained the final estimate of $\boldsymbol{\psi}$ provided by our CL-EM algorithm, namely $\hat{\boldsymbol{\psi}}$, we recover the final clustering partition using a maximum a posteriori probability (MAP) rule. It means that the estimated labels $\hat{\boldsymbol{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$ are given by:

$$\hat{z}_i = \underset{1 \le k \le K}{\arg\max}\, \hat{\pi}_k \phi(\boldsymbol{x}_i;\, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \qquad i = 1, \ldots, n.$$

We highlight this algorithm involves only D binned vectors of dimension $B_d = R_d + 1$, $d = 1, \ldots, D$ and only univariate integrals. Thus, our proposal is able to solve our initial issues linked to storage and complexity.

5 Numerical experiences on simulated data

In this section we apply the methodology to different simulated datasets in order to show in controlled frameworks its ability to recognize the minority class.

Our second aim is also to compare it to two possible competitors: classic estimation with the full dataset and a subsampling strategy. We will evaluate their performances in terms of clustering quality, measured by the ARI score (Hubert and Arabie 1985), and also in terms of both time and memory consumption. In particular, the full dataset will be our benchmark in terms of clustering quality, but it will be discarded as it is too much burdensome. The subsampling will prove to cope with our computational constraints, but resulting usually in bad clustering performances or in, even, estimation failures.

5.1 Experimental settings

Simulation analyses are conducted on datasets with 1 million data generated from several 3-dimensional two classes mixtures, different in proportions assigned to the minority class and also in means, while both covariance matrices remain equal to the identity matrix. These differences are crucial because lowering proportion of the smallest class corresponds to more difficulties in detecting it and changing means helps us in controlling classes separation and, thus, clustering complexity.

We divide our simulations into two main parts: in the first one, cluster separation is equal for all axes, while, in the second one, clusters are well separated only on one axis, while on the other two they are not. This is useful to understand the degree of separation needed by our technique. In particular, in the first part, we gradually increase the small class proportion three times from 10^{-4} to 10^{-2} and we also propose four separation degrees for cluster means, equal to $8, 6, 4, 2$ in terms of absolute difference between them. Their combination results in twelve different scenarios. Each scenario is named by using two letters: the first one (H, M, L, V) refers to the degree of separation of the scenario (respectively: high, medium, low and very low); the second one (H, M, L) refers to the imbalance of the dataset (high, medium and low). Three additional scenarios consist in a variation of scenarios HH-HM-HL where the first two dimension have the lowest separation degree, while there is a high separation on the third axis. Their names are 1HH-1HM-1HL, reminding that here high separation is present only on one axis. Table 1 details all these fifteen settings.

Regarding the three analyzed methods, we decide to compare subsampling and our bin-marginal proposal under the same memory constraints. Bin marginal uses a grid refinement R, leading to use a $2R$ memory space (binned data itself and grid); hence, subsampling is conducted with a subsample of size $2R$ to be fair. At the same time, we also analyze the influence of the grid refinement on the binned estimation and, consequently, the effect of the subsample size on the subsampling performance. In practice, the refinement can be fixed to $50, 100, 200$ and, consequently, subsample sizes can be $100, 200$ or 400. For each scenario, we simulated 20 different

Table 1: Description of the fifteen scenarios. Covariance matrices $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ are equal to the identity matrix $\boldsymbol{I}_3$ and $\pi_2 = 1 - \pi_1$.

Scenario	Separation	Imbalance	Small class proportion (π_1)	Means
HH	High	High	10^{-4}	$\boldsymbol{\mu}_1 = (-4,-4,-4)$ $\boldsymbol{\mu}_2 = (4,4,4)$
HM		Medium	10^{-3}	
HL		Low	10^{-2}	
MH	Medium	High	10^{-4}	$\boldsymbol{\mu}_1 = (-3,-3,-3)$ $\boldsymbol{\mu}_2 = (3,3,3)$
MM		Medium	10^{-3}	
ML		Low	10^{-2}	
LH	Low	High	10^{-4}	$\boldsymbol{\mu}_1 = (-2,-2,-2)$ $\boldsymbol{\mu}_2 = (2,2,2)$
LM		Medium	10^{-3}	
LL		Low	10^{-2}	
VH	Very low	High	10^{-4}	$\boldsymbol{\mu}_1 = (-1,-1,-1)$ $\boldsymbol{\mu}_2 = (1,1,1)$
VM		Medium	10^{-3}	
VL		Low	10^{-2}	
1HH	One separated component	High	10^{-4}	$\boldsymbol{\mu}_1 = (-1,-1,-4)$ $\boldsymbol{\mu}_2 = (1,1,4)$
1HM		Medium	10^{-3}	
1HL		Low	10^{-2}	

datasets of equal size (1 million) to have consistent results. To evaluate its variability, subsampling performances are evaluated on 100 different subsamples. Practical implementation, both of simulations and real application as well, was done in the R environment (R Core Team 2021). More precisely, we used the routines of the R package `Rmixmod` (Lebret et al. 2015) for the two competitors and a self-written code for our bin-marginal technique. We have chosen `Rmixmod` because its initialization phase is stochastic, enabling a better exploration of the parametric space. For this reason `Rmixmod` was preferred to concurrency, notably to `mclust` (Fraley et al. 2012), whose initialization is based on a deterministic hierarchic clustering.

5.2 Results

Clustering quality and memory

Figures 3a-3o depict the results of the simulations. Mostly, our proposal outperforms subsampling in all the settings with good performance even with the coarser grid. It encounters some difficulties only in very hard scenarios where separation and proportion are very small. Generally, it approaches with a low consumption the results obtained with the full dataset, which, on the contrary, uses a huge amount of memory.

Failures

There is another virtue in binned strategy: in fact, subsampling can fail, as reported in Figure 4. It appears the probability of failure increases if separation increases and imbalance ratio decreases. This is quite astonishing, as we expected more failures in a more imbalanced dataset, but it is not completely incoherent: in fact, results show that if subsampling does not fail (high imbalancement) it works badly; if on the contrary it can provide good results, it is prone to failures (low imbalancement). In most of scenarios, failures surprisingly increase according to subsample size. At this moment, we are not able to explain exactly the reason of this unusual behaviour, that probably could be resolved by changing the initial settings of EM (implemented in `Rmixmod`). But, fortunately, this does not affect directly our proposal based on binned data.

Time

Finally, Figure 5 shows time performances for the three strategies. Our CL-EM algorithm does not

outperform subsampled EM in execution time, while it is faster than full dataset EM. This result is coherent with our expectations. Indeed, even if both CL-EM and classic EM are linear with respect to input size (R and n respectively), the operations executed by CL-EM are more complex due to the presence of integrals (see Algorithm 1). Thus, if R and n are comparable (subsampling case), CL-EM is slower than classic EM, while it is faster if $R \ll n$ (full dataset case). In analyzing Figure 5 we also have to point out that the `Rmixmod` package is well-optimized and written using Rcpp, which enables integration between R and C++, while our code is completely written in R and it may be improved even in the phase of binning. According to Aruoba and Fernández-Villaverde (2015), Rcpp is faster than R about 100 times, so our time performances has to be scaled of at least a factor 100. The figure itself pictures our predicted performance after code optimization (blue boxplots), showing a remarkable improvement relatively to full dataset analysis.

6 Real datasets

The presented methodology is now applied to several real imbalanced datasets. Here we show three applications from different fields of interest, which are image segmentation, fraud detection and recognition of potentially hazardous asteroids. In the last two cases, we have considered a subset of three variables for each dataset. We have chosen those ones whose histograms visually resulted to be close to GMM hypotheses and with a low percentage of missing values (less than the 5% of the original data). A comprehensive view of the used datasets is given in Table 2.

Table 2: Real datasets description.

Dataset	n	D	Small class proportion
Cell-1	101,430	3	unknown
Cell-2	65,536	3	unknown
Cell-3	685,020	3	unknown
Comet	1,083,681	3	unknown
Asteroids	932,341	3	0.002
Credit card	284,807	3	0.0014

6.1 Datasets and methods

Image segmentation

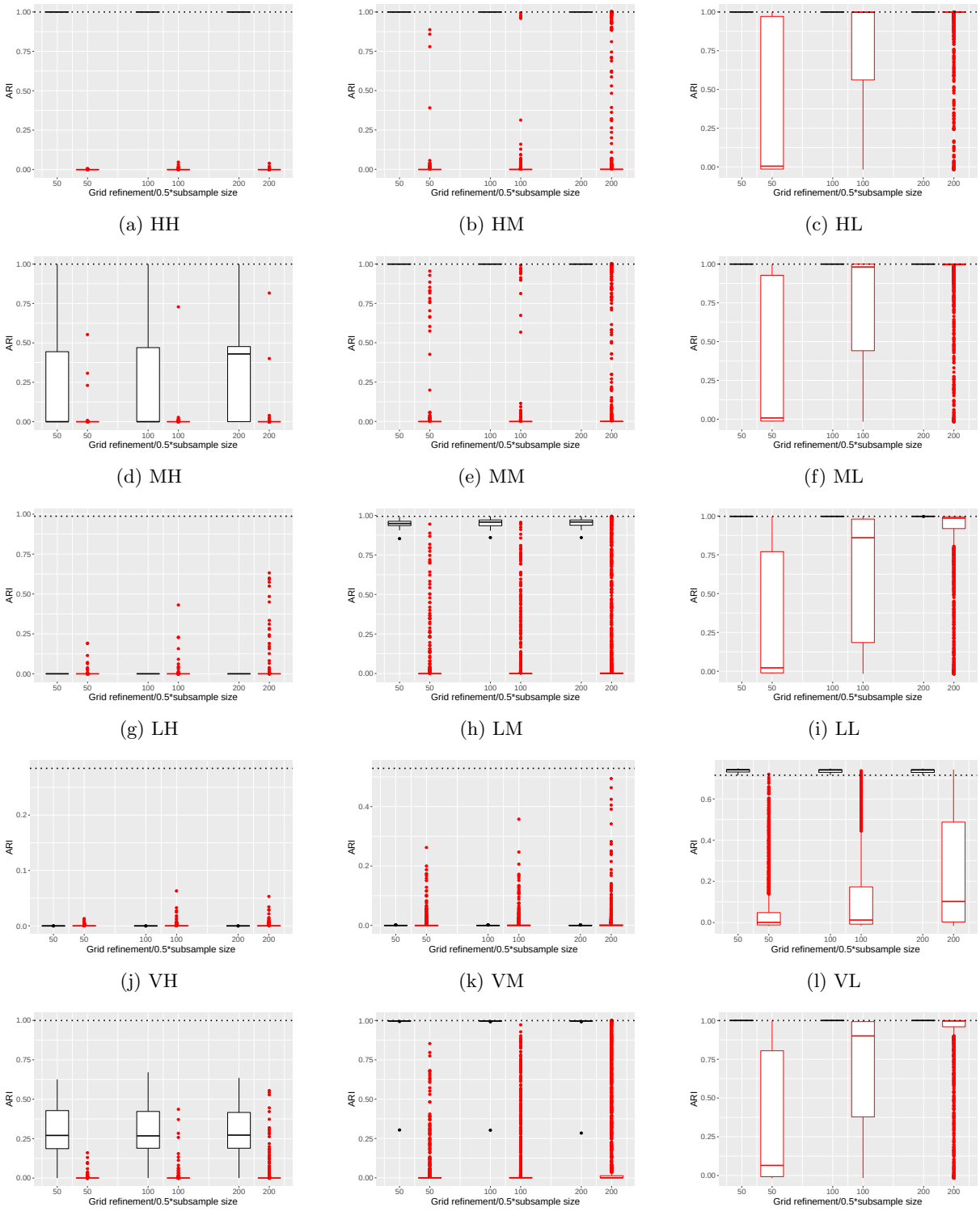
Image segmentation (Pal and Pal 1993) consists in partitioning an image into homogeneous parts and it is useful to detect and locate objects. Here we focus on those images where there are very tiny objects: for this purpose we segment three cell images available on Kaggle (To 2021) and an image picturing a distant active comet observed by NASA's Hubble Space Telescope (NASA 2017). After a brief pre-processing phase, these images result in 3-dimension datasets with a number of records ranging from 65,536 to 1,083,681. The lines of these datasets correspond to RGB pixels, that could be analyzed with our method.

Asteroids

Asteroid dataset is a collection of information about asteroids available on Kaggle (Hossain 2020). It consists in 958,524 records of 45 variables. The purpose of the analysis is to detect potentially hazardous asteroids (PHAs), which are those asteroids approaching very close to the Earth. In particular, an asteroid with small magnitude (H) and Earth minimum orbit intersection distance (moid) is considered a PHA (Quarta and Mengali 2010). We use only a subset of the features contained in this dataset, using these two variables and adding information regarding orbit eccentricity in order to remain in a more interesting 3-dimensional problem where our method has already been tested in the simulation phase. Due to the presence of missing values, the analyzed dataset contains now 932,341 records out of 958,524. The rest of the variables were discarded because they contain too many missing values (less than the 5% of the original data) and their histograms were judged not to be close to GMM hypothesis.

Credit card fraud detection

Kaggle credit card dataset (ULB 2018) is a public repository which was massively analyzed in literature (Dal Pozzolo et al. 2017, 2014; Niu et al. 2019) to detect frauds. This dataset contains 284,807 transactions, of which 492 are frauds, made by credit cards in September 2013 by European cardholders. All information given by 31 variables are anonymized and they are the result



(a) HH (b) HM (c) HL

(d) MH (e) MM (f) ML

(g) LH (h) LM (i) LL

(j) VH (k) VM (l) VL

(m) 1HH (n) 1HM (o) 1HL

Figure 3: Clustering performances for subsampled EM (red boxplots) and bin-marginal CL-EM (black boxplots), expressed in terms of ARI in dependence on grid refinement/subsample size under condition of equal memory occupancy. Dotted lines represent full dataset performances. Imbalance is decreasing from left to right and separation is decreasing from top to bottom.

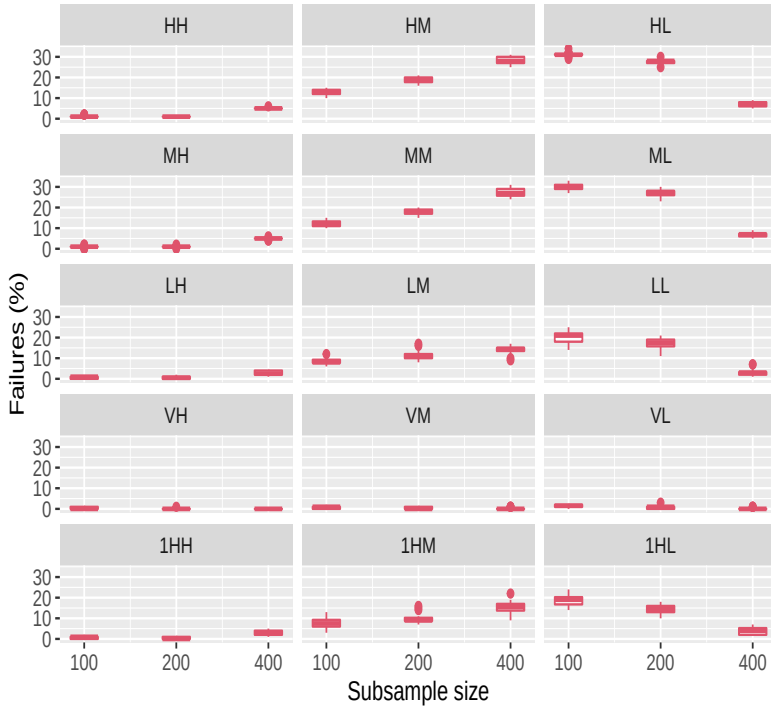


Figure 4: Percentage of subsampled EM failures.

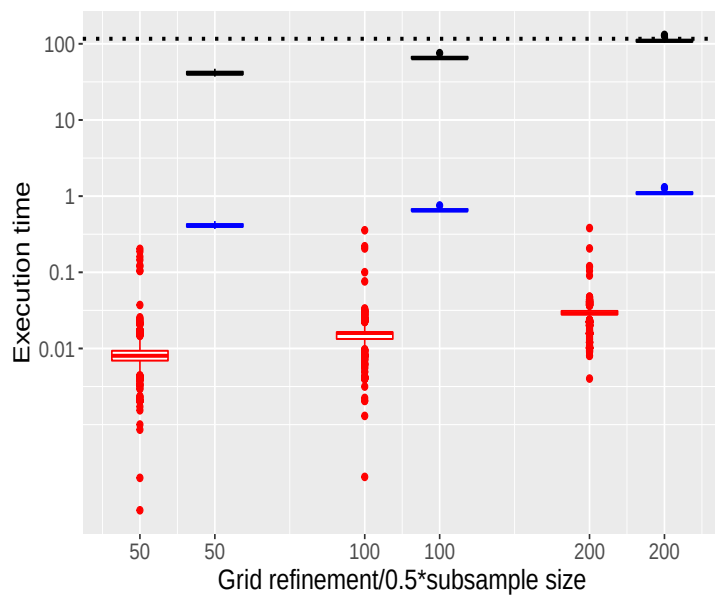


Figure 5: Scenario HH: execution time (in s) comparison between subsampled EM (red boxplots) and bin-marginal CL-EM (black boxplots) in dependence on grid refinement/subsample size in condition of equal memory occupancy. Blue boxplots show expected CL-EM time after optimization in language C++, while dotted line represents time performance for the full dataset analysis.

of a PCA transformation, so the original meaning of the variables is missed. Following the same ideas of the previous dataset, we kept only three variables (V10-V14-V17), selecting those whose histograms seemed to be closer to Gaussian assumptions.

Methods

For image segmentation, we will simply use the K-class partitions obtained with both our proposal and subsampling. For Asteroids and Credit Card datasets we perform a two-classes clustering comparing our method to both subsampled EM and full dataset EM. Actually, true classification labels are provided by the original datasets but we use them only as a benchmark, as we want to follow a completely unsupervised approach. In particular we will employ them to rank results based on ARI score (Hubert and Arabie 1985). Similarly to simulations, we used our self-written R code for bin-marginal CL-EM and `Rmixmod` for all versions of classical EM .

6.2 Results and discussion

Image segmentation

Figures 6-9 synthesize results obtained for the image segmentation of the four images. Figures marked with (a) represent the true images and those denoted with (b) the segmentation obtained with binned data. Finally, figures (c)-(d) are the best and worst (respectively associated to the full dataset likelihood of the estimated parameter) segmentation obtained with classical subsampling in condition of equal memory occupancy. It can be seen that our method successfully detects the objects, while subsampling results in very noisy segmentations. Regarding the binning grid employed, we used marginal grids of refinement 20 for all Cell images and a finer ones with 400 intervals for Comet. In addition for Cell images we selected $K = 4$, where 4 colours are recognizable, and $K = 3$ for Comet, as in this image there is a consistent group of noise (represented in our segmentation by black points)

Asteroids

Figure 10a reports the result of the comparison between our bin-marginal CL-EM and classical EM with both subsampling and full dataset. In absolute terms, generically low ARI scores suggest that a total unsupervised approach could be very risky in this case. However, our objective is to analyze the results of our proposal relatively to our competitors. Concerning this, Figure 10a shows that, despite the loss of information, bin-marginal method (black circle) has globally better performances than both subsampling (red boxplot) and full dataset EM (blue circle). Moreover, bin-marginal CL-EM is not prone to the variability of subsampling, whose result highy depends on subsample choice.

Credit card fraud detection

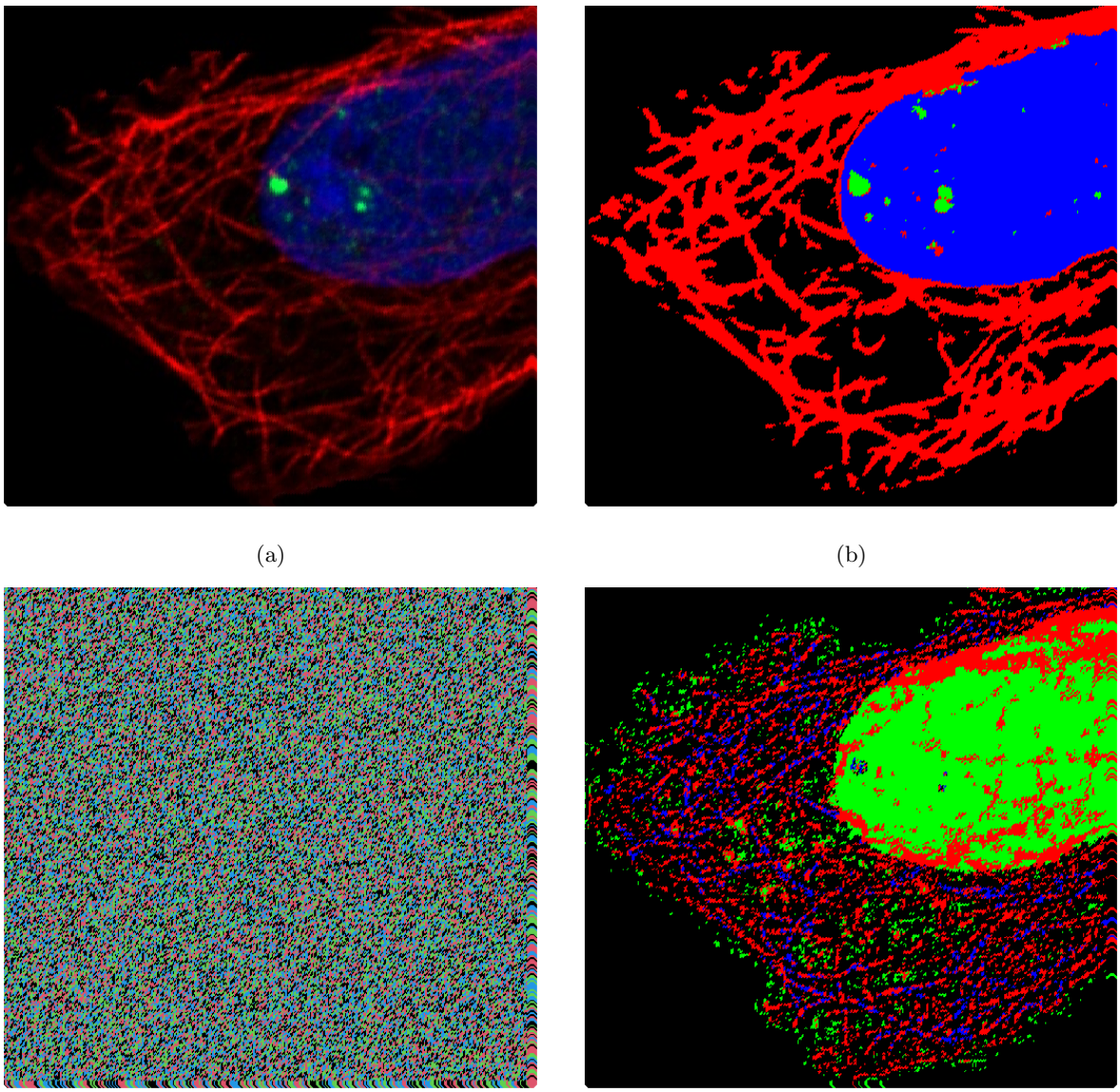
Following the same strategy used for Asteroids dataset, we build a two-classes partition using our bin-marginal technique to detect frauds among the set of credit card transactions. Based on Figure 10b, similar comments could be made. Our method seems to be globally better with our direct competitors, avoiding the high variability of subsampling.

7 Conclusion

This work has introduced a method based on Gaussian mixture models combining binned data with marginalization, which is able to detect, in an unsupervised way, imbalanced classes on large datasets under hard memory constraints. The theoretical results presented in this paper have shown that the model and the proposed estimation procedure have good statistical properties, such as identifiability and consistency, despite the huge loss of statistical information caused by our heavy bin-marginal data compression.

Both simulations and real applications have proved the competitiveness of our method with respect to the traditional subsampling method, in those cases where a full dataset clustering is out of reach. In particular, it has revealed a great potential in the context of image segmentation when very tiny objects have to be detected.

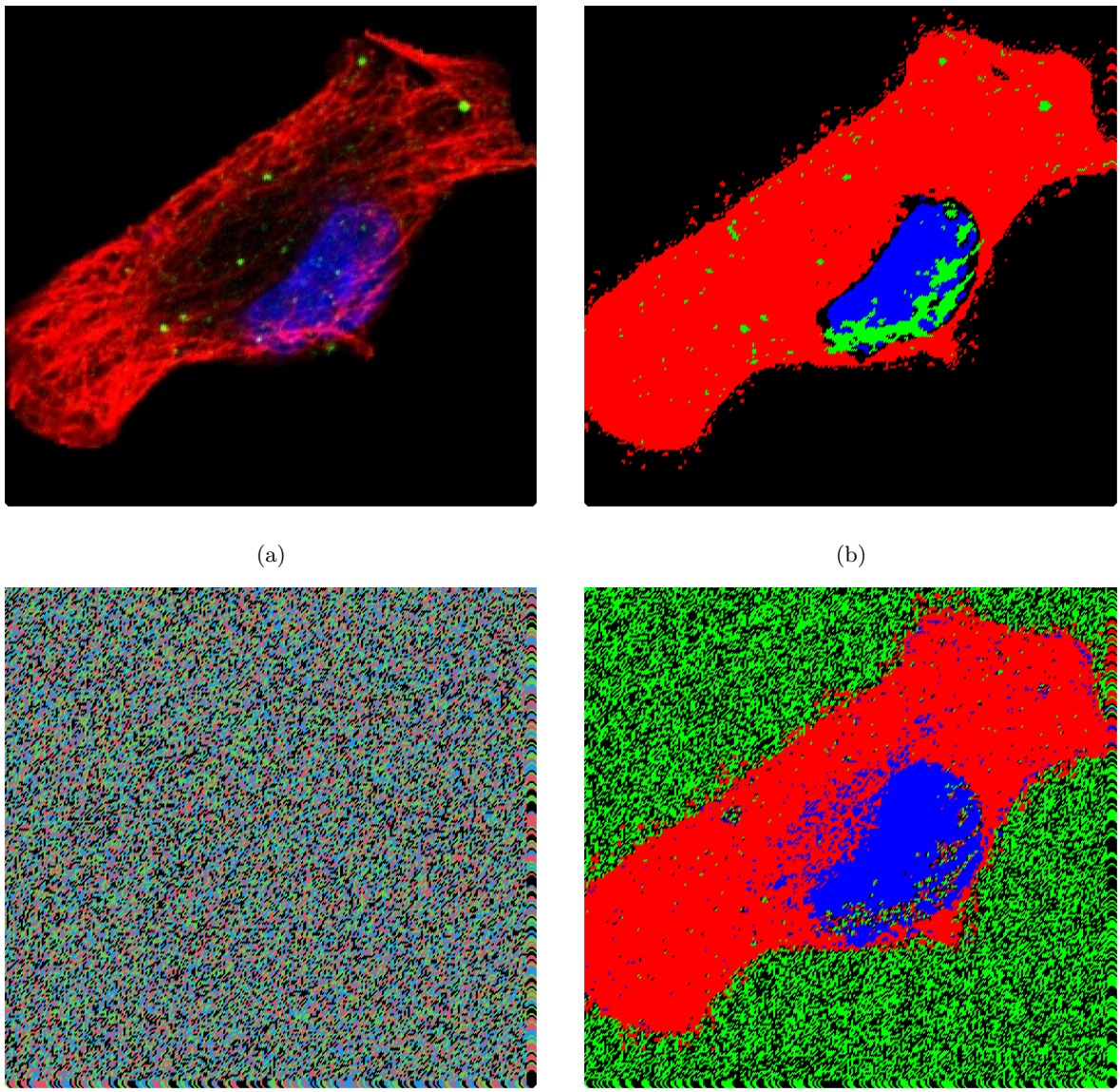
These very encouraging results have to be more developed in the future, proposing methods for model choice and optimal strategies for binning grids construction. The formulation of a model choice criterion is important, as it allows the complete automation of our technique and a more precise clustering. Optimal binning strategies are



(a) (b)

(c) (d)

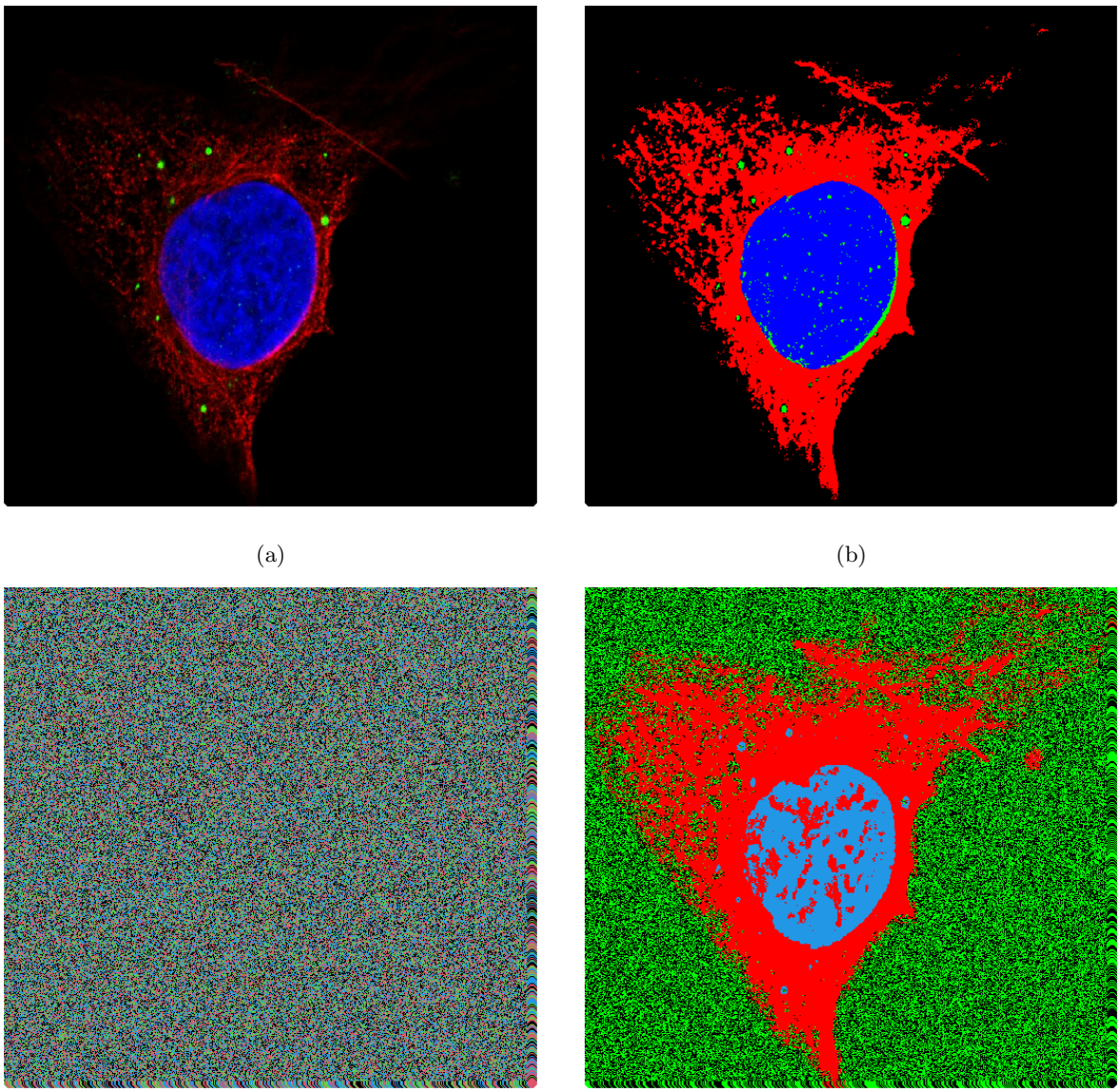
Figure 6: Cell-1 segmentation: a) Original image; b) Segmentation obtained with bin-marginal CL-EM; c-d) Worst and best segmentation obtained with two subsampled EM.



(a) (b)

(c) (d)

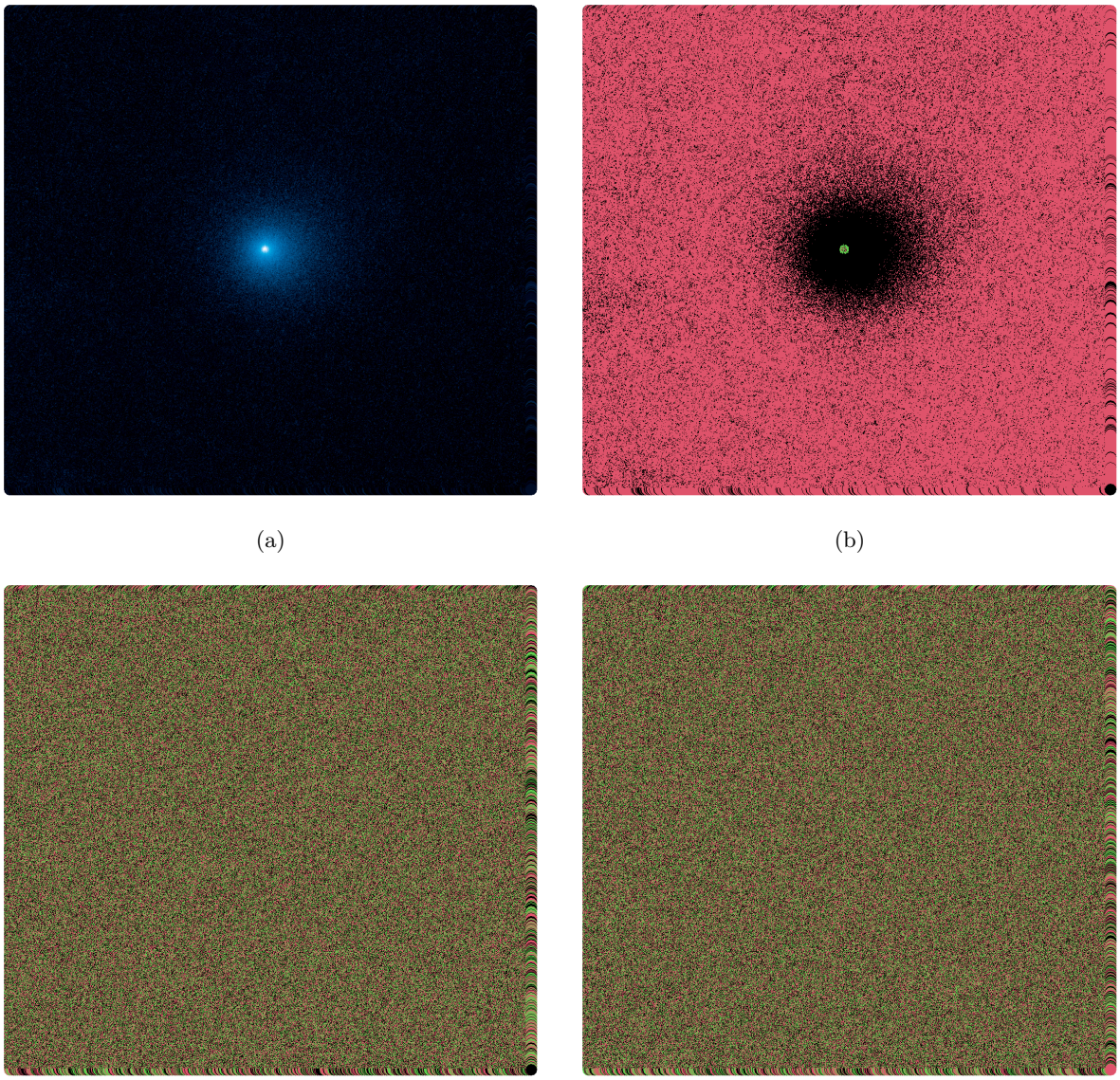
Figure 7: Cell-2 segmentation: a) Original image; b) Segmentation obtained with bin-marginal CL-EM; c-d) Worst and best segmentation obtained with two subsampled EM.



(a) (b)

(c) (d)

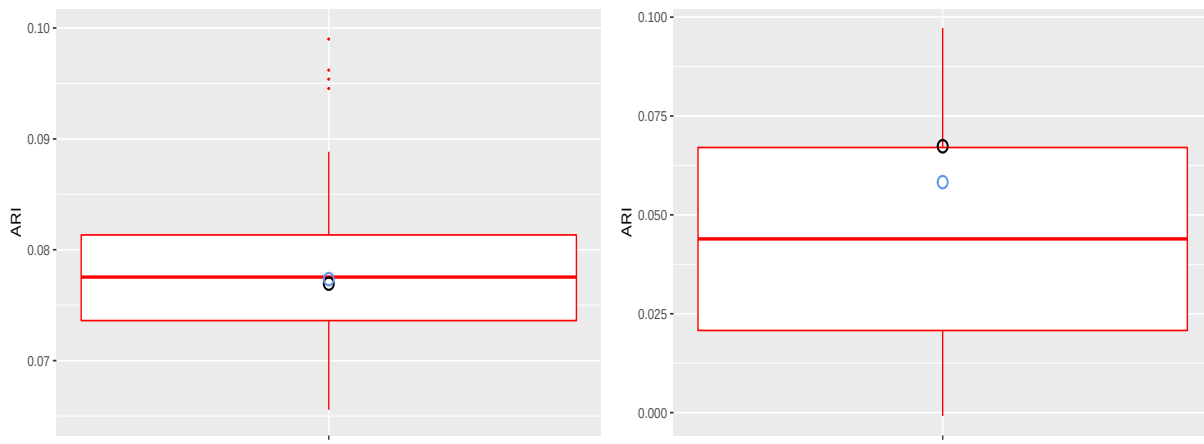
Figure 8: Cell-3 segmentation: a) Original image; b) Segmentation obtained with bin-marginal CL-EM; c-d) Best and worst segmentation obtained with two subsampled EM.



(a) (b)

(c) (d)

Figure 9: Comet image segmentation: a) Original image; b) Segmentation obtained with bin-marginal CL-EM; c-d) Worst and best segmentation obtained with two subsampled EM.



(a) (b)

Figure 10: Two-classes clustering performances in terms of ARI for subsampled EM (red boxplots), bin-marginal CL-EM (black circle) and full dataset EM (blue circle). Datasets: (a) Asteroids; (b) Credit card fraud detection.

as much crucial, because they could contribute to design highly efficient frugal grids and to select variables. Indeed, we can reasonably suppose that the selection of a certain grid refinement degree for a variable is correlated to the degree of importance of the same variable for the clustering. According to this heuristic, we could infer that uninformative variables are associated to very coarse grids. Indeed, at the limit case when a marginal grid is restricted to a single bin, we can recognize an outcome equivalent to variable selection. Thus, such an approach could provide a very appealing halfway strategy instead of "hard" classical variable selection.

Optimal strategies are also required in bin-marginal CL-EM initialization. For this reason, it is necessary to investigate the theoretical properties of bin-marginal composite log-likelihood, studying in particular its local maxima and the rate of convergence of the related estimator towards the true parameter. Indeed, we could expect the particularly high information compression of our method to have consequences on that.

In this paper we have not considered those high-dimensional situations where several small classes might appear. Actually, in a preliminary simulated example not displayed here, we studied a related problem where our technique encountered some issues, solved after having increased the size of the simulated dataset from 1 million to 10 million. Probably, this is caused by the fact that our data-reduction is really extreme and, maybe, a softer compression is needed to save enough multivariate statistical information to accomplish more complex tasks. Thus, possible solutions could be either an intelligent and frugal usage of bivariate grids or hybrid methods involving both bin-marginal and raw data. In all of these cases, we will have to remain inside the strong computational constraints of our context of reference.

We are conscious that our proposal can be used in several fields and for this reason we aim to develop deeply our technique, proposing at the same time practical applications of it in our future works. We will also provide shortly an R package that could be used by both experts and practitioners.

Acknowledgements

This research is supported by Inria and Direction Générale de l'Armement (DGA) through the Agence de l'innovation de Défence (AID).

A Appendix

Preliminary results

In this paragraph we present two results necessary for the proofs of the propositions contained in the main text. The first proposition is proved in Valiant (2012) and it helps to recover the subsequent proposition, which assesses the identifiability for univariate binned mixture models.

Proposition A.1 (**Proposition 11.5** in Valiant (2012))**.** *Given the linear combination of K univariate Gaussian densities $f(x) = \sum_{k=1}^{K} \pi_k \phi(x;\ \mu_k, \sigma_k^2)$, such that either $\mu_{k_1} \neq \mu_{k_2}$ or $\sigma_{k_1}^2 \neq \sigma_{k_2}^2$ for $k_1 \neq k_2$ and for all k $\pi_k \in \mathbb{R}^*$, the number of solutions to $f(x) = 0$ is at most $2(K-1)$.*

Proposition A.2. *Binned univariate mixtures of K Gaussian distributions are identifiable if the binning grid has $R > 4K - 3$ cut points.*

Proof If $\mathcal{X} = \mathbb{R}$, the considered probability mass functions reduces to $p(\boldsymbol{n}, \boldsymbol{\psi})$, thus it is to demonstrate that statement

$$\forall \boldsymbol{\psi}, \boldsymbol{\psi}^* \in \Psi :\ \ p(\boldsymbol{n};\ \boldsymbol{\psi}) = p(\boldsymbol{n};\ \boldsymbol{\psi}^*)\ \forall G, \boldsymbol{n}\ \Rightarrow \boldsymbol{\psi} = \boldsymbol{\psi}^* \tag{10}$$

hold almost everywhere expect for a set $\Psi^* \subset \Psi$ whose Lebesgue's measure is zero, respectively to the dimension of the original space.

Denoting with $\Phi(\cdot)$ the cumulative density function of a standard Gaussian, if G has R cut points $(a_1, \ldots, a_R)$ then it is sufficient to prove that the system

$$\begin{cases} \sum_{k=1}^{K} \pi_k \Phi(\frac{a_1 - \mu_k}{\sigma_k}) = \sum_{k=1}^{K} \pi_k^* \Phi(\frac{a_1 - \mu_k^*}{\sigma_k^*}) \\ \sum_{k=1}^{K} \pi_k \Phi(\frac{a_2 - \mu_k}{\sigma_k}) = \sum_{k=1}^{K^*} \pi_k^* \Phi(\frac{a_2 - \mu_k^*}{\sigma_k^*}) \\ \vdots \\ \sum_{k=1}^{K} \pi_k \Phi(\frac{a_R - \mu_k}{\sigma_k}) = \sum_{k=1}^{K} \pi_k^* \Phi(\frac{a_R - \mu_k^*}{\sigma_k^*}) \end{cases}$$

has only the trivial solution $\boldsymbol{\psi} = \boldsymbol{\psi}^*$ whatever the grid is. Hence, the non-zero subset of non identifiability is the one of the possible permutation of $\boldsymbol{\psi}$.

It is also equivalent to discover how many zeros can have the difference between the cumulative density functions of two different Gaussian mixtures. If this

number is a certain Z, identifiability is assured for $R > Z$.

Again, considering the difference between two cumulative functions with Z zeros, namely $h(x)$, for continuity and for the fact that $\lim_{x\to\infty} h(x) = \lim_{x\to_\infty} h(x) = 0$, it is necessary that this function has at least $Z+1$ critical points, i.e. the difference of the two respective density functions has at least $Z+1$ zeros. So it is possible to formulate the problem in the terms of maximum number of zeros of the difference between the densities of two different mixtures.

Valiant's theorem states that this maximum number is $4K-2$. Thus, if $R > 4K-3$, identifiability holds. □

Proofs of main text propositions

Proposition 1 The proof mostly relies on the two previous prepositions. In dimension D, considering a grid with $\prod_{d=1}^{D} R_d$ cut points and $B = \prod_{d=1}^{D}(R_d+1)$ bins as defined in Section 2, the statement (10) holds if the system

$$\begin{cases} \sum_{k=1}^{K} \pi_k \int_{\mathcal{B}_b} \phi(\boldsymbol{x}, \boldsymbol{\mu}_k, \Sigma_k) d\boldsymbol{x} \\ = \sum_{k=1}^{K} \pi_k^* \int_{\mathcal{B}_b} \phi(\boldsymbol{x}, \boldsymbol{\mu}_k^*, \Sigma_k^*) d\boldsymbol{x} \\ \\ b = 1, \dots, B-1 \end{cases}$$

has only the trivial solutions $\boldsymbol{\psi} = \boldsymbol{\psi}^*$. Under hypothesis of diagonal covariance matrices it is equivalent to:

$$\begin{cases} \sum_{k=1}^{K} \pi_k \int_{\mathcal{B}_b^1} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{k1}, \sigma_{k1}^2) d\boldsymbol{x} \\ \times \dots \times \int_{\mathcal{B}_b^D} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{kD}, \sigma_{kD}^2) d\boldsymbol{x} \\ = \sum_{k=1}^{K} \pi_k^* \int_{\mathcal{B}_b^1} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{k1}^*, \sigma_{k1}^{2*}) d\boldsymbol{x} \\ \times \dots \times \int_{\mathcal{B}_b^D} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{kD}^*, \sigma_{kD}^{2*}) d\boldsymbol{x} \\ \\ b = 1, \dots, B \end{cases}$$

It is clear every 1-d region $\mathcal{B}_b^d$ coincide with a certain $\mathcal{B}_{b_d}^d$, which is a bin on the d-th dimension, as the D-dimensional bins are the result of the Cartesian product of certain 1-dimensional bins. Choose every equation involving integrals on regions sharing the same projection on the first axis $\mathcal{B}_{b_1}^1$. We note this is equivalent to consider $\prod_{d=2}^{D}(R_d+1)$ systems of $B_1 = R_1 + 1$ equations for two univariate mixtures with K components. We can rewrite the previous system as:

$$\begin{cases} \sum_{k=1}^{K} \pi_{k\tilde{b}} \int_{\mathcal{B}_{b_1}^1} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{k1}, \sigma_{k1}^2) d\boldsymbol{x} \\ = \sum_{k=1}^{K} \pi_{k\tilde{b}}^* \int_{\mathcal{B}_{b_1}^1} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{k1}^*, \sigma_{k1}^{2*}) d\boldsymbol{x} \\ \\ b_1 = 1, \dots, B_1 \\ \tilde{b} = (b_2, \dots, b_D) \in \prod_{d=2}^{D} \{1, \dots, B_d\} \end{cases}$$

where for each $\tilde{b}$:

$$\pi_{k\tilde{b}} = \pi_k \prod_{d=2}^{D} \int_{\mathcal{B}_{b_d}^d} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{kd}, \sigma_{kd}^2) d\boldsymbol{x}$$

$$\pi_{k\tilde{b}}^* = \pi_k^* \prod_{d=2}^{D} \int_{\mathcal{B}_{b_d}^d} \phi(\boldsymbol{x}, \boldsymbol{\mu}_{kd}^*, \sigma_{kd}^{2*}) d\boldsymbol{x}$$

As two or more components can share the same projection on a univariate space, the true number of components for each mixture will be lower or equal to K. In addition, it is a priori different for each mixture. Let name the number of components for the two projected mixtures K_1 and K_1^*. Let consider a partition I_1 of K_1 elements of the set $\{1, \dots, K\}$. Each element of I_1 represents the subset of components sharing the same projection on the first axis for the first mixture. Similarly, each element of the K_1^*-partition I_1^* of $\{1, \dots, K\}$ represents the totality of components having the same projection for the second mixture. Actually, the same argument of Proposition A.2 demonstrates that the two projected mixtures must have the same number of components, thus $K_1 = K_1^*$, and the same partition, thus $I_1 = I_1^*$. Moreover, as one-dimensional identifiability holds:

$$\begin{cases} \sum_{k' \in I_{1k}} \pi_{k'\tilde{b}} = \sum_{k' \in I_{1k}} \pi_{k'\tilde{b}}^* \\ \mu_{k'1} = \mu_{k'1}^* \qquad \sigma_{k'1}^2 = \sigma_{k'1}^{2*} \\ k' \in I_{1k}, \\ k = 1, \dots, K_1 \\ \tilde{b} = (b_2, \dots, b_D) \in \prod_{d=2}^{D} \{1, \dots, B_d\} \end{cases}$$

Using the definition of $\pi_{k\tilde{b}}$ and $\pi_{k\tilde{b}}^*$, the entire collection of the equation $\sum_{k' \in I_{1k}} \pi_{k'\tilde{b}} = \sum_{k' \in I_{1k}} \pi_{k'\tilde{b}}^*$ forms a system of identifiability for a mixture of dimension $D-1$ and number of components given by the cardinality I_k, which is lower than K. We can then iterate the same procedure, until we obtain only one-dimensional equations of identifiability for which Proposition A.2 is valid. In this way we obtain:

$$\begin{cases} \pi_k = \pi_k^* \\ \mu_{kd} = \mu_{kd}^* \\ \sigma_{kd}^2 = \sigma_{kd}^{2*} \\ \\ k = 1, \dots, K \quad d = 1, \dots, D \end{cases}$$

This completes the proof. □

Proposition 2 Let consider two probability mass functions $p_m(\boldsymbol{m};\ \boldsymbol{\psi})$ and $p_m(\boldsymbol{m};\ \boldsymbol{\psi}^*)$. Our aim is to demonstrate

$$\forall \boldsymbol{\psi}, \boldsymbol{\psi}^* \in \Psi: \quad p_m(\boldsymbol{m};\ \boldsymbol{\psi}) = p_m(\boldsymbol{m};\ \boldsymbol{\psi}^*)\ \ \forall G, \boldsymbol{m} \\ \Rightarrow \boldsymbol{\psi} = \boldsymbol{\psi}^*.$$

We can consider a grid of dimension $R_1 \times \dots \times R_D$ as defined in Section 3 and the vectors $\boldsymbol{m}^{\boldsymbol{b}} = (\boldsymbol{m}_1^{b_1}, \dots, \boldsymbol{m}_D^{b_D})$, where $\boldsymbol{b} = (b_1, \dots, b_d) \in \prod_{d=1}^{D} \{1, \dots, B_d\}$. Each vector $\boldsymbol{m}_d^{b_d}$ is defined as

$$\boldsymbol{m}_d^{b_d} = \begin{cases} n & \text{for an index } \ b_d \ \in \{1, \dots, B_d\} \\ 0 & \text{otherwise} \end{cases}$$

So each $\boldsymbol{m}_d^{b_d}$ is a vector of counts representing the situation in which observation are concentrated in the b_d-th bin on the d-th dimension. Moreover, for each possible $\boldsymbol{m}^{\boldsymbol{b}}$ we have:

$$p_m(\boldsymbol{m}^{\boldsymbol{b}};\ \boldsymbol{\psi}) = \sum_{\boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{m}^{\boldsymbol{b}}}} p(\boldsymbol{n}';\ \boldsymbol{\psi}) = k(\boldsymbol{m}^{\boldsymbol{b}}) P_{\boldsymbol{b}}$$
$$p_m(\boldsymbol{m}^{\boldsymbol{b}};\ \boldsymbol{\psi}^*) = \sum_{\boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{m}^{\boldsymbol{b}}}} p(\boldsymbol{n}';\ \boldsymbol{\psi}^*) = k(\boldsymbol{m}^{\boldsymbol{b}}) P_{\boldsymbol{b}}^*$$

where $k(\boldsymbol{m}^{\boldsymbol{b}})$ is a constant and $P_{\boldsymbol{b}}$ (and $P_{\boldsymbol{b}}^*$) is the probability for the bin whose marginal bin on the d-th is indexed by the d-th element of $\boldsymbol{b}$. Choosing every possible value for $\boldsymbol{b}$ we obtain the same system of identifiability equation for a multivariate binned mixture model. There are no other equation to satisfy because the other probabilities for other vectors $\boldsymbol{m}$ are combinations of $P_{\boldsymbol{b}}$ (or $P_{\boldsymbol{b}}^*$). Thus if multivariate binned mixture models are identifiable the binned marginal-conjoint model is identifiable. Moreover, under the hypothesis of Proposition 1 diagonal binned conjoint-marginal multivariate mixtures are identifiable. □

Proof of Proposition 3 Let $\mathbf{X} = (X_1, \dots, X_D)$ be a mixture random variable with pdf $f(\boldsymbol{x}, \boldsymbol{\psi})$ and define the $\sum_d B_d$–dimensional random variable $\mathbf{M}$ with components $(1_{a_{d(b_d-1)} \le X_d < a_{db_d}})_{d=1,\dots,D;\ b_d=1,\dots B_d}$, margins of the raw observation $\mathbf{X}$ on the D-dimensional grid. Then $\boldsymbol{m}$ is the sum of n outcomes of i.i.d. random variables having $\mathbf{M}$ law. Hence, $\frac{1}{n}\tilde{\ell}_m(\boldsymbol{\psi};\ \boldsymbol{m})$ converges in probability to the contrast function $F(\boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi};\ \mathbf{M})]$ when $n \to \infty$, uniformly in the parameter.

We have to show that the following inequality holds:

$$\mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi};\ \mathbf{M})] \ge \mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi}^*;\ \mathbf{M})] \quad \forall \boldsymbol{\psi} \ne \boldsymbol{\psi}^* \tag{11}$$

In this case we will say that there is asymptotic identifiability. Suppose there is a point $\boldsymbol{\psi} \ne \boldsymbol{\psi}^*$ such that $\mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi};\ \mathbf{M})] = \mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi}^*;\ \mathbf{M})]$. Then we have:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi}^*;\ \mathbf{M})] - \mathbb{E}_{\boldsymbol{\psi}^*}[\tilde{\ell}_m(\boldsymbol{\psi};\ \mathbf{M})] \\ &= \mathbb{E}_{\boldsymbol{\psi}_1^*}[\ell_1(\boldsymbol{\psi}_1^*;\ \mathbf{M}_1)] - \mathbb{E}_{\boldsymbol{\psi}_1^*}[\ell_1(\boldsymbol{\psi}_1;\ \mathbf{M}_1)] + \dots \\ &+ \mathbb{E}_{\boldsymbol{\psi}_D^*}[\ell_D(\boldsymbol{\psi}_D^*;\ \mathbf{M}_D)] - \mathbb{E}_{\boldsymbol{\psi}_D^*}[\ell_D(\boldsymbol{\psi}_D;\ \mathbf{M}_D)] = 0 \end{aligned}$$

where $\mathbf{M}_d$ is a B_d dimensional random variable with components $(1_{a_d(b_d-1) \le X_d < a_d b_d})_{b_d=1,\dots,B_d}$. For all log-likelihoods ℓ_d, $d = 1, \dots, D$, inequality (11) holds. Thus, for all $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\psi}_1^*}[\ell_1(\boldsymbol{\psi}_1^*;\ \mathbf{M}_1)] \ge \mathbb{E}_{\boldsymbol{\psi}_1^*}[\ell_1(\boldsymbol{\psi}_1;\ \mathbf{M}_1)] \\ &\vdots \\ &\mathbb{E}_{\boldsymbol{\psi}_D^*}[\ell_2(\boldsymbol{\psi}_D^*;\ \mathbf{M}_D)] \ge \mathbb{E}_{\boldsymbol{\psi}_D^*}[\ell_D(\boldsymbol{\psi}_D;\ \mathbf{M}_D)] \end{aligned}$$

and equality holds for $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_1^*, \dots, \boldsymbol{\psi}_D = \boldsymbol{\psi}_D^*$.

As it is well-known for mixtures, each equality hold up to a permutation: so we can define a set of D permutations named $\rho_1, \dots, \rho_D$. In the hypothesis of our proposition, which assures that the marginal mixtures have the same number of components of the original ones and different proportions, we can match uniquely the two components thanks to proportions matching. Therefore, the D permutations reduce to only one (named ρ) and $\boldsymbol{\psi}^*$ is equal to $\boldsymbol{\psi}$ after ρ. So, in this case, asymptotic identifiability is fulfilled. □

References

S. B. Aruoba and J. Fernández-Villaverde. A comparison of programming languages in macroeconomics. *Journal of Economic Dynamics and Control*, 58:265–273, 2015.

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

I. V. Cadez, P. Smyth, G. J. McLachlan, and C. E. McLaren. Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1):7–34, 2002.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern recognition*, 28(5): 781–793, 1995.

P. Chan and S. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. 07 1998.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, 41(10):4915–4928, 2014.

A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–14, 09 2017. doi: 10.1109/TNNLS.2017.2736643.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

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.

T. Fawcett and F. Provost. Adaptive fraud detection. *Data mining and knowledge discovery*, 1 (3):291–316, 1997.

A. Fernández, S. del Río, N. V. Chawla, and F. Herrera. An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2):105–120, 2017.

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.

C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca. mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical report, 2012.

X. Gao and P. X.-K. Song. Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica*, pages 165–185, 2011.

M. S. Hossain. Asteroid dataset, 2020. https://www.kaggle.com/sakhawat18/asteroid-dataset.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert. Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software*, 67(6):1–29, 2015. doi: 10.18637/jss.v067.i06.

J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5 (1):1–30, 2018.

B. G. Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

G. McLachlan and P. Jones. Fitting mixture models to grouped and truncated data via the em algorithm. *Biometrics*, pages 571–578, 1988.

P. D. McNicholas and T. B. Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

G. Molenberghs and G. Verbeke. Models for discrete longitudinal data. 2005.

NASA. Nasa's hubble observes the farthest active inbound comet yet seen. https://hubblesite.org/contents/news-releases/2017/news-2017-40.html, 2017. Accessed: 2021-08-10.

X. Niu, L. Wang, and X. Yang. A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*, 2019.

N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern recognition*, 26 (9):1277–1294, 1993.

A. A. Quarta and G. Mengali. Electric sail missions to potentially hazardous asteroids. *Acta Astronautica*, 66(9-10):1506–1519, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

M. Ranalli and R. Rocci. Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, 26(1-2):529–547, 2016.

M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan. A multiple expert approach to the class imbalance problem using inverse random under sampling. In *International workshop on multiple classifier systems*, pages 82–91. Springer, 2009.

H. Q. To. Single cell images fold 0 [hpa]. https://www.kaggle.com/quochungto/cells-fold0, 2021. Accessed: 2021-08-10.

N. Tsapanos, A. Tefas, N. Nikolaidis, and I. Pitas. Efficient mapreduce kernel k-means for big data clustering. In *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*, pages 1–5, 2016.

M. L. G. ULB. Credit card fraud detection. https://www.kaggle.com/mlg-ulb/creditcardfraud, 2018. Accessed: 2021-08-10.

G. J. Valiant. *Algorithmic approaches to statistical questions*. PhD thesis, UC Berkeley, 2012.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949. ISSN 00034851. URL http://www.jstor.org/stable/2236315.

T. Whitaker, B. Beranger, and S. A. Sisson. Composite likelihood methods for histogram-valued random variables. *Statistics and Computing*, pages 1–19, 2020.

H. Xia, W. Huang, N. Li, J. Zhou, and D. Zhang. Parsuc: A parallel subsampling-based method for clustering remote sensing big data. *Sensors*, 19(15):3438, 2019.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

H. Yu, J. Ni, Y. Dan, and S. Xu. Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets. *Tsinghua Science and Technology*, 17(6):666–673, 2012. doi: 10.1109/TST.2012.6374368.