



HAL
open science

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

Filippo Antonazzo, Christophe Biernacki, Christine Keribin

► To cite this version:

Filippo Antonazzo, Christophe Biernacki, Christine Keribin. Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach. Working Group - Model-based Clustering, Oct 2021, Athens, Greece. hal-03505673

HAL Id: hal-03505673

<https://hal.science/hal-03505673v1>

Submitted on 31 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

F. ANTONAZZO ^{1,2}, Ch. BIERNACKI ^{1,2}, Ch. KERIBIN ^{1,3}

¹Inria

²Laboratoire de mathématiques Painlevé, Université de Lille, Villeneuve d'Ascq, France

³Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, Orsay, France

WG-MBC-2021, Athens, October 25-29, 2021

Plan

- 1 Introduction
- 2 Model
- 3 Estimation
- 4 Experiments
- 5 Discussion

Motivation: huge and imbalanced data sets

- ▶ huge in the sense tall data
 - ↪ number of observations (high dimension setting out of scope)
 - ↪ out of computer limits
 - ↪ or within computer limits but with frugal resource consumption (*green computing*)

- ▶ discover new information
 - ↪ more and more clusters: not the focus of this talk
 - ↪ reveal (valuable) tiny clusters: imbalanced data sets
a few *abnormal* objects have to be recognized among a large amount of *normal* ones
credit card fraud detection [Chan and Stolfo 1998], cancer recognition [Yu et al. 2012], fraudulent calls [Fawcett and Provost 1997]

Approaches

- ▶ **supervised** approach (classification) with imbalanced data sets
 - ↔ create **artificial balanced** data sets:
 - oversampling the minority class [*Chawla et al. 2002*],
 - undersampling the majority class [*Tahir et al. 2009*]
 - ↔ labeling could be difficult when sample size is very large

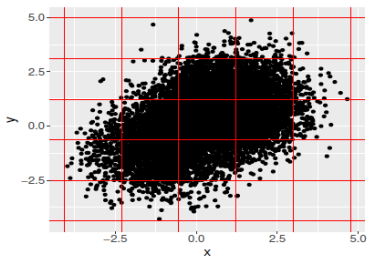
- ▶ **unsupervised** approach (clustering) with sample size is very large
 - ↔ **subsampling** [*Fraley and Raftery 2002, Xia et al. 2019*]
 - ↔ difficult to detect very tiny clusters

 - ↔ **computer science solutions**
 - powerful computers or distributed architectures (MAP-reduce, ...)
 - ↔ not frugal

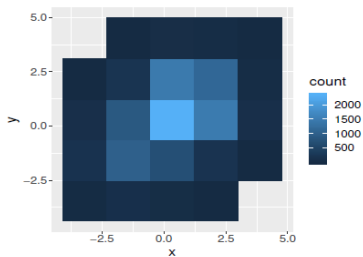
- ▶ our aim : **clustering** of **huge** and **imbalanced** datasets under **memory constraints**

Another way for data reduction

- ▶ **unsupervised** approach (clustering)
 - ↳ from **raw** to **binned** data



(a) Raw data



(b) Binned data

Our bin-marginal approach in a nutshell

Frugal unsupervised MBC (D-dimensional Gaussian mixtures) using marginal binned data:

1. from raw to binned data

↪ particular version of the EM algorithm [*McLachlan and Jones 1998; Cadez et al. 2002*]

↪ another dimensionality pb

2. from binned data to (1D-)marginal counts

↪ need to design a new EM algorithm but computationally intractable

3. optimization of a composite likelihood (CL) [*Lindsay 1988; Whitaker et al. 2020*] instead of the full one

↪ for diagonal GMM

► CL + GMM + 2D-bin [*Ranalli and Rocci 2016*]

new in our approach: harder data reduction (1D-bin)

Plan

- 1 Introduction
- 2 Model**
- 3 Estimation
- 4 Experiments
- 5 Discussion

Model Based Clustering with finite GMM

Observations $\mathbf{x} = \{\mathbf{x}_i \in \mathbb{R}^D, i = 1, \dots, n\}$ are i.i.d. according to a D -dimensional Gaussian mixture with K components:

$$f(\mathbf{x}; \psi) = \sum_{k=1}^K \pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\sum_k \pi_k = 1, \quad \pi_k > 0 \quad (k = 1, \dots, K)$$

where $\psi = (\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K)$ and $\phi(\cdot)$ is the D -variate Gaussian density

Binned data

unobservable or too many **raw** data \mathbf{x}_i
↪ vector of **binned** data $\mathbf{n} = (n_1, \dots, n_B)$

- ▶ the original sample space is divided into a partition $\{\mathcal{B}_b \subset \mathbb{R}^d, b = 1, \dots, B\}$
- ▶ $n_b = \#\{\mathbf{x}_i \in \mathcal{B}_b\}$

\mathbf{n} arises from a multinomial model with pmf [Cadez et al. 2002]¹

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

- ▶ trick for **sample size reduction**: select $B \ll n$

¹also provide an estimate of ψ with a binned version of EM

Curse of dimensionality for binned data

- ▶ in our case: **Cartesian grid** $G = G_1 \times \dots \times G_D$ where G_d is a univariate grid with $R_d + 2$ cut points
 $\hookrightarrow B = \prod_{d=1}^D (R_d + 1)$ **bins**, representing the grid's **coarseness**
- ▶ works well if $B \ll n$ and **univariate** context

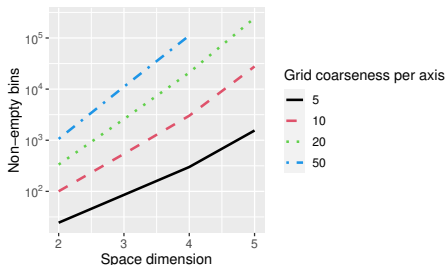
▶ when D increases

the number of non-empty bins depends exponentially on the dimension D

\hookrightarrow impossible to obtain a **manageable** amount of binned data

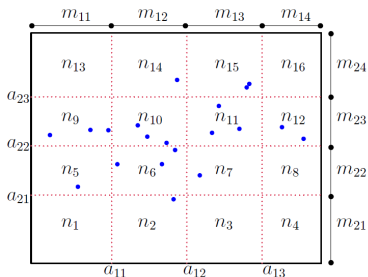
\hookrightarrow several D -dimensional **numerical integrations**.

\hookrightarrow vanishes any kind of gain



Marginal binned data

- ▶ work with the **1-D binned** data on each direction **separately**
- ▶ **marginal counts**: $\mathbf{m} = \{\mathbf{m}_1, \dots, \mathbf{m}_D\}$
 for each direction $d = 1, \dots, D$, $\mathbf{m}_d = (m_{d1}, \dots, m_{dB_d})$,
 component m_{db_d} is the count of observations x_{id} in the b_d -th bin
 of the d -th dimension



↪ store $\sum_{d=1}^D B_d$ values instead of $\prod_{d=1}^D B_d$

Bin-marginal model

► bin-marginal pdf

$$p_m(\mathbf{m}; \psi) = \sum_{\mathbf{n}' \in \mathcal{F}_m} p(\mathbf{n}'; \psi),$$

where \mathcal{F}_m is the set of tables \mathbf{n}' sharing the same marginals \mathbf{m} .

► issues

- ↔ identifiability
- ↔ mathematical complexity of the likelihood
- ↔ optimization of the likelihood

Identifiability

- ▶ GMM identifiable up to a label permutation [*Yakowitz and Spragings 1968*] (raw data)
- ▶ as so far, no reference for the binned case

Proposition (Full binned Gaussian diagonal mixtures - ABK 2021)

Under hypothesis of *diagonal covariance* matrices, *binned D -variate mixtures* of at most K_{max} components are identifiable if $R_d > 4K_{max} - 3$, $d = 1, \dots, D$.

Identifiability

- ▶ GMM identifiable up to a label permutation [*Yakowitz and Spragings 1968*] (raw data)
- ▶ as so far, no reference for the binned case

Proposition (Full binned Gaussian diagonal mixtures - ABK 2021)

Under hypothesis of *diagonal covariance* matrices, *binned D -variate mixtures* of at most K_{max} components are identifiable if $R_d > 4K_{max} - 3$, $d = 1, \dots, D$.

- ▶ the proof relies on an existing result

Proposition (11.5 - 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 $\mu_i \neq \mu_j$ or $\sigma_i^2 \neq \sigma_j^2$ for $i \neq j$ and for all k $\pi_k \neq 0$, the number of solutions to $f(x) = 0$ is at most $2(K - 1)$.

Identifiability

- ▶ GMM identifiable up to a label permutation [*Yakowitz and Spragings 1968*] (raw data)
- ▶ as so far, no reference for the binned case

Proposition (Full binned Gaussian diagonal mixtures - ABK 2021)

*Under hypothesis of **diagonal covariance** matrices, **binned D -variate** mixtures of at most K_{max} components are identifiable if $R_d > 4K_{max} - 3, d = 1, \dots, D$.*

Proposition (Marginal-binned Gaussian diagonal mixtures - ABK 2021)

***Bin-marginal D -variate** mixtures of at most K_{max} components are identifiable if **binned D -variate** mixtures are identifiable. So, under diagonal covariance matrices hypothesis, identifiability is achieved if $R_d > 4K_{max} - 3, d = 1, \dots, D$.*

Plan

- 1 Introduction
- 2 Model
- 3 Estimation**
- 4 Experiments
- 5 Discussion

EM algorithm for bin-marginal model

► complete log-likelihood

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

where \mathbf{z} gathers all $z_{ik} = \mathbb{1}_{\text{observation } i \text{ in cluster } k}$

EM algorithm for bin-marginal model

E-step

- ▶ expectation respectively to $p(\mathbf{x}, \mathbf{z} | \mathbf{m}; \psi^{(j)})$

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

$$= \sum_{\mathbf{n} \in \mathcal{F}_m} \alpha^{(j)}(\mathbf{n}) \sum_{k=1}^K \sum_{b=1}^B n_b \int_{\mathcal{B}_b} \tau_k^{(j)}(\mathbf{x}) \log[\pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] d\mathbf{x}$$

- ▶ $\alpha^{(j)}(\mathbf{n}) = \frac{p(\mathbf{n}; \psi^{(j)})}{\sum_{\mathbf{n}' \in \mathcal{F}_m} p(\mathbf{n}'; \psi^{(j)})}$ and $\tau_k^{(j)}(\cdot) = \frac{\pi_k^{(j)} \phi(\cdot; \boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)})}{f(\cdot; \psi^{(j)})}$.

M-step

- ▶ $\pi_k^{(j+1)} = \frac{1}{n} \sum_{\mathbf{n} \in \mathcal{F}_m} \alpha^{(j)}(\mathbf{n}) \sum_{b=1}^B n_b \int_{\mathcal{B}_b} \tau_k^{(j)}(\mathbf{x}) d\mathbf{x}$

↔ both steps involve the computation of all **crossed** tables (\mathcal{F}_m):
intractable

↔ alternative: use of **marginal composite likelihood**

Marginal Composite Likelihood

Let \mathbf{x} be a D -dimensional sample with n observations

$\mathbf{x}_i = (x_{i1}, \dots, x_{iD})$, $i = 1, \dots, n$, generated by a GMM with parameter ψ

- ▶ pseudo-likelihood only relying on the likelihood of the marginals

$L_d(\psi_d; \mathbf{x}_d)$

↪ $\mathbf{x}_d = (x_{1d}, \dots, x_{nd})$ the component d of the dataset

↪ with parameter $\psi_d = (\pi_1, \dots, \pi_K, \mu_{1d}, \dots, \mu_{Kd}, \sigma_{1d}^2, \dots, \sigma_{Kd}^2)$

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

- ▶ the estimator $\tilde{\psi}$ maximizing $\tilde{L}(\psi; \mathbf{x})$ is consistent and asymptotically normal [*Molenberghs and Verbeke 2005*]
- ▶ ↪ EM algorithm with CL for HMM [*Gao and Song 2011*]
- ↪ CL on bivariate-binned data [*Ranalli and Rocci 2016*]

Bin-marginal Composite Likelihood (bmCL)

- ▶ **our proposal**: combine memory reduction (bin-marginal)

$$\log p_m(\mathbf{m}; \psi) = \log \sum_{\mathbf{n}' \in \mathcal{F}_m} p(\mathbf{n}'; \psi)$$

and computational advantages of 1D-marginal CL

↪ we aim at maximizing the **bin-marginal composite** log-lik.:

$$\begin{aligned} \tilde{\ell}_m(\psi; \mathbf{m}) &= \sum_{d=1}^D \ell_d(\psi_d; \mathbf{m}_d) \\ &= \sum_{d=1}^D \sum_{b_d=1}^{B_d} m_{db_d} \log \left(\int_{B_{b_d}^d} f_d(x_d; \psi_d) dx_d \right). \end{aligned}$$

- ↪ diagonal mixtures only...
- ↪ identifiability ?

Bin-marginal CL: generic identifiability

A case of non identifiability

▶ blue mixture:

$$0.5\mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix}\right) +$$

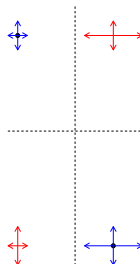
$$0.5\mathcal{N}\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}\right)$$

▶ red mixture:

$$0.5\mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & w_2 \end{pmatrix}\right) +$$

$$0.5\mathcal{N}\left(\begin{pmatrix} \nu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} w_1 & 0 \\ 0 & v_2 \end{pmatrix}\right)$$

$$\Leftrightarrow \mathbb{E}_{\psi^*}[\tilde{\ell}_m(\psi^*; \mathbf{M})] = \mathbb{E}_{\psi^*}[\tilde{\ell}_m(\psi; \mathbf{M})]$$



Identifiability **except on** the set of **null measure** composed by mixtures having **two equal proportions** with two components sharing the **same projection**

\Leftrightarrow **generic** identifiability, then consistency [Whitaker et al. 2020]

A naive EM algorithm for bin-marginal CL

on each direction d : work with m_d

- ▶ associate the **missing** vectors $(\mathbf{x}_d, \mathbf{z}_d)$, where \mathbf{z}_d is $n \times K$ indicator membership matrix for \mathbf{x}_d .
- ▶ run 1D EM algorithm separately
- ▶ how two **conciliate** the partitions from each direction ?
 - ↪ use the same π_1, \dots, π_K on each direction, in a global EM

EM algorithm for bin-marginal CL (bmCL)

With $\psi_d = (\pi_1, \dots, \pi_K, \mu_{1d}, \dots, \mu_{Kd}, \sigma_{1d}^2, \dots, \sigma_{Kd}^2)$

► bmCL E-step

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

► bmCL M-step straightforward

$$\tau_{kd}^{(j)}(\cdot) = \frac{\pi_k^{(j)} \phi(\cdot; \mu_{kd}^{(j)}, \sigma_{kd}^{2(j)})}{f(\cdot; \psi_d^{(j)})}$$

$$\pi_k^{(j+1)} = \frac{\sum_{d=1}^D \sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}_{b_d}^d} \tau_{kd}^{(j)}(x_d) dx_d}{Dn}; \quad \mu_{kd}^{(j+1)} = \frac{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}_{b_d}^d} x_d \tau_{kd}^{(j)}(x_d) dx_d}{\sum_{b_d=1}^{B_d} m_{db_d} \int_{\mathcal{B}_{b_d}^d} \tau_{kd}^{(j)}(x_d) dx_d}$$

► estimated partition: $\hat{\mathbf{z}} = \arg \max_{\mathbf{z}_k} \frac{\hat{\pi}_k \phi(\cdot; \hat{\mu}_k, \hat{\Sigma}_k^2)}{f(\cdot; \hat{\psi})}$

Plan

- 1 Introduction
- 2 Model
- 3 Estimation
- 4 Experiments**
- 5 Discussion

Numerical experiment on simulated data

- ▶ ability to recognize the minority class
- ▶ comparison with two competitors (estimation with `Rmixmod`)
 - ↪ classic estimation with the full dataset
 - ↪ a subsampling strategy
- ▶ clustering quality measured by the ARI score and time, under same memory constraints:
 - ↪ bin marginal: grid coarseness $R \leftrightarrow 2R$ memory space
 - ↪ subsampling: 100 different subsamples of size $2R$
 - ↪ $R=50, 100, 200$

Experimental settings: 1M obs from 3D 2-classes mixtures

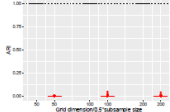
Scenario	Separation	Imbalance	Small class proportion (π_1)	Means
HH	High	High	10^{-4}	$\mu_1 = (-4, -4, -4)$ $\mu_2 = (4, 4, 4)$
HM		Medium	10^{-3}	
HL		Low	10^{-2}	
MH	Medium	High	10^{-4}	$\mu_1 = (-3, -3, -3)$ $\mu_2 = (3, 3, 3)$
MM		Medium	10^{-3}	
ML		Low	10^{-2}	
LH	Low	High	10^{-4}	$\mu_1 = (-2, -2, -2)$ $\mu_2 = (2, 2, 2)$
LM		Medium	10^{-3}	
LL		Low	10^{-2}	
VH	Very low	High	10^{-4}	$\mu_1 = (-1, -1, -1)$ $\mu_2 = (1, 1, 1)$
VM		Medium	10^{-3}	
VL		Low	10^{-2}	
1HH	One separated component	High	10^{-4}	$\mu_1 = (-1, -1, -4)$ $\mu_2 = (1, 1, 4)$
1HM		Medium	10^{-3}	
1HL		Low	10^{-2}	

20 replications of each scenario

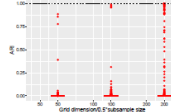
Results

quality vs memory

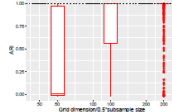
- ▶ **bmCL** (black) mostly outperforms **subsampling** (red), even with coarser grid,
- ▶ some difficulties only with very little separation and small proportion
- ▶ in general, **bmCL** approaches full data set results (dotted), with drastically less amount of memory



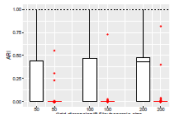
(a) HH



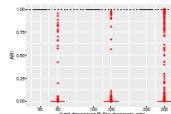
(b) HM



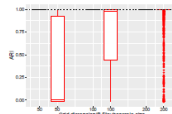
(c) HL



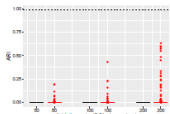
(d) MH



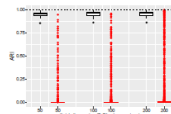
(e) MM



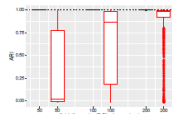
(f) ML



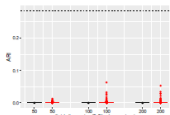
(g) LH



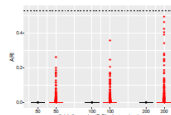
(h) LM



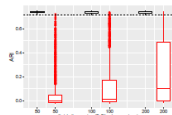
(i) LL



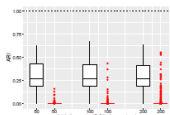
(j) VH



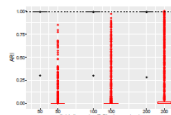
(k) VM



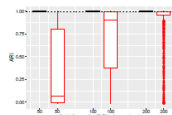
(l) VL



(m) 1HH



(n) 1HM

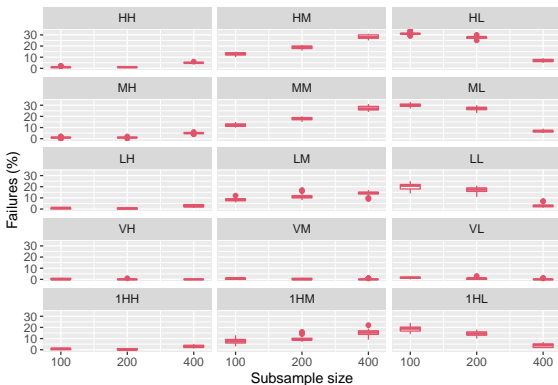


(o) 1HL

Results

subsampling failures

- ▶ probability of failure ↗ if separation ↗ and if imbalance ratio ↘
- ▶ astonishing... but
- ▶ if subsampling does not fail, it works badly

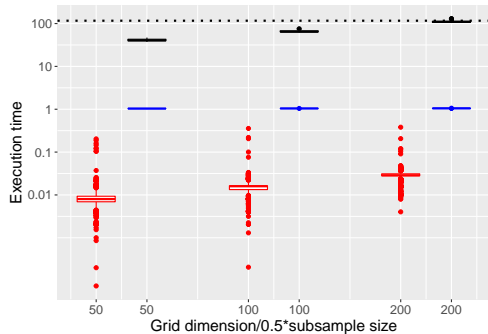


Results

time vs grid/subsample size

equal memory occupancy

- ▶ - **subsampling** EM (red)
 - bin-marginal CL-EM (black)
 - expected CL-EM time **after optimization in language C++** (blue)
 - full dataset (dotted line)
-
- ▶ remarkable improvement relatively to full data set



Real imbalanced datasets

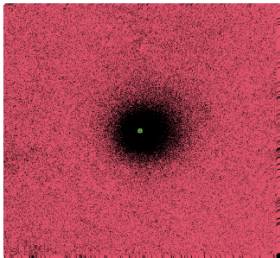
- ▶ image segmentation, fraud detection, hazardous asteroid detection
- ▶ three variables

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

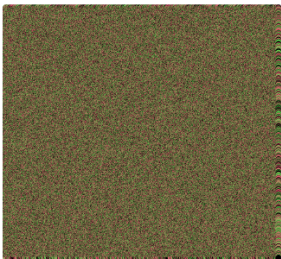
Results: Image segmentation Comet (R=400, K=3)



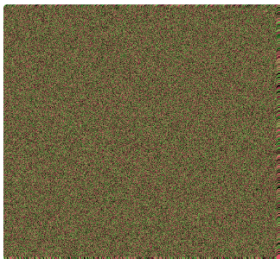
(a)



(b)

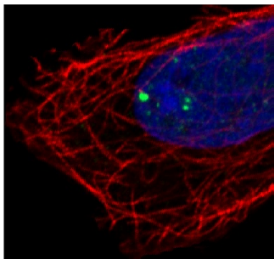


(c)

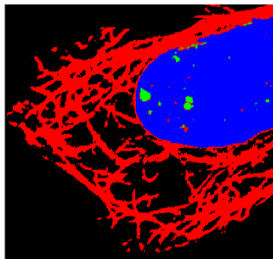


(d)

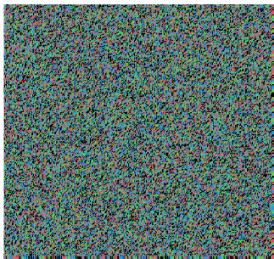
Results: Image segmentation Cell-1 (R=20, K=4)



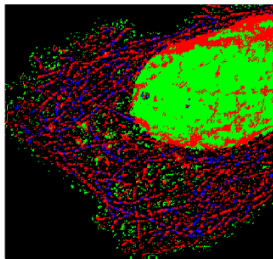
(a)



(b)

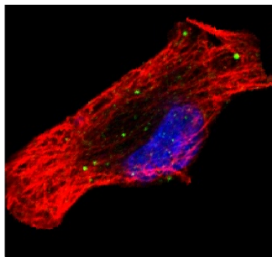


(c)

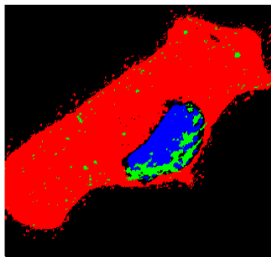


(d)

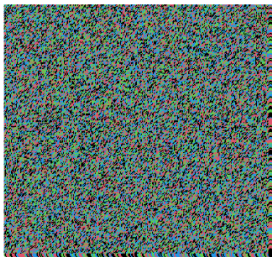
Results: Image segmentation Cell-2 (R=20, K=4)



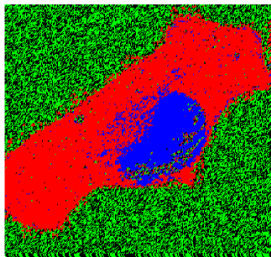
(a)



(b)

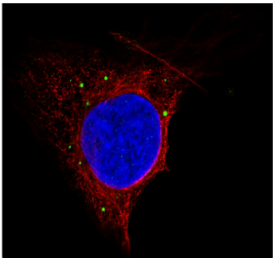


(c)

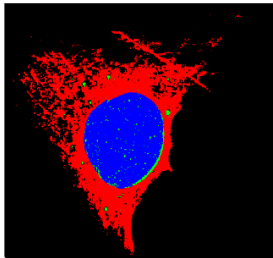


(d)

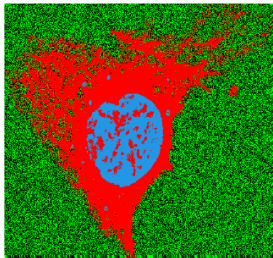
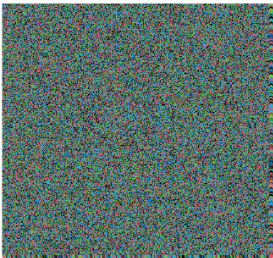
Results: Image segmentation Cell-3 (R=20, K=4)



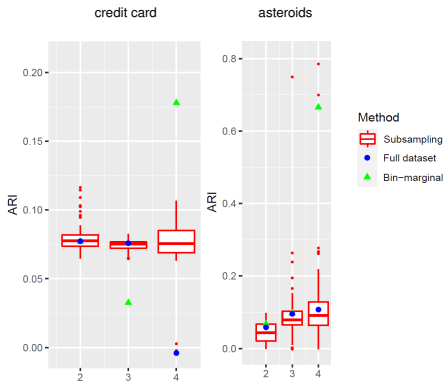
(a)



(b)



Results: credit card and asteroids



- ▶ two known clusters
- ▶ Clustering for $K = 2, 3, 4$, then clusters of highest proportions are grouped (mixture of mixtures)
- ▶ despite the loss of information, binned method seems to behave better than full GMM and subsampling
- ▶ the tiny class contains the abnormal objects (no false negative) but low ARI (false positive)

Plan

- 1 Introduction
- 2 Model
- 3 Estimation
- 4 Experiments
- 5 Discussion

Sum-up

clustering of huge and imbalanced datasets under memory constraints:

- ▶ **bin marginal composite likelihood (bmCL)** approach allows to answer:
 - ↪ memory requirements
 - ↪ tractability of EM algorithm
 - ↪ recovery of tiny classes
 - ↪ not very sensitive to grid coarseness

- ▶ **subsampling**
 - ↪ easy to implement
 - ↪ pb to recover tiny clusters
 - ↪ high variability
 - ↪ number of subsamples?

Discussion

- ▶ **bmCL** clearly outperforms **subsampling** under same memory constraint, and is frugal compared to full sample but
 - ↪ generates a lot of missing data
 - ↪ prone to **slow** convergence, open algorithmic question
 - ↪ hybrid method bmCL / subsampling?
- ▶ **preliminary** study, seminal for further researches
 - ↪ how to deal with frugality while **increasing** number of clusters
 - ↪ strategy when many (tiny) clusters
 - ↪ **grid** definition as a **model choice**?
 - ↪ specific criterion for selecting the number of clusters and grid definition

Thank you for your attention!

Identifiability (main steps)

- ▶ work with binned univariate mixtures of at most K_{max} components: pmf reduces to

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

- ▶ if G has R cut points, (a_1, \dots, a_R) then it is needed to prove that the system has only the trivial solution $\psi = \psi^*$ at a up to a relabeling whatever the grid is

$$\left\{ \begin{array}{l} \pi \sum_{k=1}^K \Phi\left(\frac{a_1 - \mu_k}{\sigma_k}\right) = \sum_{k=1}^{K^*} \pi^* \Phi\left(\frac{a_1 - \mu_k^*}{\sigma_k^*}\right) \\ \pi \sum_{k=1}^K \Phi\left(\frac{a_2 - \mu_k}{\sigma_k}\right) = \sum_{k=1}^{K^*} \pi^* \Phi\left(\frac{a_2 - \mu_k^*}{\sigma_k^*}\right) \\ \vdots \\ \pi \sum_{k=1}^K \Phi\left(\frac{a_R - \mu_k}{\sigma_k}\right) = \sum_{k=1}^{K^*} \pi^* \Phi\left(\frac{a_R - \mu_k^*}{\sigma_k^*}\right) \end{array} \right.$$

- ▶ deduce with [Prop. 11.5 - Valiant 2012] that binned univariate mixtures of at most K_{max} Gaussian distributions are identifiable if the binning grid has $R > 4K_{max} - 3$ cut points.
- ▶ induction for D-variate mixtures

EM algorithm for bin-marginal data

- ▶ complete log-likelihood

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

where $z_{ik} = \mathbb{1}_{\text{observation } i \text{ in cluster } k}$

- ▶ E-step

$$\begin{aligned} Q_m(\boldsymbol{\psi}, \boldsymbol{\psi}^{(j-1)}) &= \mathbb{E}_{\boldsymbol{\psi}^{(j-1)}}[\ell^c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Z}) | \mathbf{m}] \\ &= \sum_{\mathbf{n} \in \mathcal{F}_m} \rho(\mathbf{n} | \mathbf{m}; \boldsymbol{\psi}^{(j-1)}) \mathbb{E}_{\boldsymbol{\psi}^{(j-1)}}[\ell^c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Z}) | \mathbf{n}] \\ &= \sum_{\mathbf{n} \in \mathcal{F}_m} \alpha^{(j-1)}(\mathbf{n}) \mathbb{E}_{\boldsymbol{\psi}^{(j-1)}}[\ell^c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Z}) | \mathbf{n}] \\ &= \sum_{\mathbf{n} \in \mathcal{F}_m} \alpha^{(j-1)}(\mathbf{n}) \sum_{k=1}^K \sum_{b=1}^B n_b \int_{\mathcal{B}_b} \tau_k^{(j-1)}(\mathbf{x}) \\ &\quad \times \log[\pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] d\mathbf{x} \end{aligned}$$