Traitement statistique des donn ees manquantes-Part IV Binned data for big data analysis

Christophe Biernacki

Traitement statistique des données manquantes

Atelier Statistique de la SFdS – 10 et 11 mars 2021

–

Part IV
Binned data for big data analysis

C. Biernacki

Laboratoire P. Painlevé, UMR CNRS 8524 & Université de Lille & Inria

Merci à : F. Antonazzo, C. Boyer, G. Celeux, Q. Grimonprez, J. Jacques, J. Josse, C. Keribin, V. Kubicki, F. Laporte, M. Marbac, A. Sportisse, J. Vandaele, V. Vandewalle

It is an ongoing research work

(with its own notations, possibly differing from Part I/II)

Outline

Context and objectives of scalable clustering

Scalable clustering:

- Context
  - Clustering
  - Arising of Big Data
  - Large sample size n
  - Large dimension d
  - Green computing
- Objectives
  - Reduction of time, memory, energy, technology requirements

State of the art

Basic idea	Specific	Example
Data-reduction	Subsampling Compressing statistics Representative points Grid reduction	BRF (Bradley, 1998) BIRCH (Zhang, 1996) CURE (Guha, 1998) CLIQUE (Agrawal, 1998)
Reduce operations	Data structure Pruning clusters	CLARANS (Kaufmann, 1990) ENCLUS (Cheng, 1999)
Transform space		WaveCluster (Sheikholeslami, 1998)
Subspace clustering		SUBCLU (Kailing, 2004)
New technologies	Map-Reduce Spark	PKMeans (Zhao, 2009) SOKM (Zayani, 2016)

None of them satisfies our requests.

Model-based clustering with GMM

- We have a sample $\underline{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, realizations of a real r.v. X.
- Data come from K different sub-populations Ω_k, $k = 1, \ldots, K$.
- X follows a **Gaussian mixture model** (McLachlan & Peel, 2000) indexed by $\boldsymbol{\psi} \in \boldsymbol{\Psi}$: its p.d.f has the form

$$f(\boldsymbol{x}; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) \quad \boldsymbol{x} \in \mathbb{R}^d,$$

$$\sum_{k=1}^{K} \pi_k = 1, \qquad \pi_k > 0 \quad (k = 1, \ldots, K),$$

where $\boldsymbol{\psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K)$ and $\phi(\cdot)$ are Gaussian densities.

Model-based clustering with GMM

Model-based clustering

1. ML estimation: find $\hat{\boldsymbol{\psi}}$ via EM algorithm maximizing $\ell(\boldsymbol{\psi}; \underline{\mathrm{x}}) = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k))$.
2. Clustering phase: assign $\boldsymbol{x}_i$ to Ω_{k^*} if $k^* = \text{argmax}_{k=1,\ldots,K} \hat{\pi}_k \phi(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$.
3. Evaluate clustering accuracy.

EM algorithm: brief description (Dempster et al., 1977)

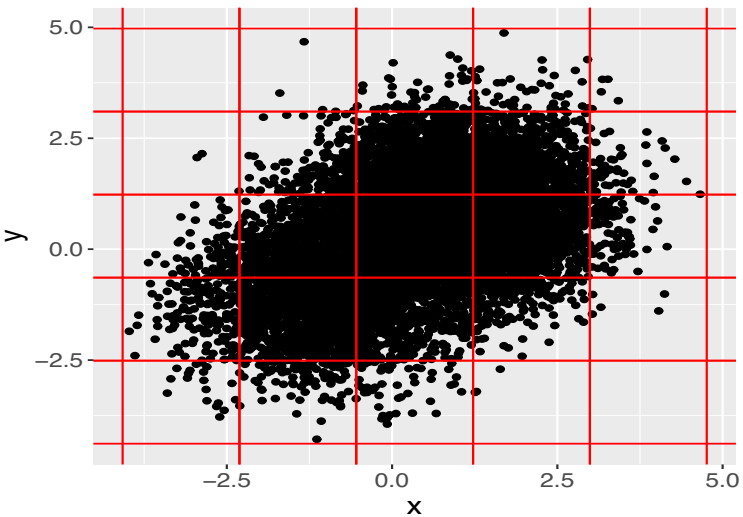
1. Introduce class labels $\underline{\mathrm{z}}$ and fix starting value $\hat{\boldsymbol{\psi}}^{(0)}$.
2. Work with $\ell_c(\boldsymbol{\psi}; \underline{\mathrm{x}}, \underline{\mathrm{z}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k))$.
3. At iteration j, select $\hat{\boldsymbol{\psi}}^{(j+1)} = \text{argmax}_{\boldsymbol{\psi} \in \Psi} \mathbb{E}_{\hat{\boldsymbol{\psi}}^{(j)}}[\ell_c(\boldsymbol{\psi}; \boldsymbol{Z}, \underline{\mathrm{x}}) | \underline{\mathrm{x}}]$,
4. Repeat 3 until $|\ell(\boldsymbol{\psi}^{(j+1)}; \underline{\mathrm{x}}) - \ell(\boldsymbol{\psi}^{(j)}; \underline{\mathrm{x}})| < \epsilon$ (or up to J iterations).

If n and d too large, it requests too much time and memory.

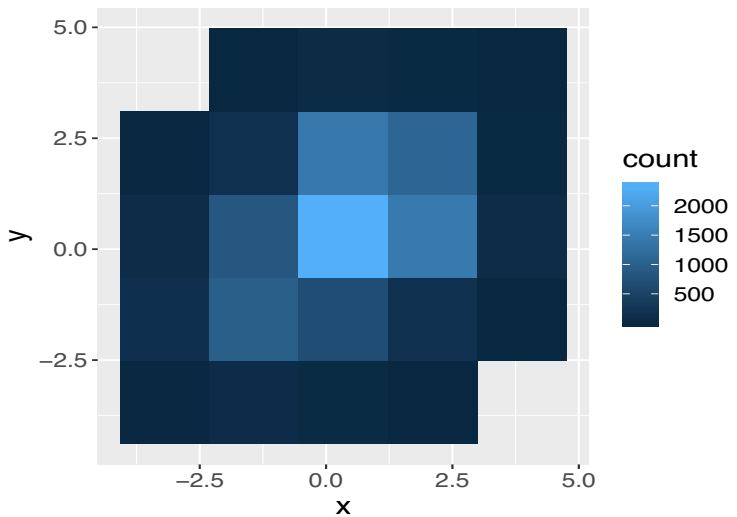
Outline

Binned data: ideas

Binned data appears when it is impossible to collect data with infinite precision (censoring, truncation,...).
Instead of knowing a complete sample $\underline{\mathrm{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \subset \mathbb{R}^d$ we know a vector $\boldsymbol{n}$ of counts of observations lying in some regions of the space delimited by a grid G.



(a) Raw data



(b) Binned data

Binned data: formalization

Suppose:

- $\underline{x} \subset \mathbb{R}^d$ a d-dimensional real sample with p.d.f. $f(\boldsymbol{x}; \boldsymbol{\psi})$,
- $\mathbb{R}^d$ is divided into $R = \prod_{r=1}^{d} R_r$ regions $\{\mathcal{S}_1, \ldots, \mathcal{S}_R\}$ by a Cartesian grid $G = G_1 \times \ldots \times G_d$ of dimension $R_1 \times \ldots \times R_d$.

Binned data $\boldsymbol{n}$

- $\boldsymbol{n} = (n_1, \ldots, n_R)$,
- $n_i = \#\{\boldsymbol{x}_j \in \mathcal{S}_i\} \quad i = 1, \ldots, R$,
- $\boldsymbol{n}$ follows a multinomial model $M(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \ldots, p_R)$,
- $p_i = P(X \in \mathcal{S}_i; \boldsymbol{\psi}) = \int_{\mathcal{S}_i} f(x; \boldsymbol{\psi}) dx \quad i = 1, \ldots, R$,
- Trick for sample size reduction: select $R << n$.

Univariate mixture models with binned data: Estimation

Identifiability of univariate mixture binned models

Identifiability is not trivial. It is satisfied for $R > 4K - 3$.

Aim: ML estimation

Optimize $\ell(\boldsymbol{\psi}; \boldsymbol{n}) = \sum_{i=1}^{R} n_i \log(\int_{\mathcal{S}_i} \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2) dx)$.

Missing data EM algorithm

1. Class labels: $\underline{z}$,
2. Raw data among each bin: $\underline{x}$-.

EM algorithm for univariate binned data: update formulas

For $k = 1, \ldots, K$

$$\hat{\pi}_k^{(j+1)} = \frac{\sum_{i=1}^{R} n_i \hat{A}_{ik}^{(j)}}{n},$$

$$\hat{\mu}_k^{(j+1)} = \frac{\sum_{i=1}^{R} n_i \hat{B}_{ik}^{(j)}}{\sum_{i=1}^{R} n_i \hat{A}_{ik}^{(j)}},$$

$$\hat{\sigma}_k^{2(j+1)} = \frac{\sum_{i=1}^{R} n_i \hat{C}_{ik}^{(j)}}{\sum_{i=1}^{R} n_i \hat{A}_{ik}^{(j)}}.$$

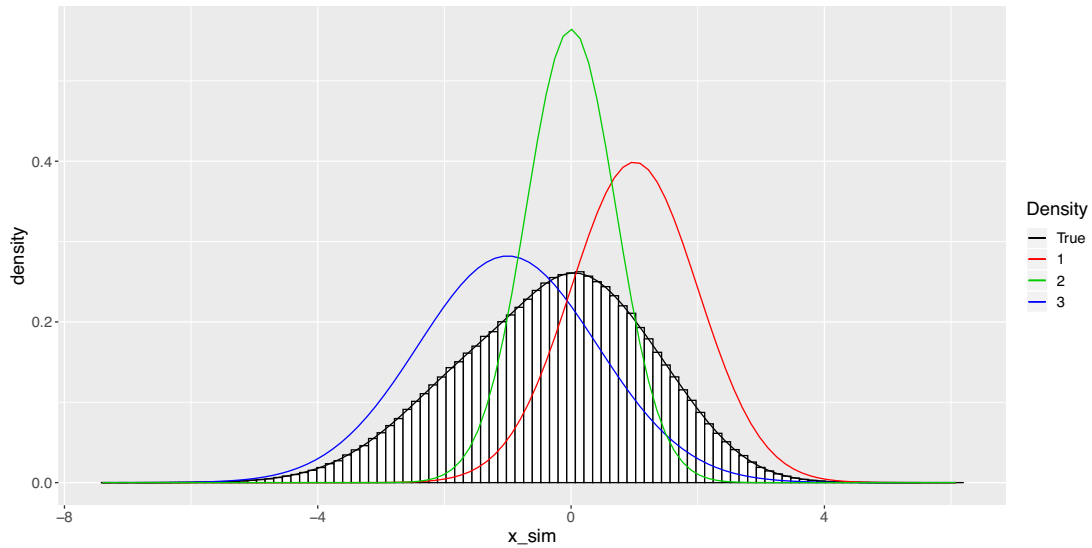
Remark: It could be shown (not displayed) all $\hat{A}_{ik}^{(j)}, \hat{B}_{ik}^{(j)}, \hat{C}_{ik}^{(j)}$ depend only on Gaussian pdfs and cdfs.

$\Downarrow$

Calculation is easy and fast.

Simulation - Raw vs Binned

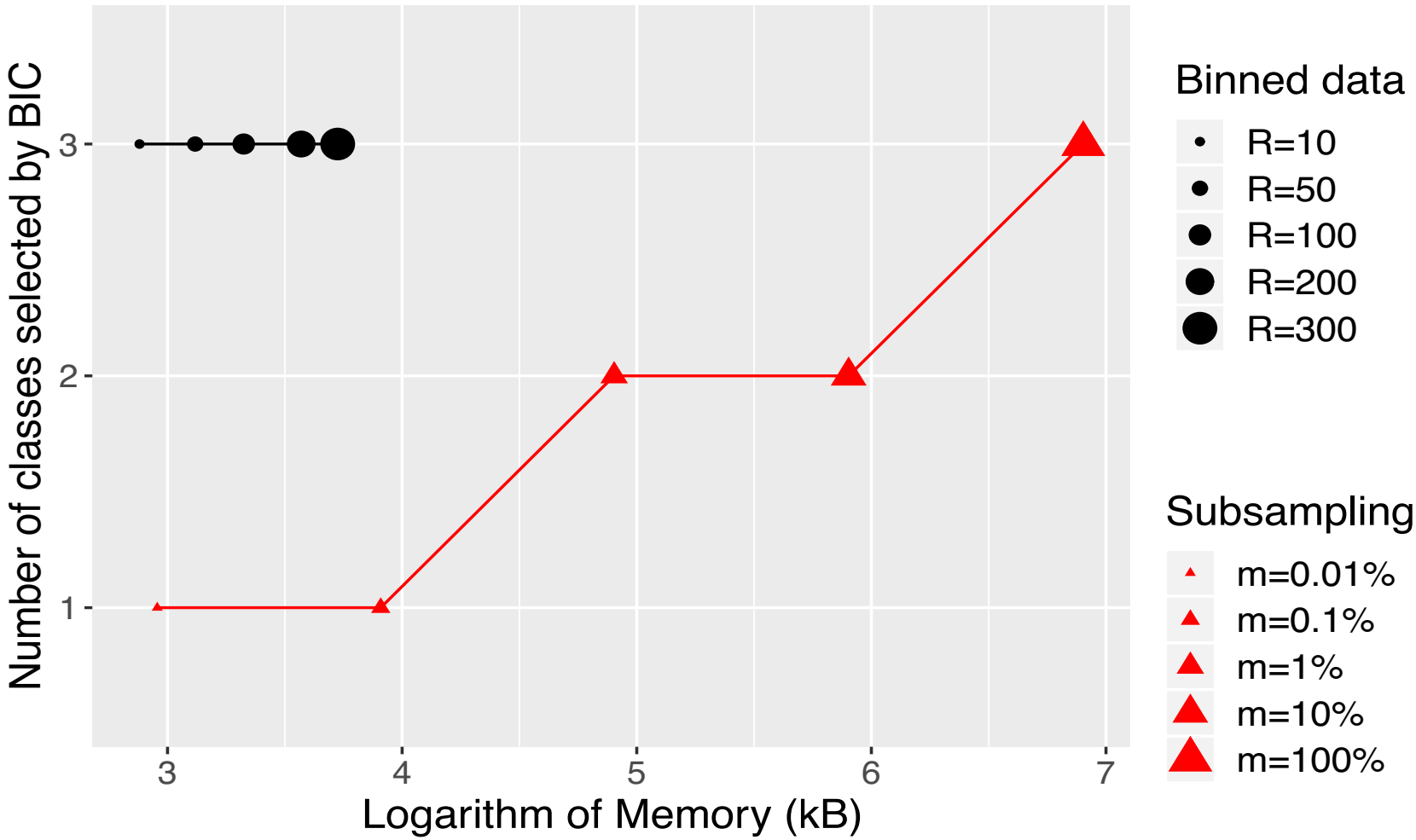
10^6 data from $f(x;\boldsymbol{\psi}) = 0.6\phi(x;-1,2) + 0.3\phi(x;1,1) + 0.1\phi(x;0,0.5)$.



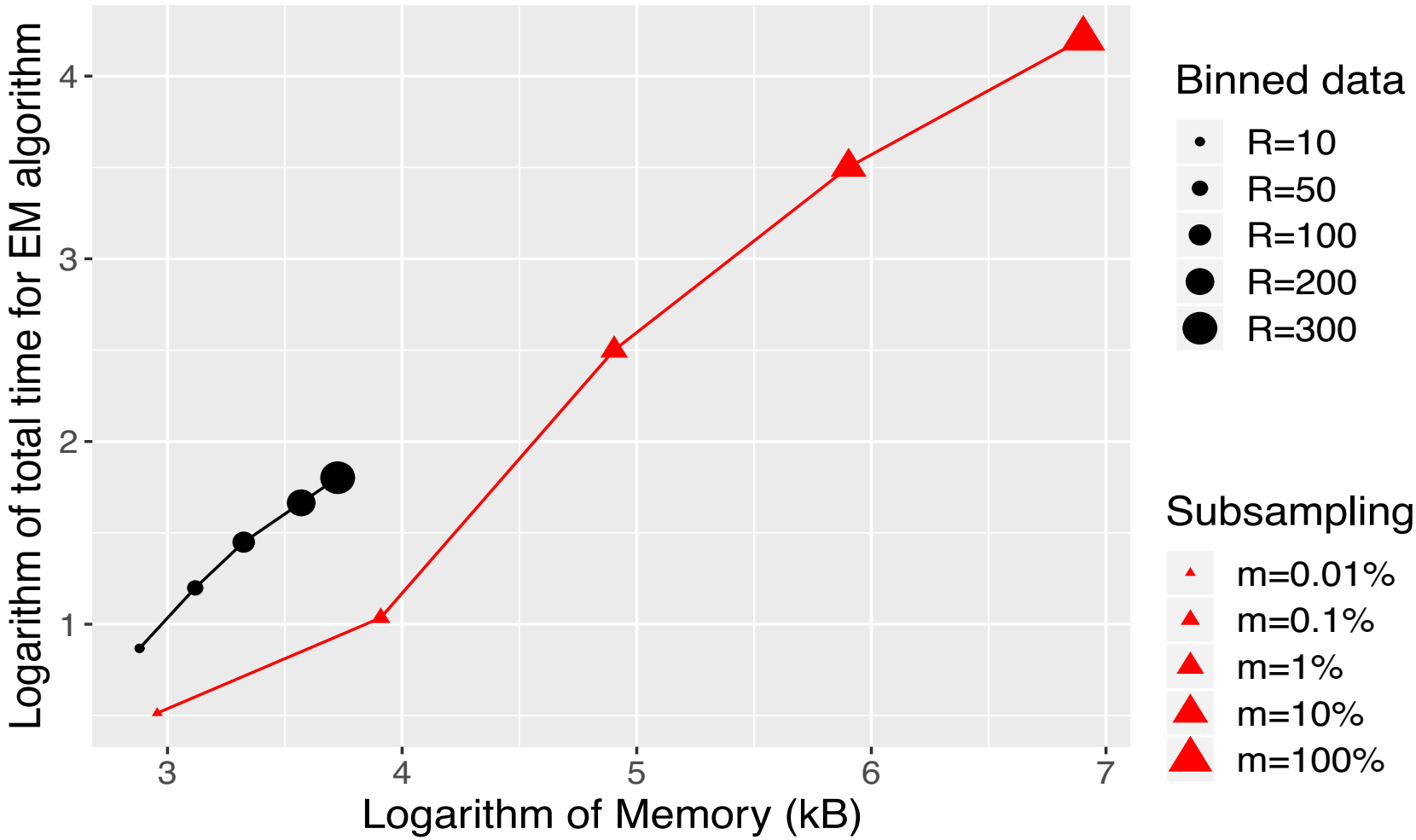
Compare our method with the subsampling one (for different R and subsample percentage m) in terms of:

1. Classes selected by BIC;
2. Time and memory used;
3. Quality of estimation.

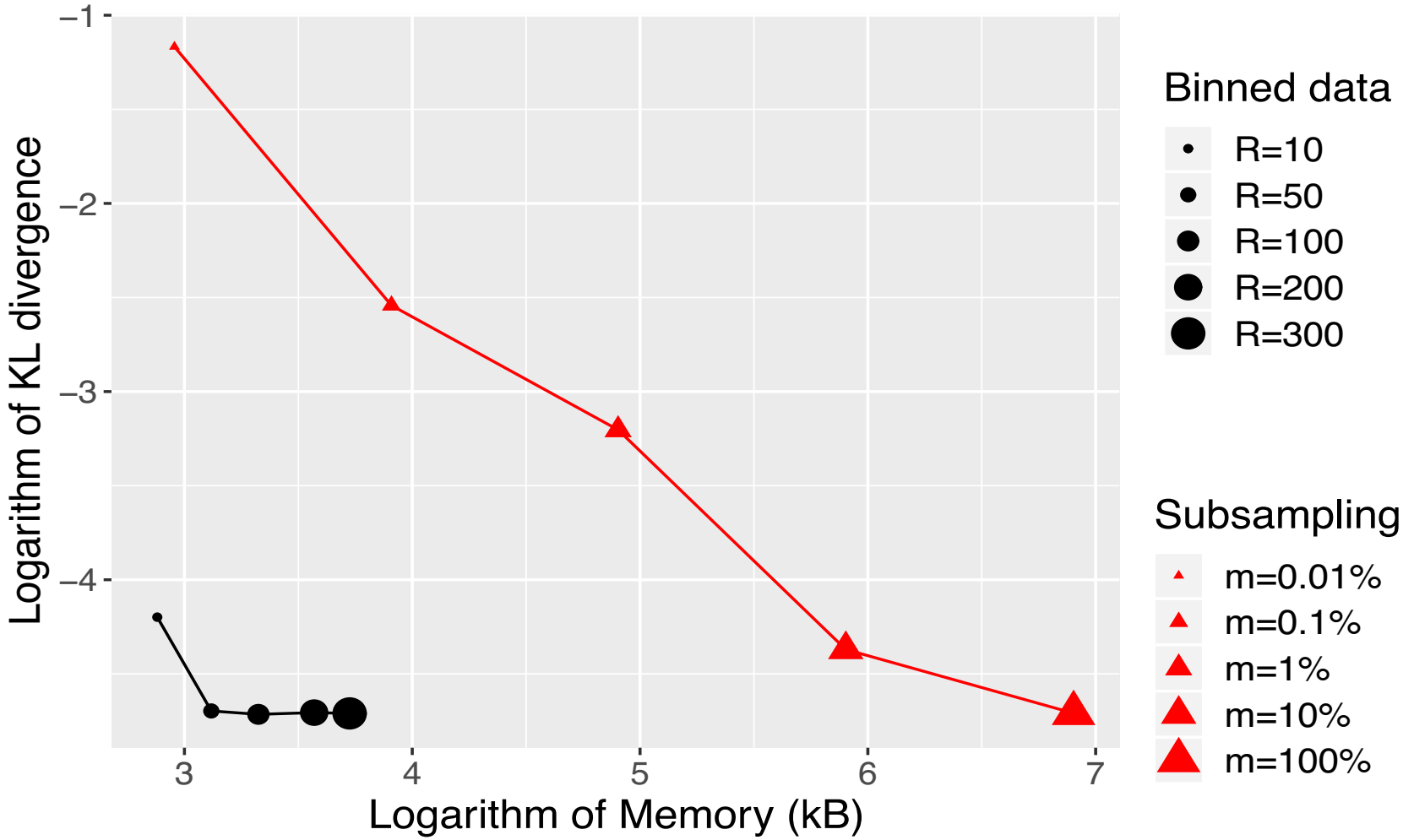
Classes selected by BIC



Total time EM algorithm

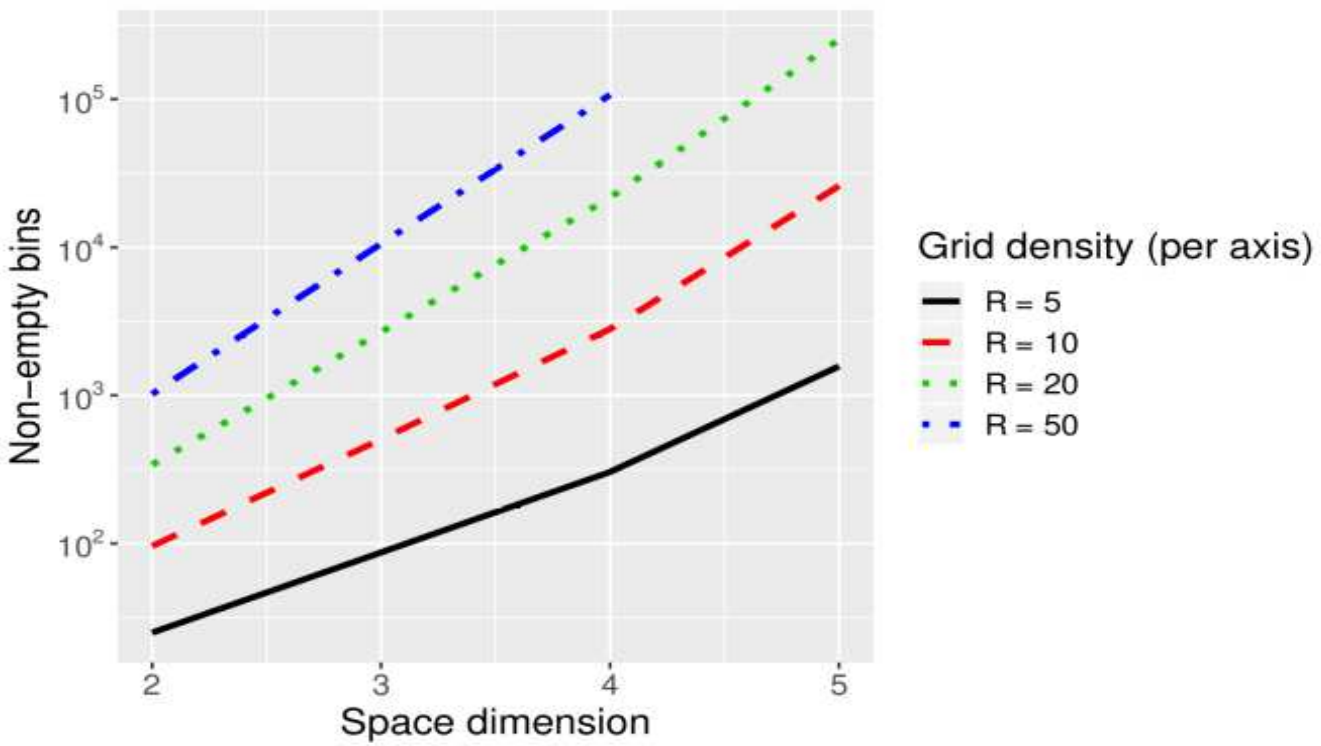


Kullbach-Leibler divergence



Outline

Multivariate bins: dimensional issues



Issue: Non-empty binned data size grows exponentially with the space dimension. $\Rightarrow$ Huge amount of memory is demanded.

Multivariate EM: numerical issues

In a multivariate context, EM updates have this form (Cadez et al., 2002):
For $k = 1, \ldots, K$:

$$\tau_k^{(j)}(\boldsymbol{x}) = \frac{\hat{\pi}_k^{(j)} \phi(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_k^{(j)}, \hat{\boldsymbol{\Sigma}}_k^{(j)})}{f(\boldsymbol{x}; \hat{\boldsymbol{\psi}}^{(j)})},$$

$$\hat{\pi}_k^{(j+1)} = \frac{\sum_{i=1}^R n_i \int_{\mathcal{S}_i} \tau_k^{(j)}(\boldsymbol{x}) d\boldsymbol{x}}{n},$$

$$\hat{\boldsymbol{\mu}}_k^{(j+1)} = \frac{\sum_{i=1}^R n_i \int_{\mathcal{S}_i} \boldsymbol{x} \tau_k^{(j)}(\boldsymbol{x}) d\boldsymbol{x}}{\sum_{i=1}^R n_i \int_{\mathcal{S}_i} \tau_k^{(j)}(\boldsymbol{x}) d\boldsymbol{x}},$$

$$\hat{\boldsymbol{\Sigma}}_k^{(j+1)} = \frac{\sum_{i=1}^R n_i \int_{\mathcal{S}_l} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k^{(j+1)})(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k^{(j+1)})^t \tau_k^{(j)}(\boldsymbol{x}) d\boldsymbol{x}}{\sum_{i=1}^R n_i \int_{\mathcal{S}_i} \tau_k^{(j)}(\boldsymbol{x}) d\boldsymbol{x}}.$$

Issue: d-dimensional numerical integration is burdensome. $\Rightarrow$ Too complex computation.

Bivariate diagonal mixtures

Let consider a bivariate mixture with **diagonal** variances and $K = 2$ (known). Thus, the parameters to estimate are:

$$\boldsymbol{\psi} = (\pi, \underbrace{\mu_{11}, \mu_{21}, \sigma^2_{11}, \sigma^2_{21}}_{\boldsymbol{\psi_1}}, \underbrace{\mu_{12}, \mu_{22}, \sigma^2_{12}, \sigma^2_{22}}_{\boldsymbol{\psi_2}}).$$

1. How to save memory?
2. How to avoid numerical integration?

Marginal counts

If $d = 2$ we have a grid $G = G_1 \times G_2$ of dimension $R_1 \times R_2$.
Denoting with $\underline{x}_1$ and $\underline{x}_2$ the first and second component of $\underline{x}$, we can define:

- $\boldsymbol{n}_1$: binned data R_1-vector of $\underline{x}_1$ under G_1.
- $\boldsymbol{n}_2$: binned data R_2-vector of $\underline{x}_2$ under G_2.

$\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are the marginal counts of $\boldsymbol{n}$.

Trick to save memory: try to work with marginal counts $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ instead of $\boldsymbol{n}$.

Joint EM for marginal counts

Having only marginal counts $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ we have to obtain $\hat{\psi}$ maximizing $\mathcal{L}(\psi; \boldsymbol{n}_1, \boldsymbol{n}_2)$. A classical EM needs numerical integrations and $\mathcal{L}(\psi; \boldsymbol{n}_1, \boldsymbol{n}_2)$ is difficult to calculate, having only this mathematical relation:

$$\mathcal{L}(\psi; \boldsymbol{n}_1, \boldsymbol{n}_2) = \sum_{\left\{\boldsymbol{n}' : \begin{smallmatrix} \boldsymbol{n}'_1 = \boldsymbol{n}_1, \\ \boldsymbol{n}'_2 = \boldsymbol{n}'_2 \end{smallmatrix}\right\}} \mathcal{L}(\psi; \boldsymbol{n}').$$

Proposal: an "alterned" EM to estimate ψ completely based on univariate calculations.

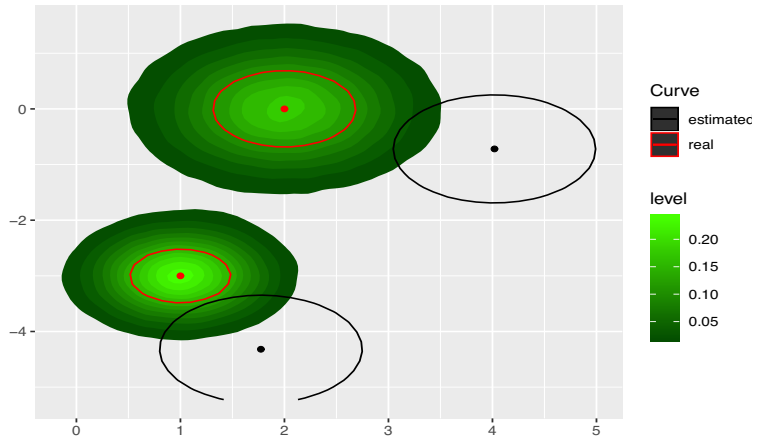
Alterned EM

Starting point: $\hat{\psi}^0 = (\hat{\pi}^0, \hat{\psi}_1^0, \hat{\psi}_2^0)$.

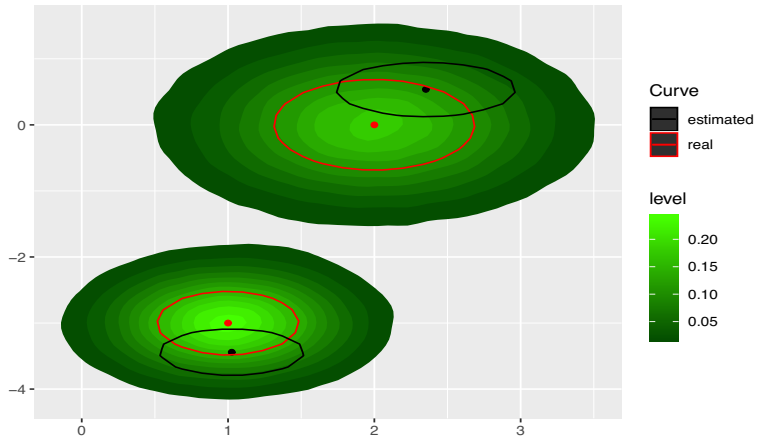
For $j = 0, \ldots$:

1. EM algorithm (1 iteration) on dimension 1 using $\boldsymbol{n}_1$ and $(\hat{\pi}^j, \hat{\psi}_1^j)$ as starting point $\rightarrow (\hat{\pi}^{j+\frac{1}{2}}, \hat{\psi}_1^{j+1})$.
2. EM algorithm (1 iteration) on dimension 2 using $\boldsymbol{n}_2$ and $(\hat{\pi}^{j+\frac{1}{2}}, \hat{\psi}_2^j)$ as starting point $\rightarrow (\hat{\pi}^{j+1}, \hat{\psi}_2^{j+1})$.
3. Current estimate: $\hat{\psi}^{j+1} = (\hat{\pi}^{j+1}, \hat{\psi}_1^{j+1}, \hat{\psi}_2^{j+1})$.
4. Repeat 2-4 until $||\hat{\psi}^{j+1} - \hat{\psi}^j|| < \epsilon$ or up to J iterations.

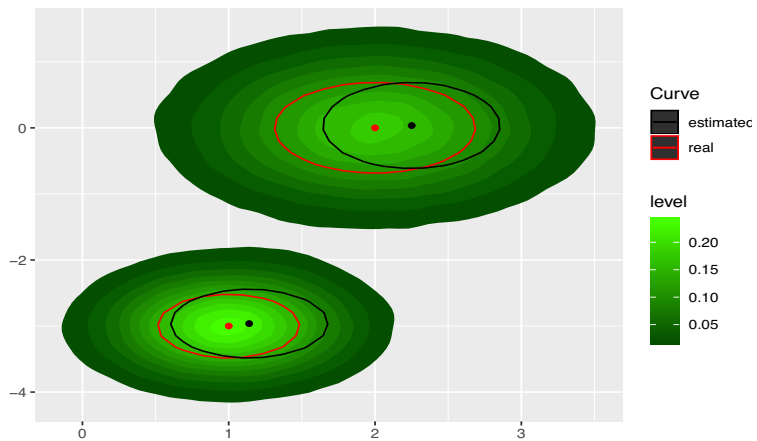
Results



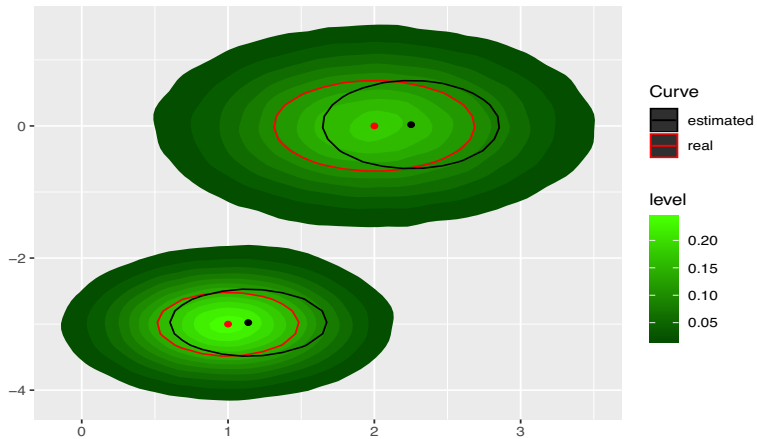
(a) Initial estimation



(b) Grid with 9 bins



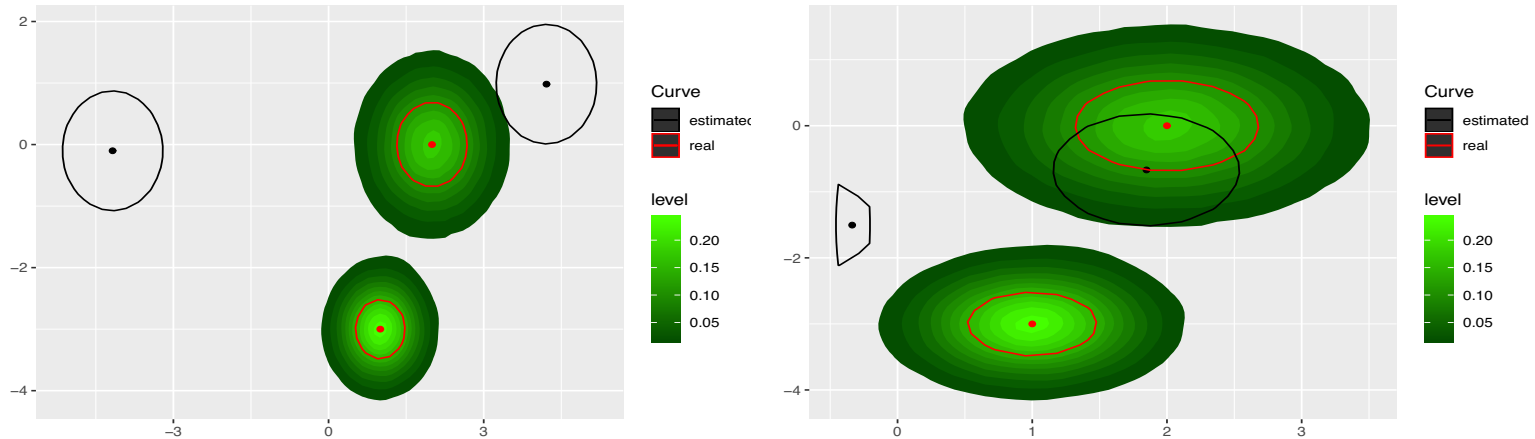
(c) Grid with 100 bins



(d) Grid with 10^4 bins

Results

Estimation from a further initial point.



(a) Initial estimation

(b) Grid with 9 bins

(c) Grid with 100 bins

(d) Grid with 10^4 bins

Future works

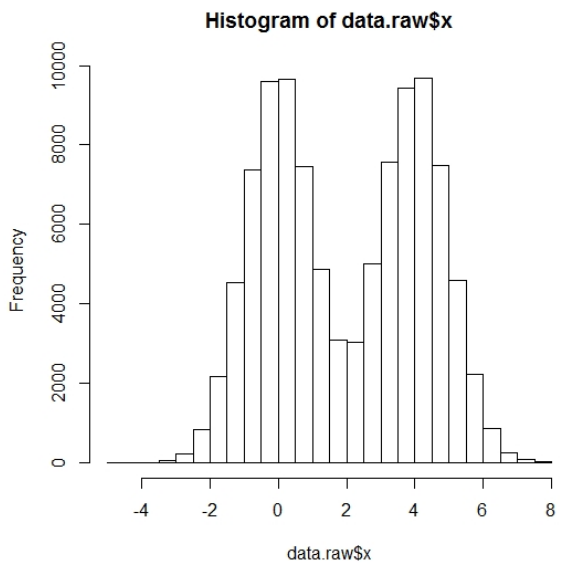
- Assessment of theoretical properties alterned EM algorithm and $\hat{\psi}$.
- Computation of $\mathcal{L}(\psi, \boldsymbol{n}_1, \boldsymbol{n}_2)$.
- Study of local maxima.
- Choice criterion for the binning grid.
- Propose grid candidates.

Outline

Run provided R files (1/2)

binned_pack.R & run_bin.txt

```
# draw raw data
n=100000
p=c(0.5,0.5)
mu=c(0,4)
v=c(1,1)
data.raw=gen_mixt(n,p,mu,v)
```



Run provided R files (2/2)

binned_pack.R & run_bin.txt

```
# EM on binned data
pi0=runif(2)
pi0=pi0/sum(pi0)
m0=runif(2)
s0=runif(2)
tim=1000
tol=1e-6
par.hat=em_bin(data.bin, pi0, m0, s0,tim,tol)
```


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


McLachlan, G. J. & Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 571-578.


Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.


Pandove, D., Shivani, G., & Rani, R. (2018). Systematic Review of Clustering High-Dimensional and Large Datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68.