Estimation of univariate Gaussian mixtures for huge raw datasets by using binned datasets

Filippo Antonazzo[1], Christophe Biernacki[2] & Christine Keribin[3]

[1] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, filippo.antonazzo@inria.fr*
[2] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, christophe.biernacki@inria.fr*
[3] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay 91405 Orsay, France, christine.keribin@universite-paris-saclay.fr*

Résumé. L'intérêt de l'apprentissage non supervisé est magnifié par la croissante constante du nombre d'individus dans les échantillons. C'est en effet l'opportunité de découvrir des informations autrefois inaccessibles. Néanmoins, une importante volumétrie de données pose des difficultés relatives à des temps de calculs rapidement prohibitifs et à la grande consommation d'énergie et des ressources matérielles. L'usage de données regroupées (ou *binned data*) sur une grille adaptative pourrait répondre à ces questions ayant trait à ce qu'on qualifierait aujourd'hui de *green computing*, sans pour autant nuire à la qualité des estimations. Une 1ère approche est menée dans le cadre des mélanges gaussiens univariés, comprenant une illustration empirique et des avancées théoriques.

Mots-clés. Apprentissage non supervisé, données regroupées, big data, green computing.

Abstract. Popularity of unsupervised learning is magnified by the regular increase of sample sizes. Indeed, it provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to prohibitive calculation times and also to high energy consumption and the need of high computational ressources. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the related estimation issues. A first attempt is conducted in the context of univariate Gaussian mixtures, included a numerical illustration and some theoretical advances.

Keywords. Unsupervised learning, binned data, big data, green computing.

1 Introduction

Assuming observations with values belonging to a real space $\mathcal{X}$, binned data correspond to a reduced dataset only containing the counts of observations in given regions of $\mathcal{X}$. In practice, binned data usually appear as soon as it is impossible to collect data with

infinite precision. Thus, such regions are often imposed by the data collecting process itself.

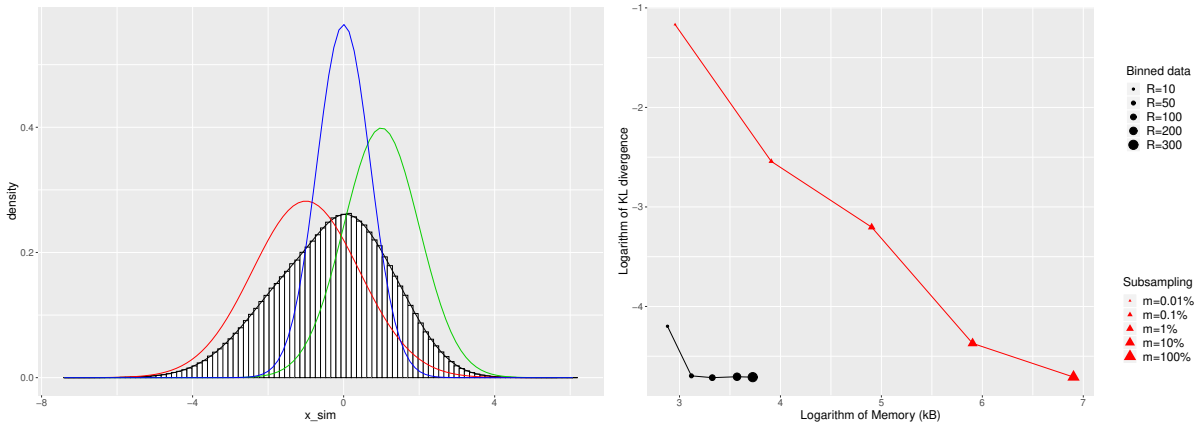
Binned data are so frequent that specific data analysis procedures are designed for them, in particular when regions are too wide to neglect uncertainty they introduce in comparison to the raw (but unavailable) dataset. For instance, in the univariate case ($\mathcal{X} = \mathbb{R}$), McLachlan & Jones (1988) introduced a binned version of the EM algorithm for estimating a univariate Gaussian mixture, whose employment was motivated by an application on red blood cells where only binned and truncated data were available. Induced by a similar problem, Cadez et al. (2002) finally extended this algorithm to the multivariate case.

In this work, we propose to use binned data with a different point of view. We suppose to have a huge amount of raw data and our challenge is to save resources (usually in terms of energy, time and computer memory) while preserving accuracy of the targeting estimation process. The key idea we defend is to group original data in order to obtain *artificially* binned ones. In this way, the size of the resulting dataset is automatically reduced, avoiding too many computing efforts. We focus our attention on the univariate Gaussian mixture estimation in this preliminary work, as a first important step to address more complex situations in the future.

Here is an early numerical example to motivate our proposed "binned" strategy. It illustrates the gain that could be expected in comparison to the classical subsampling strategy usually used for reducing the data size. In this simulation a sample of $n = 10^6$ raw data i.i.d. arises from a univariate Gaussian mixture of three components (Figure 1a), with density

$$f(x;\boldsymbol{\theta}) = 0.6\phi(x;-1,2) + 0.3\phi(x;1,1) + 0.1\phi(x;0,0.5),$$

where $\phi(\cdot;\mu,\sigma^2)$ indicates the univariate Gaussian pdf with mean μ and variance σ^2. Binned data are created through a grid for which a tuning parameter corresponds to its number of finite intervals limits, denoted here by R (more details on the grid will be given later). An EM algorithm was performed respectively with different values of R (thus different candidate binned datasets) and different values of m (thus different candidate subsample datasets).



(a) (b)

Figure 1: (a) Density simulated (black line) with the ones of the three components (red, green and blue lines); (b) Logarithm of Kullback-Leibler divergence from the true parameters for different values of R and m in function of the required computer memory (logarithmic scale).

In Figure 1b it is possible to note that the loss of information (measured by the Kullback-Leibler divergence) induced by binning is much lower than that obtained with subsampling, even negligible if we use a grid moderately dense. This is in addition accompanied by an evident gain in terms of computer memory. Such promising results could be also obtained (but not displayed here) concerning gain in terms of algorithm running time or model selection behaviour.

The outline of the paper focuses on theoretical questionings to be addressed on univariate binned data. It concerns essentially the grid properties: (1) model identifiability, (2) estimates properties and (3) grid selection. We gradually consider these questions firstly in the simplified univariate no mixture Gaussian case (Section 2) and secondly in the univariate mixture Gaussian case (Section 3).

2 Preliminary work: a single univariate Gaussian

2.1 Notations

In the general case, we denote by $\boldsymbol{x} = (x_1, \ldots, x_n)$, with $x_i \in \mathcal{X}$, a raw sample of n observations and by G a grid that divides the space $\mathcal{X}$ into R regions $\mathcal{R}_j$, $j = 1, \ldots, R$. We denote also the resulting binned data vector by $\boldsymbol{y} = (y_1, \ldots, y_R)$, where each component is defined as

$$y_j = \#\{x_i \in \mathcal{R}_j\}, \quad j = 1, \ldots, R.$$

In addition, it is assumed also that the raw sample arises from n continuous i.i.d. random variables with parametric density $f(x;\boldsymbol{\theta})$, $x \in \mathcal{X}$, indexed by a vector of parameter $\boldsymbol{\theta}$, and that $\boldsymbol{y}$ follows a multinomial distribution $M(n,\boldsymbol{p})$ with $\boldsymbol{p} = (p_1,\ldots,p_R)$, where $p_l = \int_{R_l} f(x;\boldsymbol{\theta})dx$.

In the specific case of this section, we suppose that the sample $\boldsymbol{x} \in \mathbb{R}^n$ arises from n i.i.d. univariate Gaussians $N(\mu,\sigma^2)$ with density $\phi(\cdot;\mu,\sigma^2)$. In any univariate context like this, we consider a grid G composed by R points $a_1,\ldots,a_R$ such that we obtain a vector $\boldsymbol{y}$ of $R+1$ binned data where every observation y_j is defined as $(j = 0,\ldots,R)$

$$y_j = \#\{x_i : a_j \leq x_i < a_{j+1}\},$$

while setting $a_0 = -\infty$ and $a_{R+1} = \infty$.

We make also two additional hypotheses in Section 2.2.2 and 2.2.3. First, the variance σ^2 is known and equal to 1. Second, the grids considered are equispaced and symmetric around μ. With these last regularity assumptions, the grids will be simply indexed by two parameters which are the number of points R and the "starting" point a_1. Consequently, each grid will be denoted by $G(a_1,R)$.

2.2 Theoretical results

2.2.1 Identifiability

As discussed in Section 1, first of all, we are interested by a fundamental probabilistic property which is identifiability of the Gaussian distribution, related to the binned nature of available information. In that case, thanks to the monotonicity of the Gaussian cdf, it is possible to prove the following proposition, that ensures identifiability under a slight condition on R.

Proposition 2.1 *Binned univariate normal models are identifiable for* $R \geq 2$.

2.2.2 Estimates properties

The second property is statistical. We note $\hat{\mu}^b_{a_1,R}$ the binned maximum likelihood estimate (MLE) of μ obtained from the binned dataset $\boldsymbol{y}$ with an equispaced grid $G(a_1,R)$ symmetric around μ, and $\hat{\mu}^{MLE}$ the MLE of μ obtained from the raw dataset $\boldsymbol{x}$. Good statistical properties of $\hat{\mu}^b_{a_1,R}$ are assured by the following proposition:

Proposition 2.2 $\hat{\mu}^b_{a_1,R}$ *is asymptotically unbiased and* $\lim_{a_1,R\to\infty} Var(\hat{\mu}^b_{a_1,R}) = Var(\hat{\mu}^{MLE})$.

2.2.3 Grid selection

The question of grid selection is fundamental in our work since its main originality is to estimate an optimal one. In this purpose, we are first interested to access the relative

values of the two tuning parameters of the grid (a_1 and R) to obtain the optimal grid from a variance estimates point of view. The next proposition states that a_1 should decrease at least at logarithmic rate with regards to R (and vice versa). Figure 2 graphically illustrates this fact by comparing the established lower bound and the true optimal value.

Proposition 2.3 *The sequence* $a_1^{(R)} = \max_{a_1<\mu} \frac{Var(\hat{\mu}^{MLE})}{Var(\hat{\mu}^b_{R,a_1})}$ *is bounded below by the sequence* $a^{(R)} = -2\log R + \mu$.

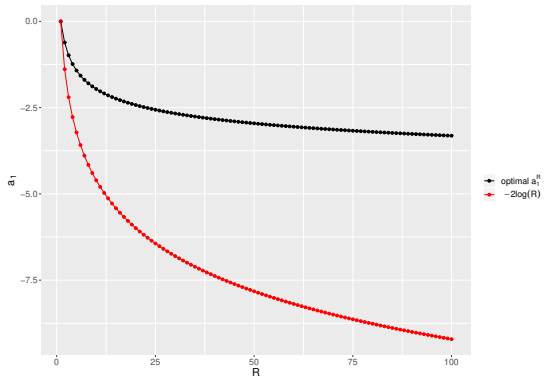


Figure 2: Lower bound for the sequence $a_1^{(R)} = \max_{a_1<0} \frac{Var(\hat{\mu}^{MLE})}{Var(\hat{\mu}^b_{R,a_1})}$ when $\mu = 0$.

The previous statement will be useful to propose sensible grid candidates but the question to select the best grid candidate is open. The following criterion, denoting by $VC^R_{a_1}$, is able to provide the best estimate $\hat{\mu}^b_{a_1,R}$ from the variance point of view, among all the equispaced grids $G(a_1, R)$ symmetric around $\frac{\min(\boldsymbol{x})+\max(\boldsymbol{x})}{2}$ (which is asymptotically equal to μ). Namely, the $VC^R_{a_1}$ criterion is defined by

$$\text{maximize}_{a_1,R} VC^R_{a_1} = \text{maximize}_{a_1,R} \sum_{i=0}^{R} \frac{(\phi(a_i, \hat{\mu}^b_{R,a_1}, 1) - \phi(a_{i-1}, \hat{\mu}^b_{R,a_1}, 1))^2}{\Phi(a_i, \hat{\mu}^b_{R,a_1}, 1) - \Phi(a_{i-1}, \hat{\mu}^b_{R,a_1}, 1)}$$

and its asymptotic property is expressed in the following proposition:

Proposition 2.4 *$VC^R_{a_1}$ criterion is consistent, i.e. the probability of selecting the best $G(a_1, R)$ grid tends to 1 when $n \to \infty$.*

3 Ongoing work: univariate Gaussian mixtures

3.1 Notations

After having considered a single Gaussian, the next step is to consider the more complex case where univariate Gaussian mixtures are involved. Thus, we assume now that each observation

$x_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ arises from a univariate K-Gaussian mixture of density

$$f(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2), \;\; \pi_k > 0, \;\; \sum_{k=1}^{K} \pi_k = 1,$$

in which μ_k denotes the mean of the k-th component, σ_k^2 is its variance and $\boldsymbol{\theta}$ is the vector that contains all the parameters, thus $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. Moreover, as the observations have real values like in the previous case, we can adopt the same notation for the grids considered.

3.2 Theoretical results

3.2.1 Identifiability

In this general case we are able to set a sufficient condition that assures identifiability, which is a consequence of Proposition 11.5 contained in Valiant (2012). It leads to the following proposition.

Proposition 3.1 *Mixtures of K Gaussian distributions for binned data are identifiable for $R > 4K - 3$.*

3.2.2 Other properties as future work

The previous proposition is only a starting point for our research in this context. In fact we are investigating the theoretical properties of the MLE for binned data and we are researching some criteria allowing to select a grid candidate among a family of sensible grids candidates. We expect that the estimates will have the same behaviour of those founded for the single Gaussian situation, but, due to the more complex form of the densities involved, the mathematical tools to be employed may be more advanced. Finally, once resolved this univariate case we will pass to the multivariate one, where new challenges will appear. In particular, the question of the number of non-empty bins when increasing the dimension will be addressed as a solution for limiting the computer memory impact of binned data even in the multidimensional case.

Bibliography

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.

Valiant, G. J. (2012). Algorithmic approaches to statistical questions (Doctoral dissertation, UC Berkeley).