https://www.scilook.eu/index.php/slif/article/view/slif18-01-027
DOI: 10.30888/2415-7538.2020-18-01-027

УДК 519.237

THE SEPARATION OF FINITE MIXTURES OF DISTRIBUTIONS

РАЗДЕЛЕНИЕ КОНЕЧНЫХ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ

Pakhomova A.A. / Пахомова А.А.
student / студентка
ORCID: 0000-0001-5104-1134
Li A.D. / Ли А.Д.
student / студентка
ORCID: 0000-0001-6714-9246
Saint Petersburg State University,
Saint Petersburg, Universitetskii prospect 35, 198504
Санкт-Петербургский государственный университет,
Санкт-Петербург, Университетский проспект 35, 198504

Abstract. *The work is devoted to the study of various methods for constructing estimates of parameters of finite mixtures of distributions, classical algorithms and their applications. The main focus was on maximum likelihood estimation, the Expectation-Maximization (EM) algorithm, and its further testing on modeled data. For conducting research and experiments, a program was written that was used to consider various scenarios of the method behavior depending on the input parameters. It is shown that the EM algorithm can give more accurate estimates with both strongly overlapping and well-separated components for different initial value selections.*

Ключевые слова: *Finite mixtures of distributions, estimates of parameters, Expectation-Maximization algorithm, maximum likelihood estimation*

Introduction.

Numerical data is deeply embedded in our lives, and it is impossible to imagine any industry without studying their features. Many works are devoted to the problems of data analysis, but the potential of working with them is unimaginable, because every day is not like the previous one, and all the data is different from each other.

Primary data analysis begins with descriptive statistics, where a suitable distribution is usually selected for existing observations. But in many areas of research, it is often not enough to "fit" just one classical distribution to the base data set. In particular, if the collected data can be considered as coming from two or more subpopulations, then you need to choose the composition of distributions.

Such compositions are referred to as distributions for the mixtures or mixtures models. They are determined by the parameters of each component and the mixing proportions in which the components of the mixture are located, and will be investigated in this work.

Mixtures of distributions are widely used for mathematical modeling of numerous phenomena in various fields: from biology to Economics and from physics to financial analysis. Therefore, they are a significant and powerful tool for modeling heterogeneous data. And the decomposition of this mixed distribution into its components, which is accompanied by the division of the main groups, can give completely new estimates of parameters and, as a result, new conclusions.

The model of finite mixtures of distributions

Mix distributions are increasingly used to model heterogeneous data in various important practical situations where data can be considered as arising from two or more subpopulations (components). The problem of decomposing a mixture into its components, i.e. estimating the parameters of a mixed distribution, has a long history and goes back to Pearson [1], who considered a mixture of two components with equal variances using the method of moments.

The mix distribution is the addition of distributions that occurs when a sample is taken from a heterogeneous population with different density functions. The final mixture can be defined more formally as follows (distribution of a random variable X with a density function of the form):

$$g(x,\Psi) = \sum_{j=1}^{k} \pi_j f(x,\theta_j)$$

final distribution of the mixture with k components and $\Psi = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)$.

Thus, $f(x,\theta_j),\ j = 1,\dots,k,$ denote the density of the mixture components with the parameter θ_j and the mixing ratio or mixing weight $\pi_1, \dots, \pi_k$ – are positive and $\sum_{i=1}^{k} \pi_i = 1$.

It is not necessary that the density of components $f(x,\theta_j),\ j = 1,\dots,k$, belong to the same parametric family, but throughout the work we will assume that they are Gaussian. Random variable X with density function:

$$f(x,\theta) = \phi(x,\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp(-\tfrac{1}{2}(x-\mu)^2/\sigma^2),$$

where $\theta = (\mu,\sigma^2)$, is called a normal or Gaussian distribution with parameters μ and σ^2.

Two-component Gaussian mixtures

Let's start with two-component Gaussian mixtures, so here are some useful properties and examples. A distribution with a density function

$$g(x,\Psi) = \pi_1\phi(x,\mu_1,\sigma_1^2) + \pi_2\phi(x,\mu_2,\sigma_2^2)$$

it is called a two-component Gaussian mixture, where $\phi(\cdot)$ is the density of the Gaussian mixture and $\Psi = (\pi_1,\pi_2,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)$. To ensure identifiability Ψ, the components of the averages are assumed to be in ascending order, i.e. $\mu_1 < \mu_2$.

Separation of components of a two-component Gaussian mixture with $\sigma_1^2 = \sigma_2^2 = 1$ it can be expressed as the difference between the average components, which is equal to $\Delta = \mu_1 - \mu_2$. Mixtures with $\Delta \leq 2$ are called strongly overlapping mixtures, while mixtures with $\Delta > 2$ are referred to as well-separated mixtures. This classification is approximately equal to the definition of unimodality of a mixture.

There is a fairly sufficient condition that the mixture is unimodal if $\Delta^2 < \frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2+\sigma_1^2)}$. According to this condition, the mixture with $\sigma_1^2 = \sigma_2^2 = 1$ is unimodal for $\Delta > 1.84$. Behboodian [2] also considered this problem and derived the following sufficient condition for a mixture of two Gaussian distributions to be unimodal: $\Delta \leq 2min(\sigma_1^2,\sigma_2^2)$.

Let's assume that *X* is a random variable with a two-component distribution of the Gaussian mixture. The average value of μ_m and the variance σ_m^2 of such a mixture are given by the formulas:

$$\mu_m = \pi_1\mu_1 + \pi_2\mu_2,$$

$$\sigma_m^2 = \pi_1(\sigma_1^2 + \mu_1^2) + \pi_2(\sigma_2^2 + \mu_2^2) - \mu_m^2.$$

Here are some examples of two-component Gaussian mixtures. The first figure shows mixtures with standard deviations $\sigma_1^2 = \sigma_2^2 = 1$, mixing proportions $\pi_1 = \pi_2 = 0.5$, and various average components. Starting with a strongly overlapping mixture, where $\Delta = 1$, the second component is displaced three times until a well-separated mixture with $\Delta = 4$ is formed.

The following fig. 1 shows the change in mixing proportions. Thus, the standard deviations of the components are again chosen to be 1, the average values of the components are μ_1 = 5 and μ_2 = 9, and the mixing proportions are now equal to π_1 = 0.7 and π_2 = 0.8, respectively.

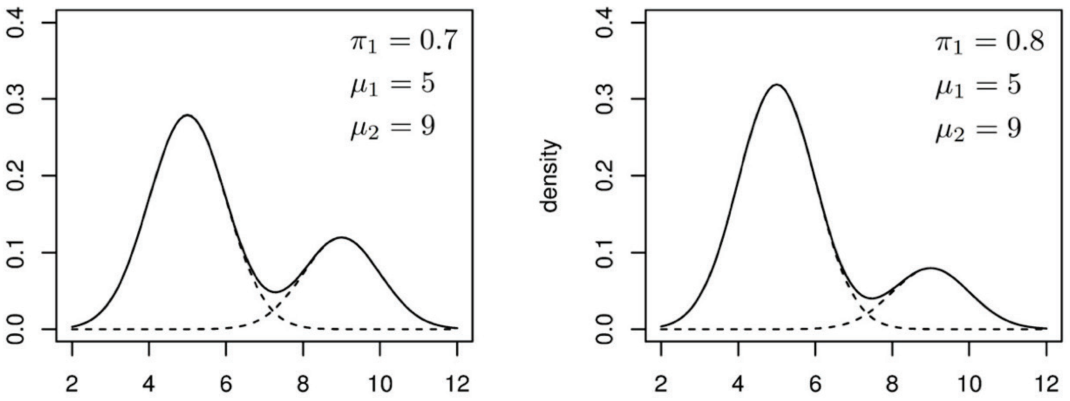


Fig. 1: densities of two-component Gaussian mixtures with equal dispersions and different mixing proportions.

Maximum likelihood estimates

There are a large number of estimation methods, including moment methods, graphical estimation procedures, maximum likelihood estimation, minimum χ^2 estimation, and Bayesian estimation. The question arises as to which estimation method should be used when estimating the distribution parameters of the mixture. A partial answer can be found, for example, in the book [3]. The authors made a comparison between the moment method and the maximum likelihood estimation method and showed that the maximum likelihood estimation is higher. Also, other authors compared several estimation methods and similarly concluded that the best fit is achieved using maximum likelihood estimation.

Therefore, in my work, I use exactly the maximum likelihood estimation, which includes maximizing the likelihood function or, equivalently, maximizing the logarithmic likelihood function.

Let $x = (x_1, \dots, x_T)$ denote independent observations from a random sample of size n of a random variable X with a density function $f(x, \theta)$, where θ is the parameter vector that we want to estimate. Then

$$L(\theta, x) = \prod_{s=1}^{n} f(x_s, \theta)$$

denotes the likelihood function, abbreviated as probability.

It is often more convenient to use the logarithm of the likelihood function instead of the likelihood itself, which is the probability logarithm given as

$$\log L(\theta, x) = \sum_{s=1}^{n} \log f(x_s, \theta)$$

This makes sense, since the logarithm is a monotone transformation and L will take its maxima at the same parameter values as L in the logarithm. The maximum likelihood estimates $\hat{\theta}$ for θ is the value that maximizes the probability $L(\theta, x)$, which is equal to

$$\hat{\theta} = \arg\max_{\theta} L(\theta, x)$$

This definition allows for the possibility of more than one maximum likelihood estimates. In fact, multiple maximums may occur in several practical applications. However, for many important models, the maximum likelihood estimate is unique and, in addition, the likelihood function itself is differentiable and top-bounded. In such cases, a solution can be found by solving the corresponding estimation equation.

Probability of distribution of the mixture

In the case of distribution of the mixture, everything is somewhat more complicated. The probability can be written as

$$L(\Psi, x) = \prod_{s=1}^{n} g(x_s, \Psi) = \prod_{s=1}^{n}\left(\sum_{j=1}^{k} \pi_j f(x_s, \theta_j)\right)$$

$$\log L(\Psi, x) = \sum_{s=1}^{n} \log g(x_s, \Psi) = \sum_{s=1}^{n} \log \sum_{j=1}^{k} \pi_j f(x_s, \theta_j)$$

Corresponding equation:

$$\frac{\partial}{\partial \Psi} \sum_{s=1}^{n} \log \sum_{j=1}^{k} \pi_j f(x_s, \theta_j) = 0$$

they do not have any analytical solutions, so they require a numerical procedure.

There are many different iterative methods to solve this problem, including the NR (Newton-Raphson) algorithm, the estimation method, and the EM algorithm [4]. They all have three main General requirements: (1) choosing reasonable initial values, (2) an iterative algorithm that determines new estimates, and (3) an appropriate stop criterion. However, the differences between these algorithms are huge. While the NR algorithm, depending on the initial values, converges to a solution very quickly, the EM algorithm is much slower, but less sensitive to the choice of initial values.

Everitt [5] compared six algorithms for estimating the parameters of a mixture of two Gaussian distributions and concluded that the NR algorithm and the EM algorithm lead to the most satisfactory results. In this paper, the study of the EM algorithm is chosen because, among other good properties, this algorithm has a convenient implementation, a more accurate calculation, since it does not require the calculation of second derivatives, and, in particular, the most reliable convergence. In addition, the advantage of fast NR algorithm consistency disappears if the separation between components is poor.

Experiment: comparing different samples

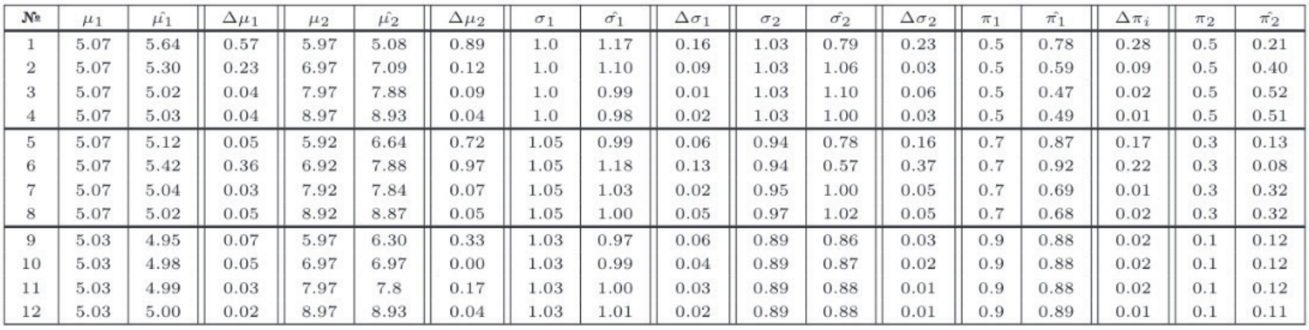
Most problems with finding parameters occur when components are not well separated. We decided to test this and conduct an experiment in which we investigate several models of finite mixtures with a total volume of $N = 400$ artificially modeled observations with different degrees of separation.

Throughout the simulation, the parameters of the first component are fixed with an average value of five and a standard deviation of one. The standard deviation of the second Gaussian distribution is also chosen to be 1, while the average value changes. Thus, an increasing separation between these two components is considered. That is, the difference Δ between the average values is equal to one, two, three, and four.

To start, the mixing proportions are set to $\pi_1 = \pi_2 = 0.5$ and the average value of the second distribution is changed. Starting with highly overlapping components, where the average values are $\mu_1 = 5$ and $\mu_2 = 6$, the second component is shifted three steps until two well-separated components appear.

In addition to well-balanced components, two other cases are investigated. In the first case, the first mixing weight is chosen to be 0.7, and the second is 0.3. The latter case involves a further imbalance of components. Here we select one main component with $\pi_1 = 0.9$ and a small second component with $\pi_2 = 0.1$.

To visualize the numerical comparison of the estimates after applying the algorithm and actual values for different mixtures, the results were furnished in fig. 2.



№	μ_1	$\hat{\mu}_1$	$\Delta\mu_1$	μ_2	$\hat{\mu}_2$	$\Delta\mu_2$	σ_1	$\hat{\sigma}_1$	$\Delta\sigma_1$	σ_2	$\hat{\sigma}_2$	$\Delta\sigma_2$	π_1	$\hat{\pi}_1$	$\Delta\pi_i$	π_2	$\hat{\pi}_2$
1	5.07	5.64	0.57	5.97	5.08	0.89	1.0	1.17	0.16	1.03	0.79	0.23	0.5	0.78	0.28	0.5	0.21
2	5.07	5.30	0.23	6.97	7.09	0.12	1.0	1.10	0.09	1.03	1.06	0.03	0.5	0.59	0.09	0.5	0.40
3	5.07	5.02	0.04	7.97	7.88	0.09	1.0	0.99	0.01	1.03	1.10	0.06	0.5	0.47	0.02	0.5	0.52
4	5.07	5.03	0.04	8.97	8.93	0.04	1.0	0.98	0.02	1.03	1.00	0.03	0.5	0.49	0.01	0.5	0.51
5	5.07	5.12	0.05	5.92	6.64	0.72	1.05	0.99	0.06	0.94	0.78	0.16	0.7	0.87	0.17	0.3	0.13
6	5.07	5.42	0.36	6.92	7.88	0.97	1.05	1.18	0.13	0.94	0.57	0.37	0.7	0.92	0.22	0.3	0.08
7	5.07	5.04	0.03	7.92	7.84	0.07	1.05	1.03	0.02	0.95	1.00	0.05	0.7	0.69	0.01	0.3	0.32
8	5.07	5.02	0.05	8.92	8.87	0.05	1.05	1.00	0.05	0.97	1.02	0.05	0.7	0.68	0.02	0.3	0.32
9	5.03	4.95	0.07	5.97	6.30	0.33	1.03	0.97	0.06	0.89	0.86	0.03	0.9	0.88	0.02	0.1	0.12
10	5.03	4.98	0.05	6.97	6.97	0.00	1.03	0.99	0.04	0.89	0.87	0.02	0.9	0.88	0.02	0.1	0.12
11	5.03	4.99	0.03	7.97	7.8	0.17	1.03	1.00	0.03	0.89	0.88	0.01	0.9	0.88	0.02	0.1	0.12
12	5.03	5.00	0.02	8.97	8.93	0.04	1.03	1.01	0.02	0.89	0.88	0.01	0.9	0.89	0.01	0.1	0.11

Fig. 2: Table with numerical results of the algorithm (*Author's development)*

Conclusion and conclusions

Can see how with increasing separation between the two components of $\Delta = 1$ to $\Delta = 4$, the method established for 25 iterations in almost all cases is coming to all real parameter values.

The paper focuses on two-component Gaussian mixtures. In practice, however, the composition of arbitrary mixtures of distributions is possible. Thus, if two clearly well-separated peaks can be observed, or if there is an approximate knowledge of the two components, the proposed procedure will lead to adequate evaluation results. If more than two peaks appear, the method presented can be easily adapted.

References

1. Pearson, K. (1894) Contributions to the Mathematical Theory of Evolution., 185, 71-110.

2. Behboodian, J. On the modes of a mixture of two normal distributions. *Technometrics*, 12(1), pp. 131–139, 1970.

3. Eisenberger, I. Genesis of bimodal distributions. *Technometrics*, 6(4), pp. 357–363, 1964.

4. Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), pp. 1–38, 1977.

5. Everitt, B. and Hand, D. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

Аннотация. *Работа посвящена изучению различных методов построения оценок параметров конечных смесей распределений, классическим алгоритмам и сферам их применения. Основное внимание уделялось оценке максимального правдоподобия, алгоритму Expectation-Maximization (EM) и его дальнейшему апробированию на смоделированных данных. Для проведения исследования и экспериментов написана программа, с помощью которой были рассмотрены различные сценарии поведения метода в зависимости от входных параметров. В ходе работы показано, что алгоритм EM при разных выборах начальных значений может давать более точные оценки как с сильно перекрывающимися, так и хорошо разделенными компонентами.*

Ключевые слова: *Конечная смесь распределений, алгоритм EM, оценка параметров, оценка максимального правдоподобия*

Статья отправлена: 21.06.2020 г.