UDK004.021

O.V. Kasitskij, P.I. Bidyuk, O.P. Gozhij

APPLICATION OF EXPECTATION MAXIMIZATION THEORY TO SOLVING THE PLOBLEM OF SEPARATION THE MIXTURE OF GAUSSIANS

Abstract.The paper is directed to the study of computationally effective algorithm of modeling and forecastingof optimization type. An analysis is given for the method of Expectation Maximization (EM algorithm), its advantages and disadvantages are considered. A derivation of the algorithm and its detailed description are provided. Some recommendations are given regarding parameter tuning for the algorithm developed. The work highlights a technique for separating the Gaussian mixture using iterative algorithm based on the EM-theory. The results of computing experiments for the EM-algorithm are presented using as example Gaussian mixture separation for two random variables. The conclusions are made regarding the possibilities of application the technique in different conditions.

Keywords. Gaussian mixture, expectation maximizationchoice of the algorithm parameters

Introduction

An expectation maximization theory and respective algorithms (EM algorithm) are used in mathematical statistics for computing of maximum likelihood estimates of probabilistic model parameters in cases when the models depend on some non-measurable variables and incomplete data. The EM algorithm is functioning iteratively and each iteration includes two basic operations. The expectation step (E-step) is used for computing an expected value of a likelihood function using current approximation for non-measurable variables. The maximization step (M-step) is used for selected model parameter estimation that maximize the likelihood computed at the previous step (i.e., at E-step).

The EM algorithm is often used for data clustering, machine learning and in computer vision systems. In the natural language processing systems the Baum-Welch algorithm is often used which is a special case of generalized EM algorithm. Thanks to the possibility of its functioning in conditions of data loss for some variables the EM algorithm also became useful for portfolio risks estimation. Also this theory is used in medical image recognition, especially in the positronemissiontomography and the single-photonemissioncomputer tomography.

For the first time the iterative procedure like EM algorithm, that

 ISSN 1562-9945

provided a possibility for numerical solution of the likelihood function maximization in the problem of probabilistic distributions separation, was proposed in the study [1]. Later on the idea was exploited in the works [2, 3, 4, 5, 6, 7]. After that it was systematically studied in the work [8]. The name of EM algorithm was proposed in the work [7], devoted to the application of the maximum likelihood approach to the statistical parameter estimation in conditions of incomplete data.

In the study [7] the concept of EM algorithm was proposed as a technique for incomplete data processing. This concept is very handy from methodological point of view and provides a good explanation for an idea of the method. The concept itself has been accepted in further analysis of the algorithm.

As a rule EM algorithms are hired for finding solutions of the problems of two types. To the first type belong statistical problems that are directed towards analysis of incomplete data, i.e. when some statistical data cannot be accessed due to quite definite reasons. Another type of problems create statistical problems that are related to such likelihood functions that do not allow an application of handy analytical research techniques but allow serious simplifications if we can add to the problem additional “non-measurable” (unobserved, hidden, latent) values. Some examples of the second type problems create the problems of image recognition and picture reconstruction. A mathematical core of these applied problems create cluster analysis techniques, classification tasks and the problems of probabilistic mixtures.

The method of sliding separation of the mixtures is at the basis of the proposed lately approach to the study of stochastic structure of chaotic informational streams in complex telecommunication nets [1, 2]. This approach is based on the stochastic model of the telecommunication net in the frames of which it is represented in the form of superposition of some simple series and parallel structures. The principle of maximum entropy in combination with the limit theorems from probability theory are naturally leading to the state that the model generates the mixtures of the gamma type distributions for a parameter that reflects the execution time (processing time) of a request from the net. The parameters of the mixture of gamma distributions generated characterize the stochastic structure of informational streams in the net. To solve the problem of statistical parameter estimation for the mixtures of exponential and

gamma distributions (in the problem of mixture separation) the EM algorithm modification is used.

To study the changes of stochastic structure of informational streams in time the EM algorithm is used in the mode of sliding window. It is very important in the frames of this approach to select an appropriate version of the EM algorithm that provides high execution rate and handy interpreting of the results achieved. This study considers in detail some properties of the EM algorithm and its frequently used modifications and a new approach is proposed directed towards enhancement of precision and stability of the EM algorithm and improvement of interpreting of its functioning results when solving the problem of mixture separation.

The main focus is made here to application of the EM algorithm to the problem of separation of normal mixtures. The problems of studying of such mixtures comprise a kernel for the method of volatility decomposition for financial indexes [3, 4] and turbulent plasma study [5].

In the probability theory the mixture of random variables is defined as a probabilistic distribution of random variable the values of which can be extracted from one of the subordinated probabilistic distributions.

The mixtures of distributions allow to represent complex distributions in the form of simpler ones, and they are used thanks to the fact that they describe well a large number of data samples from real life problems, and thanks to the easy processing of the mixture components.

Consider a set of points in the plain presented in Fig. 1. For simplicity of representation the points are shown in the plain though the theory given below is consistent with the sets of points of any finite dimensionality.

The points in the picture seem to be grouped in clusters. One cluster to the right is noticeably separated from the others. Two more clusters to the left are disposed close to each other and it is not quite clear if it is possible to correctly put a separating line between them.

The problem statement for separating a mixture of distributions

The problem of distribution mixture separation could be defined as follows. Consider a set of N points in D-dimensional space, $x_1, x_2, \ldots, x_N$, and the family F of probabilistic densities in the space ; it is necessary to find a probabilistic density $f(x) \in F$ such, that

 ISSN 1562-9945

the probability of generating the set of points, $x_1, x_2, ..., x_N$, from this density will be maximum. One of the often used approaches to defining a family of distributions is in providing all of its members with the same mathematical form, and to distinguish them with different values of parameters θ.

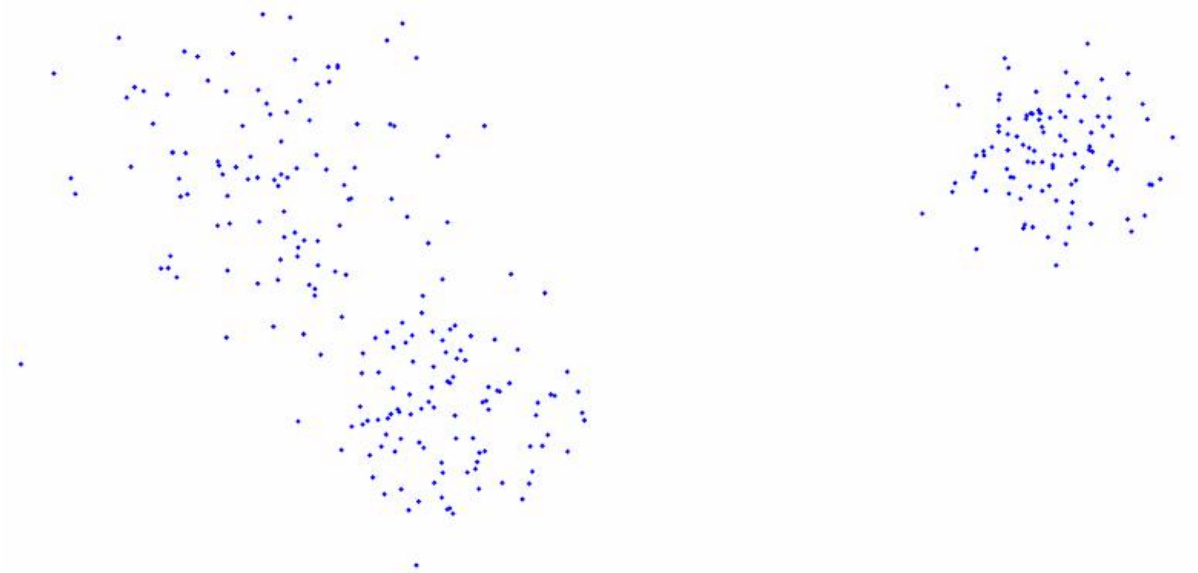

Figure 1 –300 points on the plain

Parametric model

In the following we will be considering the functions, $f(x$, that represent the mixtures of normal distributions:

$$f(x,\theta) = \sum_{k=1}^{K} p_k g(x; m_k \sigma_k),$$

where

$$g(x; m_k \sigma_k) = \frac{1}{(\sqrt{2\pi}\sigma_k)^D} e^{-\frac{1}{2}\left(\frac{\|x-m_k\|}{\sigma_k}\right)^2}$$

is a density of distribution for normal D –dimensional isotopic Gaussian random value; $\theta = (\theta_1, \theta_2, ..., \theta_K) = ((p_1, m_1, \sigma_1), ..., (p_K, m_i)$ is $K(D$-dimensional vector that includes the probabilities of mixing p_k, mathematical expectations m_k, and standard deviations σ_k, that belong to the K Gaussian distributions.

The density of each distribution, integrated over the space R^D, gives a unity:

$$\int_{R^D} g(x; m_k \sigma_k) dx = 1;$$

Here is a density function of probabilistic distribution, that is why should also integrate to unity:

;

$$\int_{R^D} f(x, \theta) dx = \int_{R^D} \sum;$$

$$\int_{R^D} \sum_{k=1}^{K} p_k g(x; m_k \sigma_k)\, dx = \sum_{k=1}^{K} p_k \int_{R^D};$$

$$\sum_{k=1}^{K} p_k \int_{R^D} g.$$

Thus, sum of the numbers , is a unity. It should also be noticed that the numbers are nonnegative (because the function is nonnegative). This fact explains why the numbers p_k are called the probabilities of mixing.

The generative model

The Gaussian mixtures have been studied well enough in direction of modeling the cluster points: each cluster is assigned to Gaussian with mathematical expectation somewhere in the middle of a cluster, and standard deviation that in some manner describes divergence of a cluster points.

Another view on this modeling problem is that the points in Fig. 1 could be generated through repetitive execution of the two-step procedure given below N times, with one run for each point, x_n:

1. Generate a random value from the set $\{1,2,\ldots,K\}$ in a way that the probability of getting k-th value is p_k. This provides a possibility to select a Gaussian from which will be generated the point x_n.

2. Generate random vector x_n from the k-th Gaussian distribution that is defined by the function $g(x; m_k \sigma_k)$.

 ISSN 1562-9945

Due to the fact that the defined above family of Gaussian mixtures is parametric, the problem of density estimation could be defined more exactly as a problem of finding the parameter vector θ such, that the mixture function $f(x,\theta)$ is generating the set of points x_n with maximum probability.

It is still necessary to establish what means "with maximum probability". That is, it is necessary to find the function $L($, that measures the likelihood of some definite model with condition that the set of random values is available.

Maximum likelihood approach

Now apply the method of maximum likelihood. The probability of getting the point (value) in a small volume of dx near the point x is equal to the value of $f(x,\theta)dx$. If the points x_n are generated independently, the probability of getting N points will be defined by the expression: $f(x_1,\theta)dx \ldots f(x_N,$. The volume dx is a constant, so it can be ignored in the process of maximizing the probability. Thus, the likelihood function can be written as follows:

.

The parameter estimation problem is formulated in the following way:

$$\hat{\theta} = arg\max_{\theta} L(X;\theta).$$

Determining the probabilities

To continue the problem solving it is useful to introduce the following function:

.

This expression is needed for determining the mixture density. It was assumed in the definition of the likelihood function, $L(X;\theta)$, that generating of the k-th component from generating model is independent on generation of the value x_n from definite component. It follows from this fact that $q(k,n)dx$ is full probability of generating the component k; and the value of x_n is generated by making use of this component.

The problem solution

As it is shown in [6], the problem under consideration has the following solution:

; (1)

((2)

(3)

The first two expressions we can understand on intuitive level as far as and are sample mean and standard deviation, respectively. They are weighted with conditional probabilities that the points (values) were received with the model k. The third equation for mixing the probabilities is not so obvious though not complicated for understanding as far as p_k could be found as a sample mean for conditional probabilities, $p(k|n)$.

An iterative procedure construction

The equations (1) – (3) are closely connected with each other due to the fact that conditions $p(k|n)$ in the right hand side depend on all variables in the left hand side of other equations. Because of this reason the system (1) – (3) cannot be solved directly. However, the EM algorithm provides a possibility to construct an iterative procedure for solving the problem.

The algorithm implementation

To perform the computational experiments the EM algorithm was implemented on the basis of equations (1) – (3). As an estimate for clustering quality the following quadratic criterion was hired:

$$\boldsymbol{Eps} = \frac{1}{K}\sum_{k=1}^{K}(m_k - m_k^*).$$

Consider the quality of separation of two Gaussians with unity standard deviation, $\sigma_1 = \sigma_2$, and with mathematical expectations, $m_1 = -m_2$the sample size was selected at 1000000.

 ISSN 1562-9945

Table 1

Results of EM algorithm application to mixture separation

№	a	Eps
1	1024	1.36e-6
2	256	3.99e-6
3	64	5.09e-6
4	16	6.82e-7
5	4	1.80e-6
6	1	4.04e-6
7	0.25	6.21e-5

Asitcanbeseenfromthetable 1, thequalityoftheEMalgorithmapplication for separating the mixtures of random variables remains quite acceptable even in the case when the distance between mathematical expectations is less than standard deviation. Now consider in some detail the values of a that are less than 4.

Table 2

EM algorithm application in cases when $a \le 4$

№	a	Eps
1	4	1.37e-6
2	2	1.13e-5
3	1	1.39e-5
4	S	1.13e-4
5	j	1.34e-3
6	1/8	3.66e-3
7	1/16	3.15e-3

The computational results, given in table 2, show that the algorithm used is performing worse when the distance between mathematical expectations of the distributions selected is less than three standard deviations what is easily explained by the three sigma rule. However, even in these cases the quality criterion is changing slowly and the convergence time of the algorithm is growing exponentially.

Conclusions

The paper provides a theoretical analysis of the target sphere of application of the EM theory. The formal problem statement was per-

formed and the method of its solving was considered in necessary detail. The ideas behind the EM theory are presented and a theoretical substantiation for the EM algorithm is given, including the problem of its convergence.

The notion of random variables mixture was introduced and the problem of the mixture separating was formulated. The parametric model for the problem is presented. Also an equivalent generative model is given that was used for constructing the algorithm for generating the Gaussian mixture of appropriate dimensionality. An iterative procedure for solving the problem of Gaussian separation was presented on the basis of EM theory.

The computational experiments performed showed that the EM algorithm was functioning a little worse when the distance between the mathematical expectations of the distributions used was less than three standard deviations what is explained easily by the three sigma rule. At the same time the quality of the distributions separation remained at acceptable level.

In the future research it is necessary to study the possibilities for effective implementation of the EM algorithm based on modern distributed and multiprocessor computer systems.

REFERENCES

1. Batrakova D.A., Korolev V.Yu. A probabilistic and statistical analysis of chaotic informational streams in telecommunication nets using the method of sliding mixture separation // The Systems and Means of Informatics. Special issue. IPIRAN, Moscow, 2006, pp. 183-209. (in Russian)
2. Batrakova D.A., Korolev V.Yu., Shorgin S.Ya. A New method for probabilistic and statistical analysis of informational streams in telecommunication nets // Informatics and its Applications, 2007, №1, pp. 13-22. (in Russian)
3. Korolev V.Yu. A New approach to determining and analysis of volatility components for financial indexes // Actuary, 2007, № 1, pp. 47-49. (in Russian)
4. Korolev V.Yu., Lomskoj V.A., Presnyakov R.R., Rey M. Analysis of volatility components by the method of sliding separation of mixtures // The Systems and Means of Informatics. Special issue. IPIRAN, Moscow, 2005, pp. 180-206. (in Russian)

5. Korolev V.Yu., Skvortsova N.N. A New method of probabilistic and statistical analysis for the processes of plasma turbulence // The Systems and Means of Informatics. Special issue. IPIRAN, Moscow, 2005, pp. 126-179. (in Russian)
6. Dellaert F. The expectation maximization algorithm. http://www.cc.gatech.edu/~dellaert/em-paper.pdf
7. McLachlan G., Peel D. Finite Mixture Models. – New York: John Wiley &Sons, 2000, p. 32-40.
8. Borman S. The expectation maximization algorithm – a short tutorial. http://www.seanborman.com/publications/EM_algorithm.pdf