Expectation–Maximization for a Linear Combination of Gaussians

Georgy Gimel’farb
Computer Science Dept.
Tamaki Campus, University of Auckland
Auckland, New Zealand
g.gimelfarb@auckland.ac.nz

Aly A. Farag, Ayman El-Baz
Electrical and Computer Engineering Dept.
University of Louisville
Louisville, KY 40292, USA
{farag,elbaz} @cvip.uofl.edu

Abstract

We propose a modified Expectation–Maximization algorithm that approximates an empirical probability density function of scalar data with a linear combination of Gaussians (LCG). Due to both positive and negative components, the LCG approximates inter-class transitions more accurately than a conventional mixture of only positive Gaussians. Experiments in segmenting multi-modal medical images show the proposed LCG-approximation results in more adequate region borders.

1. Introduction

Approximation of an empirical relative frequency distribution of scalar data with a particular probability density function is widely used in pattern recognition and image processing, e.g., for data clustering or image segmentation [3, 4, 8]. The basic problem is to accurately approximate, to within the data range, not only the peaks, or modes of the probability density function for the measurements but also its behaviour between the peaks. This is most essential for a precise data classification because borders between data classes are usually formed by intersecting tails of the class distributions. Of course, generally no accurate classification can be achieved by using only a mixed marginal probability distribution by itself. Nonetheless such rough data classification or clustering techniques are of practical interest in many important application problems, e.g., for automated screening and analysis of images obtained by computer tomography, magnetic resonance imaging, or magnetic resonance angiography.

We propose a simple modification of the well-known Expectation–Maximization (EM) algorithm in order to approximate an empirical relative frequency distribution of the scalar data with a linear combination of Gaussians (LCG). The LCG has both positive and negative components so that it approximates empirical data more accurately than a conventional mixture of only positive Gaussians [5, 9, 11]. Although the LCG is not strictly a probability density function over the whole infinite signal range due to possible negative values, the higher approximation accuracy to within a limited signal range is sometimes more important in practice.

The EM-algorithm for estimating parameters of mixed probability distributions was first proposed in the late nineteen sixties both in the general form [12] (see also [13]) and for the normal mixtures [1]. But it became popular only after the pivotal paper [2] a decade later extended this technique onto a general problem of parameter estimation from the incomplete data sets. Today a variety of the EM-algorithms exist to find the maximum likelihood parameter estimates for mixtures of probability distributions [6, 7, 10]. Our modification extends the conventional EM-scheme onto a more general approximation with the LCG.

Section 2 below derives the modified EM algorithm and discusses its pros and cons. Section 3 presents a sequential initializing scheme that produces by itself a close LCG-approximation. Some experimental results and concluding remarks are given in Section 4.

2. Likelihood maximization with an LCG

Let \( F = \{ f(y) : y = 0, \ldots, Y - 1 \} \) be an empirical relative frequency distribution representing an unknown probability density function \( \psi(y) : \int_{-\infty}^{\infty} \psi(y)dy \equiv \sum_{y=0}^{Y-1} f(y) = 1 \). Let the distribution \( f(y) \) be approximated by the LCG with \( K_p \) positive and \( K_n \) negative components \( \phi_p(y) \) where \( \theta = [\mu, \sigma] \) denotes the mean \( \mu \) and standard deviation \( \sigma \):

\[
\rho \theta(y) = \sum_{p=1}^{K_p} \alpha_p \phi_p(y) \theta_p,k - \sum_{l=1}^{K_n} \alpha_n \phi_l(y) \theta_n,l \quad (1)
\]

In line with Eq. (1), the positive weights \( \alpha \) are such that

\[
\sum_{p=1}^{K_p} \alpha_p - \sum_{l=1}^{K_n} \alpha_n = 1 \quad (2)
\]
The probability densities form a proper subset of the set of the LCGs due to the additional restriction \( p_{\mathbf{A}, \Theta}(y) \geq 0 \) which holds for the mixtures without negative components. Below this special feature is ignored because our goal is to closely approximate the empirical data within the limited range \([0, Y - 1]\). The density in Eq. (1) is assumed strictly positive only in the points \( y = 0, 1, \ldots, Y - 1 \). We also assume in this section that the numbers \( K_p \) and \( K_n \) of the components of each type are known after an initialization and do not change during the EM-process. The initialization provides also the starting parameter values \( \mathbf{A}^0 \) and \( \Theta^0 \).

An initializing algorithm outlined below in Section 3 considers the LCG-approximation of a given \( K \)-modal empirical distribution as a refinement of a conventional \( K \)-component normal mixture. The \( K_{p, \text{ref}} \) positive and \( K_{n, \text{ref}} \) negative components refining the mixture approximate the empirical data and the dominant mixture, so that \( K_p = K + K_{p, \text{ref}} \) and \( K_n = K_{n, \text{ref}} \).

The LCG providing a local maximum of the log-likelihood of the empirical data:

\[
L(\mathbf{A}, \Theta) = \sum_{y=0}^{Y-1} f(y) \log p_{\mathbf{A}, \Theta}(y)
\]

(3)

can be found using the iterative block relaxation process extending a conventional EM scheme.

Let \( \tau \) indicate an iteration and let

\[
p_{\mathbf{A}, \Theta}^{[\tau]}(y) = \sum_{k=1}^{K_p} \alpha_p^{[\tau]}(y) \varphi\left( \frac{y - \mu_p}{\sigma_p} \right) + \sum_{l=1}^{K_n} \alpha_n^{[\tau]}(y) \varphi\left( \frac{y - \mu_n}{\sigma_n} \right)
\]

be the LCG at that step. Conditional weights

\[
\pi_p^{[\tau]}(i|y) = \frac{\alpha_p^{[\tau]}(y) \varphi\left( \frac{y - \mu_p}{\sigma_p} \right)}{p_{\mathbf{A}, \Theta}^{[\tau]}(y)}; \quad \pi_n^{[\tau]}(j|y) = \frac{\alpha_n^{[\tau]}(y) \varphi\left( \frac{y - \mu_n}{\sigma_n} \right)}{p_{\mathbf{A}, \Theta}^{[\tau]}(y)}
\]

(4)

\[
\sum_{k=1}^{K_p} \pi_p^{[\tau]}(k|y) - \sum_{l=1}^{K_n} \pi_n^{[\tau]}(l|y) = 1; \quad y = 0, \ldots, Y - 1
\]

(5)

specify, respectively, relative contributions of each data item \( y = 0, \ldots, Y - 1 \) into each positive and negative Gaussian at the step \( \tau \). Using these variables, the log-likelihood function of Eq. (3) can be rewritten in the equivalent form:

\[
L(\mathbf{A}^{[\tau]}, \Theta^{[\tau]}) = \sum_{y=0}^{Y-1} f(y) \left[ \sum_{k=1}^{K_p} \pi_p^{[\tau]}(k|y) \log p_{\mathbf{A}, \Theta}^{[\tau]}(y) \right] - \sum_{y=0}^{Y-1} f(y) \left[ \sum_{l=1}^{K_n} \pi_n^{[\tau]}(l|y) \log p_{\mathbf{A}, \Theta}^{[\tau]}(y) \right]
\]

(6)

where the same term \( \log p_{\mathbf{A}, \Theta}^{[\tau]}(y) \) in the first and second brackets has to be replaced with the equivalent terms:

\[
\log \alpha_p^{[\tau]} + \log \varphi\left( \frac{y - \mu_p}{\sigma_p} \right) - \log \pi_p^{[\tau]}(i|y) \quad \text{and} \quad \log \alpha_n^{[\tau]} + \log \varphi\left( \frac{y - \mu_n}{\sigma_n} \right) - \log \pi_n^{[\tau]}(l|y),
\]

respectively.

The block relaxation scheme for the function in Eq. (6) converges to a local maximum of the likelihood function by iteratively repeating the following two steps:

1. E-step \([\tau + 1]\): find the parameters \( \mathbf{A}^{[\tau+1]} \), \( \Theta^{[\tau+1]} \) by maximizing \( L(\mathbf{A}, \Theta) \) under the fixed conditional weights of Eq. (4) for the step \( \tau \), and

2. M-step \([\tau + 1]\): find these weights by maximizing \( L(\mathbf{A}, \Theta) \) under the fixed parameters \( \mathbf{A}^{[\tau+1]} \), \( \Theta^{[\tau+1]} \) until the changes of all the parameters become small.

The E-step performs the Lagrange maximization of the likelihood function of Eq. (6) under the condition of Eq. (2) yielding the following weights estimates:

\[
\alpha_p^{[\tau+1]} = \frac{1}{\sum_{y=0}^{Y-1} \pi_p^{[\tau]}(i|y)} \sum_{y=0}^{Y-1} f(y) \pi_p^{[\tau]}(i|y)
\]

\[
\pi_p^{[\tau+1]} = \frac{1}{\sum_{y=0}^{Y-1} \pi_p^{[\tau]}(i|y)} (\pi_p^{[\tau]}(i|y) - \alpha_p^{[\tau+1]})
\]

and

\[
\alpha_n^{[\tau+1]} = \frac{1}{\sum_{y=0}^{Y-1} \pi_n^{[\tau]}(l|y)} \sum_{y=0}^{Y-1} f(y) \pi_n^{[\tau]}(i|y)
\]

\[
\pi_n^{[\tau+1]} = \frac{1}{\sum_{y=0}^{Y-1} \pi_n^{[\tau]}(l|y)} (\pi_n^{[\tau]}(l|y) - \alpha_n^{[\tau+1]})
\]

where \( \cdot \) stands for \( \cdot \)p or \( \cdot \)n, respectively.

The M-step performs the Lagrange maximization of the log-likelihood of Eq. (6) under the \( Y \) conditions of Eq. (5). It results in the conditional weights \( \pi_p^{[\tau+1]}(i|y) \) and \( \pi_n^{[\tau+1]}(l|y) \) of Eq. (4) for all \( i = 1, \ldots, K; j = 1, \ldots, L; \) and \( y = 0, \ldots, Y - 1 \). The modified EM-algorithm is valid until these weights are strictly positive, and the initial LCG-approximation should comply to this limitation. The iterations have to be terminated when the log-likelihood of Eq. (6) begins to decrease. Generally, if the initial initialization is incorrect, this algorithm may diverge from the very beginning. Thus the initial LCG has to closely approximate the empirical distribution.

3. Sequential EM-based initialization

The search for a number of Gaussians in a mixture is based on an integral absolute deviation between the empirical and model densities. The number is sequentially increasing while the deviation decrement is above a given threshold. Using such a search, a close initial estimate of the LCG is obtained by the following algorithm:

1. Find the number \( K \) of the dominant modes of the empirical distribution \( \mathbf{F} \) by sequentially approximating it with the mixtures \( \mathbf{P}_k \) of \( k = 1, \ldots, K + 1 \) Gaussians using the conventional EM-algorithm.
2. Split the absolute deviation between \( F \) and \( P_K \) into the additive and subtractive parts.

3. Find separately the numbers \( k_{\text{add}} \) and \( k_{\text{sub}} \) of Gaussians for each part by approximating it with the scaled-down normal mixtures using the conventional EM-algorithm.

The initial LCG consists of the \( K_P = K + k_{\text{add}} \) positive and \( K_n = k_{\text{sub}} \) negative Gaussians.

Figure 1 shows a typical CT slice (a) and its empirical marginal grey level distribution \( E \) approximated with the dominant normal mixture \( M \) (b). The two dominant modes represent the darker lung area and its brighter background, respectively. After the additive and subtractive parts of the absolute deviation (c) are approximated with the normal mixtures, the initial LCG-model consists of 2 dominant, 6 additive, and 4 subtractive Gaussians, that is, \( K_P = 8 \) and \( K_n = 4 \). The LCG-approximation of each class (d) corresponds to the threshold, \( t = 108 \), ensuring the best class separation.

Figure 2 presents the final LCG (a) obtained by refining the above initial one using the modified EM-algorithm, the successive changes of the log-likelihood (b) at the iterations, the 12 components of the final LCG (c), the final LCG-approximation of each class for the best separation threshold \( t = 109 \) (d), the segmentation map (e) for the CT slice in Fig. 1,a obtained for these two classes (d), and the like map (f) produced by a radiologist. The segmentation has an error of 2.91% with respect to the expert’s map. First five iterations of the algorithm increase the log-likelihood of Eq. (6) from \(-5.00\) to \(-4.21\), the refinement process having terminated then since the log-likelihood begins to decrease.

Our experiments with different CT slices have shown that the proposed LCG-approximation combined with post-processing based on a Gibbs random field (GRF) region model results in small errors of 0.9–1.1% with respect to the expert’s segmentation. The conventional normal mixtures of the same size and under the same post-processing yield about 9.5% error because of the components approximating the interval between the classes. In contrast to the LCG, these components almost always combine tails of both the class distributions so that the accurate thresholding becomes hardly possible.

Figure 3 shows an MRA (magnetic resonance angiography) image (a) and its tri-modal empirical grey level distribution (b) approximated with the dominant 3-component normal mixture (M). The classes represent dark bones and fat, brain tissues, and bright blood vessels, respectively. The goal is to separate this latter class in spite of its large intersection with the second one and very low prior probability.
The number $K = 3$ of the dominant components is obtained by thresholding the decrement of the integral absolute deviation between the empirical and mixture densities: 0.5477 ($k = 1$), 0.1739 ($k = 2$), 0.1596 ($k = 3$), and 0.1551 ($k = 4$). The parameters $(\alpha_k, \mu, \sigma)$ of the components are $(0.518, 24.7, 12.6)$, $(0.456, 105.7, 31.8)$, and $(0.026, 210.7, 25.0)$, respectively. The initialization returns nine components of the LCG, and seven first iterations of the refinement before the process terminates increase the log-likelihood from $-5.70$ for the initial LCG (c) to $-4.90$ for the final one (d,f) with the components (e). The segmentation map (g) of the image (a) obtained for the final LCG-models (d) of the classes has the error only about 1.01% with respect to the expert’s map (h).

These and other experiments with multi-modal medical images show that the modified EM algorithm produces accurate LCG-models of empirical relative frequency distributions of scalar signals, provided that the proposed sequential initial approximation results in proper numbers of the additive and subtractive components. The computations are as simple as in the conventional EM-techniques. In principle, this approach can be extended onto the LCG-based cluster analysis of multivariate empirical data containing both the valid class samples and outliers.

Acknowledgements

This work was partially supported by the Royal Society of New Zealand under the Marsden Fund Grant UOA122.

References