DENSITY ESTIMATION USING MODIFIED EXPECTATION–MAXIMIZATION ALGORITHM FOR A LINEAR COMBINATION OF GAUSSIANS

Aly A. Farag, Ayman El-Baz

CVIP, University of Louisville
Louisville, KY 40292
{farag,elbaz}@cvip.uofl.edu

Georgy Gimel'farb

CITR, University of Auckland
Auckland, New Zealand
g.gimelfarb@aukland.ac.nz

ABSTRACT

In this paper we present a new approach for density estimation. The proposed approach is based on modifying Expectation-Maximization (EM) algorithm to approximate an empirical probability density function of scalar data with a linear combination of Gaussians (LCG). We also propose a novel EM-based sequential technique to get a close initial LCG approximation the modified EM algorithm should start with. Due to both positive and negative components, the LCG approximates inter-class transitions more accurately than a conventional mixture of only positive Gaussians. Experiments on simulated images demonstrate the accuracy of our approach.

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, 5, 7]. 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 behavior 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 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 [4, 8, 10].

The EM-algorithm for estimating parameters of mixed probability distributions was first proposed in the late nineteen sixties both in the general form [11] (see also [12]) and for the normal mixtures [1]. But it became popular only after the pivotal paper [2] a decade later extended this technique into 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, 9]. 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 $\mathbf{F} = [f(q) : q = 0, \ldots, Q]$ be an empirical relative frequency distribution representing an unknown probability density function $\psi(q)$: $\int_{-\infty}^{\infty} \psi(q)dq \equiv \sum_{q=0}^{Q} f(q) = 1$. Let the distribution $f(q)$ be approximated by the LCG with K_{p} positive and K_{n} negative components $\varphi(q|\theta)$ where $\theta = [\mu, \sigma]$ denotes the mean μ and standard deviation σ:

$$p_{\mathbf{A},\boldsymbol{\Theta}}(q) = \sum_{k=1}^{K_{\mathrm{p}}} \alpha_{\mathrm{p},k}\varphi(q|\theta_{\mathrm{p},k}) - \sum_{l=1}^{K_{\mathrm{n}}} \alpha_{\mathrm{n},l}\varphi(q|\theta_{\mathrm{n},l}) \quad (1)$$

In line with Eq. (1), the weights $\mathbf{A}$ are such that

$$\sum_{k=1}^{K_{\mathrm{p}}} \alpha_{\mathrm{p},k} - \sum_{l=1}^{K_{\mathrm{n}}} \alpha_{\mathrm{n},l} = 1 \quad (2)$$

The probability densities from a proper subset of the set of the LCGs due to the additional restriction $p_{\mathbf{A},\boldsymbol{\Theta}}(q) \geq 0$ 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 $[0, Q]$. The density in Eq. (1) is assumed strictly positive only in the points $q = 0, 1, \ldots, Q$. We also assume in

0-7803-8554-3/04/$20.00 ©2004 IEEE.

this section that the numbers $K_{\rm p}$ and $K_{\rm n}$ 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 $\mathbf{\Theta}^{[0]}$.

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 $K_{\rm p,ref}$ positive and $K_{\rm n,ref}$ negative components refining the mixture approximate the difference between the empirical data and the dominant mixture, so that $K_{\rm p} = K + K_{\rm p,ref}$ and $K_{\rm n} = K_{\rm n,ref}$.

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

$$L(\mathbf{A}, \mathbf{\Theta}) = \sum_{q=0}^{Q} f(q) \log p_{\mathbf{A},\mathbf{\Theta}}(q) \tag{3}$$

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

Let τ indicate an iteration and let

$$p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q) = \sum_{k=1}^{K_{\rm p}} \alpha^{[\tau]}_{{\rm p},k} \varphi(q|\theta^{[\tau]}_{{\rm p},k}) - \sum_{l=1}^{K_{\rm n}} \alpha^{[\tau]}_{{\rm n},l} \varphi(q|\theta^{[\tau]}_{{\rm n},l})$$

be the LCG at that step. Conditional weights

$$\pi^{[\tau]}_{\rm p}(k|q) = \frac{\alpha^{[\tau]}_{{\rm p},k} \varphi(q|\theta^{[\tau]}_{{\rm p},k})}{p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q)};\ \pi_{\rm n}(l|q) = \frac{\alpha^{[\tau]}_{{\rm n},l} \varphi(q|\theta^{[\tau]}_{{\rm n},l})}{p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q)} \tag{4}$$

$$\sum_{k=1}^{K_{\rm p}} \pi^{[\tau]}_{\rm p}(k|q) - \sum_{l=1}^{K_{\rm n}} \pi^{[\tau]}_{\rm n}(l|q) = 1;\ q = 0, \ldots, Q \tag{5}$$

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

$$\begin{aligned} &L(\mathbf{A}^{[\tau]}, \mathbf{\Theta}^{[\tau]}) = \\ &\sum_{q=0}^{Q} f(q) \left[\sum_{k=1}^{K_{\rm p}} \pi^{[\tau]}_{\rm p}(k|q) \log p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q) \right] - \\ &\sum_{q=0}^{Q} f(q) \left[\sum_{l=1}^{K_{\rm n}} \pi^{[\tau]}_{\rm n}(l|q) \log p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q) \right] \end{aligned} \tag{6}$$

where the same term $\log p^{[\tau]}_{\mathbf{A},\mathbf{\Theta}}(q)$ in the first and second brackets has to be replaced with the equivalent terms: $\log \alpha^{[\tau]}_{{\rm p},k} + \log \varphi(q|\theta^{[\tau]}_{{\rm p},k}) - \log \pi^{[\tau]}_{\rm p}(k|q)$ and $\log \alpha^{[\tau]}_{{\rm n},l} + \log \varphi(q|\theta^{[\tau]}_{{\rm n},l}) - \log \pi^{[\tau]}_{\rm n}(l|q)$, 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 $[\tau+1]$: to find the parameters $\mathbf{A}^{[\tau+1]}$, $\mathbf{\Theta}^{[\tau+1]}$ by maximizing $L(\mathbf{A}, \mathbf{\Theta})$ under the fixed conditional weights of Eq. (4) for the step τ, and
2. M-step $[\tau+1]$: to find these weights by maximizing $L(\mathbf{A}, \mathbf{\Theta})$ under the fixed parameters $\mathbf{A}^{[\tau+1]}$, $\mathbf{\Theta}^{[\tau+1]}$

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:

$$\alpha^{[\tau+1]}_{{\rm p},k} = \sum_{q=0}^{Q} f(q) \pi^{[\tau]}_{\rm p}(k|q);\ \alpha^{[\tau+1]}_{{\rm n},l} = \sum_{q=0}^{Q} f(q) \pi^{[\tau]}_{\rm n}(l|q)$$

The parameters of each Gaussian are obtained by the unconditional maximization as in the conventional scheme:

$$\begin{aligned} &\mu^{[\tau+1]}_{{\rm c},k} = \tfrac{1}{\alpha^{[\tau+1]}_{{\rm c},k}} \sum_{q=0}^{Q} q \cdot f(q) \pi^{[\tau]}_{\rm c}(k|q) \\ &(\sigma^{[\tau+1]}_{k,{\rm c}})^2 = \tfrac{1}{\alpha^{[\tau+1]}_{{\rm c},k}} \sum_{q=0}^{Q} \left(q - \mu^{[\tau+1]}_{{\rm c},k} \right)^2 \cdot f(q) \pi^{[\tau]}_{\rm c}(k|q) \end{aligned}$$

where "c" stands for "p" or "n", respectively.

The M-step performs the Lagrange maximization of the log-likelihood of Eq. (6) under the Q conditions of Eq. (5). It results in the conditional weights $\pi^{[\tau+1]}_{\rm p}(k|q)$ and $\pi^{[\tau+1]}_{\rm n}(l|q)$ of Eq. (4) for all $k = 1, \ldots, K_p$; $l = 1, \ldots, K_n$; and $q = 0, \ldots, Q$. 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 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 $k = 1, \ldots, K$ Gaussians using the conventional EM-algorithm.
2. Split the absolute deviation between $\mathbf{F}$ and $\mathbf{P}_K$ into the additive and subtractive parts.
3. Find separately the numbers $k_{\rm add}$ and $k_{\rm sub}$ 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 $K_{\rm p} = K + k_{\rm add}$ positive and $K_{\rm n} = k_{\rm sub}$ negative Gaussians.

4. EXPERIMENTS AND CONCLUSIONS

To assess robustness and computational performance, the proposed segmentation techniques have been tested on a synthetic bi-modal image. A synthetic bimodal checkerboard image was used to compare the estimated and the original known parameters of the model including the number of positive and negative Gaussians, their mean values, variances, and the mixing proportions. Each "ideal" class consists of one dominant and four subordinate components shown in Fig. 1(a). By an inverse transformation of each class in the ideal region map in Fig. 1(b), gray levels in the range $[0, 255]$ are generated randomly according to their class distributions in Fig 1(c). The resulting bimodal gray level image is shown in Fig. 1(d).

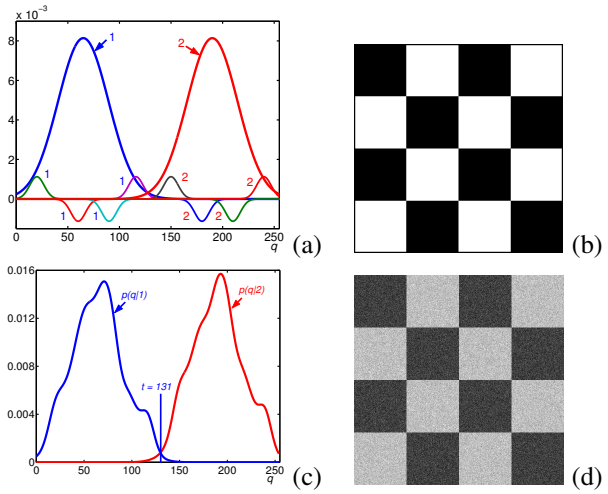

Fig. 1. Components of the LCG-models of the two classes (a) specified by the checkerboard map (b), the whole LCG-models of each class (c), and the checkerboard image having the chosen gray level distributions (d).

Table 1. Initial and final estimated parameters for the LCG model

Parameters	α_1	μ_1	σ_1^2	α_2	μ_2	σ_2^2
Original	0.50	65.0	600.0	0.50	190.0	600.0
Initial values	0.59	66.1	725.9	0.41	198.7	789.1
Final values	0.51	65.3	581.4	0.49	189.7	581.4
Error, %	2.0	0.42	3.1	2.0	0.16	3.1

Results of the sequential EM-based initialization for the simulated image in Fig. 1(d), are shown in Fig. 2. Figure 2(a), presents the two estimated dominant components for the two classes. Parameters of this initial model are given in Table 1. The Levy distance [13] of 0.1 between these two distributions indicates a large mismatch between dominant components and the empirical density. Figure 2(b),

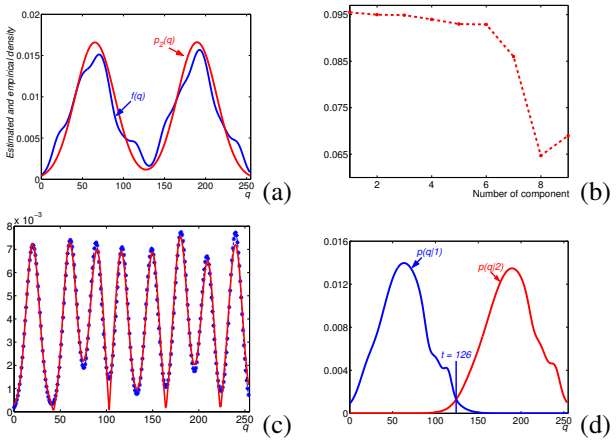

Fig. 2. Initial LCG-approximation of the empirical grey level distribution.

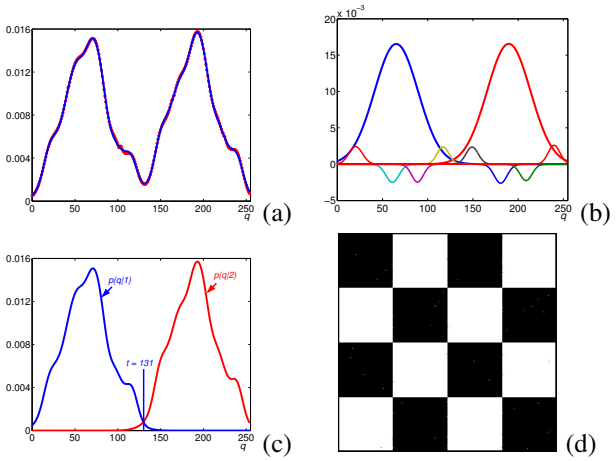

Fig. 3. Final 2-class LCG-approximation

illustrates the estimation of the number of the subordinate Gaussians, the minimum error having been obtained for the eight components. Figure 2(c), presents the initial LCG-model built with the sequential EM-based algorithm. The final mixed LCG-model P has to be split into K LCG-submodels, one per class, by associating each subordinate component with a particular dominant term in such a way as to minimize the expected misclassification rate. For the image shown in Fig. 1(d) the initial LCG-model corresponds to the minimum classification error of 0.0028 between the first and the second class for the threshold $t = 126$. The estimate for each class is shown in Fig. 2(d).

Figure 3 presents the final LCG (a) obtained by refining the above initial one using the modified EM-algorithm, the 10 components of the final LCG (b), the final LCG-approximation of each class for the best separation thresh-

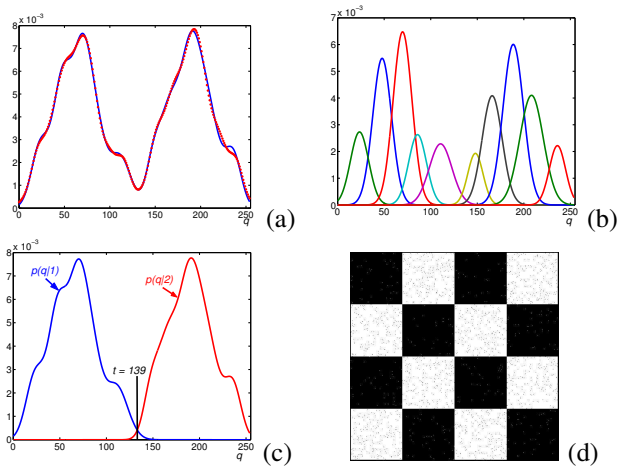

Fig. 4. Density models of the lung and chest tissues in Fig. 1(d) estimated with the mixture of 10 positive Gaussians (the minimum misclassification rate 0.013 for the threshold $t = 139$).

old $t = 131$ (c), the segmentation (d) for the checkerboard image in Fig. 1(d) obtained for these two classes. The segmentation has an error of 0.007% with respect to the the ground truth shown in Fig. 1(b). First five iterations of the algorithm increase the log-likelihood of Eq. (6) from -5.00 to -4.21, the refinement process is terminated since the log-likelihood begins to decrease. After we use the modified EM algorithm the resulting Levy distance [13] between the empirical distribution and estimated distribution becomes 0.0012 is notably smaller than before (0.1) indicating the close approximation. The final parameters of the two dominant components are given in Table 1

The approximation of the same empirical density with a conventional mixture of 10 positive Gaussians highlights advantages of using the LCG. The resulted mixture model in Fig. 4(a)–(d), has more than three times higher misclassification rate (0.013) because one of its components combines former separate tails of both the classes and actually cannot be related to only one mode. This is why our LCG-segmentation produces much more accurate segmentation. The Levy distance [13] between the estimated density using ten positive component and empirical density is 0.021 which is notably bigger than the Levy distance [13] obtained by modified EM algorithm.

These simulated images and other experiments with multimodal medical images presented in [14] 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 number 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.

5. REFERENCES

[1] N.E.Day, "Estimating the components of mixture of normal distributions" Biometrika, Vol. 56, pp. 463-474, 1969.

[2] A.P.Dempster, N. M.Laird, and D. B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Vol. 39B, no. 1, pp. 1-38, 1977.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. 2nd ed., Wiley: New York, 2001.

[4] A.Goshtasby and W.D.O'Neill, "Curve fitting by a sum of Gaussians," CVGIP: Graphical Models and Image Processing, Vol.56, no.4, pp.281-288, 1999.

[5] A.K.Jain and R.C.Dubes, "Random field models in image analysis," J. of Applied Statistics, Vol. 16, no. 2, 1989.

[6] T.Moon, "The Expectation - Maximization algorithm," IEEE Signal Processing Magazine, pp. 47-60, Nov. 1996.

[7] N.R.Pal and S.K.Pal, "A review on image segmentation techiniques," Pattern Recognition, Vol. 26, no. 9, pp.1277-1294, 1993.

[8] T.Poggio and F.Girosi, "Networks for approximation and learning," Proc. IEEE, Vol.78, no. 9, pp. 1481-1497, 1990.

[9] R.Redner and H.Walker, "Mixture densities, maximum likelihood and the EM algorithm (review)," SIAM Review, Vol. 26, no. 2, pp. 195-237, 1984.

[10] H.W.Sorenson and D.L.Alspach, "Recursive Bayesian estimation using Gaussian sums," Automatica, Vol.7, pp. 465-479, 1971.

[11] M.I.Schlesinger, "A connection between supervised and unsupervised learning in pattern recognition" Kibernetika, no. 2, pp. 81-88, 1968 [In Russian].

[12] M.I.Schlesinger and V.Hlavac, "Ten Lectures on Statistical and Structural Pattern Recognition" Kluwer Academic: Dordrecht, 2002.

[13] J. W. Lamperti, *Probability*, Wiley, New York, 1996.

[14] G. Gimel'farb, A. A. Farag, and A. El-Baz, "Expectation-Maximization for a linear combination of Gaussians", in *Proc. IAPR Int. Conf. Pattern Recognition (ICPR 2004)*, Cambridge, UK, August 2004 [in print].