Pattern Recognition 77 (2018) 188–203

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures

Jian Yu[a], Chaomu Chaomurilige[a], Miin-Shen Yang[b,*]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[b] Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan

ARTICLE INFO

Article history:
Received 13 November 2016
Revised 6 December 2017
Accepted 17 December 2017
Available online 19 December 2017

Keywords:
Expectation & maximization (EM) algorithm
Deterministic annealing EM (DA-EM)
GAUSSIAN mixtures
Self-annealing
Convergence
Parameter selection

ABSTRACT

The expectation & maximization (EM) for Gaussian mixtures is popular as a clustering algorithm. However, the EM algorithm is sensitive to initial values, and so Ueda and Nakano [4] proposed the deterministic annealing EM (DA-EM) algorithm to improve it. In this paper, we investigate theoretical behaviors of the EM and DA-EM algorithms. We first derive a general Jacobian matrix of the DA-EM algorithm with respect to posterior probabilities. We then propose a theoretical lower bound for initialization of the annealing parameter in the DA-EM algorithm. On the other hand, some researches mentioned that the EM algorithm exhibits a self-annealing behavior, that is, the equal posterior probability with small random perturbations can avoid the EM algorithm to output the mass center for Gaussian mixtures. However, there is no theoretical analysis on this self-annealing property. Since the DA-EM will become the EM when the annealing parameter is 1, according to the Jacobian matrix of the DA-EM, we can prove the self-annealing property of the EM algorithm for Gaussian mixtures. Based on these results, we give not only convergence behaviors of the equal posterior probabilities and initialization lower bound of the temperature parameter of the DA-EM, but also a theoretical explanation why the EM algorithm for Gaussian mixtures exhibits a self-annealing behavior.

1. Introduction

Since Dempster et al. [1] proposed the EM algorithm to deal with incomplete data, the EM and its extensions have been widely studied and applied in various areas (see [2,3]). In the literature, there are many researches in developing the EM algorithm and its variants for Gaussian mixtures, such as Ueda and Nakano [4], Figueiredo and Jain [5], Zhang et al. [6], Yang et al. [7], Gkalelis et al. [8], Tao et al. [9], Gao et al. [10]. It is always interesting to study the convergence properties of the EM algorithm. Boyles [11] first claimed that the generalized EM sequence will converge to a compact connected set of local maxima of the likelihood function. At the same time, Wu [12] proved that the EM algorithm converges to the stationary points of the log likelihood function. In other words, the EM may converge to a local maximum or a saddle point of the log likelihood function. Furthermore, Meng and Rubin [13] considered a supplemented EM (SEM) algorithm and then used the SEM as a tool for monitoring whether the EM has converged to a (local) maximum. To analyze the convergence rate of the EM algorithm for Gaussian mixtures, Xu and Jordan [14] and Ma et al. [15] presented the Hessian matrix of the log-likelihood function for Gaussian mixtures with respect to the collection of mixture parameters, Ma and Fu [16] proved that, if the initial parameters are set within the neighborhood, the EM algorithm will always converge to the consistent solution, and Roche et al. [17] considered the convergence property of the three EM-like algorithms for Markov random field segmentation.

In general, the EM for Gaussian mixtures is popular as a clustering algorithm. Since the performance of the EM algorithm for Gaussian mixtures heavily depends on initializations, Ueda and Nakano [5] proposed the deterministic annealing EM (DA-EM) algorithm to improve it. The basic idea of the DA-EM is to begin at a high temperature β, and then decreases the temperature to zero according to some cooling strategy to avoid poor local optima. The DA-EM algorithm had been studied and applied in various areas, such as Shoham [18], Itaya et al. [19], Guo and Cui [20] and Okamura et al. [21]. On the other hand, Figueiredo and Jain [6] pointed out that, the heuristic behind the deterministic annealing is to force the entropy of the assignments to decrease slowly for avoiding poor local optima. They also mentioned that the EM algorithm itself has a self-annealing behavior without a cooling strategy and just set an uninformative initialization of the posterior probabilities to be $1/c$ with small random perturbations.

[*] Corresponding author.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

https://doi.org/10.1016/j.patcog.2017.12.014
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

Such an observation suggests that, the equal posterior probability with small random perturbations is not an asymptotically stable fixed point of the EM algorithm for Gaussian mixtures, and this can actually interpret why the EM algorithm for Gaussian mixtures exhibits a self-annealing property. However, as we know, no theoretical analysis and proving were proposed in the literature.

In this paper, we investigate convergence behaviors of DA-EM, especially for the annealing parameter in the DA-EM algorithm, and give theoretical results for self-annealing behavior of the EM algorithm. We first construct a new Jacobian matrix of the DA-EM for Gaussian mixtures with respect to posterior probabilities. We then derive a theoretical rule for the valid temperature parameter initialization bound of the DA-EM algorithm to avoid the equal posterior probabilities to be an asymptotically stable fixed point of the DA-EM by using the Jacobian matrix analysis. That is, we give the annealing parameter selection for the DA-EM algorithm. In general, the self-annealing behavior of EM is highly related to the annealing parameter β of DA-EM where, if the annealing parameter β is equal to 1, then DA-EM becomes EM. Therefore, we can also prove that the EM algorithm always satisfies the necessary condition for the equal posterior probability not being an asymptotically stable fixed point of EM for Gaussian mixtures. That is, we prove that the EM algorithm for Gaussian mixtures can exhibit a self-annealing property. The remainder of the paper is organized as follows. In Section 2, we review the EM and DA-EM algorithms for Gaussian mixtures with problem descriptions. In Section 3, we first construct the Jacobian matrix of the DA-EM with respect to posterior probabilities. We then give some convergence theorems and also the theoretical rule for the valid initialization bound of the annealing parameter β. And so, we give a theoretical result for the self-annealing behavior of EM. In Section 4, we carry out several experiments to demonstrate our theoretical results. We also use an example to show how to apply the Jacobian matrix of the DA-EM for analyzing its convergence rate. Finally, we give the conclusions in Section 5.

2. The EM and DA-EM algorithms for Gaussian mixtures

In this section, we first give a brief description of the EM algorithm for Gaussian mixtures. A Gaussian mixture can be described as follows:

$$f(x|\Theta) = \sum_{i=1}^{c} \alpha_i f(x|\Theta_i) \tag{1}$$

where $\alpha_i > 0$, $\sum_{i=1}^{c} \alpha_i = 1$, $\Theta_i = (\mu_i, \sum_i)$, $x \in R^s$ is a column vector, and $f(x|\Theta_i)$ is defined as an s-variate Gaussian distribution with $f(x|\Theta_i) = (2\pi)^{-s/2} |\sum_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu_i)^T (\sum_i)^{-1} (x-\mu_i)\}$. Therefore, a Gaussian mixture $f(x|\Theta)$ can be parameterized by the set $\Theta = \{\Theta_1, ..., \Theta_c, \alpha_1, ..., \alpha_c\}$, where c represents the number of components in the Gaussian mixture $f(x|\Theta)$. Let the data set $X = \{x_1, ..., x_n\}$ be a random sample of size n drawn from the distribution $f(x|\Theta)$. Then its log-likelihood function can be written as follows:

$$\log f(X|\Theta) = \log \prod_{k=1}^{n} f(x_k|\Theta) = \sum_{k=1}^{n} \log \sum_{i=1}^{c} \alpha_i f(x_k|\Theta_i) \tag{2}$$

Obviously, the parameter $\Theta = \{\Theta_1, ..., \Theta_c, \alpha_1, ..., \alpha_c\}$ can be estimated as

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \{\log f(X|\Theta)\} \tag{3}$$

For a finite mixture, the EM algorithm considers $X = \{x_1, ..., x_n\}$ as observations. A label set $L = \{l_1, ..., l_n\}$ is considered as a missing part corresponding to the given data $X = \{x_1, ..., x_n\}$ in which $l_k \in \{1, 2, \ldots, c\}$. If $l_k = i$, it means that the $k$thdata point belongs to the ith class. That is, we have $g_k = [l_{1k}, \cdots, l_{ck}]^T$, where $l_{ik} = 1, if l_k = i$; 0, otherwise. Therefore, the complete log-likelihood can be represented as follows:

$$\begin{aligned} \log f(X, L|\Theta) &= \sum_{k=1}^{n} \log\left(\prod_{i=1}^{c} [\alpha_i f(x_k|\Theta_i)]^{l_{ik}} \right) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{c} l_{ik} \log[\alpha_i f(x_k|\Theta_i)] \end{aligned} \tag{4}$$

Assumed that $\Theta = \{\Theta_1, ..., \Theta_c, \alpha_1, ..., \alpha_c\}$ is given, the conditional expectation z_{ik} of l_{ik} can be given by the following Eq. (5).

$$z_{ik} = E[l_{ik}|X, \Theta] = Pr[l_{ik} = 1|x_k, \Theta] = \frac{\alpha_i f(x_k|\Theta_i)}{\sum_{i=1}^{c} \alpha_i f(x_k|\Theta_i)} \tag{5}$$

For the Gaussian mixture models, after z_{ik} is obtained, the maximum likelihood (ML) estimates of $\log f(X|\Theta)$for the parameter $\Theta = \{\Theta_1, ..., \Theta_c, \alpha_1, ..., \alpha_c\}$ are as follows:

$$\alpha_i = \frac{1}{n} \sum_{k=1}^{n} z_{ik} \tag{6}$$

$$\mu_i = \frac{\sum_{k=1}^{n} z_{ik} x_k}{\sum_{k=1}^{n} z_{ik}} \tag{7}$$

$$\Sigma_i = \frac{\sum_{k=1}^{n} z_{ik} (x_k - \mu_i)(x_k - \mu_i)^T}{\sum_{k=1}^{n} z_{ik}} \tag{8}$$

According to the theorem of Wu [12] for the EM sequence, the EM algorithm for Gaussian mixtures converges to the stationary points of $\log f(X|\Theta)$. When Θ is a stationary point of $\log f(X|\Theta)$, Ma et al. [15] gave the Hessian matrix of $\log f(X|\Theta)$ with respect to Θ and then gave the result about the asymptotic convergence rate of the EM algorithm for Gaussian mixtures.

Since the EM algorithm for Gaussian mixtures is sensitive to initials, Ueda and Nakano [4] proposed the deterministic annealing EM (DA-EM) algorithm to improve it. The DA-EM algorithm introduces a parameter β with its reciprocal corresponding to the "temperature". The only difference between the DA-EM and EM algorithms is that the DA-EM adds the annealing parameter β in the posterior probability z_{ik} as:

$$z_{ik} = \frac{(\alpha_i f(x_k|\Theta_i))^{\beta}}{\sum_{i=1}^{c} (\alpha_i f(x_k|\Theta_i))^{\beta}} \tag{9}$$

It is obvious that the DA-EM will become the EM when $\beta = 1$. The DA-EM algorithm starts $\beta^{(0)}$ at a small enough value (i.e. high temperature) and slowly increases β up to 1. Thus, the DA-EM algorithm (Ueda and Nakano [4]) can be rewritten as follows:

DA-EM algorithm (Ueda and Nakano [4])

1. Initialize
   - Set $\beta \leftarrow \beta^{(0)} (0 < \beta^{(0)} < < 1)$.
   - Set $\Theta^{(0)}$ using k-means algorithm for better results.
2. Iterate until convergence
   - E-step: estimate posterior probabilities by Eq. (9).
   - M-step: estimate $\Theta^{(new)}$ by Eqs. (6), (7) and (8).
3. Increase β.
4. If $\beta < 1$, go back to step 2;

Else stop the procedure.

Note that, in this paper we increase the value of parameter β with 1.01 times, i.e. $\beta^{(new)} \leftarrow \beta^{(old)} \times 1.01$. In fact, how much increasing in β can be determined by users. We will discuss the influence of the increasing factor in numerical examples and experiments of Section 4. In the DA-EM algorithm, the problem of maximizing the log-likelihood function is reformulated as the problem of minimizing a free energy function. The algorithm begins at high temperature corresponding to high entropy that the initial