SCIENCE DIRECT®

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

correct convergence of the EM algorithm for Gaussian mixtures☆

nwen Ma[a,*], Shuqun Fu[b]

and LMAM, Peking University, Beijing 100871, China
hangning District, Shanghai 200050, China

form 7 March 2005; accepted 7 March 2

ct

owever, there have
the true p ong as the overlap of
convergence pr symptotically on the EM
becomes a contraction mapping of the paramet
elihood when the measu f average overlap ng
samples is large enough f the initial p meters
converge to the consisten expected result.
orrect convergence neigh s the average

© 20 attern Recognition Society. Published by Elsevier Ltd. All rights reserv

Keywo EM algorithm; Gaussian mixture; Maximum li imate; Overlap ct co

1. Introduction

The expectation-maximization (EM) algorithm is a general methodology for maximum likelihood (ML) or maximum a posteriori (MAP) estimation [1]. Its convergence has been studied by many researchers (e.g., [2–7]). Since the EM algorithm is generally considered as a first-order or linearly convergent algorithm, several acceleration methods for the EM algorithm have been further proposed, e.g., Aitken acceleration [8], conjugate gradient acceleration [9], qua onian acc [10], parameter expansion acceleration [11] and "working parameter" approach [6].

However, recent studies have found that the EM algorithm can be asymptotically superlinear as the overlap measure of components in the original mixture tends to zero. Actually, Xu and Jordan [12] constructed a relation between the EM algorithm for Gaussian mixtures and the gradient algorithm of the maximum likelihood showing that the EM algorithm owns a quasi-Newton behavior as it nears an ML or MAP solution while the mixture components are well separated. Based on this relation and one of its intermediate results on the convergence rate in [12], Ma et al. [13] further proved that the asymptotic convergence rate of the EM algorithm locally around the true solution is a higher-order infinitesimal than a positive order power of an average overlap measure of component densities in the mixture as this measure tends to zero. That is, the large sample local convergence

☆ work was supported b e Natural Science Foun
na or Projects 600710 and 60471054.
* Corresponding author el.: +86 10 62758101
fax: +86 1 62751801.
E-mail address: lwma@math.pku.edu.cn

0031-3203/$30.00 © 2005
doi:10.1016/j.patcog.2

rate of the EM algorithm tends to be asymptotically superlinear when the overlap of densities in the mixture tends to zero. Recently, Ma and Xu [14] have generalized this result to the mixture of densities from exponential families.

But there has still been an important but unsolved problem of whether the EM algorithm can converge to the correct solution, i.e., the consistent solution of the true parameters of the mixture from which the sample data come. Clearly, this correct convergence problem is key to the usefulness of the EM algorithm. According to the classical convergence theory, the EM algorithm only converges to a local maximum solution of the likelihood function and cannot be guaranteed to converge to the correct solution. However, in practical applications and experiments, we often find that the EM algorithm always converges correctly when the overlap of Gaussians in the sample data or original mixture becomes small enough. This fact reveals that the correct convergence of the EM algorithm is related to the overlap of the Gaussians in the mixture of the sample data and thus we can study the correct convergence of the EM algorithm for Gaussian mixtures from the change of the overlap of Gaussians in the original mixture.

In this paper, we study the correct convergence problem of the EM algorithm for Gaussian mixtures under the theoretical framework of [13]. It is proved that the EM algorithm becomes a contraction mapping of the parameters in a neighborhood of the consistent solution of the maximum likelihood when the measure of average overlap among Gaussians in the original mixtures is small enough and the sample size is large enough. That is, when the initial parameters are given within the neighborhood, the EM algorithm will always converge to this consistent solution, i.e., the expected result. Moreover, the simulation results further demonstrate that the correct convergence neighborhood of the EM algorithm increases as the measure of average overlap among Gaussians in the original mixtures decreases to zero.

In the sequel, we introduce the Gaussian mixture model and give some definitions and a lemma in Section 2. In Section 3, we present the main results. Moreover, we substantiate them by the simulation experiments in Section 4. Finally, we conclude in Section 5.

2. Gaussian mixture, definitions and lemma

We consider the following Gaussian mixture model:

$$P(x|\Theta) = \sum_{j=1}^{K} \alpha_j$$

the correct convergence problem of the EM algorithm from the point of view of the contraction mapping.

In order to analyze the correct convergence problem of the EM algorithm, we give some definitions and a lemma related to the Gaussian mixture model which were firstly introduced in [13].

We begin to introduce the measure for the overlap of Gaussians in the mixture. We consider the following posterior densities for the Gaussian mixture Eq. (1) with the true parameters Θ^* of the sample data set:

$$h_i(x) = \frac{\alpha_i^* P(x|m_i^*, \Sigma_i^*)}{\sum_{j=1}^K \alpha_j^* P(x|m_j^*, \Sigma_j^*)} \quad \text{for } i = 1, \ldots, K. \tag{8}$$

We let

$$\gamma_{ij}(x) = (\delta_{ij} - h_i(x))h_j(x) \quad \text{for } i, j = 1, \ldots, K, \tag{9}$$

where δ_{ij} is the Kronecker function. Then, we define a group of quantities on the overlap of Gaussians as follows:

$$e_{ij}(\Theta^*) = \int_{R^d} |\gamma_{ij}(x)| P(x|\Theta^*)\, \mathrm{d}x,$$

for $i, j = 1, 2, \ldots, K$, where $e_{ij}(\Theta^*) \leqslant 1$ since $|\gamma_{ij}(x)| \leqslant 1$.

For $i \neq j$, $e_{ij}(\Theta^*)$ can be considered as a measure of the average overlap between Gaussians i and j in the mixture. When $P(x|m_i^*, \Sigma_i^*)$ and $P(x|m_j^*, \Sigma_j^*)$ have a high overlap at a point x, $h_i(x)h_j(x)$ takes a large value; otherwise, $h_i(x)h_j(x)$ takes a small value. When they are well separated at x, $h_i(x)h_j(x)$ becomes zero. Thus, the product $h_i(x)h_j(x)$ represents the degree of overlap between $P(x|m_i^*, \Sigma_i^*)$ and $P(x|m_j^*, \Sigma_j^*)$ at x in the mixture, and the above $e_{ij}(\Theta^*)$ is an average overlap measure between the Gaussians i and j in the mixture.

As a whole, we consider the worst case and define

$$e(\Theta^*) = \max_{i \neq j} e_{ij}(\Theta^*) \tag{10}$$

as an average overlap of Gaussians in the original mixture. Obviously, $0 \leqslant e(\Theta^*) \leqslant 1$.

We further introduce three kinds of special polynomial functions which we often meet in the following analysis.

Definition 1. $g(x, \Theta^*)$ is called a regular function if it satisfies:

(i) If Θ^* is fixed, $g(x, \Theta^*)$ is a polynomial function of the component variables $x_1, \ldots, x_d$ of x.
(ii) If x is fixed, $g(x, \Theta^*)$ is a polynomial function of the elements of $m_1^*, \ldots, m_K^*, \Sigma_1^*, \ldots, \Sigma_K^*, \Sigma_1^{*-1}, \ldots, \Sigma_K^{*-1}$, as well as $\mathscr{A}^* = [\alpha_1^*, \ldots, \alpha_K^*]^{\mathrm{T}}, \mathscr{A}^{*-1} = [\alpha_1^{*-1}, \ldots, \alpha_K^{*-1}]^{\mathrm{T}}$.

Definition 2. $g(x, \Theta^*)$ is called a balanced function if it satisfies (i) and the following:

(iii) If x is fixed, $g(x, \Theta^*)$ is a polynomial function of the elements of $\mathscr{A}^*$, $\mathscr{A}^{*-1}$, $m_1^*, \ldots, m_K^*$, $\Sigma_1^*, \ldots, \Sigma_K^*$, $\lambda(\Theta^*)\Sigma_1^{*-1}, \ldots, \lambda(\Theta^*)\Sigma_K^{*-1}$, where

$$\lambda(\Theta^*) = \max_{i,k} \lambda_{ik},$$

where λ_{ik} is the kth eigenvalue of the covariance matrix Σ_j^*.

Definition 3. $g(x, \Theta^*)$ is called a convertible function if it is regular and there is a nonnegative number q such that $\lambda^q(\Theta^*)g(x, \Theta^*)$ is converted into a balanced function.

Furthermore, we give certain assumptions on Θ^* that regularize the manner of $e(\Theta^*)$ tending to zero.

- We assume that Θ^* satisfies the first condition that

  (1) $\alpha_i^* \geqslant \alpha$, for $i = 1, \ldots, K$,

  where α is a positive number.
- Our second assumption is that the eigenvalues of all the covariance matrices satisfy

  (2) $\beta\lambda(\Theta^*) \leqslant \lambda_{ik} \leqslant \lambda(\Theta^*)$, for $i = 1, \ldots, K$, $k = 1, \ldots, d$,

  where β is a positive number.
- The third assumption is that the mean vectors of the Gaussians in the mixture satisfy

  (3) $vD_{\max}(\Theta^*) \leqslant D_{\min}(\Theta^*) \leqslant \|m_i^* - m_j^*\| \leqslant D_{\max}(\Theta^*)$, for $i \neq j$,

  where $D_{\max}(\Theta^*) = \max_{i \neq j} \|m_i^* - m_j^*\|$, $D_{\min}(\Theta^*) = \min_{i \neq j} \|m_i^* - m_j^*\|$, and v is a positive number.

With the above preparations, we now introduce the following lemma.

Lemma 1. *Suppose that Θ^* satisfies Conditions* (1–3) *and that $e(\Theta^*) \to 0$ is considered as an infinitesimal. If $g(x, \Theta^*)$ is a regular and convertible function, we have*

$$\int g(x, \Theta^*)\gamma_{ij}(x)P(x|\Theta^*)\, \mathrm{d}x = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)), \tag{11}$$

where $\varepsilon > 0$ is an arbitrarily small number, and $o(x)$ means that it is a higher-order infinitesimal as $x \to 0$.

The proof is given in [13].

3. Main results

We now consider the parameter mapping of the EM iteration $\Theta^{(k+1)} = M_N(\Theta^{(k)})$, which is explicitly expressed by Eqs. (3), (4) and (7). For mathematical analysis, we need to represent Θ by a set of independent variables. In order to

do so, we introduce the following subspace:

$$\mathscr{R}_1 = \left\{ \Theta : \sum_{j=1}^{K} \alpha_j = 0, \sigma_{pq}^{(j)} = \sigma_{qp}^{(j)} \text{ for all } j, p, q \right\},$$

which is obtained from

$$\mathscr{R}_2 = \left\{ \Theta : \sum_{j=1}^{K} \alpha_j = 1, \sigma_{pq}^{(j)} = \sigma_{qp}^{(j)} \text{ for all} j, p, q \right\}$$

by the constant shift Θ_0. For the Gaussian mixture, the constraint that all Σ_j are positive definite should also be added to $\mathscr{R}_2$ and thus $\mathscr{R}_1$. It can be easily verified that with this constraint, $\mathscr{R}_1$ becomes an open convex set within the original subspace. Since we will only consider the local differential properties of the parameter mapping at an interior point of the open convex set, we can set a new coordinate system for the parameter vector Θ via a set of the unit basis vectors $E = [e_1, \ldots, e_m]$, where m is the dimension of $\mathscr{R}_1$. In this way, the independent parameters of the Gaussian mixture become $\hat{\Theta} = E^{\mathrm{T}}\Theta$ and thus $\hat{\Theta}^{(k+1)} = E^{\mathrm{T}}\Theta^{(k+1)} = E^{\mathrm{T}} M_N(\Theta^{(k)})$. Certainly, this compact representation of the parameters is equivalent to the natural representation of the parameters for Gaussian mixture. However, it is convenient for mathematical analysis. We will use the two parametric representations equivalently in this paper. Hereafter, the parameter $\hat{\Theta}$ denotes the compact parameter representation with E.

Based on the relation between the two parametric representations, we have

$$\begin{aligned}\frac{\partial \hat{\Theta}^{(k+1)}}{\partial (\hat{\Theta}^{(k)})^{\mathrm{T}}} &= \frac{\partial E^{\mathrm{T}} M_N(\Theta^{(k)})}{\partial (\hat{\Theta}^{(k)})^{\mathrm{T}}} = E^{\mathrm{T}} \frac{\partial M_N(\Theta^{(k)})}{\partial (\Theta^{(k)})^{\mathrm{T}}} \frac{\partial \Theta^{(k)}}{\partial (\hat{\Theta}^{(k)})^{\mathrm{T}}} \\ &= E^{\mathrm{T}} \frac{\partial M_N(\Theta^{(k)})}{\partial (\Theta^{(k)})^{\mathrm{T}}} \frac{\partial E\hat{\Theta}^{(k)}}{\partial (\hat{\Theta}^{(k)})^{\mathrm{T}}} = E^{\mathrm{T}} \frac{\partial M_N(\Theta^{(k)})}{\partial (\Theta^{(k)})^{\mathrm{T}}} E. \end{aligned} \tag{12}$$

Using Θ instead of $\Theta^{(k)}$ in the parameter mapping or the EM iteration, we introduce the following two notations:

$$DM_N(\Theta) = \frac{\partial M_N(\Theta)}{\partial \Theta^{\mathrm{T}}}, \tag{13}$$

$$\begin{aligned} DM_N(\hat{\Theta}) &= \left. \frac{\partial \hat{\Theta}^{(k+1)}}{\partial (\hat{\Theta}^{(k)})^{\mathrm{T}}} \right|_{\hat{\Theta}^{(k)} = \hat{\Theta} = E^{\mathrm{T}}\Theta} \\ &= E^{\mathrm{T}} DM_N(\Theta) E. \end{aligned} \tag{14}$$

Therefore, we have

$$\begin{aligned} \|DM_N(\hat{\Theta}^{(k)})\| &= \|E^{\mathrm{T}} DM_N(\Theta^{(k)}) E\| \\ &\leqslant \|E\|^2 \|DM_N(\Theta^{(k)})\|, \end{aligned} \tag{15}$$

where $\|\cdot\|$ denotes the Euclidean norm for a matrix.

Since the norm $\|E\|$ is a positive constant, we only need to prove that $\|DM_N(\Theta^{(k)})\|$ can be small enough so that the iteration mapping becomes a contraction mapping via the mean value theorem. Suppose that Θ^N is a consistent solution of the maximum likelihood on the sample data set $\mathscr{S} = \{x_t\}_{t=1}^N$ and thus the EM algorithm can converge to it, that is, Θ^N is a fixed point of the parameter mapping $M_N(\Theta)$. We can analyze $\|DM_N(\Theta^{(k)})\|$ around Θ^N asymptotically, that is, we study it via $DM(\Theta^*) = \lim_{N\to\infty} DM_N(\Theta^N)$.Before doing so, we give the partial derivatives of $DM_N(\Theta)$ at Θ^* in the block forms (refer to [16] for derivation).

$$\left. \frac{\partial \alpha_j^{(k+1)}}{\partial \alpha_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} = \frac{1}{N} \sum_{t=1}^{N} \frac{\gamma_{ij}(t)}{\alpha_i^*}, \tag{16}$$

$$\left. \frac{\partial \alpha_j^{(k+1)}}{\partial m_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} = \frac{1}{N} \sum_{t=1}^{N} \gamma_{ij}(t) \Sigma_i^{*-1}(x^{(t)} - m_i^*), \tag{17}$$

$$\left. \frac{\partial \alpha_j^{(k+1)}}{\partial \Sigma_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} = -\frac{1}{2N} \sum_{t=1}^{N} \gamma_{ij}(t) U_i(t), \tag{18}$$

$$\begin{aligned} \left. \frac{\partial m_j^{(k+1)}}{\partial \alpha_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} &= \frac{\sum_{t=1}^{N} \gamma_{ij}(t) x^{(t)}}{\alpha_i^* \sum_{t=1}^{N} h_j(t)} \\ &\quad - \frac{\sum_{t=1}^{N} h_j(t) x^{(t)} \sum_{t=1}^{N} \gamma_{ij}(t)}{\alpha_i^* (\sum_{t=1}^{N} h_j(t))^2}, \end{aligned} \tag{19}$$

$$\begin{aligned} &\left. \frac{\partial m_j^{(k+1)}}{\partial m_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} \\ &= \frac{\sum_{t=1}^{N} \gamma_{ij}(t) x^{(t)} \otimes [\Sigma_i^{*-1}(x^{(t)} - m_i^*)]^{\mathrm{T}}}{\sum_{t=1}^{N} h_j(t)} \\ &\quad - \frac{\sum_{t=1}^{N} h_j(t) x^{(t)} \otimes \sum_{t=1}^{N} \gamma_{ij}(t) [\Sigma_i^{*-1}(x^{(t)} - m_i^*)]^{\mathrm{T}}}{(\sum_{t=1}^{N} h_j(t))^2}, \end{aligned} \tag{20}$$

$$\begin{aligned} &\left. \frac{\partial m_j^{(k+1)}}{\partial \Sigma_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} \\ &= \frac{\sum_{t=1}^{N} h_j(t) x^{(t)} \otimes \sum_{t=1}^{N} \gamma_{ij}(t) U_i(t)}{2(\sum_{t=1}^{N} h_j(t))^2} \\ &\quad - \frac{\sum_{t=1}^{N} \gamma_{ij}(t) x^{(t)} \otimes U_i(t)}{2 \sum_{t=1}^{N} h_j(t)} \end{aligned} \tag{21}$$

$$\begin{aligned} \left. \frac{\partial \Sigma_j^{(k+1)}}{\partial \alpha_i^{(k)}} \right|_{\Theta^{(k)} = \Theta^*} &= \frac{\sum_{t=1}^{N} \gamma_{ij}(t) R_j(t)}{\alpha_i^* \sum_{t=1}^{N} h_j(t)} \\ &\quad - \frac{\sum_{t=1}^{N} h_j(t) R_j(t) \sum_{t=1}^{N} \gamma_{ij}(t)}{\alpha_i^* (\sum_{t=1}^{N} h_j(t))^2}, \end{aligned} \tag{22}$$

$$\left.\frac{\partial \Sigma_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = \frac{\sum_{t=1}^N \gamma_{ij}(t) R_i(t) \otimes [\Sigma_i^{*-1}(x^{(t)} - m_i^*)]}{\sum_{t=1}^N h_j(t)} - \frac{\sum_{t=1}^N h_j(t) R_i(t) \otimes \sum_{t=1}^N \gamma_{ij}(t) \Sigma_i^{*-1}(x^{(t)} - m_i^*)}{(\sum_{t=1}^N h_j(t))^2} - \frac{\delta_{ij} \sum_{t=1}^N h_j(t)[I \otimes (x^{(t)} - m_i^*) + (x^{(t)} - m_i^*) \otimes I]}{\sum_{t=1}^N h_j(t)}, \tag{23}$$

$$\left.\frac{\partial \Sigma_j^{(k+1)}}{\partial \Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = \frac{\sum_{t=1}^N h_j(t) R_i(t) \otimes \sum_{t=1}^N \gamma_{ij}(t) U_i(t)}{2(\sum_{t=1}^N h_j(t))^2} - \frac{\sum_{t=1}^N \gamma_{ij}(t) R_i(t) \otimes U_i(t)}{2\sum_{t=1}^N h_j(t)}, \tag{24}$$

where

$$\gamma_{ij}(t) = (\delta_{ij} - h_i(t)) h_j(t),$$
$$h_i(t) = \frac{\alpha_i^* P(x^{(t)} | m_i^*, \Sigma_i^*)}{\sum_{t=1}^N \alpha_i^* P(x^{(t)} | m_i^*, \Sigma_i^*)},$$
$$R_i(t) = (x^{(t)} - m_i^*)(x^{(t)} - m_i^*)^{\mathrm{T}},$$
$$U_i(t) = \Sigma_i^{*-1} - \Sigma_i^{*-1}(x^{(t)} - m_i^*)(x^{(t)} - m_i^*)^{\mathrm{T}} \Sigma_i^{*-1},$$

and δ_{ij} is the Kronecker function.

With the above preparations, we are ready to give our main theorem.

Theorem 1. *Given i.i.d. samples* $\{x^{(t)}\}_1^N$ *from a mixture of K Gaussian distributions of the parameters* Θ^* *that satisfies conditions* (1–3)*, when* $e(\Theta^*)$ *is considered as an infinitesimal, as it tends to zero, we have almost surely*:

$$\lim_{N\to\infty} \|DM_N(\Theta^*)\| = \|DM(\Theta^*)\| = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)), \tag{25}$$

where $\varepsilon > 0$ *is an arbitrarily small number.*

Proof. Under the law of large number, we have almost surely:

$$\lim_{N\to\infty} \|DM_N(\Theta^*)\| = \| \lim_{N\to\infty} DM_N(\Theta^*)\| = \|DM(\Theta^*)\|. \tag{26}$$

According to the definition of the Euclidean norm, if each element of $DM(\Theta^*)$ is a higher-order infinitesimal quantity of $\mathrm{e}^{0.5-\varepsilon}(\Theta^*)$, $\|DM(\Theta^*)\|$ is also a higher-order infinitesimal quantity of $\mathrm{e}^{0.5-\varepsilon}(\Theta^*)$. Thus, we only need to prove that each element of $DM(\Theta^*)$ is a higher-order infinitesimal quantity of $\mathrm{e}^{0.5-\varepsilon}(\Theta^*)$. In the following, we will consider the elements of $DM(\Theta^*)$ in different block forms.

We begin with the partial derivative $\partial\alpha_j^{(k+1)}/\partial\alpha_i^{(k)}$. According to Eq. (16), for each pair of i and j, we have

$$\lim_{N\to\infty} \left.\frac{\partial \alpha_j^{(k+1)}}{\partial \alpha_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} \leqslant \frac{1}{\alpha} \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^N |\gamma_{ij}(t)| \leqslant \frac{1}{\alpha} e(\Theta^*) = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

As to the block form $\partial\alpha_j^{(k+1)}/\partial m_i^{(k)}$, according to Eq. (17) we have

$$\lim_{N\to\infty} \left.\frac{\partial \alpha_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^N \gamma_{ij}(t) \Sigma_i^{*-1}(x^{(t)} - m_i^*) = \int \gamma_{ij}(x) \Sigma_i^{*-1}(x - m_i^*)\, \mathrm{d}x.$$

If $g(x, \Theta^*)$ is any element of the matrix $\Sigma_i^{*-1}(x - m_i^*)$, it is a regular and convertible function with $q = 1$. According to Lemma 1, we have

$$\lim_{N\to\infty} \left.\frac{\partial \alpha_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

That is, each element of $\lim_{N\to\infty} \partial\alpha_j^{(k+1)}/\partial m_i^{(k)}|_{\Theta^{(k)}=\Theta^*}$ is $o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*))$.

Similarly, according to Eq. (18) we can prove:

$$\lim_{N\to\infty} \left.\frac{\partial \alpha_j^{(k+1)}}{\partial \Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = -\frac{1}{2} \int \gamma_{ij}(x) U_i(x)\, \mathrm{d}x = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)),$$

where

$$U_i(x) = \Sigma_i^{*-1} - \Sigma_i^{*-1}(x - m_i^*)(x - m_i^*)^{\mathrm{T}} \Sigma_i^{*-1}.$$

We turn to the block form $\partial m_j^{(k+1)}/\partial\alpha_i^{(k)}$ given by Eq. (19) and have

$$\begin{aligned}
&\lim_{N\to\infty}\left.\frac{\partial m_j^{(k+1)}}{\partial\alpha_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}\\
&=\lim_{N\to\infty}\frac{\sum_{t=1}^N\gamma_{ij}(t)x^{(t)}}{\alpha_i^*\sum_{t=1}^N h_j(t)}\\
&\quad-\lim_{N\to\infty}\frac{\sum_{t=1}^N h_j(t)x^{(t)}\sum_{t=1}^N\gamma_{ij}(t)}{\alpha_i^*[\sum_{t=1}^N h_j(t)]^2}\\
&=\lim_{N\to\infty}\frac{(1/N)\sum_{t=1}^N\gamma_{ij}(t)x^{(t)}}{\alpha_i^*[\sum_{t=1}^N h_j(t)/N]}-\lim_{N\to\infty}\frac{1}{\sum_{t=1}^N h_j(t)}\\
&\quad\times\frac{\sum_{t=1}^N h_j(t)x^{(t)}(1/N)\sum_{t=1}^N\gamma_{ij}(t)}{\alpha_i^*[\sum_{t=1}^N h_j(t)/N]}\\
&=\frac{1}{\alpha_i^*\alpha_j^*}\int\gamma_{ij}(x)x\,\mathrm{d}x-\frac{1}{\alpha_i^*\alpha_j^*}\int h_j(x)x\,\mathrm{d}x\int\gamma_{ij}(x)\,\mathrm{d}x\\
&=\frac{1}{\alpha_i^*\alpha_j^*}\int\gamma_{ij}(x)x\,\mathrm{d}x-\frac{1}{\alpha_i^*\alpha_j^*}m_j^*e_{ij}(\Theta^*).
\end{aligned}$$

Letting $g(x,\Theta^*)$ be any x_i and by Lemma 1, we have

$$\lim_{N\to\infty}\left.\frac{\partial\Sigma_j^{(k+1)}}{\partial\alpha_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}=o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

For the block form $\frac{\partial m_j^{(k+1)}}{\partial m_i^{(k)}}$ given in Eq. (20), we have

$$\begin{aligned}
&\lim_{N\to\infty}\left.\frac{\partial m_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}\\
&=\frac{1}{\alpha_j^*}\left(\int\gamma_{ij}(x)x\otimes[\Sigma_i^{*-1}(x-m_i^*)]^{\mathrm{T}}\,\mathrm{d}x\right.\\
&\quad\left.-\int\gamma_{ij}(x)(x)m_j^*\otimes[\Sigma_i^{*-1}(x-m_i^*)]^{\mathrm{T}}\,\mathrm{d}x\right).
\end{aligned}$$

Letting $g(x,\Theta^*)$ be any element of $x\otimes[\Sigma_i^{*-1}(x-m_i^*)]^{\mathrm{T}}$ or $m_i^*\otimes[\Sigma_i^{*-1}(x-m_i^*)]^{\mathrm{T}}$ and by Lemma 1, we have

$$\lim_{N\to\infty}\left.\frac{\partial m_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}=o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

We further consider the block form $\partial m_j^{(k+1)}/\partial\Sigma_i^{(k)}$. According to Eq. (21) we have

$$\begin{aligned}
&\lim_{N\to\infty}\left.\frac{\partial m_j^{(k+1)}}{\partial\Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}\\
&=\frac{1}{2\alpha_j^*}\int\gamma_{ij}(x)[m_j^*-x]\otimes U_i(x)\,\mathrm{d}x.
\end{aligned}$$

Letting $g(x,\Theta^*)$ be any element of $[m_j^*-x]\otimes U_i(x)$ and by Lemma 1, we have

$$\lim_{N\to\infty}\left.\frac{\partial m_j^{(k+1)}}{\partial\Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}=o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

Furthermore, according to Eq. (22) we have

$$\begin{aligned}
&\lim_{N\to\infty}\left.\frac{\partial\Sigma_j^{(k+1)}}{\partial\alpha_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}\\
&=\frac{1}{\alpha_i^*\alpha_j^*}\left(\int\gamma_{ij}(x)R_j(x)\,\mathrm{d}x-\Sigma_j^*e_{ij}(\Theta^*)\right),
\end{aligned}$$

where

$$R_j(x)=(x-m_j^*)(x-m_j^*)^{\mathrm{T}}.$$

Letting $g(x,\Theta^*)$ be any element of $R_j(x)$ and by Lemma 1, we have

$$\lim_{N\to\infty}\left.\frac{\partial\Sigma_j^{(k+1)}}{\partial\alpha_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}=o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

As to the block form $\partial\Sigma_j^{(k+1)}/\partial m_i^{(k)}$ given by Eq. (23), since

$$\lim_{N\to\infty}\frac{\sum_{t=1}^N h_i(t)[I\otimes(x^{(t)}-m_i^*)+(x^{(t)}-m_i^*)\otimes I]}{\sum_{t=1}^N h_i(t)}=0,$$

under the law of large number, we have

$$\begin{aligned}
&\lim_{N\to\infty}\left.\frac{\partial\Sigma_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}\\
&=\frac{1}{\alpha_j^*}\left(\int\gamma_{ij}(x)R_i(x)\otimes[\Sigma_i^{*-1}(x-m_i^*)]\,\mathrm{d}x\right.\\
&\quad-\int\gamma_{ij}(x)[\Sigma_i^*-(m_i^*-m_j^*)(m_i^*-m_j^*)^{\mathrm{T}}]\\
&\quad\left.\otimes[\Sigma_i^{*-1}(x-m_i^*)]\,\mathrm{d}x\right).
\end{aligned}$$

Letting $g(x,\Theta^*)$ be any element of $R_i(x)\otimes[\Sigma_i^{*-1}(x-m_i^*)]$ or $[\Sigma_i^*-(m_i^*-m_j^*)(m_i^*-m_j^*)^{\mathrm{T}}]\otimes[\Sigma_i^{*-1}(x-m_i^*)]$ and by Lemma 1, we have

$$\lim_{N\to\infty}\left.\frac{\partial\Sigma_j^{(k+1)}}{\partial m_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*}=o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

Now we turn to the last case, i.e., the block form $\partial\Sigma_j^{(k+1)}/\partial\Sigma_i^{(k)}$. By the law of large number, we have

$$\lim_{N\to\infty} \left.\frac{\partial\alpha_j^{(k+1)}}{\partial\Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = \frac{1}{2\alpha_j^*}\left(\int \gamma_{ij}(x)[\Sigma_i^* - (m_i^* - m_j^*)(m_i^* - m_j^*)^{\mathrm{T}}] \otimes U_i(x)\,\mathrm{d}x - \int \gamma_{ij}(x)R_i(x)\otimes U_i(x)\,\mathrm{d}x\right).$$

Letting $g(x, \Theta^*)$ be any element of $[\Sigma_i^* - (m_i^* - m_j^*)(m_i^* - m_j^*)^{\mathrm{T}}]\otimes U_i(x)$ or $R_i(x)\otimes U_i(x)$ and by Lemma 1, we have

$$\lim_{N\to\infty} \left.\frac{\partial\alpha_j^{(k+1)}}{\partial\Sigma_i^{(k)}}\right|_{\Theta^{(k)}=\Theta^*} = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*)).$$

Summing up all the results, we obtain that each element of $DM(\Theta^*)$ is $o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*))$. Therefore, we finally have $\|DM(\Theta^*)\| = o(\mathrm{e}^{0.5-\varepsilon}(\Theta^*))$.

The proof is completed.

According to Theorem 1, as the average overlap of Gaussians in the original mixture becomes small, or more precisely, $\mathrm{e}(\Theta^*) \to 0$, the norm of $DM(\Theta^*)$ becomes a higher-order infinitesimal of $\mathrm{e}^{0.5-\varepsilon}(\Theta^*)$. That is, the norm of $DM(\Theta^*)$ can be arbitrarily small as long as $\mathrm{e}(\Theta^*)$ is small enough. Because $\lim_{N\to\infty} DM_N$

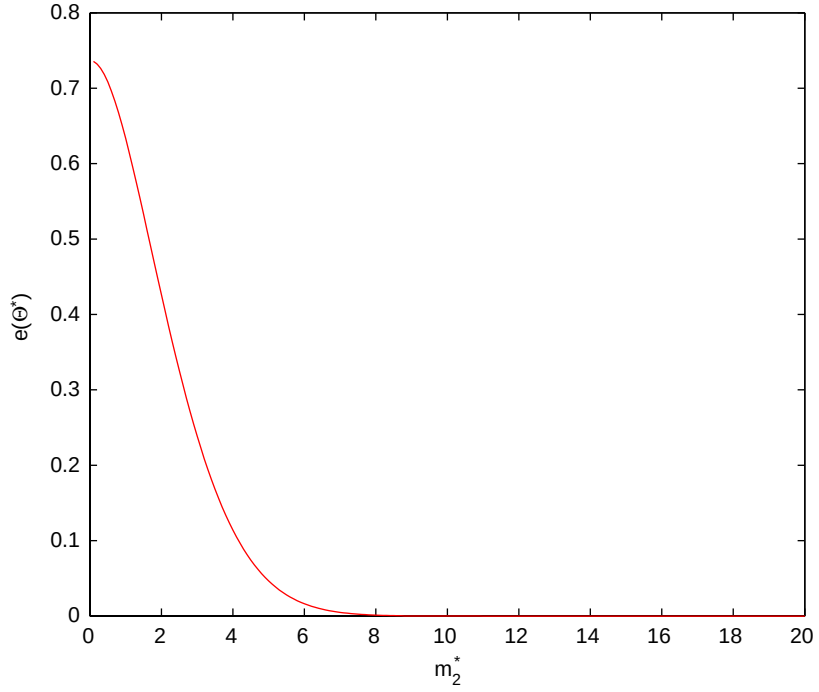

Fig. 1. The sketch of variation of $e(m^*)$ with m_2^*.

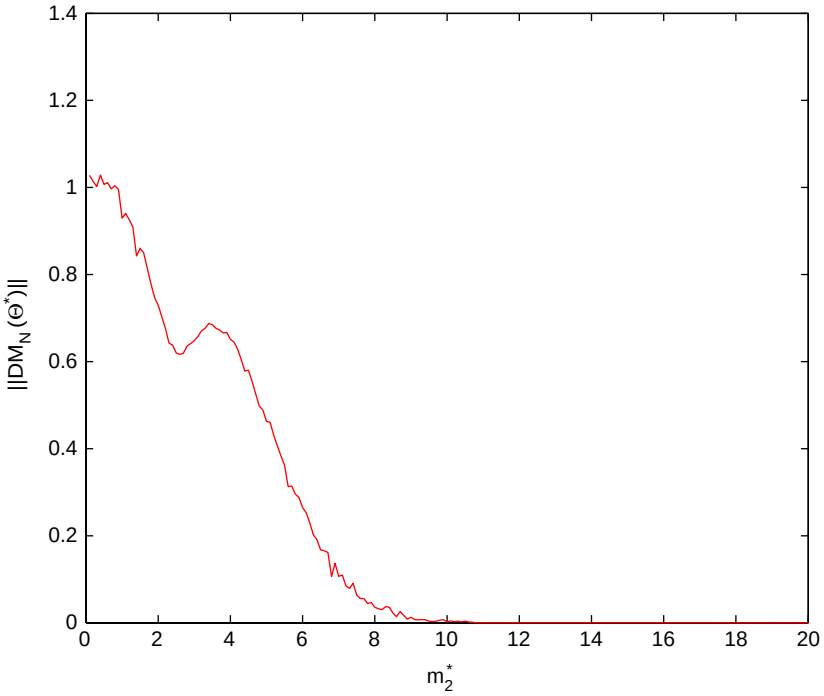

Fig. 2. The sketch of variation of $\|DM_N(m^*)\|$ with m_2^*.

Table 1
The simulation results on the correct convergence of the EM algorithm with the average overlap $e(m^*)$

m_2^*	2	4	6	10	14	20
$e(m^*)$	0.4260	0.1140	0.0164	4.5782e −014	5.7979e −036	1.7031e−057
m_1^N	0.0314	0.02426	0.007	−0.0116	0.0196	−0.0349
m_2^N	1.9531	3.9830	5.9714	9.9840	14.0066	19.9752
$\|DM_N(m^*)\|$	0.7295	0.6514	0.2647	0.0018	4.7050e−009	2.8513e−018
$\|DM_N(m^N)\|$	0.7462	0.6590	0.2581	0.0010	4.9302e−009	1.217e−023
$r(m^N)$	0.25	0.93	1.75	3.87	4.78	7.43

rather complicated in details, where we have randomly selected 5000 samples from each original Gaussian mixture.

As analyzed in the previous section, when $e(\Theta^*)$ is small enough, $\|DM_N(\Theta^*)\|$ becomes small enough so that $\|DM_N(\Theta^N)\|$ is less than 1 and the EM algorithm becomes a contraction mapping within a neighborhood of Θ^N. Now, we further demonstrate these theoretical results by the simulation experiments. We select 6 typical values of m_2^* and get the corresponding average overlaps $e(m^*)$ of two Gaussians in the original mixtures. Then, for each value of m_2^*, we compute $\|DM_N(m^*)\|$ and $\|DM_N(m^N)\|$, respectively, on a set of 5000 samples from the original mixture. Finally, we get the largest radius $r(m^N)$ of the neighborhood of m^N in which $\|DM_N(m)\| < 1$. That is, the EM algorithm can converge correctly to m^N when the initial m^0 is set within it. Clearly, $r(m^N)$ denotes the largest correct convergence radius of the EM algorithm from the point of view of the contract mapping. The simulation results for the 6 values of m_2^* are listed in Table 1.

According to the simulation results given in Table 1, $\|DM_N(m^N)\|$ is approximately equal to $\|DM_N(m^*)\|$. Moreover, as the average overlap of the Gaussians in the original mixture becomes smaller, the correct convergence radius of the EM algorithm becomes larger. Therefore, it is demonstrated by the simulation results that the EM algorithm for Gaussian mixtures tends to converge correctly when the overlap of Gaussians in the original mixture becomes small.

5. Conclusions

We have presented an analysis on the correct convergence of the EM algorithm for Gaussian mixtures. Our analysis shows that when the overlap of any two Gaussian distributions is small enough, the EM algorithm becomes a contract mapping of the parameters within a neighborhood of the consistent solution of the maximum likelihood. That is, the EM algorithm can converge consistently to the true parameters when it starts within the neighborhood. Moreover, it is further demonstrated by the simulation results that the radius of this correct convergence neighborhood becomes larger as the average overlap becomes smaller.

Although our studies in this paper are purely theoretical, they are also significant to the practical applications of the EM algorithm. In fact, these results on the correct convergence of the EM algorithm as well as the previous results [12–14] on the convergence rate of the EM algorithm show that the EM algorithm is a quite efficient method for the parameter estimation when the overlap of Gaussians in the original mixture is small enough. Practically, if we can measure the average overlap of actual Gaussians from the sample data directly, we may get the condition for the EM algorithm to converge correctly on the sample data, which is valuable for the applications of the EM algorithm. Clearly, it is probable to define a measure of the average overlap of actual Gaussians from the sample data directly. However, it is still difficult to give a reasonable and computable definition for it and we will investigate this problem in our future works.

Acknowledgements

The authors wish to express their gratitude to Prof. Lei Xu and Prof. Bingyuan Cao for some helpful discussions, and also to Miss Liangliang Wang for her support of the simulation experiments.

References

[1] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Statist. Soc. B 39 (1977) 1–38.
[2] C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Statist. 11 (1983) 95–103.
[3] R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood, and the EM algorithm, SIAM Rev. 26 (1984) 195–239.
[4] X.L. Meng, On the rate of convergence of the ECM algorithm, Ann. Statist. 22 (1994) 326–339.
[5] C. Liu, D.B. Rubin, The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, Biometrika 81 (1994) 633–648.

[6] X.L. Meng, D. van Dyk, The EM algorithm—an old folk-song sung to a fast new tune, J. R. Statist. Soc. B 59 (1997) 511–567.
[7] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27 (1999) 94–128.
[8] N. Laird, N. Lange, D. Stram, Maximizing likelihood computations with repeated measures: application of the EM algorithm, J. Amer. Statist. Assoc. 82 (1987) 97–105.
[9] M. Jamshidian, R.I. Jennrich, Conjugate gradient acceleration of the EM algorithm, J. Am. Statist. Assoc. 88 (1993) 221–228.
[10] K. Lange, A quasi-Newtonian acceleration of the EM algorithm, Statistica Sinica 5 (1995) 1–18.
[11] C. Liu, D.B. Rubin, Y. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.
[12] L. Xu, M.I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, Neural Comput. 8 (1996) 129–151.
[13] J. Ma, L. Xu, M.I. Jordan, Asymptotic convergence rate of the EM algorithm for Gaussian mixtures, Neural Comput. 12 (2000) 2881–2907.
[14] J. Ma, L. Xu, Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture, Neurocomputing, in press.
[15] L. Xu, Comparative analysis on convergence rates of the EM algorithms and its two modifications for Gaussian mixtures, Neural Processing Lett. 6 (1997) 69–76.
[16] S.R. Gerald, Matrix derivatives, in: Lecture Notes in Statistics, vol. 2, Marcel Dekker, New York, 1980.

About the Author—JINWEN MA received the Master of Science degree in applied mathematics from Xi'an Jiaotong University in 1988 and the Ph.D. degree in probability theory and statistics from Nankai University in 1992. From July 1992 to November 1999, he was a Lecturer or Associate professor at Department of Mathematics, Shantou University. He has been a full professor at Institute of Mathematics, Shantou University since December, 1999. In September 2001, has was transferred to the Department of Information Science at the School of Mathematical Sciences, Peking University. During 1995 and 2003, he also visited several times at Department of Computer Science & Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow. He has published over 60 academic papers on neural networks, pattern recognition, artificial intelligence, and information theory.