LETTER Communicated by Chris Williams

Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures

Jinwen Ma
Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin Hong Kong and Institute of Mathematics, Shantou University, Shantou, Guangdong, 515063, People's Republic of China

Lei Xu
Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin Hong Kong, People's Republic of China

Michael I. Jordan
Department of Computer Science and Department of Statistics, University of California at Berkeley, Berkeley, CA 94720, U.S.A.

It is well known that the convergence rate of the expectation-maximization (EM) algorithm can be faster than those of convention rst-order iterative algorithms when the overlap in the given mixture is small. But this argument has not been mathematically proved yet. This article studies this problem asymptotically in the setting of gaussian mixtures under the theoretical framework of Xu and Jordan (1996). It has been proved that the asymptotic convergence rate of the EM algorithm for gaussian mixtures locally around the true solution Θ^* is $o(e^{0.5-\varepsilon}(\Theta^*))$, where $\varepsilon > 0$ is an arbitrarily small number, $o(x)$ means that it is a higher-order in nitesimal as $x \to 0$, and $e(\Theta^*)$ is a measure of the average overlap of gaussians in the mixture. In other words, the large sample local convergence rate for the EM algorithm tends to be asymptotically superlinear when $e(\Theta^*)$ tends to zero.

1 Introduction

The expectation-maximization (EM) algorithm is a general methodology for maximum likelihood (ML) or maximum a posteriori (MAP) estimation (Dempster, Laird, & Rubin, 1977). A substantial literature has been devoted to the study of the convergence of EM and related methods (e.g., Wu, 1983; Redner & Walker, 1984; Meng, 1994; Liu & Rubin, 1994; Lange, 1995a; Meng & van Dyk, 1997; Delyon, Lavielle, & Moulines, 1999). The starting point for many of these studies is the fact that EM is generally a rst-order or linearly convergent algorithm, as can readily be seen by considering EM as a mapping $\Theta^{(k+1)} = M(\Theta^{(k)})$, with xed point $\Theta^* = M(\Theta^*)$. Results on

the rate of convergence of EM are obtained by calculating the information matrices for the missing data and the observed data (Dempster et al., 1977; Meng, 1994).

The theoretical convergence results obtained to date are of satisfying generality but of limited value in understanding why EM may converge slowly or rapidly in a particular problem. The existing methods to enhance the convergence rate of EM are generally based on the conventional superlinear optimization theory (e.g., Lange, 1995b; Jamshidian & Jennrich, 1997; Meng & van Dyk, 1998), and are usually rather more complex than EM. However, they are prey to other disadvantages, including an awkwardness at handling constraints on parameters. More fundamentally, it is not clear what the underlying factors are that slow EM's convergence.

It is also worth noting that EM has been successfully applied to large-scale problems such as hidden Markov models (Rabiner, 1989), probabilistic decision trees (Jordan & Jacobs, 1994), and mixtures of experts architectures (Jordan & Xu, 1995), where its empirical convergence rate can be significantly faster than those of conventional first-order iterative algorithms (i.e., gradient ascent). These empirical studies show that EM can be slow if the overlap in the given mixture is large but rapid if the overlap in the given mixture is small.

A recent analysis by Xu and Jordan (1996) provides some insight into the convergence rate of EM in the setting of gaussian mixtures. For the convenience of mathematical analyses, they studied a variant of the original EM algorithm for gaussian mixtures and showed that the condition number associated with this variant EM algorithm is guaranteed to be smaller than the condition number associated with gradient ascent, providing a general guarantee of the dominance of this variant EM algorithm over the gradient algorithm. Moreover, in cases in which the mixture components are well separated, they showed that the condition number for this EM algorithm approximately converges to one, corresponding to a local superlinear convergence rate. Thus, in this restrictive case, this type of EM algorithm has the favorable property of showing quasi-Newton behavior as it nears the ML or MAP solution. Xu (1997) further showed that the original EM algorithm has the same convergence properties as this variant EM algorithm.

We have further found by experiments that the convergence rate of the EM algorithm for gaussian mixtures is dominated by a measure of the average overlap of gaussians in the mixture. Actually, when the average overlap measure becomes small, the EM algorithm converges quickly when it nears the true solution. As the measure of the average overlap tends to zero, the EM algorithm tends to demonstrate a quasi-Newton behavior.

In this article, based on the mathematical connection between the variant EM algorithm and gradient algorithms and on one of its intermediate result on the convergence rate by Xu and Jordan (1996), as well as on the same convergence result of the original EM algorithm (Xu, 1997), we present a further theoretical analysis of the asymptotic convergence rate of the EM algorithm

locally around the true solution Θ^* with respect to $e(\Theta^*)$, a measure of the average overlap of the gaussians in the mixture. We prove a theorem that shows the asymptotic convergence rate is a higher-order infinitesimal than $e^{0.5-\varepsilon}(\Theta^*)$ when $e(\Theta^*)$ tends to zero under certain conditions, where $\varepsilon > 0$ is an arbitrarily small number. Thus, we see that the large sample local convergence rate for the EM algorithm tends to be asymptotically superlinear when $e(\Theta^*)$ tends to zero.

Section 2 presents the EM algorithm and the Hessian matrix for the gaussian mixture model. In section 3, we intuitively introduce our theorem on the asymptotic convergence rate of EM. Section 4 describes several lemmas needed for the proof; the proof is contained in section 5. Section 6 presents our conclusions.

2 The EM Algorithm and the Hessian Matrix of the Log-Likelihood

We consider the following gaussian mixture model:

$$P(x|\Theta) = \sum_{j=1}^{K} \alpha_j P(x|m_j, \Sigma_j); \qquad \alpha_j \geq 0; \sum_{j=1}^{K} \alpha_j = 1; \tag{2.1}$$

where

$$P(x|m_j, \Sigma_j) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1} (x-m_j)} \tag{2.2}$$

and where K is the number of the mixture components, x denotes a sample vector, and d is the dimensionality of x. The parameter vector Θ consists of the mixing proportions α_j, the mean vectors m_j, and the covariance matrices $\Sigma_j = (\sigma_{pq}^j)_{d \times d}$, which are assumed positive definite.

Given K and given independently and identically distributed (i.i.d.) samples $\{x^{(t)}\}_1^N$, we estimate Θ by maximizing the log-likelihood:

$$l(\Theta) = \log \prod_{t=1}^{N} P(x^{(t)}|\Theta) = \sum_{t=1}^{N} \log P(x^{(t)}|\Theta). \tag{2.3}$$

This log-likelihood can be optimized iteratively via the EM algorithm as follows:

$$\alpha_j^{(k+1)} = \frac{1}{N} \sum_{t=1}^{N} h_j^{(k)}(t) \tag{2.4}$$

$$m_j^{(k+1)} = \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \sum_{t=1}^{N} h_j^{(k)}(t) x^{(t)} \tag{2.5}$$

$$\Sigma_j^{(k+1)} = \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \sum_{t=1}^{N} h_j^{(k)}(t)(x^{(t)} - m_j^{(k+1)})(x - m_j^{(k+1)})^T; \tag{2.6}$$

where the posterior probabilities $h_j^{(k)}(t)$ are given by

$$h_j^{(k)}(t) = \frac{\alpha_j^{(k)} P(x^{(t)} | m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^{K} \alpha_i^{(k)} P(x^{(t)} | m_i^{(k)}, \Sigma_i^{(k)})}. \tag{2.7}$$

This iterative procedure converges to a local maximum of the log-likelihood (Dempster, et al., 1977). We suppose that $\hat{\Theta}$ is a local solution to the likelihood equation 2.3, and the EM algorithm converges to it. We now analyze the local convergence rate around this solution.

For the convenience of mathematical analyses, Xu and Jordan (1996) studied a variant of the EM algorithm by letting equation 2.6 be replaced by

$$\Sigma_j^{(k+1)} = \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \sum_{t=1}^{N} h_j^{(k)}(t)(x^{(t)} - m_j^{(k)})(x - m_j^{(k)})^T; \tag{2.8}$$

that is, the update of $\Sigma_j^{(k+1)}$ is based on the last update $m_j^{(k)}$ instead of $m_j^{(k+1)}$ in equation 2.6. For clarity, we denote this variant of EM by VEM in this article. They showed that at each iteration, the following relationship holds between the gradient of the log-likelihood and the VEM update step:

$$\mathcal{A}^{(k+1)} - \mathcal{A}^{(k)} = P_{\mathcal{A}^{(k)}} \frac{\partial l}{\partial \mathcal{A}}|_{\mathcal{A}=\mathcal{A}^{(k)}} \tag{2.9}$$

$$m_j^{(k+1)} - m_j^{(k)} = P_{m_j^{(k)}} \frac{\partial l}{\partial m_j}|_{m_j=m_j^{(k)}} \tag{2.10}$$

$$\mathrm{vec}\left[\Sigma_j^{(k+1)}\right] - \mathrm{vec}\left[\Sigma_j^{(k)}\right] = P_{\Sigma_j^{(k)}} \frac{\partial l}{\partial \mathrm{vec}[\Sigma_j]}|_{\Sigma_j=\Sigma_j^{(k)}}; \tag{2.11}$$

where

$$P_{\mathcal{A}}(k) = \frac{1}{N}\left(\mathrm{diag}\left[\alpha_1^{(k)}, \ldots, \alpha_K^{(k)}\right] - \mathcal{A}^{(k)}\mathcal{A}^{(k)T}\right) \tag{2.12}$$

$$P_{m_j}(k) = \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \Sigma_j^{(k)} \tag{2.13}$$

$$P_{\Sigma_j}(k) = \frac{2}{\sum_{t=1}^{N} h_j^{(k)}(t)} \left(\Sigma_j^{(k)} \otimes \Sigma_j^{(k)}\right); \tag{2.14}$$

and where $\mathcal{A}$ denotes the vector of mixing proportions $[\alpha_1, \ldots, \alpha_K]^T$, j indexes the mixture components $(j = 1, \ldots, K)$, k denotes the iteration number, vec[B] denotes the vector obtained by stacking the column vectors of

the matrix B, $\mathrm{vec}[B]^T = .\mathrm{vec}[B]/^T$, and $\otimes$ denotes the Kronecker product. Moreover, given the constraints $\sum_{j=1}^{K} \alpha_j^{.k/} = 1$ and $\alpha_j^{.k/} \geq 0$; $P_{\mathcal{A}}.k/$ is a positive definite matrix and the matrices $P_{m_j}.k/$ and $P_{6_j}.k/$ are positive definite with probability one for N sufficiently large. Assembling these matrices into a single matrix $P.2/ = \mathrm{diag}[P_{\mathcal{A}}; P_{m_1}; \ldots; P_{m_K}; P_{6_1}; \ldots; P_{6_K}]$; we have the following equation for the VEM update step given by equation 3.7 in Xu and Jordan (1996):

$$2^{.k+1/} = 2^{.k/} + P\left(2^{.k/}\right) \frac{@l}{@2}|_{2=2^{.k/}}; \tag{2.15}$$

where 2 is the collection of mixture parameters:

$$2 = [\mathcal{A}^T; m_1^T; \ldots; m_K^T; \mathrm{vec}[6_1]^T; \ldots; \mathrm{vec}[6_K]^T]^T:$$

In order to represent 2 to a set of independent variables for derivation, we introduce the following subspace,

$$\mathcal{R}_1 = \left\{ 2: \sum_{j=1}^{K} \alpha_j = 0; \gamma_{pq}^{.j/} = \gamma_{qp}^{.j/}; \text{ for all } j; p; q \right\};$$

which is obtained from

$$\mathcal{R}_2 = \left\{ 2: \sum_{j=1}^{K} \alpha_j = 1; \gamma_{pq}^{.j/} = \gamma_{qp}^{.j/}; \text{ for all } j; p; q \right\}$$

by the constant shift 2_0. For the gaussian mixture, the constraint that all 6_j are positive definite should also be added to $\mathcal{R}_2$ and thus $\mathcal{R}_1$. It can be easily verified that this constraint makes $\mathcal{R}_1$ be an open convex set of it. Since we will consider only the local differential properties of log-likelihood function at an interior point of the open convex set, we can set a new coordinate system for the parameter vector 2 via a set of the unit basis vectors $E = [e_1; \ldots; e_m]$, where m is the dimension of $\mathcal{R}_1$.

In fact, for each $2' = 2 - 2_0 \in \mathcal{R}_1$, let its coordinates under the bases $e_1; \ldots; e_m$ be denoted by 2_c; we have

$$2 - 2_0 = E2_c; \text{ or } 2 = E2_c + 2_0: \tag{2.16}$$

Multiplying its both sides by E^T, it follows from $E^TE = I$ that

$$E^T2 = E^TE2_c + E^T2_0; \text{ or } 2_c = E^T2 - E^T2_0: \tag{2.17}$$

Putting it into equation 2.16, we have

$$2' = 2 - 2_0 = E^TE.2 - 2_0/ = EE^T2' \quad \text{for} \quad 2' \in \mathcal{R}_1: \tag{2.18}$$

Thus for a matrix A, let $\|A\| = \max_{\|2'\|=1;2'\in\mathcal{R}_1} \|A2'\|$ be its norm constrained on $\mathcal{R}_1$. From equation 2.18, we certainly have the equality that $\|A\| = \|AEE^T\|$. We use the Euclidean norm for vectors in this article and have that

$$\|E^T\| = \max_{\|2\|=1;2\in\mathcal{R}_1} \|E^T 2\| = \max_{\|2\|=1;2\in\mathcal{R}_1} .2^T EE^T 2/^{\frac{1}{2}}$$

$$= \max_{\|2\|=1;2\in\mathcal{R}_1} .2^T 2/^{\frac{1}{2}} = \max_{\|2\|=1;2\in\mathcal{R}_1} \|2\| = 1:$$

Following the last inequality on p. 137 of Xu and Jordan (1996), we have the local convergence rate around $\hat{2}$ is bounded by

$$r = \lim_{k\to\infty} \frac{\|2^{.k+1/} - \hat{2}\|}{\|2^{.k/} - \hat{2}\|}$$

$$\leq \|E^T.I + P.\hat{2}/H.\hat{2}//\| = \|E^T.I + P.\hat{2}/H.\hat{2}//EE^T\|$$

$$\leq \|E^T.I + P.\hat{2}/H.\hat{2}//E\|\|E^T\|: \tag{2.19}$$

By the fact $\|E^T\| = 1$, we have

$$r \leq \|I + E^T P.\hat{2}/H.\hat{2}/E\|: \tag{2.20}$$

For the original EM algorithm, Xu (1997) further showed that the convergence rate by the original EM algorithm and the VEM algorithm is the same. Therefore, equation 2.20 also holds for the original EM algorithm. As a result, the following analyses and results apply to both the EM and VEM algorithm.

Suppose that the samples $\{x^{.t/}\}_1^N$ are randomly selected from the gaussian mixture with the parameter 2^*. When the gaussian mixture model satisfies certain regularity conditions, the EM iterations arrive at a consistent solution on maximizing log-likelihood equation 2.3 (Veaux, 1986; Redner & Walker, 1984). In this article, we assume that the EM algorithm asymptotically converges to this true parameter correctly (i.e., when N is large, for the sample data $\{x^{.t/}\}_1^N$, the EM algorithm converges to $\hat{2}$ with $\lim_{N\to\infty} \hat{2} = 2^*$), and we analyze the local convergence rate around this consistent solution in the limit form. It follows from equation 2.20 that an upper bound of the asymptotic convergence rate is given by

$$r \leq \lim_{N\to\infty} \|I + E^T P.\hat{2}/H.\hat{2}/E\|$$

$$= \|I + E^T \lim_{N\to\infty} P.\hat{2}/H.\hat{2}/E\|$$

$$= \|I + E^T \lim_{N\to\infty} P.2^*/H.2^*/E\|: \tag{2.21}$$

In order to estimate this bound, the rest of this article pursues an analysis of the convergence behavior of $P.\Xi^*/H.\Xi^*/$ as N increases to in□nity.

First, we give the Hessian of the log-likelihood equation 2.3 in the block forms. The detailed derivation is omitted (for an example of such a derivation, see Xu and Jordan, 1996, and for the general methods for matrix derivatives, see Gerald, 1980):

$$H_{\mathcal{A};\mathcal{A}^T} \triangleq \frac{@^2 l}{@\mathcal{A}@\mathcal{A}^T} = -\sum_{t=1}^{N} H_A^{-1}.t/.H_A^{-1}.t//^T;$$

$$H_{\mathcal{A};m_j^T} \triangleq \frac{@^2 l}{@\mathcal{A}@m_j^T} = \sum_{t=1}^{N} R_{A_j}^{-1}.t/.x^{\cdot t/} - m_j/^T 6_j^{-1};$$

$$H_{\mathcal{A};6_j^T} \triangleq \frac{@^2 l}{@\mathcal{A}@6_j^T} = -\frac{1}{2}\sum_{t=1}^{N} R_{A_j}^{-1}.t/ \otimes \mathrm{vec}[6_j^{-1} - U_j.t/]^T;$$

$$H_{m_i;m_j^T} \triangleq \frac{@^2 l}{@m_i@m_j^T} = -6_i^{-1}\sum_{t=1}^{N} \pm_{ij} h_i.t/ + \sum_{t=1}^{N} \circ_{ij}.t/6_i^{-1}.x^{\cdot t/} - m_i/ \times .x^{\cdot t/} - m_j/^T 6_j^{-1};$$

$$H_{m_i;6_j^T} \triangleq \frac{@^2 l}{@m_i@\mathrm{vec}[6_j]^T} = -\frac{1}{2}\sum_{t=1}^{N} \circ_{ij}.t/\mathrm{vec}[6_j^{-1} - U_j.t/]^T \otimes .6_i^{-1}.x^{\cdot t/} - m_i// - \sum_{t=1}^{N} \pm_{ij} h_i.t/.6_i^{-1}.x^{\cdot t/} - m_i// \otimes \mathrm{vec}[6_i^{-1}]^T;$$

$$H_{6_i;6_j^T} \triangleq \frac{@^2 l}{@\mathrm{vec}[6_i]@\mathrm{vec}[6_j]^T} = \frac{@}{@\mathrm{vec}[6_i]} \otimes \frac{@l}{@\mathrm{vec}[6_j]^T}$$

$$= -\frac{1}{4}\sum_{t=1}^{N} \circ_{ij}.t/\mathrm{vec}[6_j^{-1} - U_j.t/]^T \otimes \mathrm{vec}[6_i^{-1} - U_i.t/] - \frac{1}{2}\sum_{t=1}^{N} \pm_{ij} h_i.t/.-6_i^{-1} \otimes 6_i^{-1} + 6_i^{-1} \otimes U_i.t/ + U_i.t/ \otimes 6_i^{-1}/;$$

where $\Xi = \{\mathcal{A}; m_j; 6_j; j = 1; \ldots; K\}$, $\pm_{ij}$ is the Kronecker function, and

$$H_A^{-1}.t/ = [h_1.t/=®_1; \ldots; h_K.t/=®_K]^T;$$

$$\circ_{ij}.t/ = .\pm_{ij} - h_i.t//h_j.t/;$$

$$R_{A_j}^{-1}.t/ = [\circ_{1j}.t/=®_1; \ldots; \circ_{Kj}.t/=®_K]^T;$$

$$U_i.t/ = U_i.x^{\cdot t/}/ = 6_i^{-1}.x^{\cdot t/} - m_i/.x^{\cdot t/} - m_i/^T 6_i^{-1}:$$

When the K gaussian distributions in the mixture are separated enough, that is, the posterior probabilities $h_j(t)$ are approximately zero or one at Θ^*, by the above expression of the Hessian matrix, it is easy to verify that $\lim_{N\to\infty} P(\Theta^*)/H(\Theta^*)$ will become to the following block diagonal matrix:

$$F = \begin{pmatrix} -I_K + \mathcal{A}^*[1;1;\cdots;1] & 0 \\ 0 & -I_{K\times(d+d^2)} \end{pmatrix}; \tag{2.22}$$

where $\mathcal{A}^* = [\alpha_1^*;\cdots;\alpha_K^*]^T$ and I_n is the nth order identity matrix.

Furthermore, we rearrange the E formed from the unit basis vectors that span $\mathcal{R}_1^1$ into the following two parts:

$$E = \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix}; \tag{2.23}$$

with its first part $E_1 = [e_1^1;\cdots;e_{K-1}^1]$ and $e_1^1;\cdots;e_{K-1}^1$ are unit basis vectors of $\mathcal{R}_1^1 = \{x \in R^K: \sum_{j=1}^K x_i = 0\}$.

From equations 2.22 and 2.23, we have

$$E^T(I + F)E = \begin{pmatrix} E_1^T\mathcal{A}^*[1;1;\cdots;1]E_1 & 0 \\ 0 & 0 \end{pmatrix} = 0; \tag{2.24}$$

since $[1;\cdots;1]E_1 = 0$.

Substituting this result into the upper bound in equation 2.21, we have:

$$r \le \|I + E^T \lim_{N\to\infty} P(\Theta^*)/H(\Theta^*)E\| = \|E^T(I + F)E\| = 0:$$

Thus, the EM algorithm for gaussian mixtures has a superlinear convergence rate. A special case of this result was first given in Xu and Jordan (1996).

Actually gaussian distributions in the mixture cannot be strictly separated. In the rest of this article, we will show that $\lim_{N\to\infty} P(\Theta^*)/H(\Theta^*)$ approximately converges to F as the overlap of any two gaussian distributions in the mixture tends to zero.

3 The Main Theorem

We begin by introducing a set of quantities on the overlap of component gaussian distributions as follows:

$$e_{ij}(\Theta^*) = \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} |\delta_{ij}(t)| = \int |\delta_{ij}(x)|P(x)|\Theta^*|\,dx; \text{ for } i,j = 1;\cdots;K;$$

where $°_{ij}.x/ = .±_{ij} - h_i.x//h_j.x/$ and

$$h_l.x/ = \frac{®_l^* P.x|m_l^*; 6_l^*/}{\sum_{k=1}^K ®_k^* P.x|m_k^*; 6_k^*/} \text{ for } l = 1; : : : ; K:$$

Since $|°_{ij}.x/| \leq 1$, $e_{ij}.2^*/ \leq 1$.

If $i \neq j$, $e_{ij}.2^*/ = \int h_i.x/h_j.x/P.x|2^*/dx$, which can be considered as a measure of the average overlap between the distributions of gaussian i and j. In fact, if $P.x|m_i^*; 6_i^*/$ and $P.x|m_j^*; 6_j^*/$ overlap in a higher degree at a point x, then $h_i.x/h_j.x/$ certainly takes a higher value; otherwise, it takes a lower 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 a degree of overlap between $P.x|m_i^*; 6_i^*/$ and $P.x|m_j^*; 6_j^*/$ at x in the mixture environment, and the above $e_{ij}.2^*/$ is an average overlap between the distributions of gaussian i and j in the mixture.

Since two gaussian distributions always have some overlap area, we always have $e_{ij}.2^*/ > 0$. Moreover, by the fact $\sum_{j=1}^N h_i.x/ = 1$, we also have

$$\begin{aligned} e_{ii}.2^*/ &= \int |°_{ii}.x/|P.x|2^*/dx = \int h_i.x/.1 - h_i.x//P.x|2^*/dx \\ &= \sum_{j\neq i} \int h_i.x/h_j.x/P.x|2^*/dx = \sum_{j\neq i} e_{ij}.2^*/ > 0: \end{aligned}$$

Clearly, $e_{ii}.2^*/$ represents the overlap of the ith gaussian distribution with all the other gaussian distributions in the mixture.

We consider the worst case and de□ne

$$e.2^*/ = \max_{ij} e_{ij}.2^*/ \leq 1: \tag{3.1}$$

By experiments, we can □nd that as the gaussian distributions in the mixture become more separated, the EM algorithm can asymptotically converge to the true parameter 2^* with a better convergence rate. Here, we mathematically study the asymptotic convergence properties of

matrix U_i such that

$$6_i^* = U_i^T \begin{pmatrix} ,_{i1} & 0 & \cdots & 0 \\ 0 & ,_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & ,_{id} \end{pmatrix} U_i; \tag{3.3}$$

where $,_{i1}; : : : ; ,_{id}$ are the eigenvalues of 6_i^*. By taking the transformation $y = U_i.x - m_i^*/$, we have from equation 3.2 that

$$\sum_{j=1}^{n} ,_{ij}^{-1} y_i^2 = 1;$$

which is a hyperellipsoid at the origin in the transformed space. It follows from the inverse transformation $x = m_i^* + U_i^T y$ that the characteristic contour is a hyperellipsoid at m_i^* in the original space. In fact, the axes of this characteristic hyperellipsoid describe the attenuating rates of the component density in the coordinate basis directions in the transformed space (or the d principal directions in the original space from m_i^*), respectively, since they are the square roots of the eigenvalues, which are actually the variances of the ith component distribution in these directions, respectively. In other words, the density will attenuate rapidly as a point goes away from the center along a basis direction if the corresponding axis is small; otherwise, it will attenuate slowly. For simplicity, we use the minimum-radius hypersphere, which includes the characteristic contour,

$$.x - m_i^*/^T.x - m_i^*/ = ,_{max}^i;$$

instead of the characteristic contour, and then the radius of this hypersphere is $.,_{max}^i/^{\frac{1}{2}}$. In the same way, we can have the characteristic contour and the mini-radius hypersphere for the jth component distribution. Since the overlap between these two densities is approximately proportional to the overlap between the two characteristic contours or minimum-radius hyperspheres, we can de□ne $´_{ij}.2^*/$ as the overlap between the densities of components i and j as follows:

$$´_{ij}.2^*/ = \frac{·,_{max}^{i\,\frac{1}{2}}.,_{max}^{j\,\frac{1}{2}}}{\|m_i^* - m_j^*\|}:$$

Therefore, for a gaussian mixture of true parameter 2^*, we de□ne

$$´.2^*/ = \max_{i \neq j} ´_{ij}.2^*/ \tag{3.4}$$

as an overall overlap degree of gaussians in the mixture from point of view of the worst case. Obviously, $\acute{}.\Phi^*) \to 0$ is equivalent to $e.\Phi^*) \to 0$.

Second, we consider some assumptions that regularize the manner of $e.\Phi^*)$ tending to zero.

We first assume that Φ^* satisfies

Condition 1: $\alpha_i^* \geq \alpha$; for $i = 1; \ldots; K$;

where α is a positive number. When some mixing proportion α_i^* tends to zero, the corresponding gaussian will withdraw from the mixture, which degenerates to a mixture with a lower number of the mixing components. This assumption excludes this degeneracy.

Our second assumption is that the eigenvalues of all the covariance matrices satisfy

Condition 2: $\bar{}_{,}.\Phi^*) \leq {}_{,ik} \leq {}_{,}.\Phi^*)$; for $i = 1; \ldots; K; \quad k = 1; \ldots; d$;

where $\bar{}$ is a positive number and ${}_{,}.\Phi^*)$ is defined to be the maximum eigenvalue of the covariance matrices $\Sigma_1^*; \ldots; \Sigma_K^*$, that is,

$${}_{,}.\Phi^*) = \max_{i;k} {}_{,ik};$$

which is always upper bounded by a positive number B. That is, all the eigenvalues uniformly attenuate or reduce to zero when they tend to zero. It follows from condition 2 that condition numbers of all the covariance matrices are uniformly upper bounded, that is,

$$1 \leq \cdot.\Sigma_i^*) \leq B'; \text{ for } i = 1; \ldots; K;$$

where $\cdot.\Sigma_i^*)$ is the condition number of Σ_i^* and B' is a positive number.

The third assumption is that the mean vectors of the component densities in the mixture satisfy

Condition 3: $\circ D_{max}.\Phi^*) \leq D_{min}.\Phi^*) \leq \|m_i^* - m_j^*\|$

$$\leq D_{max}.\Phi^*); \text{ for } i \neq j;$$

where $D_{max}.\Phi^*) = \max_{i \neq j} \|m_i^* - m_j^*\|; D_{min}.\Phi^*) = \min_{i \neq j} \|m_i^* - m_j^*\|$, and $\circ$ is a positive number. That is, all the distances between two mean vectors are the same-order infinitely large quantities when they tend to infinity. Moreover, when the overlap of densities in the mixture reduces to zero, any of two means $m_i^*; m_j^*$ cannot be arbitrarily close; there should be a positive value T such that $\|m_i^* - m_j^*\| \geq T$ when $i \neq j$. Also, it is natural to assume that the mean vectors take different directions when they diverge to infinity.

With the above preparations, we are ready to introduce 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:

$$\lim_{N\to\infty} P(\Theta^*)/H(\Theta^*) = F + o(e^{(0.5-\varepsilon')}(\Theta^*)); \tag{3.5}$$

where F is given by equation 2.22 and ε' is an arbitrarily small positive number.

According to this theorem, as the overlap of distributions in the mixture becomes small or, more precisely, $e(\Theta^*) \to 0$, the asymptotic value of $P(\Theta^*)/H(\Theta^*)$ becomes F plus a matrix in which each element is a higher-order infinitesimal of $e^{(0.5-\varepsilon')}(\Theta^*)$. It further follows from equation 2.24 that $E^T(I + F)/E = 0$. Thus, we have

$$\begin{aligned} I + E^T \lim_{N\to\infty} P(\Theta^*)/H(\Theta^*)/E &= E^T(I + \lim_{N\to\infty} P(\Theta^*)/H(\Theta^*))/E \\ &= o(e^{(0.5-\varepsilon')}(\Theta^*)). \end{aligned}$$

That is, each element of $I + E^T \lim_{N\to\infty} P(\Theta^*)/H(\Theta^*)/E$ is a higher-order infinitesimal of $e^{(0.5-\varepsilon')}(\Theta^*)$. Because the norm of a real matrix $A = (a_{ij})_{m\times m}$ is always not larger than $\sqrt{m}\max_i \sum_{j=1}^m |a_{ij}|$, the norm $\|I + E^T P(\Theta^*)/H(\Theta^*)/E\|$ is certainly a higher-order infinitesimal of $e^{(0.5-\varepsilon')}(\Theta^*)$. Therefore, it follows from equation 2.21 that the asymptotic convergence rate of the EM algorithm locally around Θ^* is a higher-order infinitesimal of $e^{(0.5-\varepsilon')}(\Theta^*)$. That is, when $e(\Theta^*)$ is small and N is large enough, the convergence rate of the EM algorithm approaches approximately zero. In other words, the EM algorithm in this case has a quasi-Newton-type convergence behavior, with its asymptotic convergence rate dominated by the infinitesimal $e^{(0.5-\varepsilon')}(\Theta^*)$.

From this theorem, we can also find that $P(\Theta^*)$ tends to the inverse of $H(\Theta^*)$ as $e(\Theta^*)$ tends to zero. Thus, when the overlap of gaussian distributions in the mixture becomes a small value, the EM algorithm approximates to the Newton algorithm. Actually, $P(\Theta^*)$ makes the EM algorithm for the gaussian mixture be different from the conventional first-order iterative algorithm such as gradient ascent. This may be the basic underlying factor that speeds up the convergence of the EM algorithm when the overlap of gaussians in the mixture is low.

4 Lemmas

In this section, we describe several lemmas that will be used for proving the main theorem. The norm for a vector or matrix is always assumed to be Euclidean norm in the following.

We now define three kinds of special polynomial functions that we often meet in the further analyses:

De□nition 1. $g.x; 2^*/$ is called a regular function if it satis□es:

(i) If 2^* is □xed, $g.x; 2^*/$ is a polynomial function of the component variables $x_1; : : : ; x_d$ of x.

(ii) If x is □xed, $g.x; 2^*/$ is a polynomial function of the elements of $m_1^*; : : : ; m_K^*$, $6_1^*; : : : ; 6^*$

As $E.\|Y\|^i|0; I_d/$ is certainly finite and $\lambda.2^*/$ is upper bounded, $E.\|X - m_j^*\|^i|m_j^*; 6_j^*/$ is upper bounded. Since $\alpha_j^* < 1$ and $m.2^*/$ is lower bounded by $\frac{T}{2}$ under the constraint $\|m_i^* - m_j^*\| \geq T$, it follows from equation 4.2 that there exist a positive number ι and a positive integer q such that

$$E.\|X\|^k|2^*/ = \int \|x\|^k P.x|2^*/\,dx \leq \iota m^q.2^*/: \tag{4.4}$$

We further get an upper bound for $E.g^2.X; 2^*//$. In the process of $e.2^*/$ tending to zero under the assumptions, the elements of $\mathcal{A}^*$ and $\mathcal{A}^{*^{-1}}$ are bounded. Since $\|6_j^*\| = \lambda_{max}^j \leq \lambda.2^*/ \leq B$, the elements of each 6_j^* are bounded. Moreover, since $\lambda_{j1}; \ldots; \lambda_{jd}$ are the eigenvalues of 6_j^* and $\lambda_{j1}^{-1}; \ldots; \lambda_{jd}^{-1}$ are the eigenvalues of $6_j^{*^{-1}}$, we have

$$\|\lambda.2^*/6_j^{*^{-1}}\| \leq \lambda.2^*/\frac{1}{\lambda.2^*/} = \frac{1}{=}:$$

That is, the elements of $\lambda.2^*/6_j^{*^{-1}}$ are also bounded. We consider that $g.x; 2^*/$ is a polynomial function of $x_1; \ldots; x_d$ and that the coefficient of each term in $g.x; 2^*/$ is a polynomial function of the elements of $\mathcal{A}^*$, $\mathcal{A}^{*^{-1}}$, $m_1^*; \ldots; m_K^*, 6_1^*; \ldots; 6_K^*, \lambda.2^*/6_1^{*^{-1}}, \ldots, \lambda.2^*/6_K^{*^{-1}}$ with constant coefficients. Because the elements of these matrices and vectors except $m_1^*; \ldots; m_K^*$ are bounded, the absolute value of the coefficient is certainly upper bounded by a positive-order power function of $m.2^*/$. Therefore, $|g.x; 2^*/|$ is upper bounded by a positive polynomial function of $\|x\|$ with its coefficients being some polynomial functions of $m.2^*/$ of positive constant coefficients. As $g^2.x; 2^*/$ is certainly a balanced function, it is also upper bounded by a positive polynomial function of $\|x\|$. According to equation 4.4 and the property that $m.2^*/$ is lower bounded by $\frac{T}{2}$, there certainly exist a positive number ι and a positive integer p such that

$$E.g^2.X; 2^*// \leq \iota^2 m^{2p}.2^*/:$$

Finally, it follows from the Cauchy-Schwarz inequality and the fact $|°_{ij}.x/| \leq 1$ that

$$\begin{aligned}\int |g.x; 2^*/°_{ij}.x/|P.x|2^*/\,dx &= E.|g.X; 2^*/||°_{ij}.X/|/\\ &\leq E^{\frac{1}{2}}.g^2.X; 2//E^{\frac{1}{2}}.|°_{ij}.X/|/\\ &\leq \iota m^p.2^*/\sqrt{e_{ij}.2^*/}\\ &\leq \iota m^p.2^*/\sqrt{e.2^*/}:\end{aligned}$$

Since $\rho(\theta^*)$ is not an invertible function, that is, there may be many θ^*s for a value of $\rho(\theta^*)$, we further define

$$f(\rho) = \sup_{\rho(\theta^*)=\rho} e(\theta^*). \tag{4.5}$$

Because $e(\theta^*)$ is always not larger than 1 by the definition, $f(\rho)$ is well defined. By this definition, we certainly have

$$e_{ij}(\theta^*) \le e(\theta^*) \le f(\rho(\theta^*)). \tag{4.6}$$

Lemma 2. Suppose that θ^* satisfies conditions 1–3. As $\rho(\theta^*)$ tends to zero, we have:

(i) $\rho(\theta^*)$, $\rho_{ij}(\theta^*)$ and $\frac{\max_i}{\|m_i^* - m_j^*\|}$ are the equivalent infinitesimals.

(ii) For $i \neq j$, we have

$$\|m_i^*\| \le T'\|m_i^* - m_j^*\|; \tag{4.7}$$

where T' is a positive number.

(iii) For any two nonnegative numbers p and q with $p + q > 0$, we have

$$\|m_i^* - m_j^*\|^p \cdot \max_i{}^{-q} \le O(\rho^{-p\vee q}(\theta^*)); \tag{4.8}$$

$$\|m_i^* - m_j^*\|^p \cdot \max_i{}^{-q} \ge O(\rho^{-p\wedge q}(\theta^*)); \tag{4.9}$$

where $p \vee q = \max\{p, q\}$; $p \wedge q = \min\{p, q\}$.

See the appendix for the proof.

Lemma 3. When θ^* satisfies conditions 1–3 and $\rho(\theta^*) \to 0$ as an infinitesimal, we have

$$f''(\rho(\theta^*)) = o(\rho^p(\theta^*)); \tag{4.10}$$

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

See the appendix for the proof.

Lemma 4. Suppose that $g(x, \theta^*)$ is a regular and convertible function and that θ^* satisfies conditions 1–3. As $e(\theta^*) \to 0$ is considered as an infinitesimal, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \phi_{ij}(t) g(x^{(t)}, \theta^*) = \int g(x, \theta^*) \phi_{ij}(x) P(x|\theta^*)\, dx$$

$$= o(e^{0.5-\varepsilon}(\theta^*)) \tag{4.11}$$

where ε is an arbitrarily small, positive number.

Proof. Because g.x; 2*/ is a regular and convertible function, there is a positive integer p such that ¸^p. 2*/g.x; 2*/ is a balanced function. By lemmas 1 and 2, we have

$$\begin{aligned}
&\left| \int g.x; 2^*/°_{ij}.x/P.x|2^*/dx \right| \\
&\quad \le \int |g.x; 2^*/°_{ij}.x/|P.x|2^*/dx \\
&\quad = ¸^{-q}.2^*/\int |¸^q.2^*[\blacksquare]2^*/°_{ij}.x/|P.x|2^*/dx \\
&\quad \le 1¸^{-q}.2^*/m^p.2^*/e^{0:5}.2^*/ = 1¸^{-q}.2^*/\|m_i^*\|^p e^{0:5}.2^*/ \\
&\quad \le 1'.¸^{-q}_{i\max}/\|m_i^* - m_j^*\|^p e^{0:5}.2^*/ \le o'^{-n}.2^*/e^{0:5}.2^*/;
\end{aligned}$$

where we let m.2*/ = $\|m_i^*\|$, 1'; o are positive numbers and n = p ∨ q.

By lemma 3, we further have

$$\begin{aligned}
&\lim_{e.2^*/\to 0} \frac{\int |g.x; 2^*/°_{ij}.x/|P.x|2^*/dx}{e^{0:5-"}.2^*/} \\
&\quad = \lim_{'.2^*/\to 0} \frac{\int |g.x; 2^*/°_{ij}.x/|P.x|2^*/dx}{e^{0:5-"}.2^*/} \\
&\quad \le o \lim_{'.2^*/\to 0} \frac{e".2^*/}{'^n.2^*/} \le o \lim_{'.2^*/\to 0} \frac{f".'.2^*/}{'^n.2^*/} = 0:
\end{aligned}$$

Therefore, we have

$$\int g.x; 2^*/°_{ij}.x/P.x|2^*/dx = o.e^{0:5-"}.2^*//: \tag{4.12}$$

Corollary 1. Suppose that 2* satisfies conditions 1–3 and that e.2*/ → 0 is considered as an infinitesimal. If g.x; 2*/ = $\sum_{l=1}^{n} g_l.x; 2^*/$ and each $g_l.x; 2^*/$ is a regular and convertible function, we have

$$\lim_{N\to\infty}$$

5 Proof of Theorem 1

We are now ready to prove the main theorem.

Proof of Theorem 1. We are interested in the matrix $P.2^*/H.2^*/$, which determines the local convergence rate of EM.

The explicit expressions of $P.2^*/$ and $H.2^*/$ allow us to obtain formulas for the blocks of $P.2^*/H.2^*/$. To clarify our notation, we begin by writing out these blocks as follows:

$$
\begin{aligned}
&P.2^*/H.2^*/ \\
&= \operatorname{diag}[P_{\mathcal{A}}; P_{m_1}; \ldots; P_{m_K}; P_{6_1}; \ldots; P_{6_K}] \\
&\times \begin{pmatrix}
H_{\mathcal{A};\mathcal{A}^T} & H_{\mathcal{A};m_1^T} & \cdots & H_{\mathcal{A};m_K^T} & H_{\mathcal{A};6_1^T} & \cdots & H_{\mathcal{A};6_K^T} \\
H_{m_1;\mathcal{A}^T} & H_{m_1;m_1^T} & \cdots & H_{m_1;m_K^T} & H_{m_1;6_1^T} & \cdots & H_{m_1;6_K^T} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
H_{m_K;\mathcal{A}^T} & H_{m_K;m_1^T} & \cdots & H_{m_K;m_K^T} & H_{m_K;6_1^T} & \cdots & H_{m_K;6_K^T} \\
H_{6_1;\mathcal{A}^T} & H_{6_1;m_1^T} & \cdots & H_{6_1;m_K^T} & H_{6_1;6_1^T} & \cdots & H_{6_1;6_K^T} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
H_{6_K;\mathcal{A}^T} & H_{6_K;m_1^T} & \cdots & H_{6_K;m_K^T} & H_{6_K;6_1^T} & \cdots & H_{6_K;6_K^T}
\end{pmatrix} \\
&= \begin{pmatrix}
P_{\mathcal{A}}H_{\mathcal{A};\mathcal{A}^T} & P_{\mathcal{A}}H_{\mathcal{A};m_1^T} & \cdots & P_{\mathcal{A}}H_{\mathcal{A};m_K^T} & P_{\mathcal{A}}H_{\mathcal{A};6_1^T} & \cdots & P_{\mathcal{A}}H_{\mathcal{A};6_K^T} \\
P_{m_1}H_{m_1;\mathcal{A}^T} & P_{m_1}H_{m_1;m_1^T} & \cdots & P_{m_1}H_{m_1;m_K^T} & P_{m_1}H_{m_1;6_1^T} & \cdots & P_{m_1}H_{m_1;6_K^T} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
P_{m_K}H_{m_K;\mathcal{A}^T} & P_{m_K}H_{m_K;m_1^T} & \cdots & P_{m_K}H_{m_K;m_K^T} & P_{m_K}H_{m_K;6_1^T} & \cdots & P_{m_K}H_{m_K;6_K^T} \\
P_{6_1}H_{6_1;\mathcal{A}^T} & P_{6_1}H_{6_1;m_1^T} & \cdots & P_{6_1}H_{6_1;m_K^T} & P_{6_1}H_{6_1;6_1^T} & \cdots & P_{6_1}H_{6_1;6_K^T} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
P_{6_K}H_{6_K;\mathcal{A}^T} & P_{6_K}H_{6_K;m_1^T} & \cdots & P_{6_K}H_{6_K;m_K^T} & P_{6_K}H_{6_K;6_1^T} & \cdots & P_{6_K}H_{6_K;6_K^T}
\end{pmatrix}:
\end{aligned}
$$

Based on the expressions of the Hessian blocks, we have:

$$
\begin{aligned}
P_{\mathcal{A}}H_{\mathcal{A};\mathcal{A}^T} &= \frac{1}{N}.\operatorname{diag}[\text{®}_1^*; \ldots; \text{®}_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/ \times \left(-\sum_{t=1}^{N} H_A^{-1}.t/.H_A^{-1}.t//^T \right) \\
&= -.\operatorname{diag}[\text{®}_1^*; \ldots; \text{®}_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/ \left(\frac{1}{N} \sum_{t=1}^{N} H_A^{-1}.t/.H_A^{-1}.t//^T \right):
\end{aligned}
$$

By taking the limit as $N \to \infty$, we obtain:

$$\lim_{N\to\infty} P_{\mathcal{A}} H_{\mathcal{A};\mathcal{A}^T} = -.\mathrm{diag}[\mathbb{R}_1^*;::::;\mathbb{R}_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/ \lim_{N\to\infty} \frac{1}{N} \times \sum_{t=1}^{N} H_A^{-1}.t/.H_A^{-1}.t//^T:$$

We further let

$$I.2^*/ = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} H_A^{-1}.t/H_A^{-1}.t/^T \tag{5.1}$$

and compute this matrix by the elements.

When $i \neq j$, we have

$$I.2^*/.i;j/ = \frac{1}{\mathbb{R}_i^*\mathbb{R}_j^*} \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} h_i.t/h_j.t/ = \frac{1}{\mathbb{R}_i^*\mathbb{R}_j^*} e_{ij}.2^*/ \leq \frac{1}{\mathbb{R}^2} e.2^*/ = o.e^{0:5-"}.2^*//:$$

When $i = j$, we further have

$$I.2^*/.i;i/ = \frac{1}{.\mathbb{R}_i^*/^2} \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} h_i.t/[1 - .1 - h_i.t//] = \frac{1}{.\mathbb{R}_i^*/^2} \int h_i.x/P.x|2^*/dx - \frac{1}{.\mathbb{R}_i^*/^2} \int |^\circ_{ii}.x/|P.x|2^*/dx \geq \mathbb{R}_i^{*^{-1}} - \frac{1}{\mathbb{R}^2} e.2^*/ = \mathbb{R}_i^{*^{-1}} + o.e^{0:5-"}.2^*//:$$

Then we have

$$I.2^*/ = \mathrm{diag}[\mathbb{R}_1^{*^{-1}};::::;\mathbb{R}_K^{*^{-1}}] + o.e^{0:5-"}.2^*//: \tag{5.2}$$

Therefore, we get

$$\begin{aligned}
&\lim_{N\to\infty} P_{\mathcal{A}} H_{\mathcal{A};\mathcal{A}^T} \\
&= -.\mathrm{diag}[\mathbb{R}_1^*;::::;\mathbb{R}_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/I.2^*/ \\
&= -.\mathrm{diag}[\mathbb{R}_1^*;::::;\mathbb{R}_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/.\mathrm{diag}[\mathbb{R}_1^{*^{-1}};::::;\mathbb{R}_K^{*^{-1}}] + o.e^{0:5-"}.2^*// \\
&= -I_K + \mathcal{A}^*[1;1;::::;1] + o.e^{0:5-"}.2^*//:
\end{aligned}$$

For $j = 1; \ldots; K$, we have

$$P_{\mathcal{A}} H_{\mathcal{A};m_j^T} = \frac{1}{N}.\mathrm{diag}[\alpha_1^*; \ldots; \alpha_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T / \sum_{t=1}^{N} R_{A_j}^{-1}.t/.x^{\cdot t/} - m_j^*/^T 6_j^{*^{-1}}$$

$$= .\mathrm{diag}[\alpha_1^*; \ldots; \alpha_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T / \frac{1}{N} \sum_{t=1}^{N} R_{A_j}^{-1}.t/.x^{\cdot t/} - m_j^*/^T 6_j^{*^{-1}}:$$

Since $R_{A_j}^{-1}.t/ = [\circ_{1j}.t/=\alpha_1^*; \ldots; \circ_{Kj}=\alpha_K^*]^T$, we let $g.x; 2^*/$ be any element of $\mathcal{A}^{*^{-1}}.x - m_j^*/^T 6_j^{*^{-1}}$. Obviously, this kind of $g.x; 2^*/$ is a regular and convertible function with $q = 1$. By lemma 4, we have:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} R_{A_j}^{-1}.t/.x^{\cdot t/} - m_j^*/^T 6_j^{*^{-1}} = o.e^{0:5-"}.2^*//: \tag{5.3}$$

Because the elements of the matrix $.\mathrm{diag}[\alpha_1^*; \ldots; \alpha_K^*] - \mathcal{A}^*.\mathcal{A}^*/^T/$ are bounded, we further have

$$\lim_{N\to\infty} P_{\mathcal{A}} H_{\mathcal{A};m_j^T} = o.e^{0:5-"}.2^*//: \tag{5.4}$$

Because the Hessian matrix has the following property,

$H_{m_j;\mathcal{A}}$

Next, we consider $P_{m_i} H_{m_i;m_j^T}$ and have

$$P_{m_i} H_{m_i;m_j^T} = -\pm_{ij} I_d + \qquad 1$$

On the other hand, we have

$$
\begin{aligned}
P_{6_i} H_{6_i;m_j^T} &= \frac{2}{\sum_{t=1}^{N} h_i.t/} 6_i^* \otimes 6_i^* H_{m_j;6_i^T}^T \\
&= -6_i^* \otimes 6_i^* \left\{ \frac{1}{\sum_{t=1}^{N} h_i.t/} \sum_{t=1}^{N} {}^\circ_{ij}.t/\mathrm{vec}[6_i^{*^{-1}} - U_i.t/] \right. \\
&\qquad \left. \otimes \{6_j^{*^{-1}}.x^{\cdot t/} - m_j^*/\}^T \right\} \\
&\quad - 26_i^* \otimes 6_i^* \left\{ \frac{1}{\sum_{t=1}^{N} h_i.t/} \sum_{t=1}^{N} \pm_{ij} h_i.t/\{6_i^{*^{-1}}.x^{\cdot t/} - m_i^*/\}^T \right. \\
&\qquad \left. \otimes \mathrm{vec}[6_i^{*^{-1}}] \right\};
\end{aligned}
$$

and by the same argument used for $P_{m_i} H_{m_i;6_j^T}$, we have

$$
\lim_{N\to\infty} P_{6_i} H_{6_i;m_i^T} = 0.e^{0:5-"}.2^*//: \tag{5.10}
$$

Now we turn to the remaining case, $P_{6_i} H_{6_i;6_j^T}$, and have:

$$
\begin{aligned}
P_{6_i} H_{6_i;6_j^T} &= -\frac{1}{2} 6_i^* \otimes 6_i^* \left\{ \frac{1}{\sum_{t=1}^{N} h_i.t/} \sum_{t=1}^{N} {}^\circ_{ij}.t/\mathrm{vec}[6_j^{*^{-1}} - U_j.t/] \right. \\
&\qquad \left. \otimes \mathrm{vec}[6_i^{*^{-1}} - U_i.t/] \right\} \\
&\quad - \pm_{ij}.6_i^* \otimes 6_i^*/\{-6_i^{*^{-1}} \otimes 6_i^{*^{-1}} + 6_i^{*^{-1}} \\
&\quad \otimes \left\{ \frac{1}{\sum_{t=1}^{N} h_i.t/} \sum_{t=1}^{N} h_i.t/U_i.t/ \right\} \\
&\quad + \left\{ \frac{1}{\sum_{t=1}^{N} h_i.t/} \sum_{t=1}^{N} h_i.t/U_i.t/\} \otimes 6_i^{*^{-1}} \right\}:
\end{aligned}
$$

By corollary 1 and the law of large number, we have

$$
\begin{aligned}
\lim_{N\to\infty} P_{6_i} H_{6_i;6_j^T} &= 0.e^{0:5-"}.2^*// - \pm_{ij}.6_i^* \otimes 6_i^*/.6_i^{*^{-1}} \otimes 6_i^{*^{-1}}/ \\
&= -\pm_{ij} I_{d^2} + 0.e^{0:5-"}.2^*//:
\end{aligned}
$$

Summing up all the results, we obtain:

$$\lim_{N\to\infty} P.2^*/H.2^*/ = \begin{pmatrix} \\ 0 & -I_{K\times.d+d^2/} \end{pmatrix}$$

$$+ o.e^{0:5-"}.2^*//$$

$$= F + o.e^{0:5-"}.2^*//;$$

and the proof is completed.

8 Conclusions

We have presented an analysis of the asymptotic convergence rate of the EM algorithm for gaussian mixtures. Our analysis shows that when the overlap of any two gaussian distributions is small enough, the large sample local convergence behavior of the EM algorithm islargelar

two

$\|m_i^*\|$ or $\|m_j^*\|$ is upper bounded. Otherwise, if $\|m_i^*\|$ and $\|m_j^*\|$ both increase to infinity, since m_i^* and m_j^* take different directions, the order of the infinitely large quantity $\|m_i^*\|$ is certainly lower than or equal to that of $\|m_i^* - m_j^*\|$, which also leads to equation 4.7. Therefore, (ii) holds under the assumptions.

Finally, we turn to (iii). For the first inequality, equation 4.8, we consider three cases, as follows.

In the simple case $p = q > 0$, according to (i), we have

$$\|m_i^* - m_j^*\|^p \cdot \epsilon_{\max}^i / {}^{-q} = \|m_i^* - m_j^*\|^p \cdot \epsilon_{\max}^i / {}^{-p}$$

$$= \left(\frac{\epsilon_{\max}^i}{\|m_i^* - m_j^*\|} \right)^{-p} = O(\delta^{-p}) = O(\delta^{-p \vee q}).$$

If $p > q$, since $\epsilon_{\max}^i$ is upper bounded and according to (i), we have

$$\|m_i^* - m_j^*\|^p \cdot \epsilon_{\max} / {}^{-q} \le O(\delta^{-p}) = O(\delta^{-p \vee q}).$$

If $p < q$, as $\|m_i^* - m_j^*\| \ge \Delta$, we can have

$$\|m_i^* - m_j^*\|^p \cdot \epsilon_{\max}^i / {}^{-q} \le O(\delta^{-q}) = O(\delta^{-p \vee q}).$$

Summing up the results on the three cases, we obtain:

$$\|m_i^* - m_j^*\|^p \cdot \epsilon_{\max}^i / {}^{-q} \le O(\delta^{-p \vee q}).$$

In a similar way as above, we can prove the second inequality, equation 4.9.

Proof of Lemma 3. We first prove that

$$f(\delta) = o(\delta^{-p}),$$

as $\delta \to 0$, where p is an arbitrarily positive number p.

We consider the mixture of K gaussian densities of the parameter θ^* under the relation $\delta(\theta^*) = \delta$. When $i \ne j$, for a small enough δ, there is certainly a point m_{ij}^* on the line between m_i^* and m_j^* such that

$$\alpha_i^* P_i(m_{ij}^* | m_i^*, \Sigma_i^*) = \alpha_j^* P_j(m_{ij}^* | m_j^*, \Sigma_j^*).$$

We further define

$$E_i = \{x : \alpha_i^* P_i(x | m_i^*, \Sigma_i^*) \ge \alpha_j^* P_j(x | m_j^*, \Sigma_j^*)\},$$

$$E_j = \{x : \alpha_j^* P_j(x | m_j^*, \Sigma_j^*) > \alpha_i^* P_i(x | m_i^*, \Sigma_i^*)\}.$$

As $\epsilon(2^*)$ tends to zero, $\frac{\max_i}{\|m_i^* - m_j^*\|}$ and $\frac{\max_i}{\|m_i^* - m_j^*\|}$ are the same-order infinitesimals. Moreover, $\cdot(\theta_i^*)$ and $\cdot(\theta_j^*)$ are both upper bounded. Thus, there certainly exist some neighborhoods of m_i^* (or m_j^*) in E_i (or E_j). For clarity, we let $\mathcal{N}_{r_i}(m_i^*)$ and $\mathcal{N}_{r_j}(m_j^*)$ be the largest neighborhood in E_i and E_j, respectively, where r_i and r_j are their radiuses. Obviously r_i and r_j are both proportional to $\|m_i^* - m_j^*\|$ when $\|m_i^* - m_j^*\|$ either tends to infinity or is upper bounded. So there exist a pair of positive numbers b_1 and b_2 such that

$$r_i \geq b_i \|m_i^* - m_j^*\|; \text{ and } r_j \geq b_j \|m_i^* - m_j^*\|:$$

We further define

$$\mathcal{D}_i = \mathcal{N}_{r_i}^c(m_i^*) = \{x: \|x - m_i^*\| \geq r_i\};$$
$$\mathcal{D}_j = \mathcal{N}_{r_j}^c(m_j^*) = \{x: \|x - m_j^*\| \geq r_j\};$$

and thus

$$E_i \subset D_j; \qquad E_j \subset D_i:$$

Moreover, from the definitions of $e_{ij}(2^*)$ and $h_k(x)$, we have

$$\begin{aligned} e_{ij}(2^*) &= \int h_i(x) h_j(x) P(x|2^*)\, dx \\ &= \int_{E_i} h_i(x) h_j(x) P(x|2^*)\, dx + \int_{E_j} h_i(x) h_j(x) P(x|2^*)\, dx \\ &\leq \int_{\mathcal{D}_j} h_i(x) h_j(x) P(x|2^*)\, dx + \int_{\mathcal{D}_i} h_i(x) h_j(x) P(x|2^*)\, dx \\ &= \alpha_j^* \int_{\mathcal{D}_j} P_j(x|m_j^*, \theta_j^*)\, dx + \alpha_i^* \int_{\mathcal{D}_i} P_i(x|m_i^*, \theta_j^*)\, dx: \end{aligned}$$

We now consider $\int_{\mathcal{D}_i} P_i(x|m_i^*, \theta_i^*)\, dx$. Since $r_i \geq b_i \|m_i^* - m_j^*\|$,

$$\int_{\mathcal{D}_i} P_i(x|m_i^*, \theta_i^*)\, dx \leq \int_{\|x - m_i^*\| \geq b_i \|m_i^* - m_j^*\|} P_i(x|m_i^*, \theta_i^*)\, dx:$$

When $\epsilon(2^*)$ is sufficiently small, by the transformation $y = U_i(x - m_i^*)/\|m_i^* -$

$m_j^*\|$, we have

$$\int_{\mathcal{D}_i} P_i.x|m_i^*; 6_i^*/\,dx \le \int_{\|y\|\ge b_i} o\,\frac{\|m_i^* - m_j^*\|}{\cdot_{,\,\max}^{i}/^{\frac{d}{2}}} e^{-\frac{1}{2}\frac{\|m_i^*-m_j^*\|^2}{\cdot_{,\,\max}^{i}/^{d}}\|y\|^2}\,dy; \tag{A.4}$$

where o is a positive number.

By lemma 2, we have

$$\|m_i^* - m_j^*\|.\,_{,\,\max}^{i}/^{-\frac{d}{2}} \le O.´^{-c_1}.2^*//;$$

$$\|m_i^* - m_j^*\|^2.\,_{,\,\max}^{i}/^{-d} \ge O.´^{-c_2}.2^*//;$$

where $c_1 = 1 \vee \frac{d}{2}$; $c_2 = 2 \wedge d$.

According to these results, we have from equation A.4 that

$$\int_{\mathcal{D}_i} P_i.x|m_i^*; 6_i^*/\,dx \le {}^{1}\int_{\mathcal{B}_i} \frac{1}{´^{c_1}} e^{-½\frac{1}{´^{c_2}}\|y\|^2}\,dy; \tag{A.5}$$

where $\mathcal{B}_i = \{y\colon \|y\| \ge b_i\}$, ^{1} and ½ are positive numbers.

Furthermore, we let

$$F_i.´/ = \int_{\mathcal{B}_i} P.y|´/\,dy; \qquad P.y|´/ = \frac{1}{´^{c_1}} e^{-½\frac{1}{´^{c_2}}\|y\|^2};$$

and consider the limit of $\frac{F_i.´/}{´^{p}}$ as $´$ tends to zero.

For each $y \in \mathcal{B}_i$, we have

$$\lim_{´\to 0} \frac{P.y|´/}{´^{p}} = \lim_{□=\frac{1}{´}\to\infty} \frac{□^{.c_1+p/}}{e^{□^{c_2}½\|y\|^2}} = 0$$

uniformly in $\mathcal{B}_i$, which leads to

$$\lim_{´\to 0} \frac{F_i.´/}{´^{p}} = \int_{\mathcal{B}_i} \lim_{´\to 0} \frac{P.y|´/}{´^{p}}\,dy = 0;$$

and thus $F_i.´/ = o.´^{p}/$: It further follows from equation A.5 that

$$\sup_{´.2^*/=´} \int_{\mathcal{D}_i} P_i.x|m_i^*; 6_i^*/\,dx = o.´^{p}/: \tag{A.6}$$

Similarly, we can also prove

$$\sup_{´.2^*/=´} \int_{\mathcal{D}_j} P_j.x|m_j^*; 6_j^*/\,dx = o.´^{p}/:$$

As a result, we have

$$\begin{aligned} f_{ij}(\theta) &= \sup_{\theta:\theta^*=\theta} e_{ij}(\theta^*) \\ &\le \sup_{\theta:\theta^*=\theta} \left(\alpha_j^* \int_{D_j} P_j(x|m_j^*, \Sigma_j^*)\, dx + \alpha_i^* \int_{D_i} P_i(x|m_i^*, \Sigma_i^*)\, dx \right) \\ &\le \sup_{\theta:\theta^*=\theta} \int_{D_j} P_j(x|m_j^*, \Sigma_j^*)\, dx + \sup_{\theta:\theta^*=\theta} \int_{D_i} P_i(x|m_i^*, \Sigma_i^*)\, dx \\ &= o(\theta^p). \end{aligned}$$

In the case $i = j$, we also have

$$\begin{aligned} f_{ii}(\theta) &= \sup_{\theta:\theta^*=\theta} \int |\Delta_{ii}(x)|P(x)|\theta^*|\, dx \\ &= \sup_{\theta:\theta^*=\theta} \int h_i(x)(1 - h_i(x))|P(x)|\theta^*|\, dx \\ &\le \sum_{j\ne i} \sup_{\theta:\theta^*=\theta} \int h_i(x)h_j(x)|P(x)|\theta^*|\, dx = \sum_{j\ne i} e_{ij}(\alpha) \\ &= o(\theta^p). \end{aligned}$$

Summing up the results, we have

$$f(\theta) \le \max_{i,j} f_{ij}(\theta) = o(\theta^p).$$

Moreover, because

$$\lim_{\theta\to 0} \frac{f''(\theta)}{\theta^p} = \lim_{\theta\to 0} \left(\frac{f(\theta)}{\theta^{\frac{p}{2}}}\right)'' = 0,$$

we finally have $f''(\theta) = o(\theta^p)$ and thus $f''(\theta^*) = o(\theta^{*p})$.

Acknowledgments

The work was supported by the HK RGC Earmarked Grant 4297/98E. We thank Professor Jiong Ruan and Dr. Yong Gao for the valuable discussions.

References

Delyon, B., Lavielle, M., & Moulines E. (1999). Convergence of a stochastic approximation version of the EM algorithm. Annals of Statistics, 27, 94–128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc., B, 39, 1–38.

Gerald, S. R. (1980). Matrix derivatives. New York: Dekker.

Jamshidian, M., & Jennrich R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. J. R. Statist. Soc., B, 59, 569–587.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. Neural Computation, 6, 181–214.

Jordan, M. I., & Xu, L. (1995). Convergence results for the EP approach to mixtures of experts architectures. Neural Networks, 8(9), 1409–1431.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. J. R. Statist. Soc., B, 57, 425–437.

Lange, K. (1995b). A quasi-Newton acceleration of the EM algorithm. Statist. Sinica, 5.

Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika, 81, 633–648.

Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. Ann. Statist., 22, 326–339.

Meng, X. L., & van Dyk D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. J. R. Statist. Soc., B, 59, 511–567.

Meng, X. L., & van Dyk D. (1998). Fast EM-type implementations for mixed effects models. J. R. Statist. Soc., B, 60, 559–578.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of IEEE, 77, 257–286.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood, and the EM algorithm. SIAM Review, 26, 195–239.

Veaux, D. De R. (1986). Parameter estimation for a mixture of linear regressions. Unpublished doctoral dissertation, Stanford University.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. Annals of the Statistics, 11, 95–103.

Xu, L. (1997). Comparative analysis